EXPLORATORY RESEARCH

D5.3 Final tool and model description, and case studies results

Deliverable 5.3

Domino

Grant:	783206
Call:	H2020-SESAR-2016-2
Topic:	SESAR-ER3-06-2016 ATM Operations, Architecture, Performance and Validation
Consortium coordinator:	University of Westminster
Edition date:	20 December 2019
Edition:	01.00.00

Authoring & Approval

Authors of the document

Name/Beneficiary	Position/Title	Date
Luis Delgado / University of Westminster	Project member	18 December 2019
Gérald Gurtner / University of Westminster	Project member	18 December 2019
Silvia Zaoli / Università di Bologna	Project member	18 December 2019
Piero Mazzarisi / Università di Bologna	Project member	18 December 2019
Damir Valput / Innaxis	Project member	18 December 2019
Andrew Cook / University of Westminster	Project coordinator	18 December 2019
Fabrizio Lillo / Università di Bologna	Project member	18 December 2019

Reviewers internal to the project

Name/Beneficiary	Position/Title	Date
Graham Tanner / University of Westminster	Project member	20 December 2019
Gérald Gurtner / University of Westminster	Project member	20 December 2019
Luis Delgado / University of Westminster	Project member	20 December 2019

Approved for submission to the SJU By — Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
Andrew Cook / University of Westminster	Project coordinator	20 December 2019

Rejected By - Representatives of beneficiaries involved in the project

Name/Beneficiary	Position/Title	Date
N/A		

Document History

Edition	Date	Status	Author	Justification
01.00.00	20 December 2019	Release	Domino Consortium	New document for review by the SJU

Domino

NOVEL TOOLS TO EVALUATE ATM SYSTEMS COUPLING UNDER FUTURE DEPLOYMENT SCENARIOS

This deliverable is part of a project that has received funding from the SESAR Joint Undertaking under grant agreement No 783206 under European Union's Horizon 2020 research and innovation programme.

Abstract

This deliverable presents the final results obtained from the Domino project. It presents the corresponding metrics, the model, and a detailed analysis of two case studies. The main modifications to the model with respect to the previous version are highlighted, including curfew management. The calibration of the model is presented, which is similar to the previous version, with more in-depth analyses and further effort dedicated to the calibration process. Two case studies are defined in this deliverable, using previous definitions of the three base mechanisms: 4D trajectory adjustments, flight prioritisation, and flight arrival coordination. The case studies are defined to have a focused insight into the efficiency of the mechanisms in specific environments. The two case studies are run by the model and analysed using metrics previously defined, including centrality and causality metrics. The results show different levels of efficiency for the three mechanisms, highlight the degree of robustness to the propagation of negative effects (such as delay) in the system, demonstrate various trade-offs between the indicators, and support a discussion of the limit of the mechanisms.

The opinions expressed herein reflect the authors' views only. Under no circumstances shall the SESAR Joint Undertaking be responsible for any use that may be made of the information contained herein.

Table of Contents

List of figures

List of tables

Executive Summary

The primary goal of Domino was to improve the state of the art regarding a methodology for analysing the architecture of, and interdependencies within, the air transportation system, by capturing different facets of causality under the impact of a selection of ATM mechanisms. In this regard, the project has made progress on three key topics to reach this goal, as listed below.

- The possibility to model, at a disaggregated level, the full gate-to-gate European air transportation system.
- The introduction of new metrics allowing the capture of some subtle network effects.
- A powerful statistical analysis, combining the capabilities of the model with the power of classical and new metrics to perform a full, network-wide assessment.

The re-implementation of the Mercury simulator, using an agent-based paradigm, is one of the most important achievements of Domino. To our knowledge, it is the only full, European Civil Aviation Conference (ECAC)-wide model able to simulate key stakeholders, such as, passengers, airlines and the network manager, in an integrated simulator. With respect to previous versions, it allows us to inject complex behavioural rules for different agents, in particular airlines.

Various metrics have been deployed by Domino. Some are classically used in air traffic flow management (ATFM) (such as average delay), some have been imported from other fields, others have been developed specifically for this project: in particular, the centrality metrics, which take into account the itineraries of passengers and the precise timing of the scheduled flights. At their core, they represent the most relevant metrics in terms of connectivity considering passengers. For their practical usage, Domino has identified some shortcomings. Causality metrics, on the other hand, have been used before in the air transportation system. Their introduction answered the necessity for decision-makers to understand causal links between subsystems, as opposed to correlations, in order to gain some high-level knowledge. The new methods introduced in Domino, allow us to capture different facets of causality, in particular with an emphasis on how rare events trigger other rare events in the system.

The model allows us to measure a large number of low-level observables, which both poses a problem and raises an opportunity. With so much data, different practical issues arise – such as storage or analysis time – but, more importantly, statistical analyses must be performed with care. Due to the number of observables, the stochastic nature of the simulator, its geographical scope, and the dynamic nature of the system, many different analyses can be performed. In this deliverable, we have focused on the variety of metrics available to the modeller, and on the possibility to restrict the analysis in scope (such as geographically or by stakeholder). Domino has shown that the model can be used to inspect, with a high level of detail, different aspects of the system. In particular, it is able to shed light on the inner functioning of different mechanisms (such as flight swapping or dynamic cost indexing), understanding under which conditions they would, or would not, provide benefits for the different stakeholders. Domino's model sheds light on the role of exogenous and endogenous noise, the behaviours of agents and the initial conditions (passengers, schedules, etc.) on the efficiency of different mechanisms.

Of the three mechanisms investigated, the 4D trajectory adjustments mechanism seems to have the greatest impact. Its application in the first case study, hub management, shows that it has some impact on delays, costs, centrality, and causality metrics. It is efficient at reducing costs for the airlines, mostly through the protection of critical flights and overall arrival delay reduction. However, whilst connecting passengers tend to gain from this mechanism, other passengers have their average arrival delay increased, which highlights an important trade-off that policymakers should take into account. Passenger centrality tends to worsen, which indicates that cost-driven airlines are negatively affecting more (low cost-impacting) itineraries than they improve (high cost-impacting) ones. Overall, the mechanism creates some buffer in the system, as shown by the causality metrics, decreasing the potential delay propagation channels.

The flight prioritisation mechanism has almost no effect at a system level, except for a tendency to decrease delay feedback loops (hence decreasing the probability that a flight is late because previous flights are delayed). This mechanism is, by its nature, limited with regard to which airlines can use it (e.g. having enough flights in an ATFM regulation at an arrival airport with expected costs sufficiently different to benefit from swapping them) and in which circumstances (e.g. at arrival airports which have issued an ATFM regulation). The impact at the network level is thus also rather limited.

The effect of the horizon of the flight arrival coordination mechanism is highly dependent on the exact optimisation algorithm of its queuing process. Overall, the larger the Extended Arrival Manager (E-AMAN) radius, the higher the uncertainty associated with the flights therein. This translates into suboptimal behaviours if the optimiser does not have proper uncertainty computation capabilities. For instance, longer holding times can be assigned to flights that previously had fuel-saving instructions, due to landing sequence-breaking uncertainties.

Looking forward, Mercury can serve as a test bed for different types of simulations. Different optimisation processes for E-AMAN, rules for flight swapping or trajectory management, levels of congestion, levels of compensation and duty of care for passengers are all examples of modifications which can be tested, relying on a realistic representation for the other components of the model. With respect to other tools, Mercury provides many advantages. For instance, whereas the EUROCONTROL RNEST tool is more advanced regarding airspace management (including, for example, explicit ATFM regulations and the Computer-Assisted Slot Allocation (CASA) algorithm implementation), Mercury takes into account behavioural (potentially sub-rational) effects from different agents, realistic, stochastic generation of delays, passenger management, and a highly detailed cost of delay model, driving the most important decisions for airlines.

1 Introduction

One of the challenges when introducing or modifying solutions in the ATM system is to understand the effect of these, not at a local level, but their repercussion in the wide-network. Domino has developed a model and a set of metrics which allow modellers to capture these interactions. Moreover, the impact of these changes is not only experienced by airlines and in terms of delay, but also by passengers and in terms of cost. Domino is able to estimate these different layers of metrics.

This deliverable presents the final results of the Domino project. This includes the development of a dedicated agent-based model which is able to estimate indicators for the entire ECAC region for a full day of operations for both flights and passengers perspective. Some relevant stakeholders are explicitly modelled.

Three different mechanisms with three different levels of implementation each have been developed in Domino. The broad analysis of these mechanisms was presented in the analysis of different case studies in D5.2 Investigative case studies results. With feedback from stakeholders gathered at different consultation activities (see D6.3 Workshop results summary), the model has been updated and more detailed case studies defined. These final case studies are presented and analysed in this deliverable.

1.1 Case studies

D5.2 Investigative case studies results presented the results of modelling the main scenarios identified in D3.2 Investigative case studies description. A total of 14 scenarios were analysed. This broad perspective allowed us to understand some of the limitations and capabilities of the suggested metrics.

In this deliverable, instead, we decided to focus on fewer scenarios and grouped them into two case studies:

- **Hub delay management**: where three hubs are modelled with high ATFM regulations on them and the impact of introducing the 4D Trajectory Adjustment (4DTA) mechanism is evaluated along with the Flight Prioritisation (FP).
- **Effect of E-AMAN scope on arrival manager**: where the planning horizon of the E-AMAN is increased and the Flight Arrival Coordination (FAC) mechanism which tries to optimise the expected cost of the airlines is applied.

The system is analysed with a reference where the mechanisms are modelled at Level 0 (trying to replicate current practices) and with advanced behaviours, at Level 2.

These case studies were reported in D3.3 Adaptive case studies description. More details on the scenarios and mechanisms modelled in this deliverable are presented in Section 1.

1.2 Domino model

The Domino model is an agent-based model (ABM) executed over a stochastic event driven simulator. A total of 8 agent types are modelled:

- Flights
- Airline operating centres (AOC)
- Ground airport
- Network manager
- E-AMAN
- DMAN
- Radar
- Flight swapper

Approximatively 32.000 instances of these agents are created during the simulations.

The full architecture of the model is presented in detail in D4.1 Initial ABM model design. In this deliverable we present the main evolutions carried out from the implementation plan described in D4.1.

The usage of a standard procedure and methodology to develop the ABM (Gaia methodology based on agents, roles and their interactions) has allowed us to create a model which is flexible and adaptable in the behaviour of the agents. Moreover, the model is integrated and run over an event-driven simulator which has been developed using standard approaches and libraries (Python SimPy). All this mean that the code is easy to follow, and it is relatively simple to introduce and/or adapt new mechanisms and events.

The model captures the interaction between the agents, and it can represent flights and passenger itineraries. Therefore, many metrics can be captured at a very low level, such as for example, departure delay for flights and passengers, missed connections, taxi times, holding fuel.

The different processes are modelled with the required level of detail to capture the required metrics and interactions for the analysis of the mechanism: from explicit modelling of messages between agents to the usage of stochastic probabilities calibrated with historical data. This implements a meso-model which balances accuracy, computation time and flexibility. The current architecture allows modellers to modify these processes as required in a simple manner, for example, replacing taxi times which are generated following stochastic distributions with explicit modelling of movements at an airport.

For a full description of the model the reader is referred to D4.1 Initial ABM model design. In D5.3, Section 3 compiles the modifications introduced in the model since its design and it provides more detail on the agents' features along with other internal modelling particularities.

1.3 Model calibration

As in the previous version of the model, some effects are explicitly modelled (e.g., propagation of delay due to reactionary delay), but others are the results of a higher abstraction (e.g., actual taxi time or turnaround times). These parameters are stochastically sampled from distributions that need to be calibrated.

Some issues identified in the analysis of previous results, along with the modifications implemented in the model since the results reported in D5.2 Investigative case studies results, means that these calibration activities needed to be performed again.

This has been done considering the baseline scenario which represents the current (2014) operations and considering the analysis from different data sources to use them as calibration targets.

The calibration activities and results are gathered in Section 4 of this deliverable.

1.4 Domino metrics

As presented previously, the ABM is able to generate very low level indicators. These need to be combined in a meaningful manner to gain understanding on the impact of introducing/modifying mechanisms in the ATM system.

A review of classical (flight and passenger) metrics along with the definition of advanced network metrics which focus on centrality and causality were reported in D5.1 Metrics and analysis approach. These metrics were tested on the investigative case studies (D5.2 Investigative case studies results) and presented to stakeholders on different consultation activities (see D6.3 Workshop results summary).

The description of the metrics used for the analysis of the results of this deliverable are detailed in Section 5.1. Finally, from the presentation of the preliminary results to stakeholders and experts, feedback was gathered that there was a need to relate the new metrics presented in Domino to operational indicators, this has been addressed in Section 5.1.2.

1.5 Feedback into this deliverable

In order to produce this deliverable, some modifications to the model, the mechanism and how the metrics are used have been performed. These come from feedback obtained from different consultation and dissemination activities carried out by Domino. In particular the participation to an airspace users' workshop organised by EUROCONTROL and a dedicated workshop on Domino with stakeholders and experts organised at the SJU premises.

The feedback obtained was consolidated in D6.3 Workshop results summary, and primary focuses on the applicability of the new network metrics, the capabilities of the agent-based model and the behaviour of some agents (e.g. how the effect of curfews are considered by airspace users (AUs) from earlier in the day).

1.6 Structure and contents of this deliverable

This document is structured as follows: Section 2 presents the scenarios that have been modelled with a description of the mechanism implemented. In Section 3 a summary of the ABM developed in Domino is presented, focusing on the changes carried out with respect to the model described in [1].

The calibration of the model is presented in Section 4 followed by the results on the case studies in Section 5. This section is formed by three parts:

- description of the metrics computed in 5.1,
- summary of key results in 5.2, and
- detailed analysis of the results for the two case studies in 5.3.

Details about the methods to test the presence of a causality relationship between two state variables are presented in Annex I (Section 9), and detail values on the scenarios results are provided in Annex II (Section 10).

The document closes with conclusions gathered in Section 6, as well as some considerations on next steps, focusing on future lines of research. References and acronyms can be found in Sections 6.1 and 8 respectively.

2 Scenarios and mechanisms modelled

2.1 Scenarios modelled

Scenarios simulated with the final version of the model are more focused on a few specific aspects of the air transportation system, as opposed to the results presented in D5.2. They are grouped in two *case studies*, as explained in D3.3.

2.1.1 Hub delay management

In this case, we are interested in disruptions at hubs in order to understand under which conditions disruptions can be alleviated by mechanisms.

We started by choosing three important hubs in Europe: Amsterdam (EHAM), Heathrow (EGLL), and Paris Charles de Gaulle (LFPG). For these airports, we considered that some major disruptions have hit them at the same time. The disruptions were created by manually defining an ATFM regulation at each airport with the following characteristics:

- Starting and finishing in the first part of the day: 06:00 - 14:00 local time.
- Reducing the capacity at the airport to half their nominal capacity: EHAM with 45 arrivals/hour, EGLL with 54 arrivals/hour and LFPG with 44 arrivals/hour.

In addition to these manually set ATFM regulations, the rest of the delay in the system is set as default (i.e., nominal day of operations) and ATFM regulations are defined at other airports based on a randomly selected nominal day.

To mitigate the effects of disruptions, two mechanisms are modelled:

- Flight Prioritisation processes (FP): based on ATFM slot swapping (inspired in UDPP concepts) and allowing inter-airline swaps.
- 4D Trajectory Adjustments (4DTA): consisting on dynamic cost indexing (allowing us to recover delay or adjust the speed to save fuel) and coupled with actively waiting-for-passenger when it is considered to be the alternative which will provide the lowest operational costs for the airline.

In total three scenarios are modelled:

- **Hub delay management baseline**: Disruptions at the three hubs without any mitigation action to be used as baseline. All mechanisms are implemented at Level 0.
- **Hub delay management FP Level 2**: Disruptions at the three hubs with FP at level 2 (flight swapping between airlines allowed).

- **Hub delay management 4DTA Level 2**: Disruptions at the three hubs with 4DTA at level 2 (full DCI and waiting for passengers with conjoint decision).

Note that even if the manually set disruptions are only defined for the three hubs, the mechanisms are implemented everywhere, and the whole network is simulated in each case.

2.1.2 Effect of E-AMAN scope on arrival manager

The second case study focuses on the implementation of the Flight Arrival Coordination (FAC) mechanisms. In this case, we are interested on analysing the impact of extending the E-AMAN horizon. A larger horizon means that flights are considered in the landing sequence earlier providing, on one hand potentially larger optimisation savings, but on the other hand, affecting flexibility and having higher uncertainty (the landing sequence is modified more times, as each time a flight enter the E-AMAN scope a new landing sequence might be produced).

In this case, to see the effect of modifying this horizon, we consider the *stressed* baseline, so that larger delays can be mitigated through the mechanism.

As a result, we simulated these four scenarios:

- **E-AMAN scope on arrival baseline:** ‘Stressed’ baseline scenario, with all mechanisms at level 0 and high delay across the system.
- **E-AMAN scope on arrival FAC Level 0 extended range:** ‘Stressed’ scenario with FAC at level 0 and 600 NM as planning horizon.
- **E-AMAN scope on arrival FAC Level 2 nominal range:** ‘Stressed’ scenario with FAC at level 2 (full cost minimisation) and 200 NM as planning horizon.
- **E-AMAN scope on arrival FAC Level 2 extended range:** ‘Stressed’ scenario with FAC at level 2 again, but 600 NM as planning horizon.

E-AMAN at level 2 considers the information from the flight (which has been relied by the AOC) on which is the expected total cost of each available landing slot. This considers the cost of fuel, but also other costs such as passenger costs, cost of delay, potential curfew at the end of the day, etc. See Section 2.2.3 for more details. Note that the change of horizon has been applied only at airports which have an E-AMAN.

Finally, in this deliverable we will focus on the scenario where the mechanism is modelled at Level 2. The extended range at Level 0 is only computed as another reference to better understand the effect of extending this horizon.

2.2 Mechanisms modelled

In this section we summarise the three mechanisms that are modelled in Domino focusing on their implementation for this deliverable. Their implementation in baseline (Level 0) and advance (Level 2) are described as these are the ones analysed. For more information on the mechanisms the reader is referred to [1] and to [2].

2.2.1 4D Trajectory Adjustments (4DTA)

4D Trajectory Adjustments is a mechanism which deals with modifications of the trajectory of flights just before push-back and once airborne in order to deal with delay. This mechanism is formed of two sub-mechanisms:

- Waiting for passengers: actively delaying outbound flights to wait for delayed connecting passengers.
- Dynamic cost indexing: modifying the speed of the flight to speed up and recover delay or, in some cases, even slow down to save some fuel.

2.2.2 Baseline - Level 0

At Level 0, we use rules of thumb that serve as an approximation of the current practices in the airline industry for the tactical management of flight delay and waiting for passengers at the hub. Two sub-mechanisms are considered:

- Determining the cost index of a flight (before take-off), i.e., an increment or not on the cruising speed.
- Deciding whether a flight waits for delayed connecting passenger.

The specific parameters of these mechanisms have been calibrated according to the feedback received from a number of experts in the industry.

Cost index is calculated before the take-off (i.e., at push back) and it is fixed throughout the flight. To decide on its speed, the flight uses the information about its departure delay. At 'pushback_ready', the departure delay is assessed by comparing estimated off-block time (EOBT) with scheduled off-block time (SOBT):

departure_delay = EOBT - SOBT

The attempted delay recovery is then performed according to the probability distribution shown in Figure 1. If the estimated departure delay is smaller than 15 minutes, the flight does not try to recover it. If it is larger than 60 minutes, the flight will try to recover as much delay as possible (up to 5 minutes) by selecting a higher speed than the planned one. Lastly, the decision on recovering any delay between 15 and 60 minutes is made stochastically according to the linear probabilistic distribution on Figure 1.

Additionally, the maximum delay is considered to be recovered is limited by the amount of extra fuel that would be required for this recovery and it is capped at 70% of the total amount of additional fuel available. Moreover, in order to make the application of this rule more aligned with the current practices, the flight never speeds up to the maximum possible speed; rather, the speeding up is capped at 90% of the maximum velocity. Finally, if after applying all of the so far mentioned constraints, the amount of delay that can be recovered is lower than 5 minutes, no recovery is performed. That decision was made upon the consultation with the experts, due to the fact that the recovery of the delays lower than 5 minutes is seldom performed.

Note that a change on cruise speed will imply also a change on the TOD, generally increasing the cruise and reducing the slower descend.

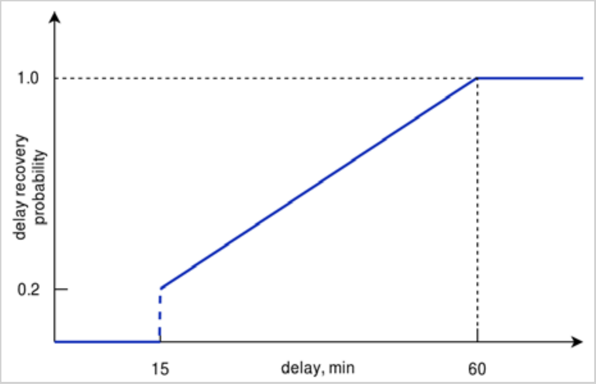


Figure 1: Probabilistic distribution for deciding on delay recovery depending on the estimated departure delay.

Wait for passengers is performed 5 minutes before the 'pushback_ready' event, and it is triggered by a special event called 'pax_check_event'. At that moment, we run a check that inspects which passengers are not at the gate ready for boarding, and we estimate how much time they need to make it to the gate. For this estimate, we use the information on their current position: If they are/were arriving on a connecting flight, the **in-block time** of their previous flight is used (real or estimated, depending on the status of the flight). In addition, the average **minimum connecting time** is taken into account for the calculation of their estimated at-gate time. The average minimum connecting time has been pre-calculated for each airport and it depends on the type of connection the passenger is making: domestic - domestic, domestic - international, etc.

Finally, the flight decides to wait for any passenger with a flexible ticket who's at-gate time is estimated to be at most 15 minutes later with respect to the flight's expected push back time.

Other approached used in previous projects include wait only if the load factor is lower than 80% of the planned one and at most 10 minutes.

2.2.3 Level 2

Level 2 couples the assessment of cost index and wait-for-passenger decision process via a unified cost function. That way, the optimisation is improved by relying on the sum of all the estimated costs, including the possibility to recover a part of delay by speeding up and spending more fuel than planned. There are two types of cost that a flight takes into account in the cost optimisation process:

- **fuel cost**: the cost of the extra fuel that would be needed in order to recover a delay (fully or partially);
- **time cost**: the cost of unrecovered delay, which includes non-passenger costs (maintenance, crew and curfew costs), as well as passenger costs (compensation, soft costs and the costs due to the effects of reactionary delays).

The delay recovery is performed with a time resolution of 1 minute (i.e., the flight can only decide to recover a rounded number of minutes). Unlike at Level 0, there is no limitation on the maximum velocity that the flight can choose in order to recover delay, as the decision is purely driven by the cost and the objective to find the optimal solution given the estimated costs.

Additionally, to make a decision on waiting for passenger, the component of cost due to waiting/not waiting for passenger is described and added to the overall cost function via the following two costs:

- **waiting cost**: the cost of waiting a passenger group p_i for n minutes with respect to EOBT. This wait would essentially delay the EOBT for n minutes, and thus this is the cost of delaying the flight (EOBT) for n minutes.
- **not-waiting cost**: the cost of not waiting a group of passengers and having to take care of them. This cost includes different types of care that the airline needs to provide to the passengers that missed their connecting flight for no fault of their own: duty of care, compensation cost and transfer cost. Transfer cost is calculated by searching for alternative itineraries for stranded passengers, estimating the cost of each itinerary and choosing the least expensive one. Additionally, this also includes soft costs - the cost airline will suffer due to a potential future loss of passengers or reputation.

There are two times during a flight when delay is assessed and 4DTA mechanisms potentially applied: at ‘pax check event’ (5 minutes before ‘pushback_ready’) and the top of climb.

At **5 minutes before a flight is ready for push back**, a joined assessment of departure delay (and its potential recovery) and wait for passenger options is performed. Similarly, as in Level 1, we assess current estimated departure delay and consider recovery options by speeding up (changing cost index before departure) through assessing the cost of those options. At the same time, the check for missing passenger is performed, and waiting costs and not-waiting costs are calculated for each passenger group. All those estimated costs are added and observed on the domain of recoverable delay (from 0 to the maximum number of minutes a flight can recover by speeding up and using the extra fuel available). The decision that minimises the total cost is taken, and according to it cost index might be changed (speeding up) and a number of passenger groups waited for. I.e., a decision to wait and recover delay is performed before push back.

Example. *Let’s assume that the currently estimated departure delay of the flight is 20 minutes, and there are 2 passenger groups estimated to be late for boarding with delays of 10 and 15 minutes (w.r.t. updated push back time, i.e., waiting 10 minutes will allow the first group of passengers arrive to the flight, waiting 15 minutes will ensure that both group of passengers can board the plane). In this case, all the costs are assessed on the domain of possible delays ranging from 0 (recovering all of the delay, which would require significant increase in velocity and thus spending a large amount of additional fuel, if all 20 minuets can be recovered by speeding up) to 35 minutes (meaning the flight is waiting for both passenger groups and deciding not to recover any delay).*

At the **top of climb**, the assessment of the expected arrival delay and potential speeding up is done. The delay recovery decision is made by observing the total cost, i.e., the sum of the fuel and time cost (naturally, no passenger costs apply here). The flight chooses the recovery time (in minutes) that expects to minimise the total cost (on the domain ranging from 0 minutes to total estimated arrival delay). In addition, a flight has the possibility to slow down at the top of climb if the (currently) expected arrival time (**EIBT**) is at least 30 minutes before the scheduled arrival time (**SIBT)**. In this case, the flight decides to slow down by adding x minutes to the flight, where $x = SIBT - EIBT - 30$. This has been done in order to save fuel and prevent potential holding times which can likely occur in case of such very early arrivals.

2.3 Flight Prioritisation (FP)

Flight Prioritisation mechanism deals with the potential swap of ATFM slots by airlines at regulations defined at arrival airports.

2.3.1 Baseline - Level 0

At level 0, no swap is performed by the airlines. The delay that is assigned due to ATFM regulations at arrival is performed by the flight who receives it.

2.3.2 Level 2

At Level 2, swaps between flights can be performed among the flights of a given airline, or between flights of different airlines. Every time a departure flight plan is submitted (whether it is the first flight plan for this flight or not), the airline estimates if a swap can be done with this flight if ATFM delay has been assigned. The conditions considered are the following:

- both flights must be in the same regulation at their arrival airport;
- the estimated cost of the swap must be negative (i.e., the swap has a positive impact overall).

The cost of the swap is estimated using the following rule: if COBT 1 is the controlled off-block time of the first flight, COBT 2 the time of the second flight, and cost1 and cost2 the delay cost function of the first and second flight, then we compute:

$$cost_1(COBT_2) + cost_2(COBT_1) - (cost_1(COBT_1) + cost_2(COBT_2))$$

i.e., the total cost of delay if the COBT are swapped minus the cost of delay if they are not. The cost function used for this mechanism is different from the one used for 4DTA at TOC, since the estimation of the cost happens on the ground, before the departure of the flight. More specifically, the cost function is evaluated using as delay the controlled off-block time minus the scheduled one, and includes the following components:

- non-pax cost (maintenance and crew);
- passenger soft cost;
- duty of care;
- passenger compensation.

Moreover, the delays for passengers are computed using the updated information on their next flights and worst case scenarios. In particular, when a passenger is expected to miss their next flight and when a flight may miss the curfew. Some network effects are also taken into consideration using, again, the worst case scenario. We gather information on the next flights using the same aircraft and consider that all of these flights will have the same cost than the current one.

Note that in theory we could estimate exactly which flight will be impacted by the propagation of delay, in terms of aircraft and/or connecting passengers, since all information is known to the airlines (or can be requested from another airline). However, this information takes too long to compute, especially for passengers, which increases by a large factor the time of the simulation. As a consequence, in this deliverable the above heuristics is used for the decision-making process of swapping the slots.

Flights can be swapped among different airlines. The mechanism works otherwise exactly the same. The airline starts by checking all flights in the same regulation at the arrival airports, and then requests some information from another airline, if needed, in order to compute the total cost of the swap. The same cost function is applied, and the swap is performed if the cost is negative.

It is clear that in reality, different airlines will never share their true cost, first because it is sensitive information, and more importantly because they have no incentive not to inflate their own reported cost. In the model, we consider that there is a market mechanism (e.g., credit system, auction) in place which allows us to have an efficient market for swaps and thus do a swap only if it is beneficial in average (for instance by giving back some money to the airline delaying its flight). This market mechanism might or might not be feasible in reality, which is another research question. Thus, the model should be considered as a best case in this regard.

2.4 Flight Arrival Coordination (FAC)

The Flight Arrival Coordination mechanism focuses on the sequencing done by the extended arrival manager at airports. This mechanism is only implemented in airports which have or are expected to operate an E-AMAN system (24 airports as according to the SESAR Pilot Common Project [3]: EBBR, EDDB, EDDF, EDDL, EDDM, EGCC, EGKK, EGLL, EGSS, EHAM, EIDW, EKCH, ENGM, ESSA, LEBL, LEMD, LEPA, LFMN, LFPG, LFPO, LIMC, LIRF, LOWW and LSZH.

For the above mentioned airports there are two moments when flights are issued delay when approaching them (see Figure 2):

1. When the flight enters the planning horizon of the E-AMAN (with a default value of 200 NM from the airport).
2. When the flight enters the tactical or execution horizon of the E-AMAN (defined at 120 NM from the airport).

The distances selected for the planning and execution horizon are in accordance with the expected extension of the arrival managers from 100-120 NM to 180-200 NM [3]. As presented in the case studies (see Section 2.1.2) we will explore the effect of increasing the planning horizon from 200 NM to 600 NM.

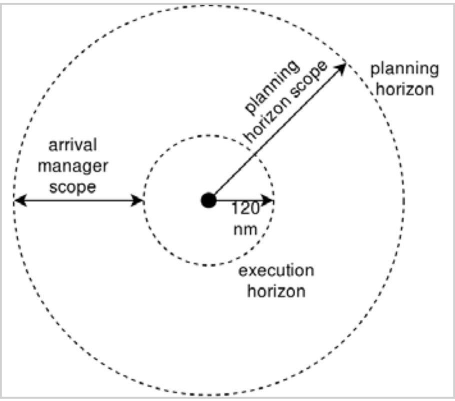


Figure 2: Flight Arrival Coordination horizons

When a flight enters the planning horizon, all the flights which are located in the scope of the arrival manager, i.e., between the planning and the execution horizon, are re- optimised, i.e., assigned to the slots which are either planned or available, considering a given optimisation function which depends on the Level of the mechanism. At the planning horizon, the flight which triggers this optimisation, i.e., the one which enters the arrival manager, receives the amount of delay that it is expected to experience and tries to absorb as much delay as possible by slowing down (saving some fuel). At the tactical horizon, the flight which exits the E-AMAN will be issued with a slot (assigned as the output of another re-optimisation) and the required delay (if any) will be performed as holding.

The optimisation of all the flights within the E-AMAN every time a flight enters or exits the system ensures that the best sequence is maintained within the arrival manager with respect to the optimisation function, and that the flight can slow down to absorb part of the delay saving some fuel if delay is expected. However, as the amount of delay that can be absorbed is very limited, only the flight which enters the arrival manager considers this speed and TOD variation. Note that only available or planned slots are considered in the optimisation and once a landing slot has been assigned to a flight which exits the execution horizon it is fixed. Note that in some cases, the slot which has been planned at the planning horizon for a given flight might not be available anymore when it reaches the execution horizon. Finally, at both horizons the arrival capacity at the airport is considered to ensure that the arrival sequence respect the airport throughput.

For the airports which are not listed above, a simple arrival manager located at 100 NM from the airport is considered, and a first-in first-out approach modelled. The assigned delay will hence be done as holding. This ensures that the arrival capacity at the airport is not exceeded.

2.4.1 Baseline - Level 0

In Level 0, current principles applied on E-AMAN systems are considered: The Flight Arrival Coordination tries to minimise the amount of holding delay that will be carried out at the TMA by minimising the total holding delay. The FAC is focused on the maximisation of the arrival throughput at the runway. No information from the airlines is taken into account when applying this mechanism. When a flight enters the planning horizon, the first slot available in the sequence from the flight estimated landing time is assigned. In a similar manner, once the flight enters the execution horizon, the first available slot is assigned, and the holding delay computed.

2.4.2 Level 2

In Level 2, the arrival manger tires to minimise the expected total cost for each flight. This includes the:

- cost of fuel: considering potential fuel savings by slowing down but also fuel cost by performing estimated holding (estimated by the flight), and
- cost of delay: considering passenger and non-passenger related costs of delays, estimated and provided by the AOC.

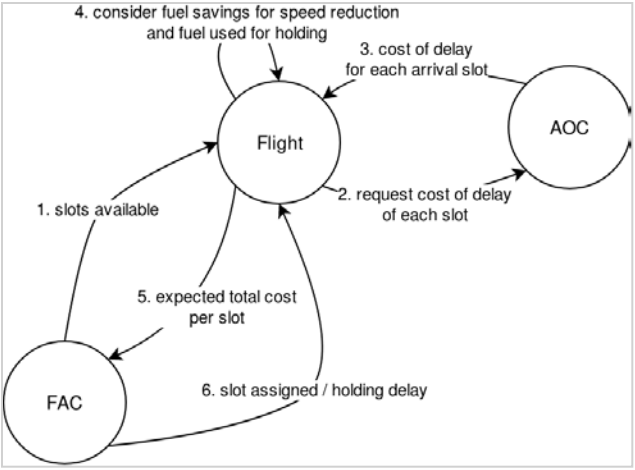


Figure 3: FAC Level 2 messages

The FAC considers the expected cost of each slot for each flight as this information is provided to the FAC by the flight. As depicted in Figure 3:

1. when the flight enters the E-AMAN, the FAC send a list of slots available to the Flight;
2. then the flight requests the expected cost of delay to its AOC;
3. the total expected cost of using each slot (in the current implementation the total cost of delay function) is returned to the flight;
4. which with information on the aircraft performance and flight details (e.g., current weight) estimates the fuel required for each slot (savings by slowing down and holding);
5. the total expected cost (including cost of delay and cost of fuel) is sent to the FAC to be used during the optimisation process to assign the slots.

When the flight enters the planning horizon, once the FAC has received this expected cost per slot, the arrival sequence, considering all the flights in the E-AMAN scope are optimised. When the flight reaches the execution horizon, the same optimisation is performed with the already provided expected cost of slots functions.

3 Domino agent-based model

This section presents the agent-based model on which the results are based. Since the design for the model has been presented in detail in D4.1, we do not reproduce all the description of the model, but rather highlight the main differences between the final implementation and the original design.

3.1 Agents

3.1.1 Flight

The flight is basic agent which is tasked with executing its planed trajectory, from push-back to arrival. The main features of the flights are the following:

- Taxi-out phase, based on taxi distribution at the airport.
- Iterative climb/cruise/descent phases, going through all the way points defined in the trajectory. A stochastic variation on the segments lengths are applied to each segment independently.
- If the flight decides to recover delay by applying DCI, the length of the rest of the flight (cruise and descend phase) are modified to consider the impact of DCI on the location of the TOD. See Section 3.1.2 for more details on the cost functions used.
- Taxi-in phase is based on distributions on stochastic arrival taxi time from the airports.
- Once the flight arrives to the gate two processes are triggered:
    - the aircraft is transferred to the ground airport for the computation of the turnaround processes
    - the passengers are hand-over to the airline for their arrival/onward connection processes.

As described in Section 3.1.2, flights can be cancelled either stochastically or based on an internal decision from the airline, because of curfews at the airports.

3.1.2 Airline operating centre

The airline is the most complex agent in the system, as it manages its own flights and passengers. Therefore, it decides between alternatives (e.g. flight plan, slow swapping, waiting-for-pax options) based on delayed rewards. Here is a list of actions potentially considered by the agent:

- Submit flight plans, deciding which one to select (there is one option per potential route between origin-destination) decisions on best one.
- Reassess flight delay before departure, potentially triggering another flight plan submission.

- Process arriving passengers:
    - Check if their destination was the final one or they are connecting. If final destination, computing delays and potential compensations.
    - Request connecting times for connecting passengers and based on this information:
        - check if they could make the connection to the next flight,
        - otherwise, consider the handling of these passengers by rebooking them, considering:
            - its own flights,
            - flights from the alliance the airline is from (if any),
            - flights outside of alliances (only for premium passengers)
            - if not possible consider associated cost for care of passengers (e.g., hotel).
    - Trigger passenger driven costs (e.g., Regulation 261 compensation, duty-of-care):
        - Based on types and magnitudes of delays,
        - Assigning cost to individual flights to trace back their costs to their flights.
- Decide if wait for connecting passenger by actively delaying outbound flights, potentially triggering another flight plan submission.
- Decide if speed-up/slow down flights to recover delay/fuel.
- Make decision on cancellation of flight if flight is expected to miss a curfew.
- Manage the cancellation of flights (based on exogenous probability), which will trigger the rebooking of passengers.
- Make decision on swapping flights inside and outside its own fleet, depending on the mechanism, based on expected cost in case of ATFM regulation at arrival airport.
- Airlines also incorporate the estimation of cost of delay at different stages, see Section 3.1.2.

3.1.3 Ground airport

The ground airport is mainly tasked with estimating the time required for operations carried out at the airport. There is a Ground airport agent per airport allowing us to consider the particularities of each airport type to produce these times. In particular, it provides:

- Averages for turnaround times and connecting times to airlines, so that they can consider these times when estimating arrival delays for their cost functions.
- Actual turnaround (based on airline and airport type).
- Actual taxi times.
- Actual passenger connecting times.

As part of the turnaround process, the Ground airport holds the aircraft (internally a resource that can be used only by one agent at a time, see Section 3.2.2), until the end of the process.

3.1.4 Network Manager

The network manager is tasked with accepting or rejecting flight plan, issuing ATFM regulations applying them to flights. Its main decisions are the following:

- Check if a received flight plan infringes a curfew. Such flight plans are rejected, because, as described in Section 3.2.3, all curfews are considered *hard* in Domino, i.e., no flight can plan to arrive after the curfew.
- Computing ATFM delay for a given flight plan:
    - Stochastically for regulations issued in the airspace. A probability (depending on the scenario) is used to draw if a given flight plan is affected by a regulation. Then, another distribution is used to draw how much delay is assigned to that flight. These probabilities and distributions are estimated based on historically analysed datasets (AIRAC 1313-1413, 1702 and 1709). Note that the stochastic ATFM regulations are also divided between regulations due to weather and regulations for other causes. This is required in order to estimate if downstream delay is potentially entitled to generate compensation for passengers as described in Regulation 261.
    - Explicitly for airports. These regulations are decided at the initialisation of the simulations (but agents (and in particular AOC) do not have this information), either from a randomly sampled day, from a fix historical one or manually defined. In this case, an explicit queue is created and managed by the network manager for the arrival airport. This queue models individual slots as defined by the regulation declared capacity. More details on regulations can be found in Section 3.2.2.
- Disseminate accepted flight plans to interested parties, in practice E-AMANs and airlines.
- Swap flights at explicit ATFM regulations when FP mechanism is used and requested by airlines.

3.1.5 E-AMAN

The role of the E-AMAN/AMAN agent is to manage the arrival sequence of flights at airports. An airport might implement an extended arrival manager (E-AMAN) or just an arrival manager (AMAN). In case of having an AMAN system (the default option), once flights enter in the horizon of the system, they are tactically sequenced in a first come first served approach. The goal is to use all arrival slots and assign required holding delay to arrival flights.

When an E-AMAN is implemented, two radii are defined, the first one is the planning horizon, and the second one is the tactical horizon. When flights enter in the planning horizon, all the flights within the E-AMAN scope (between the planning and the tactical horizon) are considered to optimise the landing sequence. Then, once flights arrive to the tactical horizon, their final slot is assigned and required delay will be performed as holding. The goal is to optimise the arrival sequence to minimise a given objective and also, allowing flights to slow down within the E-AMAN to absorb required arrival delay with the objective of reducing fuel consumption and required total holding delay.

The objective function considered when optimising the arrival sequence are:

- For Level 0: the total arrival delay. Only arrival delay at the airport is considered, providing an optimisation close to first come first served and closer to current E-AMAN objectives.
- For Level 1: the total expected delay at the airport. This considers the arrival delay of flights but also their potential reactionary delay. The objective is then to minimise the total delay including the departure delay at the airport.
- For Level 2: the E-AMAN does not focus on delay but on cost. The agent request to the flight the expected cost for using each of the available slots (this information will be provided to the flight by the AOC). Then these cost functions are used to minimise the total expected delay. When considering the expected cost different factors are taken into account: expected holding fuel required, expected cost of delay (considering both, arrival and departure (reactionary) delay), expected knock-on effect with potential missing curfew at the end of the day, saving expected by reducing speed to get a later slot).

As each time a flight enters the planning horizon the sequence might change, some sub-optimality might occur.

Finally note that only airports which according to the PCP are due, or already have, an E-AMAN implemented will have this system in the model [3].

3.1.6 DMAN

The departure manager is implemented at all the airports in the model. In this case, it provides slots to flights following a first come first served. The main goal of the DMAN is to ensure that departure runway capacity is maintained. It explicitly models a queue of slots for the departure sequence.

Flights request a slot when ready at gate, the firs slot available is assigned and the required delay computed.

3.1.7 Radar

There is a central radar in Domino, which is considered as an agent. It does not perform any decision but relies information about the position of the flights to interested stakeholders. For each trajectory, different waypoints are identified, when the flight reaches these, they are captured by the Radar agent which notifies the relevant agents.

The Radar agent is in charge of:

- Disseminating flight plans and flight plan changes to the different agents so that they can be informed on these changes (e.g., DMAN, E-AMAN).
- Augment flight plans, by adding intermediate waypoints and providing events to simulate signal broadcasting when flights reach those points defined by other users (e.g., reaching the planning horizon for E-AMAN).

3.1.8 Flight swapper

The flight swapper (Flight Prioritisation Agent in D4.1) is used when FP level 2 is implemented. In this case, flights from different airlines can swap their arrival slots due to ATFM regulations. In order to do that, an airline has to estimate the expected gain from the swap. As a consequence, it needs information from the other airlines. As explained in D5.2, we consider that a market mechanism is in place which allows the airlines to trade slots using their own intrinsic value. In other words, it can be considered that airlines reveal their true cost through the mechanism. This is why in model airlines can provide their cost functions to another airline.

However, exchanging messages with cost functions build around different estimation points is time consuming from a computational point of view. As a result, we decided to create the flight swapper which has access to the necessary internal information from all airlines to compute cost functions required to assess the swap possibilities. Direct memory accesses, instead of explicit messages between the agents, are faster. This comes at the price of breaking the full agent paradigm of communication via messages. Note that this, however, does not change the model principles, i.e. the underlying mathematical representation, it is only an implementation choice done for performance issues. The agent can:

- Compute cost function from any airline.
- Compute valid flight swaps.
- Estimate the total value of any given flight swap.

3.2 Other internal modelling details

3.2.1 Cost functions

Cost functions are the cornerstones of any agent-based models. It is the algorithmic piece which allows agents to make their decisions.

The main cost functions are computed by the airline operator centre. Different versions of them have been implemented, depending on the level detail needed and on the information available when making the decisions. In practice, the cost function is a cost of delay, which is the variable. ‘Delay’ means different things for different roles at different point in time. For instance:

- DCI at TOC is interested in the arrival delay. In this case, the delay accrued before reaching the TOC is already fixed (and some costs are already realised).
- In the case of flight swapping, departure delay is more relevant, as the flight is still on-ground and is based on the ATFM regulation slot.

In general, note that the consideration or not of buffers is relevant and that different amount of delay can be considered if departure delay or estimated arrival delay with respect to schedule are considered.

In addition, different costs should be considered for different functions. For instance:

- Duty of care should not be considered in DCI at top of climb, because duty of care happens before departure.
- Cost of fuel should not be considered for flight swapping, since the fuel used in different slots should be approximately be the same.
- In the case of the E-AMAN, some delay might even represent cost savings as a later slot might allow some fuel savings by reducing the speed of the flight.

The cost of delay function can also use various level of update information when estimating the delay and cost. For instance, one can use the most up to date EOBT or the SOBT. Different updated information will be used at different times.

Finally, there is a compromise between the accuracy of the function and computational load. Some functions have to be built (and thus information accessed) quite frequently (e.g., to estimate the cost of each possible slot at an arrival manager). For instance, one could compute exactly the estimated cost of all passengers travelling on all flights using the same aircraft for the duration of the day. However, in practice doing this is very long. In addition, airlines do compute costs with this level of detail on their practice. In Domino, we partly used some heuristics for some functions. In the last version of the model, consider two cost functions:

- A heuristic ground version, used for instance for FP: uses OBT, takes into account duty of care, compensation, soft cost, curfew cost, non-passenger cost (maintenance, crew, at gate).
- A heuristic air version, used for instance for DCI and FAC: uses IBT, takes in account compensation, soft cost, curfew cost, non-passenger cost (maintenance, crew, on cruise), and fuel.

Passenger delays (and their associated costs) are accounted explicitly by considering if passengers (from this flight) can make their connections. Both use a very conservative heuristic knock-on factor by considering that all flights after this one and using the same aircraft with incur the same cost.

Figure 4 presents two cost of delay functions for two flights. In the first one, there are two points where a miss connection of passengers happens leading to higher expected costs. In the second one, there is one point where delay will propagate leading to the potential miss of a curfew (see Section 3.2.3 for more details on curfew buffers estimation).

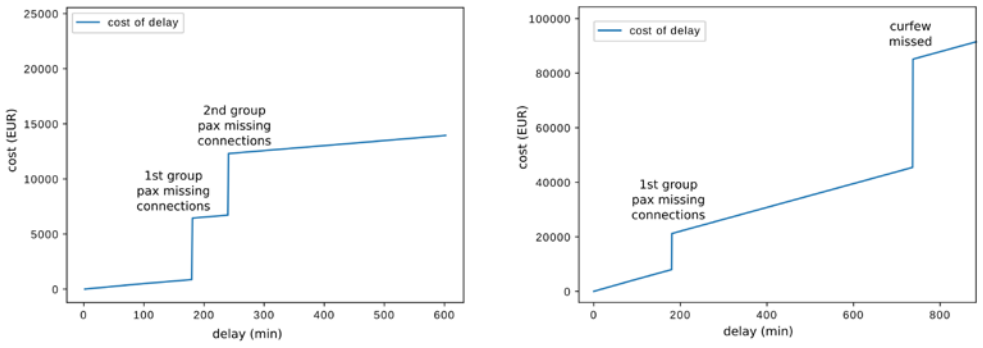


Figure 4: Examples of cost of delay functions as estimated by the E-AMAN

3.2.2 Resources

Resources in agent-based models are common. Typically, an agent needs something to make an action, and needs a resource for that. However, they are really important in concurrent implementation of agent-based models, where agents are taking actions **in parallel**, potentially at the same (internal) time. In this case, resource management becomes critical, as it avoids agents to perform actions that they should not as they don't have access to the resource which is used by another agent.

Concurrent issues can be further divided into two categories:

- Out-of-date information. Typically, an agent makes a decision based on parameters that are changing in the same time stamp due to the action of another agent. In most cases, it doesn't cause an unstable state of the system, in the sense that the first agent can be considered taking an obsolete decision, useless but unharmful. As an example, a flight could be considering waiting for passengers due to delays from an incoming flight. It can be that this flight is even more delayed (and knows it at the same time than the other flight decides to wait) and thus the second flight will wait when it shouldn't.
- Concurrent resource access race. In this case, an agent modifies its internal state based on the accessibility of some resources, which is also being accessed and/or modified by another agent. As an example, a flight could request for potential slots in a regulation as part of the slot swapping process. It then decides to use slot #1, while another flight making the same decision has already taken it (between the physical time of the access and the physical time of the booking). This kind of issues usually gives rise to serious inconsistencies in the code, and/or programming errors.

Below are explained how some concurrent issues were solved by exclusive access to dedicated resources implementing a traffic light system. The resources in the model can be *requested* by agent, i.e., when they are *released* by the current agent, the next in line in the request queue can automatically use it. Technically, it is considered as an event by the simulation engine.

3.2.2.1 Aircraft

The aircraft are considered as resources in the model, and only one agent can use the at the same time. In practice, all a flight agent requests the resource, then releases it at arrival to the ground airport, who holds it for a certain time (the turnaround time) before it is released, to be use by the next flight. In this way, reactionary delay is explicitly modelled in the system as the aircraft is not released for the next flight until the completion of the turnaround.

3.2.2.2 ATFM regulation resources

Explicit ATFM regulations are complicated to manage in a concurrent way, because some physical time elapses between the point where the airline asks for the possible slots and the point where it actually chooses a slot. In the meantime, another flight can take the slot. This is due to the fact that a slot is pre-assigned when an airline submits a flight plan but the flight plan is not confirmed at that point, as the airline might request a re-routing instead.

To avoid this effect, we build a 'booker' for each ATFM regulation queue. The booker is a resource, with only one agent able to use it at the same time. Moreover, all changes to the regulation slots (e.g., a slot assignment) have to be done by the booker, and no one else. As a consequence, only one flight at a time can access and modify the regulation slots. Note that two identical simulations can end up with difference results nevertheless, since the concurrent race for requests can happen, and will be solve internally, most likely randomly.

3.2.3 Curfew

An important addition to the model with respect to D5.2 is the presence of curfews for some airports. The details of how these curfews have been set, and for which airports, can be found in [4]. In the model, curfews have two impacts:

- The network manager agent rejects flight plans for which the flight would arrive later than the threshold;
- The airline operating centre takes this into account when computing expected costs, by adding a curfew cost to a flight which may potentially infringe the curfew.

This translates into having slightly more cancelled flights than in the baseline, but more importantly airlines are more sensitive to delays on their flights if these delays may cause the last flight of the day to infringe any curfew.

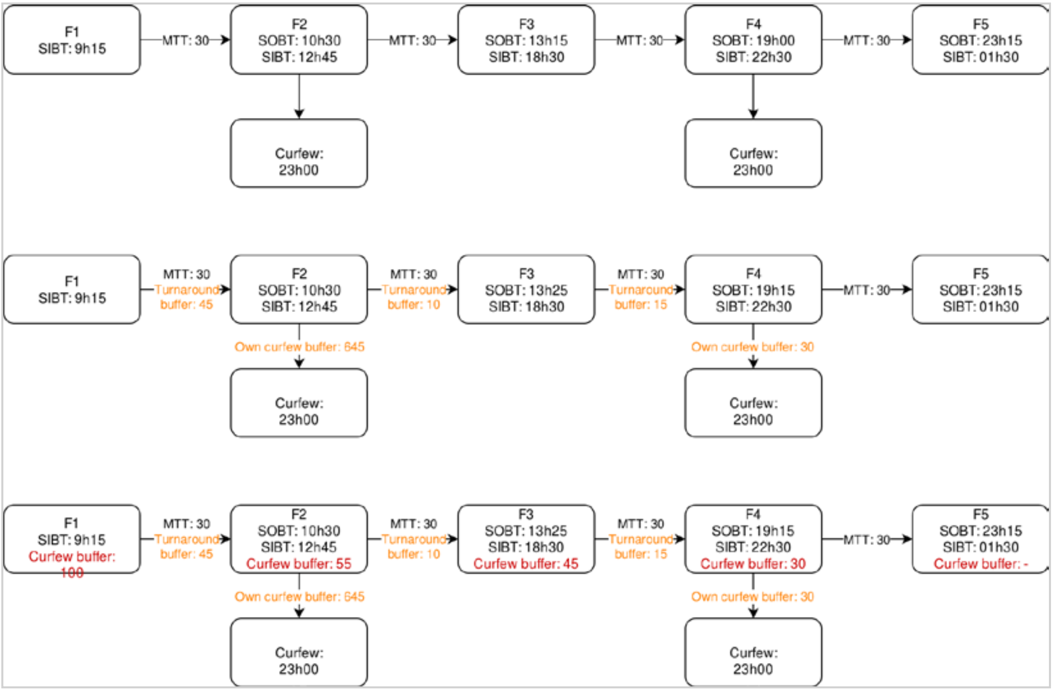


Figure 5: Example of estimation of curfew buffers

Figure 5 presents how the curfew buffers have been estimated. In the first image, five consecutive flights of the same aircraft are presented. As depicted, flight 2 and flight 4 have as destination an airport with a curfew at 23h00. Between each rotation there is an estimated minimum turnaround time required (MTT). With this information, it is possible to compute, as presented in the second image, the buffers to propagate delay between the flights (turnaround buffer) and to breach the different curfews (e.g., flight 2 should be delayed by more than 645 minutes for the flight to breach the curfew, as its SIBT is at 12h45 and the curfew at destination is at 23h00). Then, finally, with this information, it is possible to compute the curfew buffer for the different flights considering from which moment it is possible that the flights would propagate enough delay to potentially breach a curfew. For example, flight 3 has a curfew buffer of 45 minutes as if the flight is delayed over 45 minutes, it will propagate over 30 minutes to flight 4 which will then breach a curfew. In this example, even if flight 2 has 645 minutes of delay until it breaches the curfew, its curfew buffer is only 55 minutes, as after that it could potentially propagate enough delay for flight 4 to breach its curfew.

In the model, these buffers are computed using the most up-to-date available information, such as the updated EIBT of downstream flights.

3.2.4 Passengers

Passengers are not considered agents in the model, since they do not make decisions during their trips. A passenger class exists in the code, as a place holder for all passengers of the same type following the same itineraries. These groups can be split in case of rebooking. The class is also used to compute the soft cost incurred by airlines, corresponding to image deterioration due to delays.

4 Model calibration and validation

Most of the model calibration process has been explained in D5.2 Investigative case studies results. In this deliverable the same principles apply, we only highlight the differences with respect to the previously reported calibration.

4.1 Data used for calibration

Table 1 presents some of the key processes that are modelled in Domino, and how their distributions have been adjusted as reported in D5.2.

Table 1: Processes model in Domino with distributions

Process	Distribution	Based on
Taxi-in	LogNormal distribution considering mean, standard deviation and modifier to consider baseline or stressed scenarios.	IATA Summer Season 2010 from CODA [5]
Taxi-out	LogNormal distribution considering mean, standard deviation and modifier to consider baseline or stressed scenarios.	IATA Summer Season 2010 from CODA [6]
Climb uncertainty	Normal distribution minutes	Analysis DDR difference between planned and executed trajectories (m2, m3) from DCI4HD2D Project [7]
Cruise	Normal distribution NM	Analysis DDR difference between planned and executed trajectories (m2, m3) from DCI4HD2D Project [7]
Wind	Empirical probability distribution function for planned wind during the cruise. Used average wind between regions. No noise added on execution.	For each ANSP to ANSP origin and destination airport consider the difference between requested speed and observed average ground speed for cruise segments from DDR2 analysis (AIRAC1409) [8]

Turnaround time	Exponential distribution considering: - Minimum turnaround time based on airport size, aircraft wake and type of airline (REG, CHT, LCC, FSC) Lambda which depends on scenario (Default or High delay)	Analysis of turnaround times performed in POEM project and used in ComplexityCosts project [9]
Probability ATFM delay	When regulation is explicit at airport, the regulations are based on a given historical day. The days are selected based on their percentile ranked by number of regulations at airport in the day. There is a minimum and maximum percentile to be considered for baseline and stressed scenarios. For regulations in the airspace there are two probabilities one for regulations due to weather and another for regulations due to any other reason.	Based on analysis of DDR2 (AIRAC1313-1413 excluding days with industrial actions) [8]
ATFM delay	Empirical probability distribution function for regulations due to weather and regulations for other reasons.	Based on analysis of DDR2 (AIRAC1313-1413 excluding days with industrial actions) [8]
Non-ATFM delay	Exponential distribution with different lambda as a function of scenario: baseline, stressed	-
Passenger connecting times	LogNormal distribution Considering minimum connecting times per airport and type of connection (between national flights, from national to international and between international flights), sigma and percentile of passengers who connect in less than the minimum connecting time.	Based on analysis of minimum connecting times at ECAC airports originally performed in POEM project [10]

Variation of cruise length due to DCI	Normal distribution NM	Analysis of Performance using Airbus PEP [11]
Cancellation	Probability based on historical cancellation rate and explicit cancellation for missing arrival curfew at airports.	CODA 2017 report [12] and information on airports implementing curfews from EUROCONTROL.

Table 2 presents the data sources used to identify the target values for key indicators as in D5.2. Note that we are basing the scenarios on the traffic of a given day. Therefore, we have been able to reconstruct the schedules of that day with the execution of the flights from DDR2, and use this information for the validation of the model.

Table 2: Calibration parameters considered for targets

Parameter	Source
Departure delay	Reconstructed schedules compared with AOBT from DDR2 (m3)
Arrival delay	Reconstructed schedules compared with estimated AIBT from DDR2 (m3)
Delay distribution per reason (Reactionary, en-route, capacity, weather)	CODA 2017 report [12]
Flight plan length (NM)	DDR2 (m1)
Flight plan duration (min)	DDR2 (m1)
Flight execution length (NM)	DDR2 (m3)
Flight execution duration (min)	DDR2 (m3)
Taxi-in	Reconstructed schedules compared with planned taxi times from DDR2 (m1)
Taxi-out	Take off time - AOBT estimated form DDR2 (m3)
Gate-to-gate time (min)	DDR2 with estimated taxi times
Cancellation rate	CODA 2017 report [12]

Table 3 present the parameters that have been adjusted. Some of these parameters have been chosen based on expert judgment, whereas others have been calibrated using some of the above metrics.

Table 3: Parameters that have been adjusted in the model from the distributions

Process	Parameter	Possible values
Turnaround time	Lambda of exponential distribution	• Default delay scenario value • High delay scenario value
Climb uncertainty	Extra climb minutes	• Value adjusted for calibration in baseline scenario
Airport capacity modifier	Reduction of airport capacity	• Default scenario value: no reduction • High delay scenario value: half capacity reduction
Airport capacity adjustment	Adjustment of capacity at airport considering demand lower due to non-inclusion of non-passenger commercial flights	Value adjusted based on baseline scenario analysis of DDR2 flights
Non-ATFM delay	Parameter in exponential distribution, typical delay	• Default delay scenario value (calibrated on DDR2 data) • High delay scenario value
Probability of ATFM delay explicit at airport	Day selection minimum and maximum percentile considered	• Default delay scenario values (0.3-0.8) • High delay scenario values (0.8-1)
Taxi time modifier	Increment of taxi time	• Default scenario value: no increment • High delay scenario value: half increment
Passenger connecting time	Percentage of passengers who made the connection in less time than the MCT Sigma for the LogNormal distribution	Value adjusted for calibration in baseline scenario
Cancellation rate	Ratio of cancelled flights per day	Value adjusted for calibration in baseline scenario

4.2 Calibration process

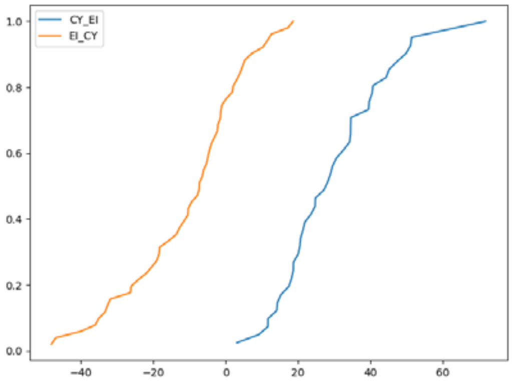
The basic principle of calibration is to measure something in output of the model, compare it to some empirical data, and adjusting some parameters so they eventually matched. The calibration on the final version has been done following a similar process than the previous version, in particular calibrating mainly average values. The calibration process is iterative by nature, since parameters usually impact several observables on the system.

4.2.1 Airborne times

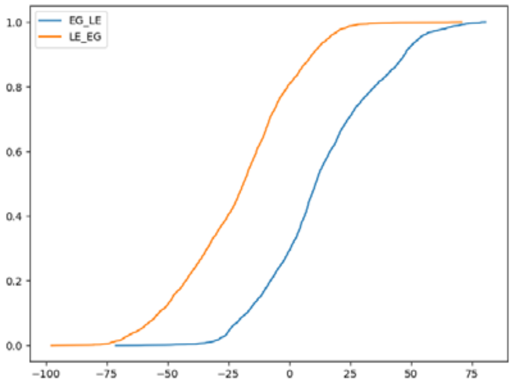
The first indicator that was adjusted was the flying time (from take-off to landing). In the model, flights select their flight plan using some pre-computed trajectories. The pre-computation is done based on routes built from a clustering analysis on DDR data, as explained in D5.2. For each possible route between each origin and destination pair, a trajectory is computed for each aircraft type which could use that route. These trajectories are estimated using aircraft performance model (BADA 4) and adding wind to the cruise phase. The output of this pre-computation phase is the pool of flight plans that will be used by the AOC when selecting the one operated by the flight.

Note that at execution, there are uncertainties added to the actually flown trajectory (time required to reach the top of climb or actual flight plan distance) as explained above. Moreover, airlines and flights might use some mechanism (such as 4DTA) to modify their trajectories, and there are also delays due to arrival queue management at airports (holdings).

Based on the work carried out in Vista (Vista Consortium, 2018), an average estimated wind can be computed as the difference between the estimated ground speed (distance of the segment divided by time required to cover the segment according to the flight plan) and the requested speed from historical DDR data. A weighted average wind is then estimated per flight considering the length of all the segments. Finally, cumulative distribution functions of average wind are produced by grouping the flights based on their origin-destination NAS. As shown in Figure 6,in this manner probability distributions of average cruise winds encountered, which differ by origin and destination, thus capturing the general weather pattern (e.g., flights from Canada to Ireland will in general have head wind (negative) while flying from Ireland to Canada tail winds are more commonly encountered (positive winds), see Figure 6 a)).



a) Canada (CY) to Ireland (EI) flights



b) United Kingdom (EG) to Spain (LE) flights

Figure 6: Examples of cumulative distribution of average wind between different NAS

In Domino, we have selected a fixed wind per origin-destination based on the distributions presented above and considering the 0.42 percentile of the wind. This provides a slightly higher head-wind than in the average case and has been adjusted as part of the calibration to ensure that the total average flying time is properly calibrated with historical data.

The re-computation of all the possible trajectories for the pool of flight plans is a very long process, several days are needed to compute all the trajectories needed for the simulated day of operations (there are a total of 42.942 trajectories available for a total of 11.344 origin-destination pairs (3.8 trajectories in average per origin-destination pair, one per possible route and aircraft type)).

4.2.2 Taxi times

Taxi times impact both arrival and departure delay (through turnaround), but are relatively independent of other factors (in the model). As a consequence, in a second step, we slightly adjust both distributions to match empirical data for them.

4.2.3 Non-ATFM delay

On top of the ATFM regulations, a non-ATFM delay is applied to the flights before their departure. This delay represents different exogenous factors, like missing crew, problem when boarding passengers, etc.

As a third step, we use this distribution of delay to match simulated departure delay to the empirical one.

4.2.4 Turnaround time

The turnaround time is the time necessary for the aircraft to be ready for next flight (including de-boarding and boarding of passengers). As a consequence, it impacts reactionary, departure, and arrival delay at the same time. In Domino we adjusted it so that reactionary delay has the right share in the delay analysis (see below), and the arrival reaches its empirical value. Note that the starting values for the turnaround are based on an analysis of turnaround times performed in POEM project and used in the ComplexityCosts project which considered aircraft, airport and airline types (SESAR, 2013).

Since this parameter also impacts departure delay, a few iterations were needed between the non-ATFM delay and this parameter to perform the calibration.

4.2.5 Curfews

Curfews are an important part of the airlines decision-making process when it comes to tactical management as even delays early in the day might propagate to eventually cause a curfew to be breached as presented in Section 3.2.3. Flights infringing curfews, i.e., landing after a certain time of the day, may face two issues:

- a fine to pay (‘soft’ curfew),
- a rejection of their flight plan, i.e., flights cannot be planned to land after a certain time in some airports (‘hard’ curfew).

Note that these terminologies are not official, standard terms. Further work across this research area is needed, and curfew data needs to be made available, to enable comparisons between projects. As reported in [4], curfews application can be very complex. For example, the curfew may be active only for a certain type of aircraft, for flights coming from a certain direction, for departures, for arrivals, etc. This complexity is driven by the fact that some of these limitations are related with environmental practices (e.g., noise pollution).

In Domino, a simplified version is implemented. We consider only curfews which prevent airlines from filing a flight plan to the destination. If the expected arrival time to the airport is after the curfew time, the flight plan will be rejected by the network manager and if no alternatives are available (e.g., using a different earlier arriving flight plan), the flight will be cancelled. This choice has two consequences for the model:

- the network manager can reject some flight plans if they infringe a curfew,

- the airline has to take into account the price of a flight that would not make it to its final destination when making different decisions (such as flight swapping or delay recovery to avoid downstream potential breach of curfews).

Concerning the second point, the team used prior EUROCONTROL knowledge of this issue through its UDPP development project. For that, EUROCONTROL developed a cost model including potential curfew costs, estimated with the help of the UDPP airline board. Domino considers the same costs:

- 40000 euros for a light or medium aircraft, as flagged in their wake turbulence category,
- 80000 euros for a heavy jet.

Airlines are aware that delays early in the day can translate into delays for the last flight, which might breach a curfew (as presented in Section 3.2.3).

In the model version used to generate the results presented in D5.2 Investigative case studies results, curfews were already modelled. However, which airports implemented curfew was based on the analysis of historical data by identifying airports which didn't have flights arriving after a certain threshold. This led to an overestimation of the airports which enforce curfews. In D5.2, if a flight breached a curfew all subsequent flights planned with that aircraft were also cancelled. In some instance, abnormally large delays earlier in the day might trigger the cancellation of many flights.

In the current version of the model, these shortcomings have been fixed. First, curfews are enforced at arrival only at airports in the ECAC region which were provided by EUROCONTROL. These airports have been considered as airports with curfews as part of the validation activities of UDPP. This reduced the number of airports which enforce curfews to 14. Secondly, only the flight which breaches the curfew is cancelled and subsequent flights with the same aircraft are considered to be possible to be operated if they won't breach a curfew themselves.

As for the time when these arrival curfews start, data provided from EUROCONTROL has been used. However, for 5 of the 14 airports, some scheduled flights in the model were planned to arrive after the curfew times provided by EUROCONTROL. For those airports, and in order to prevent flights breaching the curfew even if on-time, the curfew time has been increased to the SIBT of those late arriving flights, providing a buffer of 15 minutes, and rounding them to the next 15 minutes.

4.3 Final state of calibration

The following table shows the final values of the observables in the model and compare them to historical data. All the averages (except flying delay, which is very small) are within 5% of their empirical values. Most of them are within 1% of their targets, which ensures a good level of calibration for these average values.

Table 4: Model calibration summary

Average value of:	Simulations	Historical	Difference (mins)	Error
Departure delay	11.41	11.43	-0.02	-0.15%
Flying delay	-0.05	-0.16	0.11	-70.45%
Taxi delay	-4.78	-4.62	-0.16	3.50%
Arrival delay	6.58	6.65	-0.07	-0.99%
Arrival delay without earlies	11.32	11.57	-0.25	-2.12%
Scheduled G2G time	159.47	159.47	0	0.00%
Actual G2G time	154.65	154.69	-0.04	-0.02%
Scheduled flying time	136.37	137.28	-0.91	-0.66%
Actual flying time	136.33	137.12	-0.78	-0.57%
Scheduled flying distance	965.37	960.57	4.8	0.50%
Actual flying distance	954.33	948.51	5.82	0.61%
Actual taxi-out time	12.14	12.52	-0.38	-3.02%
Actual taxi-in time	5.6	5.73	-0.13	-2.26%

In order to further check the calibration, we also compared the modelled delays results with the type of delay experienced by flights. For this, we used a CODA report from 2017 which presents a break-down of the types of delay. This information has been used to qualitative ensure that the model is producing the type of delays that are expected and not as a quantitative target. The report from CODA used is for the whole of 2017, whereas are targets are for a specific day. Moreover, there might be some differences on how delays are labelled, e.g., whether the report computes averages only on most penalising delays on not. The following table shows the differences between the report and the simulated data.

Table 5. Distribution of delays among the main reasons of delay

Type of delay	Proportion in simulations	Proportion in CODA 2017	Minutes needed in simulation to match proportions
Reactionary	27.85%	44.50%	1.94
Turnaround	58.55%	35.80%	-2.52
En-route	7.60%	7.50%	0.00
Capacity	3.67%	7.20%	0.41
Weather	2.32%	1.90%	-0.04

4.4 Further comparison and remaining known issues

Further analyses were performed to better understand the relationship between the historical data and the model output. We show a few of them in the following.

The first important metrics is the gate-to-gate time. The distribution shown in Figure 7 is driven by the structure of the OD pair distance, but depends also more weakly on the performance data (for speed) and the various delay distributions (e.g., holdings at arrival, taxi times). The agreement is very good between simulations and empirical data.

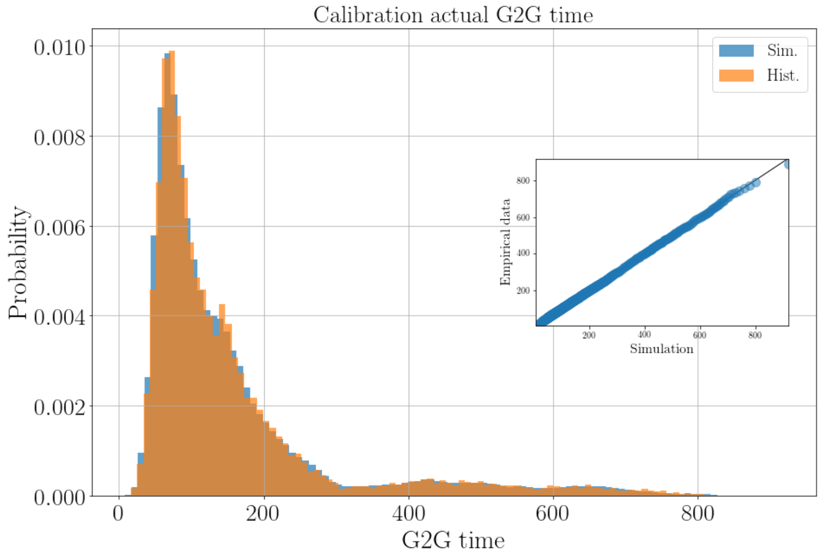


Figure 7: Probability distribution of gate-to-gate times in simulations and historical data. Inset: QQ-plot between both distributions.

Other important distributions to review are departure and arrival delay, since they structure the model and form the basis of the airlines' decisions. On the left of Figure 8, we can see that departure delays have quite a significant difference in the model with respect to historical data. In the model we assumed that no flight could depart before their schedule time. However, this is not the case in reality, as shown by the orange distribution. In the historical dataset, flights can be up to 20 minutes earlier than their schedules. The model could consider that if the aircraft is ready and there is no missing passenger, the flight could be ready to depart earlier than their schedule. However, it is not clear which parameters should be considered to properly model this behaviour. Finally, in some instances, the values observed in the historical data could also be an artefact due to some misalignment between DDR and schedule data.

The arrival delay distribution is closer to the historical one, as seen on the right of the figure. It is however interesting to note that the simulations create more delays around 20 minutes than there are in the historical data. Conversely, less flights have almost no delay (around 9 minutes). This might be due partially to the previous departure delay artefact, since early departing flights are more likely to be on time at arrival.

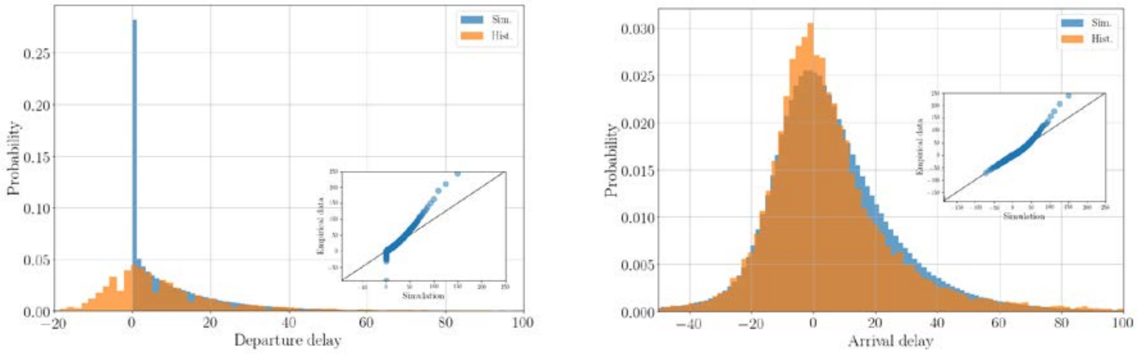


Figure 8: Probability distributions for departure (left) and arrival (right) delay in simulations and historical data. Insets: corresponding QQ-plots.

It is interesting also to compare the differences of features between the planned and the actual trajectories. In Figure 9, we show the distribution of differences in gate-to-gate times between planned and actual trajectories. In this case, the agreement between simulations and historical data is very good, even if the simulations seem to have slightly fatter tails on both sides.

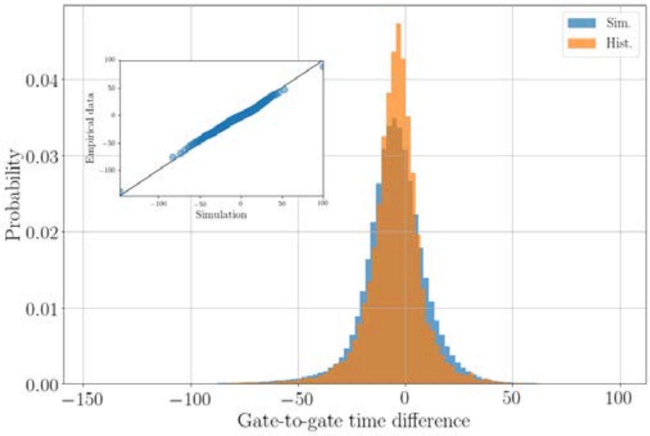


Figure 9: Probability distributions of gate-to-gate differences in simulations and historical data.

Finally, we highlight a discrepancy between empirical data and simulations which may have some long-reaching consequences. When examining passenger data, we have noted that the average delay per passenger seemed to be quite low compared to flight delay. For instance, in the baseline, the average arrival passenger delay is only 66% of the average flight arrival delay. This is counter-intuitive, since passengers can have multi-leg itineraries, miss a connection, and thus have much a much higher delay than the flight they take. However, the ratio of the average delays can be smaller than 1, as shown in Table 6, displays the values of the ratios for departure and arrival delay, in the simulations and in the historical data.

Table 6: Estimations of ratios between pax delay and flight delay

	Simulations	Historical
Ratio pax arrival delay/flight arrival delay	0.66	0.91
Ratio pax departure delay/flight departure delay	0.97	1.11

The reason for this ratio to be smaller than one is related to a simple statistical effect: if big planes are in average less delayed than smaller ones, then this drives naturally the passenger delay down. To analyse this effect, we show in Figure 10 the average flight delay as a function of the number of passengers on the flight, average per quantile (60 quantiles represented). We also show the Pearson correlation coefficients (computed on all points, not just the quantiles) in the legend. Clearly, the simulations have a different behaviour than the empirical data. Overall, it seems that the simulations tend to have lower arrival delays for large flights, and higher delays for small ones. At departure, the coefficients are not so different, and the noise is larger than the average trends, which explains why the ratios are not so different for departure in Table 6.

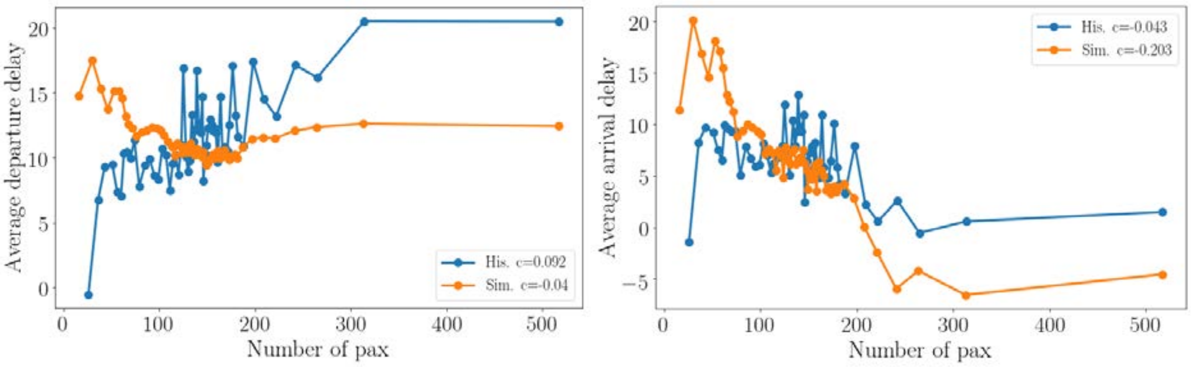


Figure 10: Average departure (left) and arrival (right) delay as a functions of the number of pax in flight. Delays are averaged over quantiles.

Note that there is a correlation between larger number of passenger and longer flight distance. There might be different reasons to explain this behaviour (buffers, effect of wind, airline behaviour). It is expected that larger aircraft carrying more passengers are more important for airlines as they might potentially incur on higher costs. For this reason, it is expected that larger flights will be more protected than small ones. We could expect this pattern to be more present in the historical/empirical data.

5 Case studies results

5.1 Metrics

5.1.1 Metrics definitions

This section is dedicated to the explanation of the metrics that are used to analyse the model's outputs and how they are computed. We first consider the same classical metrics related to delays, passengers and costs that were considered in [2], plus some additional passenger-related metrics, whose computation is made possible by the new organisation of the output. We list all the classical metrics below. Secondly, we consider the more advanced, network metrics of centrality and causality that were also used in [2], which are very briefly summarised here. As anticipated in [13], an additional measure of centrality and an alternative method to detect causality are added to the set of tools. These new methods are explained and justified in this section.

5.1.1.1 Classical metrics

a) Delay metrics

For delay, we report the following classical metrics, averaged over 50 iterations of the agent-based model, per scenario:

- Average departure delay of flights[1].
- Fraction of flights with departure delay > X minutes
- Total departure delay of flights with departure delay > X minutes
- Average departure delay of flights with departure delay > X minutes
- Average arrival delay across all flights[2]
- Average arrival delay without earlies[3]
- Fraction of flights with arrival delay > X minutes

[1] Departure delays are always positive, as in the model flights never depart before their scheduled off-block time.

[2] Early arrivals are counted as negative delays.

[3] Early arrivals are counted as 0.

- Total arrival delay of flights with arrival delay > X minutes
- Average arrival delay of flights with arrival delay > X minutes where X=15, 60, 180
- Average gate-to-gate delay of flights (obtained as the difference between the scheduled and the actual gate-to-gate time)
- Average per-passenger gate-to-gate delay (for each flight, the delay is divided by the number of passengers on the flight; the result is averaged over all flights)
- Fraction of cancelled flights

Average reactionary delay (computed as the mean delay of flights whose main reason for delay is reactionary[4])

- Fraction of flights with any reactionary delay

b) Cost metrics

For the costs, we report the following classical metrics, averaged over 50 iterations of the model:

- Average excess cost of fuel (the extra cost with respect to the planned cost of fuel, which can also be negative if fuel was saved)
- Average cost of compensation
- Fraction of flights paying compensation
- Average cost of transfer
- Fraction of flights paying transfer
- Average duty of care cost
- Fraction of flights paying duty of care
- Average soft costs
- Fraction of flights paying soft costs
- Average non-pax costs (crew + maintenance)
- Fraction of flights paying non-pax costs
- Average total excess cost relative to scheduled flight plan

All costs are in euros. Average costs are computed for all flights, including those that did not experience such costs as indicated above (counted as zero).

c) Passenger metrics

We consider the following metrics related to passengers, averaged over 50 iterations of the model:

[4] The main reason for the delay to a flight is flagged as reactionary if among all the delays experienced by the flight, the largest part is reactionary.

- Average passenger delay across all passengers[5]
- Average passenger delay without earlies[6]
- Average delay of connecting and non-connecting passengers.
- Fraction of passengers with a modified itinerary
- Fraction of passengers arriving at their final destination on the same day
- Fraction of passengers receiving compensation
- Average compensation received
- Fraction of passengers receiving duty of care
- Average duty of care received
- Fraction of passengers with delay > X minutes
- Total delay of passengers with delay > X minutes
- Average delay of passengers with delay > X minutes where X=15, 60, 180

5.1.1.2 Centrality metrics

The first centrality metric that we consider is trip centrality, which was introduced in Deliverable 5.1 [14]. The outgoing trip centrality of an airport counts all the potential itineraries having that airport as the origin, while the incoming trip centrality counts those having that airport as a destination. Potential itineraries are all the sequences of any number of flights that can be potentially taken one after the other, given their schedule. An itinerary of n legs is weighted α^n, where $\alpha < 1$, so that itineraries made of more legs are counted less. For all the results shown in this deliverable $\alpha = 0.2$. With respect to how the metric was defined in Deliverable 5.1 and computed in Deliverable 5.2 [2], we make a small modification to account for a minimum connecting time between subsequent flights used in an itinerary. Given a minimum connecting time of Δt, that we fix to 15 minutes, this modification is simply obtained by shifting forward the arrival time of each flight by Δt. Trip centrality can either count only the itineraries made of legs of the same airline or alliance, corresponding to setting $\varepsilon = 0$ (see Deliverable 5.1 [14]), or count also the itineraries using two or more airlines or alliances. For all the results presented in this deliverable we used $\varepsilon = 0$, therefore the walks counted are those within an alliance or within an airline (for airlines that do not belong to any alliance).

The loss of an airport's outgoing and incoming trip centrality between the network of scheduled flights, and the actual network, measures the loss of potential outgoing or incoming itineraries that

[5] Computed as the arrival delay at the passenger's final destination; early arrivals are counted as negative delays.

[6] The average is computed counting early arrivals as zeros. This metric is interesting because of its related metrics for the flight delays, used in several other reporting contexts, although there is a general lack of standardisation and consistency between frameworks. Note that whereas for a flight early arrival are not usually beneficial, and thus can be set to 0, they are probably more often important for passengers, whose utility always decreases with the length of the trip (this is captured in the previous average passenger delay metric).

are not feasible any more (due to delays or cancellations), therefore it quantifies the decrease in the potential to go from that airport to the rest of the airport network or *vice versa*. See Deliverables 5.1 and 5.2 for further details on the computation and interpretation of trip centrality losses. Note that for trip centrality, when the centrality loss is averaged over the entire network, the loss of incoming centrality equals exactly the loss of outgoing centrality. In fact, each loss of outgoing centrality corresponds to an equal loss of incoming centrality of another airport. Therefore, in this case, we will refer to it as 'Average trip centrality loss'. When referring to a subset of the airports, incoming and outgoing centrality losses can be different.

Secondly, we consider 'passenger centrality', which was introduced in Deliverable 5.2. In the computation of passenger centrality each itinerary contributes to the outgoing or incoming centrality of an airport an amount which corresponds to the number of passengers on that itinerary. Therefore, the outgoing passenger centrality of an airport corresponds to the number of passengers that depart from that airport (either as their first departure or taking a flight connection there) and are directed to another destination, either with a direct flight or with connections. The incoming centrality of an airport, instead, corresponds to the number of passengers that land in that airport, either as their final destination or to take a connection.

To compute the loss of passenger centrality, in the actual network we only count passengers that reach their destination using their scheduled itinerary. The actual outgoing passenger centrality of an airport corresponds to the number of passengers that were counted in the scheduled outgoing passenger centrality and that manage to follow their scheduled itinerary. If, for example, N incoming passengers miss their connection in airport i, and are rebooked to another outgoing flight, airport i will have a loss of outgoing centrality amounting to N. The same loss would apply if N passengers depart late from i and miss their next connection at another airport. While in Deliverable 5.2 we could not use the information on which itineraries had been disrupted in the simulations due to issues in how the information had been stored, in the current model output the information is available.

5.1.1.3 Centrality metrics

In addition to these metrics, we introduce another centrality metric, as mentioned in Deliverable 3.3, inspired by the concept of betweenness centrality. The purpose of this metric is to measure the potential flow of passengers through an airport (and its loss in the actual network), a quantity of particular interest for hubs. In fact, with trip centrality, the centrality of an airport does not account for itineraries passing through that airport, therefore its loss does not account for missed connections at the airport itself. The loss of outgoing passenger centrality, as explained in Deliverable 5.2, does account for missed connections in the airport itself, however it is computed using the scheduled and actual passenger itineraries: information that is not easily available outside of a modelling context. The same is true for classical metrics, such as the number of missed connections at an airport. This is why we propose a centrality metric, that we call 'trip betweenness centrality', that computes the potential passengers flow and its loss based only on scheduled and actual flights and on passenger demand for each origin-destination pair (on a specific day or averaged on a longer period). The value of trip betweenness centrality for an airport (see below for its definition) is an estimate of the number of passengers connecting at that airport, given the flight schedules and the demand for each origin-destination pair. The loss of trip betweenness is an estimate of the number of passengers connecting at that airport that had a disrupted itinerary. It therefore represents a tool to evaluate the risk of connecting in a certain airport that can be used by stakeholders interested in the passengers' point of view.

In standard betweenness centrality [15], a node is considered central if a large fraction of the shortest paths between each pair of nodes in the network pass from it. More precisely, if $\sigma_{ij}$is the number of shortest paths between i and j and$\sigma_{ij}^{k}$is the number of such shortest paths that pass by k, the betweenness centrality of k is

$$b(k) = \sum_{ij} \frac{\sigma_{ij}^{k}}{\sigma_{ij}}.$$

In this definition, only the shortest paths are considered, i.e., if a path from node i to node j which passes from node k is not the shortest path joining i and j, it would not contribute to the centrality of node k. For ATM applications, however, we deem it more realistic that passengers would not only use the shortest paths, as these could only be available at certain times of the day or could well be more expensive than a slightly longer path (where path length might be intended as number of legs, duration or a combination of the two). In addition, the standard betweenness centrality does not consider that the network is temporal, and therefore that paths should be time-ordered. Temporal generalisations of betweenness centrality have been proposed in the literature [16]. Here, we propose a temporal version of betweenness centrality tailored for ATM applications, where all the paths that satisfy the following constraints are accepted:

- the path must have no more than 2 legs;
- the path must be time-ordered, i.e. the departure time of the second leg must be later than the landing time of the first leg, and there must be a connecting time of at least Δt mins; note that this time has the meaning of the minimum connecting time in a scheduled itinerary, which might be larger than the minimum time needed to take a connection, as passengers would not choose exceedingly tight itineraries;
- the duration of the path must not exceed K times the duration of the fastest path connecting the same origin-destination pair (including directs);
- the demand for the corresponding origin-destination pair must be non-zero.

With these constraints, both the number of legs and the duration of a path are taken into consideration. Choosing K=2,Δt=45 (the value chosen looking at the connecting times in the scheduled passengers' itineraries of 12 September 2014) and using the demand of 12 September 2014, these constraints select a set of paths such that 50% of the 2-leg passengers' itineraries of 12 September are among the accepted paths, which however include many more paths that were not used on that day but are acceptable according to the constraints (a total of 170 000 accepted 2-leg itineraries, of which around 35 000 correspond to real passengers itineraries).

Note that real passengers' itineraries also include a small number (~4000) of 3-leg itineraries. However, if we also accept 3-leg paths (with the same constraints on duration) the accepted 3-leg paths are around 700 000, therefore including a large number of itineraries that were not used by passengers. We therefore decided to only accept paths up to two legs, so that the set of accepted paths is more similar to the set of paths actually used by passengers.

Additionally, we want to weight the contribution of pair (i,j) to the trip betweenness centrality by the demand d_{ij} for that origin-destination pair. As a first option, one could think of computing the trip betweenness centrality of an airport k as:

$$tb(k) = \sum_{ij} \frac{\sigma_{ij}^{k}}{\sigma_{ij}} d_{ij},$$

Where σ_{ij} is the number of accepted paths between i and j (including directs) and σ_{ij}^{k} the number of such paths that pass through k. Note that the product $\frac{\sigma_{ij}^{k}}{\sigma_{ij}} d_{ij}$ is an estimate of the number of passengers that pass from k going from i to j, assuming that the demand is split equally among the σ_{ij}accepted paths. However, this is not realistic, in fact we verified that when there is a direct between i and j, on average 94% of the demand is satisfied by the direct. Therefore, a more realistic estimate for the number of passenger connections in k is obtained reducing $d_{ij}$to 6% of the total demand for origin-destination pairs having a direct flight between them, and letting σ_{ij}be the number of accepted paths between i and j, excluding directs. In conclusion, tb(k) is computed according to (1), but where σ_{ij} and $d_{ij}$are redefined as explained. We checked that results do not change significantly if the demand reduction is done with the percentage specific for that origin-destination pair instead than using the average.

We consider a path to be disrupted if the actual connecting time is at leastΔt_2(which in principle can be smaller than Δt, as it represents the minimum time needed to take a connection). Calling ${\sigma_{ij}}'^{k}$the number of accepted paths from i to j assign from k that are not disrupted, the actual betweenness centrality is:

$$tb^{R}(k) = \sum_{ij} \frac{{\sigma_{ij}}'^{k}}{\sigma_{ij}} d_{ij},$$

For the results presented in this Deliverable, we use$\Delta t_2 = 20$. The loss of betweenness $\Delta b(k) = b(k) - b^{R}(k)$ is therefore an estimate of the number of passengers connecting in k whose itinerary is disrupted. The relative loss $\Delta b(k)/b(k)$is an estimate of the fraction of passengers connecting in k whose itinerary is disrupted, i.e. the probability to have one's itinerary disrupted if connecting at k.

The scheduled trip betweenness has, as expected, a strong linear correlation with the number of connecting passengers in the same airport (correlation coefficient=0.98). The two quantities are plotted in Figure 11 in double logarithmic scale, together with their linear fit (blue line, R^2=0.96).

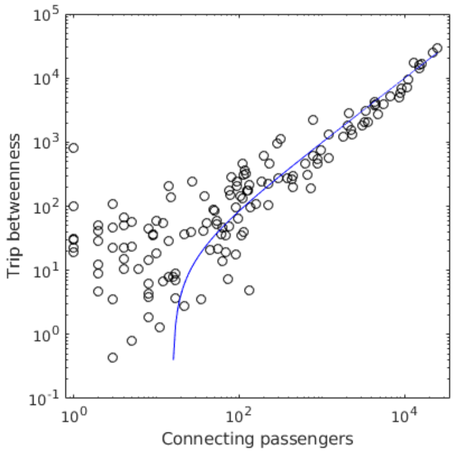


Figure 11: Number of connecting passengers in each airport vs trip betweenness estimate on 12SEP14. The plot is in double logarithmic scale, the blue line is a linear regression y=a*x+b with a=1.02±0.01. b=-15±13, R^2=0.96).

5.1.1.4 Causality metrics

In the ATM system, delays and congestion states propagate through the system due to the interactions between the flights and the environment, e.g. the network manager, the airports or arrival coordinators. The proposed causality metrics aim to detect the extent to which the congested state of an airport causes congestion in other nodes of the network, thus providing a toolbox which is able to characterise the channels of interactions between the different sub-parts of the system.

In time series analysis, a (directional) causal relationship between two systems is detected when the information on the state of one system helps in predicting the future state of the other. The presence of a causal relationship is assessed by means of statistical tests whose most well-known example is the Granger causality metric [17]. Indeed, it has been recently applied to airport networks [18] and [19].

Here, a data driven approach is adopted to identify the channels through which the delay propagates and establishes a network of causal relationships, where a link between two airports is present if delay propagates (in statistical sense) from one to the other. Causality is tested between the states of delay (or congestion) of airports in the network, measured as a given quantile of the distribution of the flight delays for that airport within one-hour window. In the following, we will consider two different definitions for the state of delay, by measuring it as either the mean or the third quartile of the distribution of delays. Finally, we consider the whole ECAC ATM system by averaging over all flights, without distinguishing between airlines.

The topology of the resulting causal network may change depending on the mechanism implemented in the system. This relates the presence of innovations at the micro level to its impact on delay dynamics and propagation at some macro level of aggregation, such as airports, airlines, or passengers. For example, a smaller number of causal links or less positive feedback subsystems can be seen as an improvement of the system, as they signal a diminished coupling of the systems' elements.

There exist several methods to detect a causality relationship, each one assessing the statistical significance of one time series in forecasting another. Granger causality (*in mean*) by [17] tests the statistical significance of the forecasting performance on average, by considering both small and large values, whereas the statistical test introduced by [20], namely Granger causality *in tail*, restricts the analysis to the prediction of extreme events, which are defined as the states falling in the tail of the distribution. When studying the propagation of congestion between airports, delays which are small with respect to flight time are probably not relevant for delay propagation, as they are typically fairly easily absorbed by buffers (or during the flight). Granger causality in tail tends to capture exactly the propagation of extreme events, thus describing the dynamics of congestion in the ATM system. However, the statistical test introduced by [20] suffers a high false positive rate when the time series describing the state of congestion of an airport displays non-zero autocorrelation. This is in effect very common for the ATM system: it is likely that a congestion lasts for many hours, thus resulting in systemic delays for flights and persistent state of congestion for the airport (i.e. *positive autocorrelation* shown by the state of delay). To solve this drawback of the test introduced in [20], we propose a novel statistical approach to identify causal relationships between the states of congestion of two airports, also in presence of non-zero autocorrelation and non-zero terms of interaction.

In the following, we introduce the definitions for both the state of delay and the state of congestion used in the causality analysis and describe the statistical tests.

a) State of delay, state of congestion

Here, we give some operative definitions for the random variables describing the states of the corresponding airports.

- **State of delay**. A random variable $X_i \in \mathbb{R}$ associated with airport i determining the level of delay of all flights departing from that airport within one hour time window, i.e. $x_{i,t}$ with t=1,...,24. We consider two different definitions:

  1. the *average* delay of all flights departing from the airport in that given hour;

  2. the 75% percentile (i.e. the third *quartile*) of the distribution of delays of the flights departing from the airport in that given hour.

  As pointed out by [18], a Z-Score standardisation procedure is applied to reduce the non-stationarity of the time series caused by daily seasonality, which may result in a biased evaluation of the Granger causality metric. The standardised time series of airport i is calculated as $\tilde{x}_{i,t} = (x_{i,t} - \bar{x}_i^t)/\sigma_i^t$ where $\bar{x}_i^t$ and σ_i^t are the mean and the standard deviation of the delay states of airport i recorded at hour t across all available days (or, equivalently, simulations of the ABM model).

- **State of congestion (or extreme delay)**. A binary random variable $Z_i \in \{0,1\}$ associated with airport i determining if the airport is congested or not, according to a threshold value Q which is determined as the 80% quantile of the distribution of the state of delay, by considering all airports and all days (simulations). In other words:

$$\begin{cases} z_{i,t} = 1 \text{ if } x_{i,t} \geq Q, \\ z_{i,t} = 0 \text{ if } x_{i,t} < Q. \end{cases}$$

When the actual the state of delay falls in the right tail of the distribution determined by the quantile Q, we say that the state of delay is *extreme*, thus identifying the airport as congested.

b) Review of causality methods

Any causality method is based on detecting a causal relationship between two time series by testing if the knowledge of past observations of one time series allows us to estimate the future observations of the other time series better than without considering it. Assume we observe the realisation of a stochastic variable X whose realisation x_t at time t represents the state of delay of an airport (according to some previous definition).

(i) Granger causality in mean

$Y \equiv \{y_t\}_{t=1,\dots,T}$ is said to Granger-cause *in mean* $X \equiv \{x_t\}_{t=1,\dots,T}$ if we reject the null hypothesis that the past values of Ydo not provide statistically significant information about future values of X by assuming a linear predictive model [17]. In other words, if Y Granger-causes X, it is possible to use the past observations of Y to improve the prediction performance (with a certain degree of confidence) of the future value of X, weighting equally small and large values in assessing the prediction performance. In the ATM application, this is equivalent to say that the delay observed in Y is 'transmitted' to X, thus the causal relationship from Y to X can be interpreted as the presence of a channel for the process of delay propagation within the ATM system.

(ii) Granger causality in tail by Hong et al.

The statistical approach introduced by [20] aims to evaluate whether the knowledge of the past extreme events for a random variable Y helps in forecasting the occurrence of future extreme events for another random variable X. With a similar spirit of [17], this Granger causality in tail test aims to evaluate whether extreme events in an airport cause extreme events in another airport, by analysing the binary time series of the states of congestion of the airports. Let us consider Z_1 and Z_2 as the states of congestion associated with the states of delay X and Y of two airports, respectively.

Y is said to Granger-cause *in tail* X *if we reject the null hypothesis that the past extreme events, i.e.* $\{Z_{2,s}\}_{s=t,t-2,\dots,t-M}$ *for given* $M > 0$, *of* Y*do not provide statistically significant information about the future extreme event, i.e.* $Z_{1,t+1}$, *of* X, *thus revealing the presence of a propagation channel for 'extreme' delays (or congestion) between two airports.*

However, the proposed statistical test is not robust to the presence of positive autocorrelation for the states of congestion, thus resulting in a very high rate of false positives (see below). For this reason, we introduce a novel method to test for Granger causality in tail.

(iii) Granger causality in tail with BiDAR

Let us consider a single binary time series $\{Z_t\}$, e.g., describing the state of congestion of an airport. We can describe Z_t by a DAR(p) (discrete auto-regressive of order p) process [21], i.e.

$$Z_t = V_t Z_{t-\tau_t} + (1 - V_t) U_t,$$

meaning that Z_t can be copied from the past ($t-\tau$ for some $\tau = 1,\dots,p$) or sampled from some Bernoulli marginalU_t, according to the Bernoulli random variable $V_t = 0{,}1$, which selects step by step what is the case.

Then, the DAR(p) process can be generalised to the case of two binary random states, i.e.Z_1andZ_2describing the states of congestion of two airports, by considering the following BiDAR(p) (binary discrete auto-regressive of order p) process,

$$\begin{cases} Z_{1,t} & = V_t((1-A_t)Z_{1,t-\tau_t^{11}} + A_t Z_{2,t-\tau_t^{12}}) + (1-V_t)U_{1,t} \\ Z_{2,t} & = S_t(B_t Z_{1,t-\tau_t^{21}} + (1-B_t)Z_{2,t-\tau_t^{22}}) + (1-S_t)U_{2,t} \end{cases}$$

where, as before, $V_t$determines for $Z_{1,t}$if copying or not from the past (similarly for S_t in relationship to $Z_{2,t}$), but the Bernoulli random variable $A_t$selects now if copying the past (extreme) values of Z_1itself or Z_2 (*vice versa* with B_t).

Hence, the off-diagonal term of interaction$A_t$determines the level of causality. In particular, $Y$is said to Granger-cause *in tail* X if we reject the null hypothesis that the past extreme events, i.e. $\{Z_{2,s}\}_{s=t,t-2,\dots,t-p}$ for given $p > 0$, of Ydo not provide statistically significant information about the future extreme events, i.e. $Z_{1,t+1}$, of X, by testing for non-zero term of interaction A_t.

For further details about the statistical tests here described, see Appendix II.

(iv) Causality network

Given a method to detect causality between two time series, we can consider the network of airports where a link $i \rightarrow j$ is present if i 'Granger causes' j. This approach has already been considered in a recent analysis of the Chinese air transportation network [18], where only Granger causality in mean has been used, and in [19] where we have used also the Granger in tail of [20].

Given N time series, representing the state of the$N$airports in the network, the (chosen) causality test is performed on all the possible $M = N(N-1)$ pairs. When performing multiple hypothesis testing, a correction to the significance level of each single test should be applied to obtain the desired overall level Γ, i.e. if we test M hypotheses simultaneously with a desired Γ, then a significance level $\Gamma' < \Gamma$ should be applied to each single test to correct for the increased chance of rare events, and therefore, the increased probability of false rejections. This has typically not been considered in the literature. However, it can have a huge impact on the number of detected causal links.

Here, we apply two types of correction:

1. the Bonferroni correction which compensates for the multiple comparisons in the most conservative way by setting $\Gamma' = \Gamma/M$;
2. false discovery rate (FDR) method which controls for the rate of false positives in the following way:
    a. consider the all the $M = N(N-1)$ hypotheses we aim to test, together with the p-values of each statistical test $\{p_i\}_{i=1,\dots,M}$
    b. sort the p-values in ascending order $p_1 \leq p_2 \leq \dots \leq p_M$

Founding Members

c. find the position k such that $k = max\{i: p_i \leq \frac{i}{M}\frac{5\%}{\sum_{j=1}^{M} 1/j}\}$

d. validate the rejections of the first k hypotheses with p-values $p_1, \dots, p_k$.

Note that the test rejections validated with the Bonferroni corrections, are validated with the FDR method, too. However, FDR could accept some rejections of the null hypothesis which were excluded by Bonferroni.

The Bonferroni correction is extremely strict and essentially aims to minimise false positives, sometimes at the cost of accepting as consistent with the null hypothesis low probability observations. FDR, on the other hand, uses a more balanced threshold both for specificity and for sensitivity.

Whatever the type of correction applied, in the causality analysis we set the overall confidence level of the statistical test at $\Gamma = 5\%$.

5.1.2 Relationship between network metrics and operational indicators

5.1.2.1 Centrality metrics

In this section, we explore the relationships between centrality losses and standard operational indicators, in particular passenger-related costs and number of passengers with modified itineraries. All the analyses shown in this section are performed on the **baseline scenario of hub delay management** (see Section 2.1).

As was explained in Section 5.1, the trip betweenness of an airport is an estimate of the flux of s passing through an airport, given the schedules and the demands for each origin-destination pair. Therefore, the loss of trip betweenness of an airport is an estimate of the number of passengers connecting there that experienced disruption, either because of a missed connection or because of the cancellation of one leg of their itinerary. If this is a good estimate, we expect the loss of trip betweenness to be correlated to the number of passengers with modified itineraries and also to the passenger-related costs, as the latter are mostly generated by modified itineraries. In fact, if we consider the total loss of trip betweenness on the entire network (sum of the loss of each airport) in the model iterations, we find that this has a strong linear correlation both with the total number of passengers with modified itineraries (ρ=0.61) and with the total passenger-related costs associated with all flights (ρ=0.77). The Pearson correlation coefficients are computed on 100 iterations, see Figure 12. Therefore, in a situation in which we do not know the actual costs and the passengers' itineraries, but we know the schedules, the delays and the demand, we can provide a (rough) estimate of the aggregate passenger-related costs for all flights and of the number of passengers with modified itineraries using trip betweenness loss.

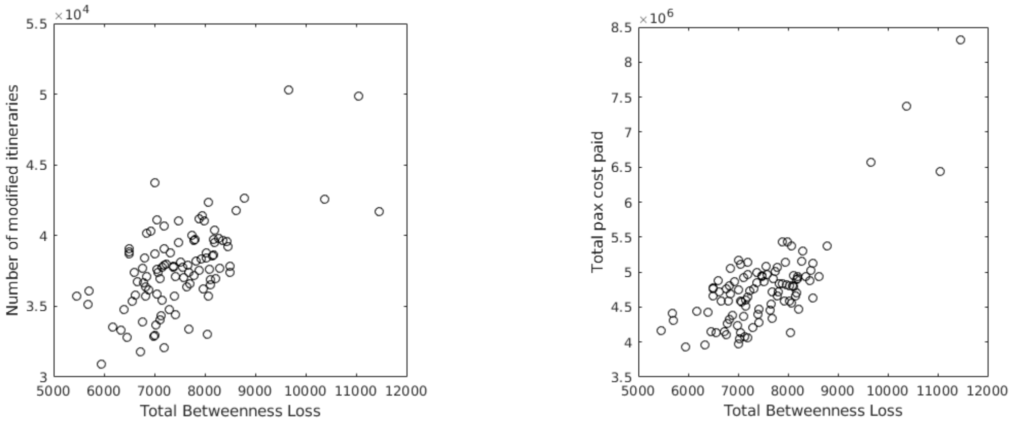


Figure 12: Trip betweenness loss. Left panel: Total trip betweenness loss on the network against total number of passengers with modified itineraries. Each dot represents one iteration. Right panel: Total trip betweenness loss on the network against total passenger-related costs associated with all flights. Each dot represents one iteration.

Note that the difference between the loss of trip betweenness and the number of passengers with modified itineraries is mainly due to the fact that the itineraries counted by trip betweenness do not completely coincide with the itineraries actually used by passengers in the simulated day, as explained in Section 5.1. Thus, the loss of trip betweenness actually measures the number of passengers with modified itineraries that we would have on a day in which passengers use all the itineraries counted by trip betweenness, i.e. all the itineraries accepted by the constraint we applied, with the approximations mentioned in Section 5.1. Additionally, the said difference is also partly due to the fact that in trip betweenness we consider an itinerary disrupted if the actual connecting time is less thanΔt_2, while in the model's simulation passenger connections are stochastic, and can require longer or shorter times than that, also depending on the airport and on the type of flights. However, we see in Figure 12 that, on at the aggregate level, the loss of trip betweenness still gives meaningful information for this particular day.

If we consider, instead, single airports, we can verify if the percentage loss of trip betweenness centrality of the airport is correlated to the fraction of passengers with modified itineraries among those having a connection there, and if the absolute loss is correlated to the cost of flights landing there. Of the 246 airports having non-zero betweenness centrality, 152 have actual passenger itineraries connecting there. For each airport of this subset, we computed the correlation between percentage trip betweenness loss and fraction of passengers with modified itineraries. We found that around 52% of these airports show a significant positive correlation (all the others have no significant correlation). The histogram of the obtained correlation coefficient is shown in Figure 13, left panel. When we consider the correlation with costs of incoming flights, instead, around 54% of these airports show a significant positive correlation (all the others have no significant correlation). The histogram of the obtained correlation coefficient is shown in Figure 13, right panel.

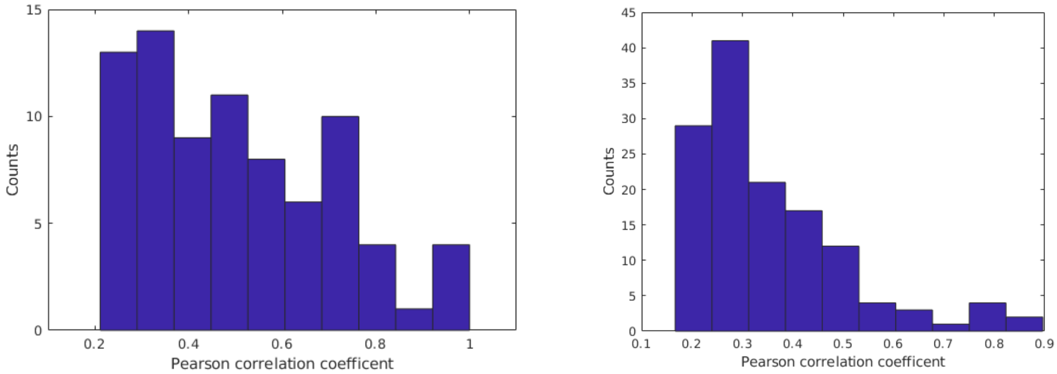


Figure 13: Pearson correlation coefficients for trip betweenness. Left panel: Histogram of the Pearson correlation coefficients between trip betweenness loss and number of passengers with modified itineraries (among those connecting there) for airports having a significant positive correlation; right panel: Histogram of the Pearson correlation coefficients between trip betweenness loss and the passenger-related costs associated to incoming flights for airports having a significant positive correlation.

Moving to outgoing passenger centrality loss, for a single airport, this value coincides exactly with the number of passengers departing or connecting there having modified itineraries on the simulated day. Therefore, its sum over all airports is clearly strongly correlated (ρ=0.98) with the total number of passengers with modified itineraries, although it is always larger because the modified itineraries of connecting passenger are counted twice when summing over all airports (see Figure 14, left panel). It is also strongly correlated with the total passenger-related costs (ρ=0.79, Figure 14, right panel). At the single airport level, for around 88% of the airports having non-zero outgoing passenger centrality, we find a positive correlation with the cost of outgoing flights, with many correlations close to 1 (see Figure 15, left panel). Interestingly, the correlations with sum of costs of outgoing and incoming flights are smaller (see Figure 15, right panel), probably because the cost of incoming flights includes costs related to non-connecting incoming passengers (e.g. in the case of the cancellation of an incoming flight) that do not affect outgoing passenger centralities (although it also includes the cost of connecting incoming passengers, which do affect outgoing passenger centrality).

Passenger centrality provides exact information on the number of modified itineraries, while trip betweenness provides only an estimate, and it also provides more precise estimates of costs with respect to trip betweenness, especially at the single airport level. However, note that it requires knowledge of the scheduled and actual passenger itineraries to be computed, differently from trip betweenness.

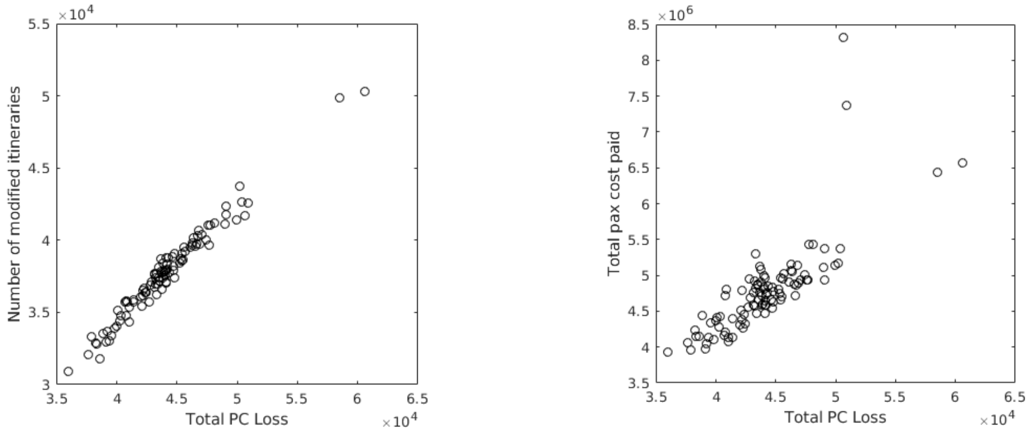


Figure 14: Passenger centrality loss.Left panel: Total passenger centrality loss in the network against total number of passengers with modified itineraries. Each dot represents one iteration; Right panel: Total passenger centrality loss in the network against total passenger-related costs associated to all flights. Each dot represents one iteration.

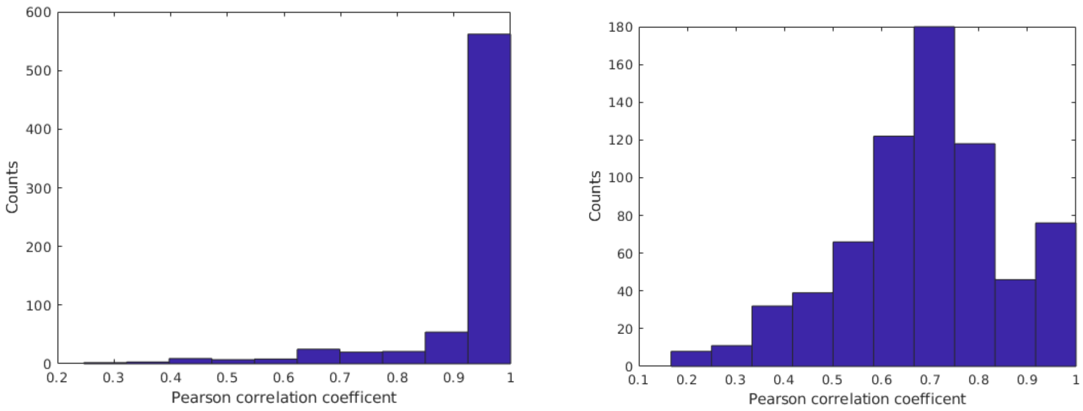


Figure 15: Pearson correlations for passenger centrality loss.Left panel: Histogram of the Pearson correlation coefficients between outgoing passenger centrality loss and cost of outgoing flights, for airports having a significant positive correlation; right panel: Histogram of the Pearson correlation coefficients between passenger centrality loss and cost of outgoing and incoming flights, for airports having a significant positive correlation.

The total trip centrality loss in the network is only weakly correlated with the total number of passengers with modified itineraries and with the total passenger-related costs (ρ=0.2 and 0.25, respectively). This is probably because the itineraries that are considered by trip centrality (i.e. all the feasible itineraries in the network, of any length, with the longer ones weighted less) have little overlap with the real passengers' itineraries. However, at the single airport level the correlations are similar to those of betweenness centrality: we computed for each of the 756 airports having non-zero outgoing trip centrality the correlation between centrality loss and the number of passengers with modified itineraries among those departing from there, and between centrality loss and the cost of flights departing from there. For modified itineraries, we find that 56% of the considered airports have a positive correlation coefficient (all the others have no significant correlation), with histogram shown in Figure 16, left panel. For costs, we find that 65% of the considered airports have a positive correlation coefficient (all the others have no significant correlation), with histogram shown in Figure 16, right panel.

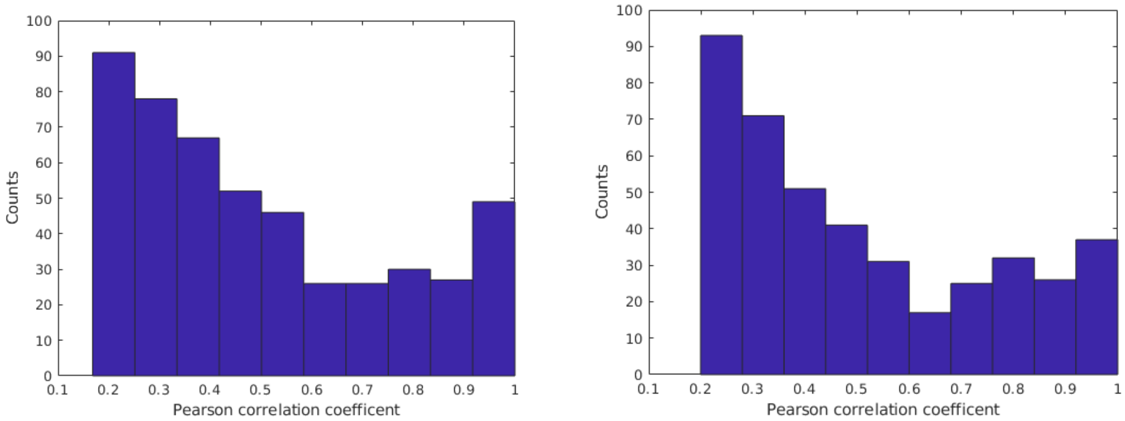


Figure 16: Pearson correlations for outgoing trip centrality loss. Left panel: Histogram of the Pearson correlation coefficients between outgoing trip centrality loss and number of passengers with modified itineraries (among those departing from there) for airports having a significant positive correlation; right panel: Histogram of the Pearson correlation coefficients between trip centrality loss and the passenger-related costs associated with outgoing flights for airports having a significant positive correlation.

5.1.2.2 Causality metrics

In this section, we show how to use the causality analysis from an operative point of view, in particular to answer some possible questions from the perspectives of stakeholders.

The detection of a causal link between two airports can be interpreted as a channel of delay propagation arising from different mechanisms which can be present at the same time: (i) one leg effects mediated by a flight connecting the two airports; (ii) more than one leg effects because of the presence of trips, possibly served by different aircrafts; (iii) some geographical effects due to the proximity of airports, e.g. the weather or the air traffic regulations.

Hence, the network of causal links gives to the *regulator* a global picture of the whole ATM system in relation to the process of delay propagation, but also amplification in the presence of feedback mechanisms represented by some subsystems in which the delay can be transmitted in a loop. Then, in the Domino project, we are interested in the impact of some innovations having on the ATM system, in particular in terms of network effects. The causality analysis can help in determining this impact in relation to the process of delay (or cost of delay) propagation, in particular by highlighting if some innovation tends to disrupt such propagation channels or feedback subsystems, thus increasing the resilience of the system.

So far, we have found that there are three classical network metrics which are of interest in the description of the causality network associated with the ATM system:

1. *link density*, namely the density of causal links with respect to all possible couples of nodes. It captures the average level of causality of the system, i.e. how many propagation channels exist;
2. *link reciprocity*, namely the ratio of the number of links pointing in both directions to the total number of causal links. It is a measure of the likelihood that node-airports in the causality network are mutually linked, thus representing minimal subsystems (i.e. formed by two nodes) of delay amplification, statistically associated with the presence of round-trip flights;
3. *number of feedback triplets*, namely the number of triangles with all links directed clockwise (or anti-clockwise). Similar to reciprocated links, they represent subsystems of three nodes where delay is transmitted in a circle, thus amplified in a loop.

Since the values of the last two network metrics depend on the link density of the network, when comparing some scenario with the baseline of the ABM model, we consider the normalised (w.r.t. the Erdos-Renyi random benchmark) values of such metrics, dividing each value by the expected value of the corresponding metric in the Erdos-Renyi random graph (with the same link density). The obtained value is defined as the *over-expression* of the network metric with respect to the Erdos-Renyi benchmark model.

Since the presence of amplifying feedback mechanisms for delay can be captured by such network metrics, i.e. reciprocated links, triangles, and feedback triplets in the network of causal links among airports, an ATM innovation which tends to disrupt such feedback effects would represent an improvement for the ATM system. Thus, we can quantify the *systemic* impact of an innovation mechanism by measuring the percentage changes of these network metrics from the baseline to the scenario where the innovation is implemented.

Table 7: Network metrics for Granger causality networks applying Bonferroni correction (i) - built with the state of delay of airports as the average delay of the departing flights.

	GC in mean net.	Erdos-Renyi benchmark (GC in mean)	GC in tail net. (Hong et al.)	Erdos-Renyi benchmark (Hong et al.)	GC in tail net. (BiDAR)	Erdos-Renyi benchmark (BiDAR)
Link density	0.004	0.004	0.225	0.225	0.004	0.004
Reciprocity	0.209	0.002	0.17	0.112	0.03	0.002
Feedback triplets	128	0.5	98107	61749	2	0.2

Table 8: Network metrics for Granger causality networks applying FDR correction (i) - built with the state of delay of airports as the average delay of the departing flights.

	GC in mean net.	Erdos-Renyi benchmark (GC in mean)	GC in tail net. (Hong et al.)	Erdos-Renyi benchmark (Hong et al.)	GC in tail net. (BiDAR)	Erdos-Renyi benchmark (BiDAR)
Link density	0.008	0.008	0.352	0.352	0.026	0.026
Reciprocity	0.217	0.004	0.229	0.176	0.04	0.013
Feedback triplets	303	3	322977	238248	282	93

Table 9: Network metrics for Granger causality networks applying Bonferroni correction (ii) - built with the state of delay of airports as the third quartile of the delay distribution of the departing flights.

	GC in mean net.	Erdos-Renyi benchmark (GC in mean)	GC in tail net. (Hong et al.)	Erdos-Renyi benchmark (Hong et al.)	GC in tail net. (BiDAR)	Erdos-Renyi benchmark (BiDAR)
Link density	0.003	0.003	0.218	0.218	0.003	0.003
Reciprocity	0.192	0.001	0.166	0.109	0,029	0.001
Feedback triplets	66	0.1	91981	56254	2	0.01

Table 10: Network metrics for Granger causality networks applying FDR correction (ii) - built with the state of delay of airports as the third quartile of the delay distribution of the departing flights.

	GC in mean net.	Erdos-Renyi benchmark (GC in mean)	GC in tail net. (Hong et al.)	Erdos-Renyi benchmark (Hong et al.)	GC in tail net. (BiDAR)	Erdos-Renyi benchmark (BiDAR)
Link density	0.004	0.004	0.346	0.346	0.024	0.024
Reciprocity	0.192	0.002	0.227	0.173	0.033	0.012
Feedback triplets	114	1	311030	225272	240	70

The introduced network metrics can be also used to characterise the Granger causality networks built with the different statistical tests: Granger Causality in mean, Granger Causality in tail by Hong et al., and Granger Causality in tail by the novel method introduced here for the first time and based on the bivariate generalisation of the DAR(p) process, namely BiDAR(p). The values of such metrics, i.e. link density, reciprocity, and the number of feedback triplets, are shown in Table 7, Table 8, Table 9, and Table 10 in the case of the baseline scenario (0) of the ABM model developed in Domino. In particular, we show the results for both the definitions of the state of delay of an airport, i.e. the average delay of departing flight within one-hour time window or the third (75%) quartile of the distribution of the delay of departing flights, and by considering both multiple testing corrections, i.e. Bonferroni and false discovery rate corrections.

The results suggest that predicting statistically if an airport is congested or not at some future time is much easier than forecasting the state of delay itself, as highlighted by the difference in link density for the networks of Granger causality in mean and in tail. Moreover, the statistical test by Hong et al. leads to a link density much higher than the one obtained with the novel test based on the BiDAR(p) process. Two effects (both present) can be responsible for this discrepancy: (i) the presence of causal relationships which are not detected by the BiDAR(p) test and (ii) a higher rate of false positives for the test by Hong et al. Since the BiDAR test is parametric, whereas the test by Hong et al. does not rely on a specification of the generative model, the latter is in effect more flexible, thus capturing potentially some causal relationships which cannot be described by the BiDAR(p) model. Nevertheless, the test by Hong et al. is not robust in the presence of autocorrelation for the binary state variables. There are several ways to show this. The simplest is to generate the binary data with the following BiDAR(1) model,

$$\begin{cases} Z_{1,t} & = V_t((1-A_t)Z_{1,t-1} + A_t Z_{2,t-1}) + (1-V_t)u_{1,t} \\ Z_{2,t} & = S_t(Z_{2,t-1}) + (1-S_t)u_{2,t}, \end{cases}$$

describing two binary state variables Z_1 and Z_2 which are both autocorrelated, i.e. both V_t and S_t having probability different from zero (randomly chosen in the unit interval), and Z_2 is 'causing' Z_1 because of the non-zero term of interaction described by A_t Bernoulli random variable with probability equal to one-half. Note that Z_1 does not 'Granger cause (in tail)' Z_2. However, let us consider the Granger in tail test by Hong et al. with confidence level equal to 5% for the null hypothesis Z_1 does not 'Granger cause (in tail)' Z_2.

Table 11: Frequency test rejection under null Z_1 does not Granger cause in tail Z_2 with BiDAR(1) as data generating process.

Hong et al. test with BiDAR(1) as generating process	M=2	M=5	M=10	M=15	M=20	M=25	M=30
T=500	0.43	0.42	0.41	0.41	0.37	0.45	0.39
T=1000	0.50	0.49	0.43	0.46	0.51	0.55	0.51
T=2000	0.57	0.54	0.56	0.52	0.47	0.52	0.65

The results are shown in Table 11 for different values of both the time scale parameter M and the length T of the times series. In each case, the frequency of test rejection, i.e. the detection of a (false) causal relationship from Z_1 to Z_2, is about one-half, much higher than the confidence level of the test (0.05), thus displaying a very high rate of false positives when the binary time series are autocorrelated in time. Figure 17 shows the average autocorrelation coefficients for the time series of the state of congestion from the ABM. It is evident that autocorrelation effects are significant and therefore we can expect that the test of Hong et al. detects a significant fraction of false positive causal links in tail.

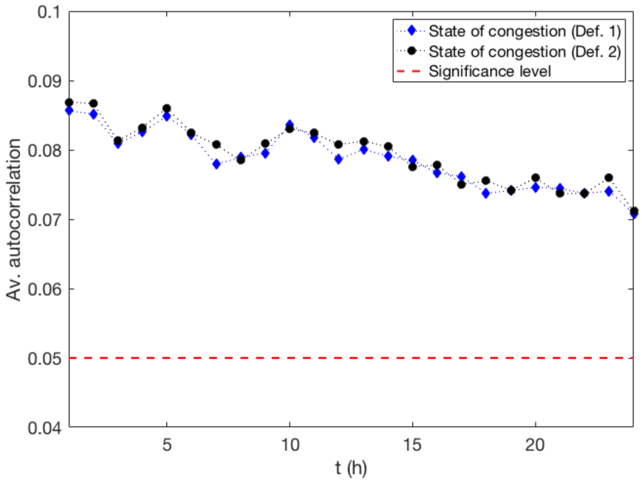


Figure 17: Average autocorrelation coefficients for the time series of the state of congestion (over all airports). The red line represents the bound of significance, meaning that a value larger than the bound is statistically significant.

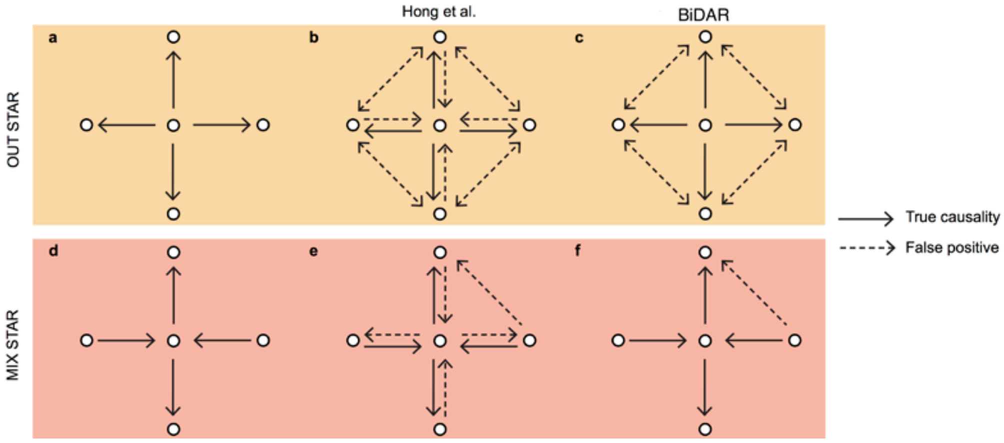


Figure 18: Pictorial representation of two instances of causality networks (left panels), how the true causal links are detected by the Hong et al. test, together with false positives because of both autocorrelation and network effects (middle panels), and similarly for the novel BiDAR causality method (right panels), which, however, solves the issue of false positives, thus decreasing the detection of false reciprocated causal links.

The high rate of false positives for the Hong et al. test can lead to both an increase of link density and an overestimation of the reciprocated links of the Granger causality network. A pictorial representation of this behaviour is shown in Figure 18 for two different types of networks of interaction. Two spurious effects may lead to the false detection of a causal interaction with the Hong et al. test: (i) because of autocorrelated time series, if there exist a directional causal relationship from A to B, it is likely that also the causal interaction in the opposite direction, i.e. from

B to A, is detected; (ii) in the presence of network effects, if A causes B and B causes C, then it is likely that the causal interaction from A to C is detected (even if it is in effect mediated by B) because of the pairwise causality analysis, i.e. we consider couples of nodes, and not the whole system at once. The novel statistical BiDAR test represents a solution for the first kind of spurious detections, see Figure 18. Furthermore, the BiDAR(p) process can be generalised to the multivariate case, thus a new statistical test can be introduced to solve the second issue about spurious detections. However, this is out of the scope of the present analysis.

Since spurious detections of causality interactions by the Hong et al. test induce overestimations of subgraphs such as reciprocated links and triangles, we focus more on the novel BiDAR causality method to detect the channels of propagation of 'extreme' delays within the ATM system. When looking at the subsystems of delay amplification, i.e. reciprocated links and feedback triplets, the causality network built with the BiDAR test displays characteristics similar to those obtained with the other two statistical tests of Granger causality, in particular the over-expression of the network metrics with respect to the Erdos-Renyi random benchmark, i.e. the expected value of such metrics in a Erdos-Renyi random graph with the same link density of the Granger causality network, see Table 7, Table 8, Table 9, and Table 10. This is a further confirmation of the importance of this kind of feedback subsystems for the ATM network. In the analysis of the Domino mechanisms (below), we study what is their impact in terms of either disrupting or preserving such delay amplification network effects.

Moving from the systemic point of view to more local aspects, Granger causality networks can be usefully applied to understand what the interacting subparts of the ATM system are and how strong the interaction is. A directional causal link $i \to j$ in the Granger causality network built with the state of delay of the airports describes the delay propagation from airport$i$ to airport j, whatever the propagation mechanism is: direct flight, two legs effect, and so on (note that the kind of propagation mechanism is not important at this stage, since the causality analysis is in effect a data-driven approach). Nevertheless, the directional causal link can bring more information than just the propagation of delays, for instance it can be correlated to the propagation of the cost of delay.

The relationship between delay and cost of delay cannot be trivially determined, because of many network effects. In fact, small delays at one hub can have a huge impact in terms of costs, e.g. because of missing connections, whereas large delays at some peripheral airports may be less important, e.g., when flights are not involved in passengers' connections. Costs related to missing connections are usually described in terms of one leg effects, i.e. the first flight involved in a connection should land at the arrival airport before the departure of the second flight. Nevertheless, network effects with more than one leg can be also important, e.g. three flights in a row representing some passengers' trip. Hence, the aggregated network of flights (link = one leg effect) between airports could not be enough to characterise the channels of delay and cost propagation. Moreover, only airlines have access to the complete information about the correlation between costs of delays, but only for their own flights. Hence, within the framework of the Domino ABM model, an interesting question from the perspective of each stakeholder without full information about the system is to understand, at least partially, the relationship between the delay propagation (reconstructed by means of DDR files) and the cost payed by airlines at the airports (which is, on the contrary, not accessible in the real world), thus revealing what are the most important nodes of the ATM network and the level of predictability, or, equivalently, how strong the interaction is between two nodes in the process of propagation.

To this end, Granger causality networks built with the states of delay of the airports can be used to test whether two airports connected in the causality network displays some significantly positive correlation for the cost payed by airlines at those airports, and vice versa, i.e. not significant correlation, if not connected. If yes, this is a signal of the causal links in the Granger causality network as channels of propagation of costs, and not only delays.

For the correlation analysis, let us associate to each airport a *cost variable* defined as the 90% percentile of the distribution of the cost of delay of departing flights obtained, within the ABM modelling framework, as the sum of the following components:

1. the cost of compensation for delayed flights at the arrival;
2. the cost of transferring connecting passengers in the case of missed connection;
3. the cost associated with the duty of care of delayed passengers;
4. soft passenger-related costs;
5. non passenger-related costs such as the crew cost.

Then, given a network of causal interactions between airports, we consider the sample Pearson correlation between the cost variables of two airports when:

1. they are *connected* by a causal link in the network;
2. they are *not connected*.

Finally, for each airport we obtain two samples of correlation coefficients, the one for neighbours in the causality network and the one for non neighbours. Thus, we can consider the average Pearson correlation for the two samples, together with the standard deviation as measure of dispersion with respect to the average correlation.

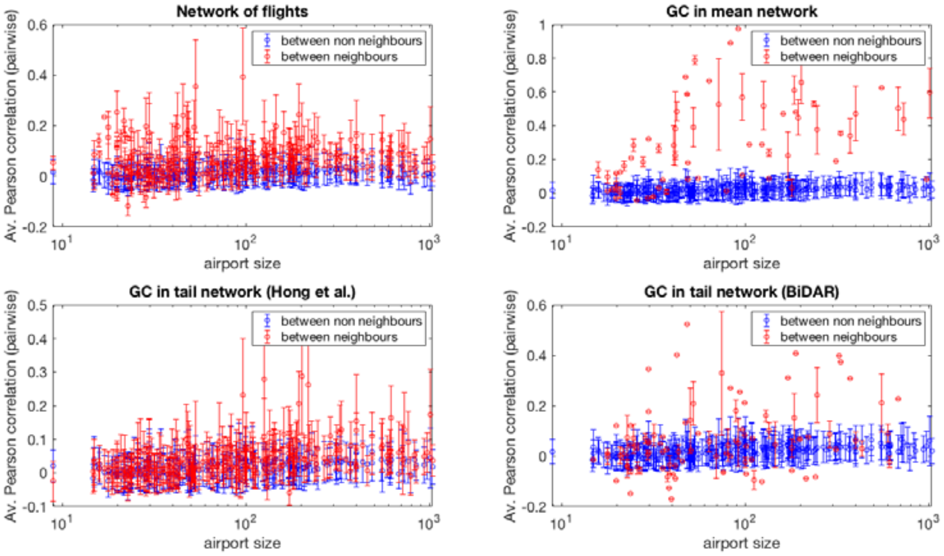


Figure 19: Average Pearson correlation coefficient between the cost variables of an airport and either its neighbour (red dots) or its non-neighbour (blue dots) in the network of flights (top left), the network of Granger causality in mean (top right), the network of Granger causality in tail built with the Hong et al. test (bottom left), and the network of Granger causality in tail built with the novel BiDAR test (bottom right). The airports are sorted according their size, measured as the average number of flights per day. The error bar is the standard deviation associated with the distribution of Pearson correlation coefficients, for the two samples of either neighbours or non-neighbours in the network of causal interactions. The Granger causality networks are built by using the state of delay of airports defined as the third quartile of the distribution of delays of departing flights and the Bonferroni correction is applied.

In Figure 19, the results of the correlation analysis are shown by ordering the airports according to their size measured as the average number of flights per day and for four networks of interactions:

1. the *network of flights*, which displays a link if there exists a direct flight connecting the two node-airports;
2. the causality network built with the *Granger causality in mean* test, applied to the states of delay of airports (defined as the third quartile of the distribution of departing delays of flights);
3. the causality network built with the *Granger causality in tail test by Hong et al.*, applied to the states of congestion of airports;
4. the causality network built with the *Granger causality in tail test based on the BiDAR process*, applied to the states of congestion of airports.

Here, for all Granger causality networks, we correct the increasing size of false positive because of multiple hypothesis testing by using the Bonferroni correction. However, similar results can be obtained by applying the FDR method.

We note that the correlation of the cost of delay payed by airlines are on average larger for connected airports than for non-connected ones, for all considered networks. However, this information cannot be used to select the channels of cost propagation in the case of either Granger causality in tail and the network of flights (one leg effects), because of partial superposition of the two samples of correlation coefficients. On the contrary, Granger causality in mean network displays a significant distinction, thus we can claim with a certain degree of confidence that the causal links are associated with the process of propagation of both delays and costs of delay. Hence, according to these findings, the causality network built with the Granger causality in mean test represents a method to correlate the *observed* delays of flights at some airports with the *unobserved* costs of delay payed by airlines (at the same airports).

5.2 Summary of key results

Most of the analyses have been performed by considering 100 simulations of the model for each scenario. For the centrality analyses, we have considered 50 simulations. The results of each metric have been averaged across these, and in the following discussion, we show these averages. Specifically, as in D5.2, we consider the baseline scenario as a reference point, and both for classical and network metrics we show their percentage change with respect to the baseline. As a robustness check, we have considered subsamples of 50 simulations and compared the results (data not shown): thus the vast majority of the results reported are those that have shown consistent results in the subsamples. When a significant variability between the subsamples is found, we discuss it in the detailed analysis of Section 5.3.

In this section, we have collected the most relevant results in three figures, one per mechanism (4DTA, FP, FAC), each of them composed of five panels. The three top panels report the results for classical metrics, namely, from left to right: delay, costs, and passengers-related metrics (delays, missed connections, re-routings, etc.). The two bottom panels show the results of centrality and causality metrics. For centrality metrics, we consider trip centrality, passenger centrality, and trip betweenness. For causality metrics, we define the state of congestion of the airport as the third quartile of the delay distributions of flights departing from a given airport. We then show metrics related to the new causality (BiDAR) approach, with the false discovery rate (FDR) correction (see Sections 5.1.1 and 5.3) and essentially confirm the main conclusions discussed here.

5.2.1 4D Trajectory Adjustment

The top-left panel of Figure 20, shows that the introduction of the 4DTA mechanism improves the airspace system by making it less affected by delays. This is true for all the displayed quantities, namely the average arrival delay for flights with more than 15 minutes of delay, their fraction, and the reactionary delay (number of flights and amount of delay). The detailed analysis shows that this is consistent across different measures of delay. The top-centre panel shows that with 4DTA, there is a sizeable reduction of excess fuel cost (up to almost 20%). This can be understood by the fact that flights use 4DTA to control, in a more efficient way, their total costs, and the cost of fuel is a significant factor driving part of the solution. Other types of cost are only marginally affected by the introduction of 4DTA, some of them are even increasing (see Section 5.3). Non-passenger delay costs display a small but sizeable decrease. Overall, the costs are reduced by more than 10% when 4DTA is introduced in the system. Also, passenger delays are reduced (see top-right panel). This is much more evident when considering connecting passengers, since they benefit more by the introduction of 4DTA. On the negative side, the fraction of modified itineraries slightly increases, but we can safely affirm that passengers are better off when 4DTA is in operation.

The causality and centrality metrics partly confirm this view. The centrality metrics display a slightly larger loss, possibly in connection with the larger number of modified itineraries. However, the causality metrics show a very significant substantial decrease of density and reciprocity, indicating that the propagation of distress from one airport to the others is much weaker, as well as the two-airport feedback effects. There is, however, a small increase in the feedback triplets. In summary, the introduction of the 4DTA mechanism makes the system better from the point of view of airlines, passengers and the environment (due to reduced fuel consumption). The system is more efficient (from the cost and delay perspective) and more robust to local shocks at airports, which propagate much less.

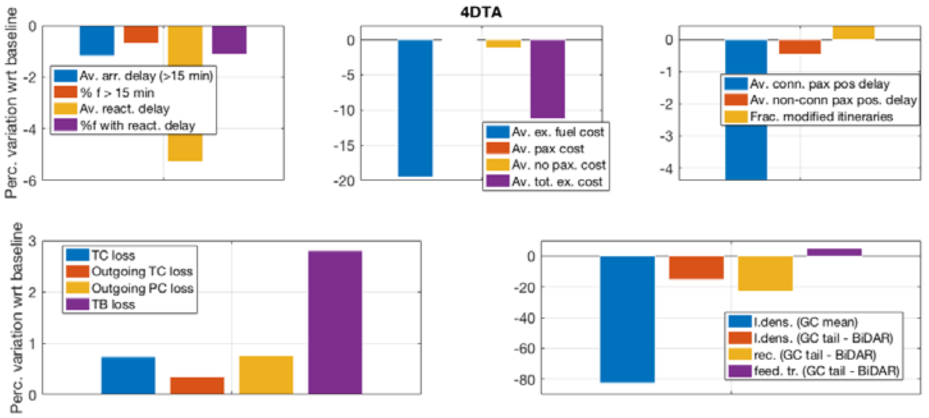


Figure 20: Summary of percentage changes of metrics in 4DTA with respect to the baseline

5.2.2 Flight Prioritisation

The top panels of Figure 21 show the percentage change of classical metrics when FP is implemented at three large airports, namely: LFPG, EGLL, EHAM, where ATFM regulations have been manually issued on the morning traffic. Note that the displayed variations are restricted to the three airports where the FP mechanism is applied during the regulation time periods. The overall picture is that the system is worse off, since all but one metric displays a worsening with respect to the baseline. It is important to note, however, that these variations are quite small (never larger than 1.5%, often much smaller). This suggests that the introduction of FP has little or no (or a slightly negative) impact, at least when measured with classical metrics, when FP is in operation. The more detailed analysis shown in Section 5.3 fully supports this conclusion.

The bottom panels of Figure 21 show centrality and causality metrics for the three airports where FP is implemented. Again, the variations are very small and their sign is not common across the airports. Possibly only EHAM displays an overall benefit through the introduction of FP, but, again, the percentage changes are very small. When the analysis is extended to all airports (i.e., not only the three where FP is implemented but all those modelled in the ECAC region), the percentage variation of all metrics becomes extremely small. In summary, the introduction of FP appears to have essentially little or no effect (or maybe slightly negative) on the system, when considering delay, cost, centrality, and causality. More surprisingly, the same conclusion holds when restricting the analysis to the airports where the FP mechanism is implemented.

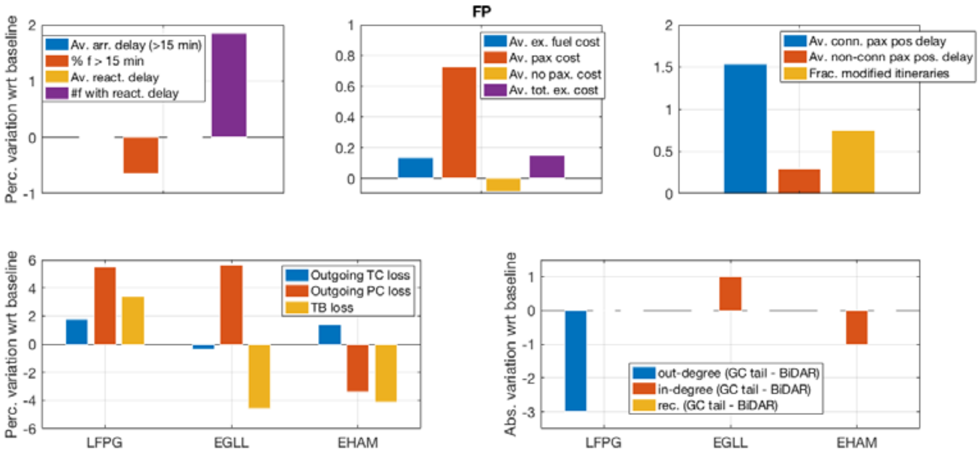


Figure 21: Summary of percentage changes of metrics in FP with respect to the baseline. The percentage variations of the classical delay and cost metrics are restricted to the three airports where the FP mechanisms are applied and we consider all flights landing at any of the three airports within the time window of active regulation.

5.2.3 Flight Arrival Coordination

Flight Arrival Coordination is tested in two different settings, one where the radius of the E-AMAN system is nominal (200 NM), and another, where it is extended (600 NM). Moreover, in the simulations, the FAC is implemented in 24 major airports. Figure 22 shows the results of our analyses. All the displayed results are obtained by restricting the study to the airports where the mechanisms are applied. In particular, we consider only flights landing at an airport of the restricted sample in computing the delay and cost metrics. The left- and right-hand panels in the top part of Figure 22 clearly show that the introduction of FAC increases the delay of flights and passengers. It appears also that the extended radius of the FAC produces larger delays than the nominal range. As a consequence, passenger costs are larger (see the top-central panel). Quite surprisingly, the excess fuel cost is only very slightly smaller (in the nominal radius FAC setting) or even larger (in the extended radius scenario) than in the baseline scenario. Thus, it seems that the introduction of the FAC mechanism makes the system less efficient. This is due to a discrepancy between the E-AMAN planned and actual holding required time, which causes the assignments of additional holding delays to respect the planned landing sequence (see Section 5.3.2). These results are highly robust, as can be seen in the detailed analysis presented in Section 5.3. Generally, the introduction of FAC makes the system worse off for almost all the classical metrics. This conclusion holds even when considering the whole ECAC space, and not only the 24 airports where the mechanism is implemented.

Centrality metrics (bottom-left panel) show small and positive variations, meaning that the introduction of the mechanism makes the centrality loss of these airports larger. This is likely due to the increase of modified itineraries and more generally to the increased delays. The causality metrics are extracted from the network of causal relationships between all the couples of the ECAC airports, but considering only the subgraphs involving at least one airport where the mechanism is implemented. Here the outcome is mixed and not clear: while the introduction of FAC in the nominal radius E-AMAN makes the system slightly worse off, in the extended radius scenario the system becomes significantly less connected, from a causal point of view, both in terms of the number of causal links and of feedback effects (reciprocated links and triplets). This could be explained by the fact that FAC increases the arrival delay of flights independently of their departure delays, thus masking the causal relationships due to network effects. In summary, the introduction of the FAC mechanism appears to make the system worse off from the point of view of airlines and passengers, as well as regarding the environment. With the exception of causality, all the metrics are worse for the extended range, than for the nominal range.

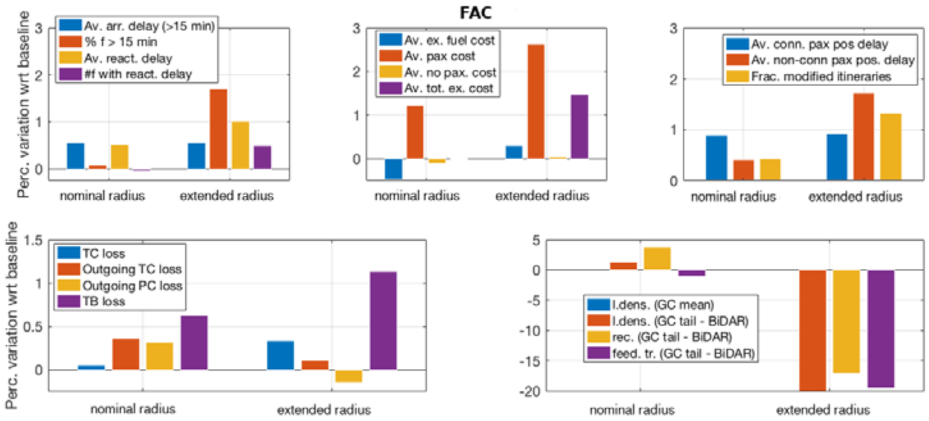


Figure 22: Summary of percentage changes of metrics in FAC with respect to the baseline. The percentage variations of both classical and centrality metrics are restricted to the airports where the FAC mechanisms are applied. In particular, we consider only flights landing at any airport of the restricted sample in computing the delay and cost metrics. The causality metrics are extracted from the network of causal relationships between any couple of the ECAC airports, but considering only the subgraphs, i.e. the reciprocated links or the feedback triangles, involving at least one airport where the FAC mechanism is implemented.

5.3 Detailed analysis

5.3.1 Hub delay management

5.3.1.1 Classical metrics: delay and costs

The percentage variations in flight delay metrics and in cost-related metrics in the advanced scenarios with respect to the baseline are shown in Figure 23 and Figure 24. The value of the metrics and their interquartile range are reported in Annex I (Section 9), Table 12 and Table 15.

For the 4DTA advances scenario, all flight delay metrics show improvements with respect to the baseline, especially with a large percentage reduction in the number of flights with large delays (>60 and >180 mins), both in departure and arrival. This suggests that the possibility to adjust speed is useful to partially absorb large delay. This might be linked to expected high cost of missed connections which might be compensated with small partial reduction of large delays. For the same scenario, we see a percentage reduction in the total excess cost of around 10% with respect to the baseline, mainly driven by a large reduction of the extra fuel costs (partly counterbalanced by an increase in curfew costs and soft costs).

The large percentage reduction (~20%) in extra fuel costs is due to the possibility to slow down to save fuel when a flight is estimated to arrive early (as estimated at the top of climb) with respect to their schedule. In summary, the 4DTA mechanism is effective in reducing both average delays and costs, by speeding up to recover part of the delay and by slowing down to save fuel when the flight is ahead of schedule.

In the FP scenario, instead, we see small and negative variations in the delay metrics (except for a small improvement in the tail of delays >180 mins), suggesting that the FP mechanism is too local to see significant changes at a whole system level on the delays. The total excess cost has a small percentage increase with respect to the baseline, driven by the increase of transfer costs, soft costs and crew costs.

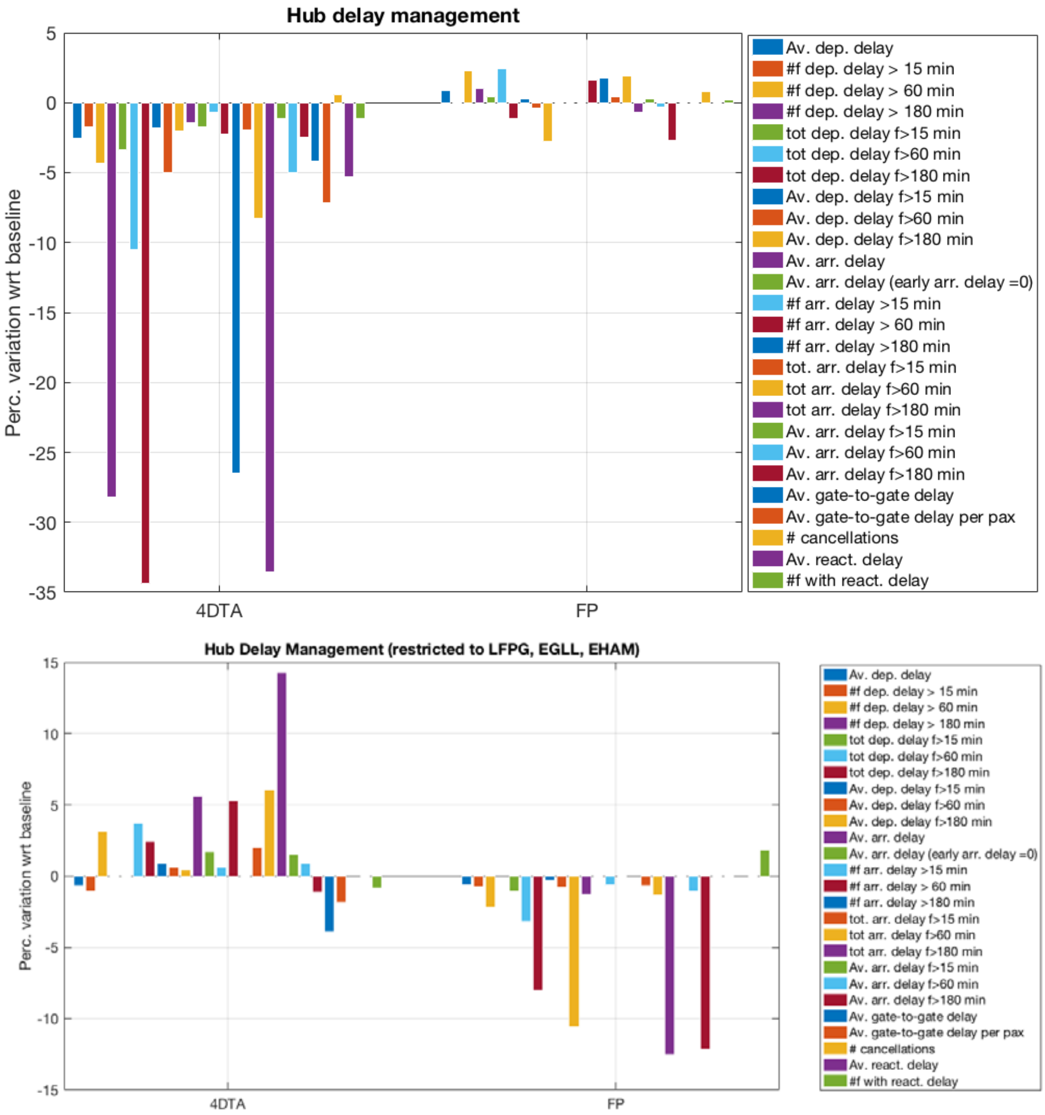


Figure 23: Change in delay metrics HDM.Top: percentage change in classical delay metrics in HDM advanced scenarios with respect to the baseline. Bottom: percentage change in classical delay metrics in HDM advanced scenarios with respect to the baseline, on a restricted sample (see text).

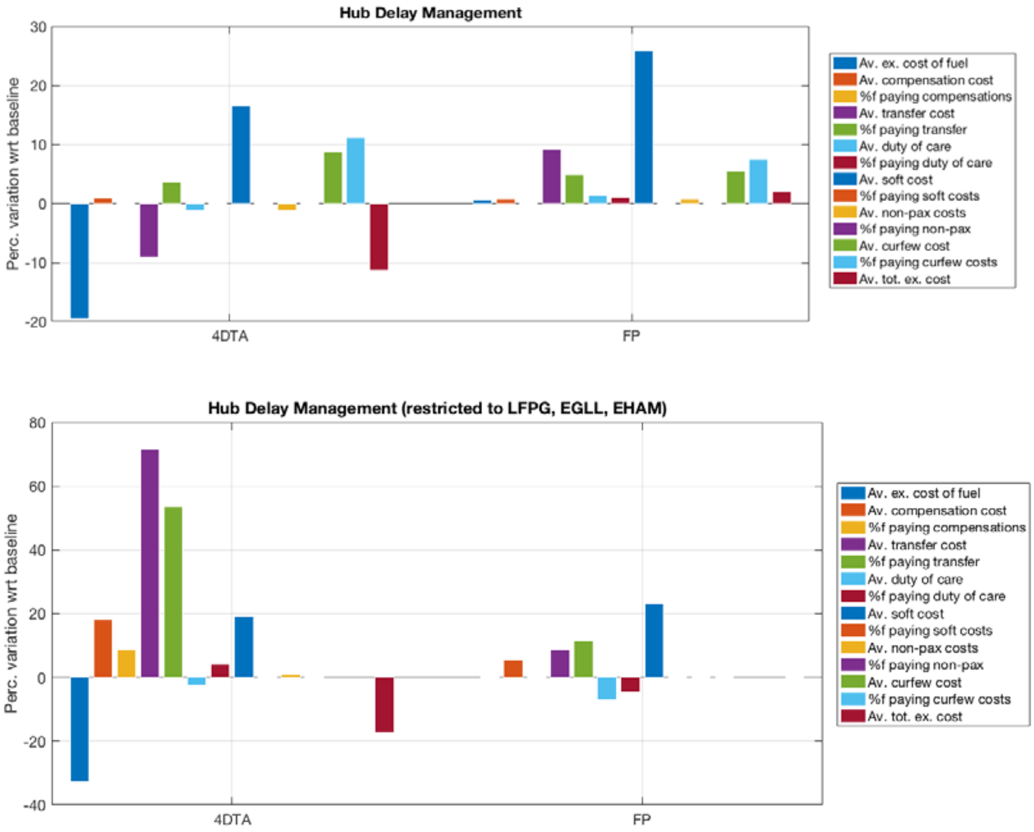


Figure 24: Change in cost metrics in HDM. Percentage Top: percentage change in classical cost metrics in HDM advanced scenarios with respect to the baseline. Bottom: percentage change in classical cost metrics in HDM advanced scenarios with respect to the baseline, on a restricted sample (see text).

5.3.1.2 Passengers-related metrics

The percentage variations in passengers-related metrics in the advanced scenarios with respect to the baseline are shown in Figure 25. The value of the metrics and their interquartile ranges are reported in Annex I (Section 9), Table 13.

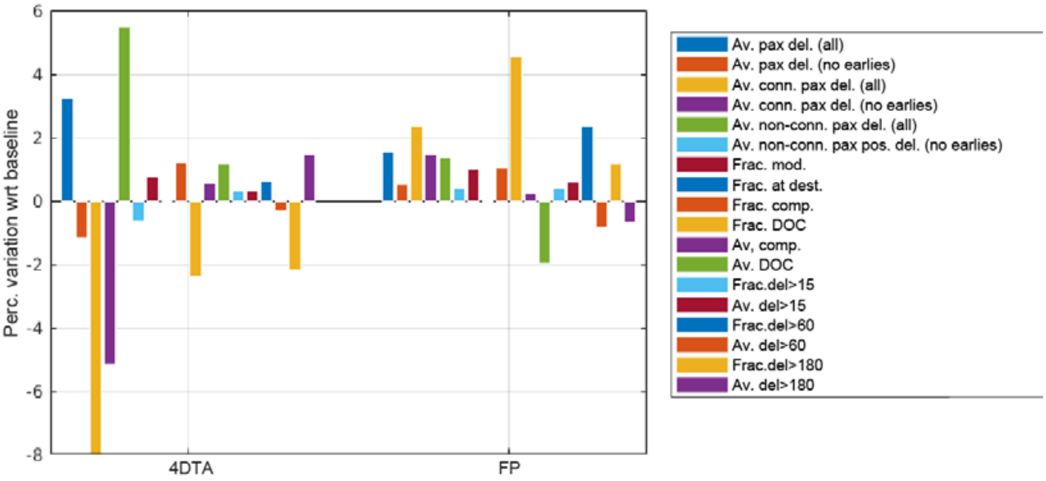


Figure 25: Change in passenger-related metrics in HDM.Percentage change in passengers-related metrics in Hub delay management advanced scenarios with respect to the baseline.

The most notable change concern the passengers' delays in the scenario with 4DTA implemented: while the average delay of all passengers (including those arriving early) increases, the average delay of passengers excluding the earlies decreases. The increase in the delay of all passengers is due to the fact that flights landing early reduce this early arrival with respect to the baseline, which also explains the similar increase in the average delay of all non-connecting passengers. When early arrivals are excluded, the reduction of the average passenger delay shows the effectiveness of the mechanism in reducing positive delays. Connecting passengers, in particular, are affected positively, with an average decrease of 8% in their average delay. In the same scenario, a reduction of the fraction of passengers receiving duty of care and a smaller increase of the average duty of care produce an overall improvement, and the same is true for the large delays (>180 min). Therefore, in this scenario we see a general improvement from the passenger point of view, especially for connecting passengers. If we restrict the analysis to passengers that are departing from or landing at one of the three manually regulated airports during a regulation (see Figure 26 and Table 14), we still find that connecting passengers are positively affected, but in general passengers' delays increase, as do non-connecting passengers delays. This is probably due to the fact that the regulations create several departure delays, and in the advanced scenario they are only recovered when it is advantageous cost-wise (e.g., if there are connecting passengers missing their connections due to the delay). As a result, connecting passengers are better off than in the baseline but non-connecting passengers are worse off.

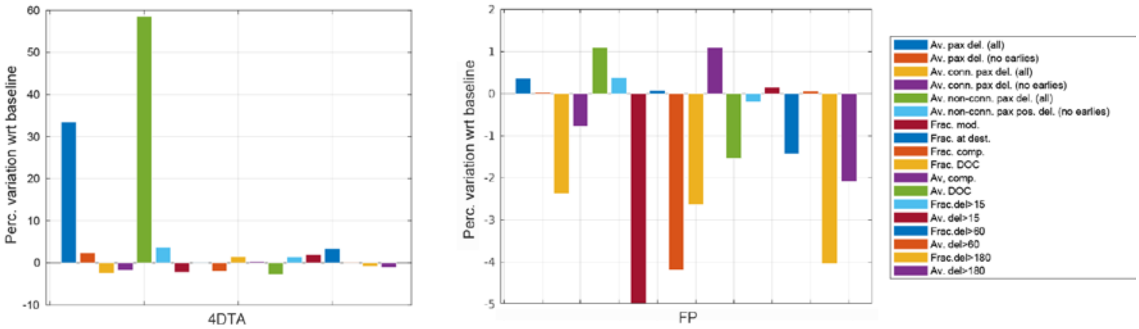


Figure 26: Change in passenger metrics in HDM restricted sample. Percentage change in passengers-related metrics in the advanced scenarios (left panel: 4DTA, right panel: FP) with respect to the baseline, on a restricted sample (see text).

The improvements for connecting passengers can be due to a more efficient use of the 'wait for passengers' option or to a speeding up of the flights with connecting passengers on board. Regarding the latter factor, we verified that the flights carrying at least one connecting passenger choose slightly higher percentage speeds than flights that contain no connecting passengers when applying 4DTA to speed up (see Figure 27), therefore this is one factor of the improvement for connecting passengers. We also note that the use of wait for passengers is quite different in the two scenarios. In fact, although a similar number of flights considers applying the wait for passengers option in the two scenarios, only 15% of them actually apply it in the advanced scenario, against 27% in the baseline, and the average wait is shorter (3.8 minutes against 6.7 in the baseline). This suggests that the strict rule used in the baseline scenario (i.e., always waiting for any flex passengers if the wait is below 15 minutes) is not efficient cost-wise when compared to the more flexible strategy of the advanced scenario. In fact, flex passengers (7.7% of the 3.4M passengers) are widely spread over flights (67.9% of flights have at least one flex passenger) and often they are connecting passengers (21.6% of them), this leads to a significant large use of wait for passengers in the baseline. This should be adjusted as part of a future calibration. The strategy used in the advanced scenario, evaluating the cost of waiting for passengers, leads to a more moderate use of the option, which however seems more efficient both for costs (as seen in the previous paragraph) and for connecting passengers.

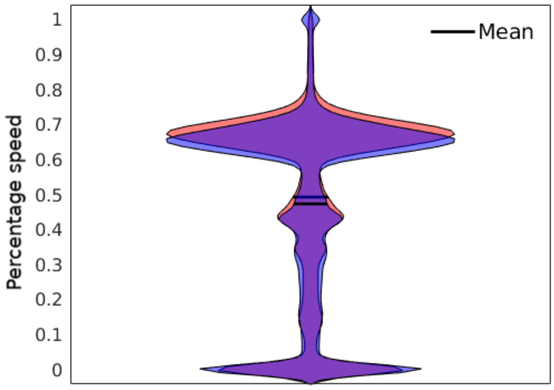


Figure 27: Violin plot of the percentage speed selected when applying 4DTA at TOC by flights with at least one connecting passenger (orange) and by flights not carrying any connecting passenger (purple).

In the FP scenario, instead, we have a small overall worsening: delays have a small increase, and more passengers receive duty of care, on average larger. If we restrict the analysis to passengers that are landing at one of the three manually regulated airports during a regulation, we see improvements in several metrics (see right panel of Figure 26 and Table 14). However, these results are not consistent if we compare two independent sets of 50 realisations, therefore they might be just random fluctuations. In fact, they are based on a set of itineraries much smaller than the full dataset (whose results are instead consistent across the two sets of 50 independent realisations).

5.3.1.3 Centrality metrics

The percentage variations in centrality metrics in the advanced scenarios with respect to the baseline are shown in Figure 28. The average losses of all three centrality metrics show a small percentage increase with 4DTA and FP. This is consistent with the increase in the fraction of passengers with modified itineraries.

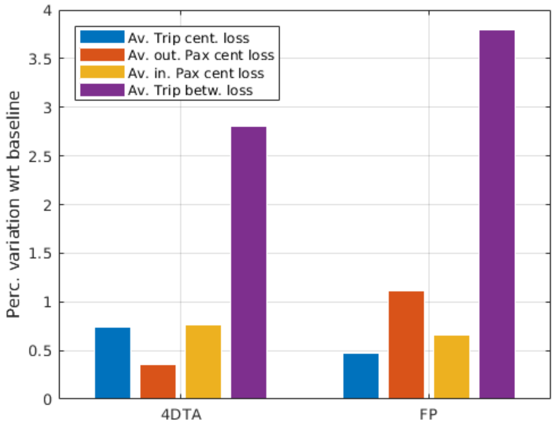


Figure 28: Percentage change in centrality metrics in hub delay management advanced scenarios with respect to the baseline.

For the FP scenario, we also focus on the three airports in which extra manual regulations are applied: EGLL, EHAM and LFPG, for which Figure 29 shows the percentage change in average centrality losses with respect to the baseline. The metrics do not offer a univocal interpretation of this change, as some show an improvement and some a worsening. For example, trip betweenness shows that, when FP is applied, there is a lower probability of having a disrupted itinerary when connecting in EGLL or EHAM, but a higher one when connecting in LFPG. As trip betweenness considers also possible itineraries that are not used by passengers on that particular day, this might not be related to what happens to the passengers travelling on September 12th. In fact, the outgoing passenger centrality, whose loss accounts for real missed connections and cancellations in the considered airport and downstream for the passengers' itineraries of September 12th, disagrees with trip betweenness in EGLL, where it shows an improvement. This is probably because the disrupted itineraries connecting in EGLL that are counted by trip betweenness were not used by passengers on that particular day, and therefore imply no loss of outgoing passenger centrality. These results on individual airports, however, cannot be used to draw general conclusions on the effects of the FP mechanism on these three airports, because they are not statistically significant. In fact, if we perform the analysis on two independent sets of 50 iterations of the model, we find results that are not consistent.

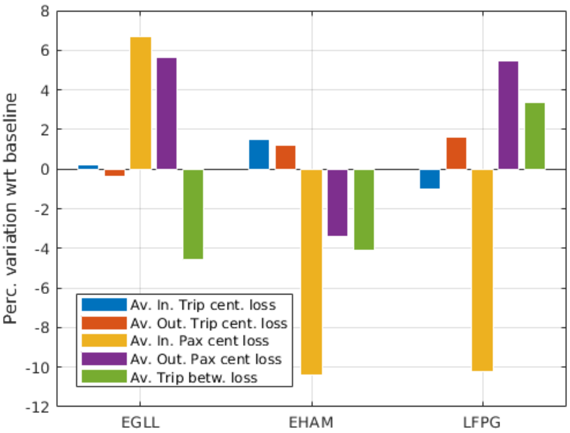


Figure 29: Percentage change in centrality metrics of the three airportssubject to manual regulations in the FP scenario with respect to the baseline.

5.3.1.4 Causality metrics

The percentage variations in causality metrics in the advanced scenarios with respect to the baseline are shown in Figure 30 and Figure 31. The percentage decrease (on average) of the link density in the Granger causality in mean network for the 4DTA scenario with respect to the baseline is consistent with the decrease of departing delays of flights. Moreover, the slight increase of the same network metric in the FP scenario is consistent with the increase of the departing delay, even if this effect is smaller if compared with the 4DTA mechanism.

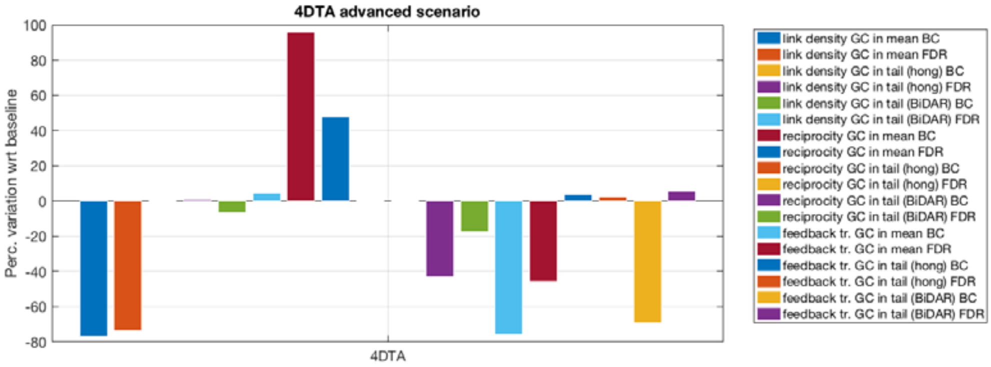


Figure 30: Percentage change of link density, and over-expression in hub delay management (i) - advanced scenario with 4DTA with respect to the baseline of both reciprocity and the number of feedback triplets of the Granger causality networks. The networks of Granger causality in mean and in tail, for both Hong et al. and BiDAR methods, are built with the states of delay of airports defined as the third quartile of the distribution of departing delays of flights. We consider both corrections for multiple hypotheses testing, namely Bonferroni correction (BC) and false discovery rate (FDR) correction.

In the 4DTA advanced scenario, the percentage variations of all network metrics, i.e., link density, reciprocity, and the number of feedback triplets, for the Granger causality in tail network built with the Hong et al. test are really small, thus we can consider them as not significant. These findings together with the significant percentage variations of the metrics for the Granger causality in tail network built with the novel BiDAR test suggest that the much larger link density of the Hong et al. case (w.r.t. the other two causality networks) can be in effect explained by the presence of a high number of false positives, which are in effect random noise. Hence, when moving from the baseline to the advanced scenario, any variation can be explained in terms of random changes because of the changing of link density, thus resulting zero or really small variations of the over-expression of network metrics with respect to the Erdos-Renyi random benchmark. Similarly, for the FP advanced scenario.

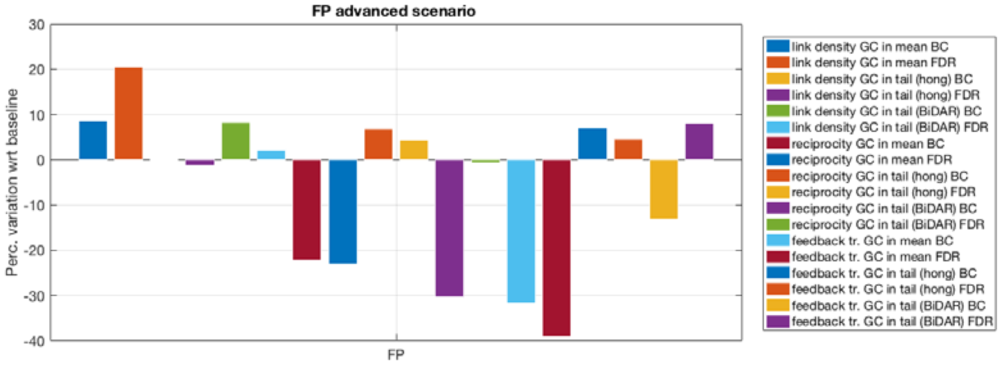


Figure 31: Percentage change of link density, and over-expression in hub delay management (ii) - advanced scenario with FP with respect to the baseline of both reciprocity and the number of feedback triplets of the Granger causality networks in Hub delay management advanced scenario with FP implemented with respect to the baseline. The networks of Granger causality in mean and in tail, for both Hong et al. and BiDAR methods, are built with the states of delay of airports defined as the third quartile of the distribution of departing delays of flights. We consider both corrections for multiple hypotheses testing, namely Bonferroni correction (BC) and false discovery rate (FDR) correction.

In the 4DTA scenario, Granger causality in mean networks (in both BC and FDR cases) display an increase of the over-expression of reciprocity, but a decrease when considering the number of feedback triplets. On the contrary, in the case of Granger causality in tail by BiDAR test, we find the significant decrease of both metrics. Hence, the 4DTA mechanism tends to reduce the average level of causality, i.e. the propagation of delays or congestions, and, at the same time, is able to disrupt the feedback subsystems of delay amplification, with the exception of reciprocated causal links in the Granger causality in mean case, which are, on the contrary, preserved by the mechanism.

Finally, in the FP advanced scenario, despite the slight increase of the level of causality (measured by Granger causality both in mean and in tail), the percentage variations of the over-expression of the network metrics are negative (if we do not consider the Hong et al. case, which is largely affected by random noise). Thus, even if the effect is smaller if compared with 4DTA, these findings support the possibility that FP tends to disrupt some feedback subsystems of delay amplification.

5.3.1.5 Conclusions regarding hub delay management

- 4D trajectory adjustments are much more efficient at a network level than flight prioritisation; they have a greater impact on flight, passenger, centrality and causality metrics.
- 4D trajectory adjustments are efficient at reducing costs for the airline, as well as the average delays for flights; connecting passengers benefit greatly from this, but non-connecting passengers see their arrival delay increase. Hence, there is a trade-off between airline economic efficiency and (some) passenger utility.
- Centrality tends to worsen at the network level with 4D trajectory adjustments, however, the airports under stress seem to have various losses or improvements, depending on their particular case.

Causality tends to decrease with 4D trajectory adjustments, with some exceptions concerning reciprocity, i.e. two-legged feedback loops. This might be due to the wait-for-passenger mechanism, impacting single aircraft that need to go back to their base.

- Causality for extreme events (in tail) is relatively unaffected by 4D trajectory adjustments in general.
- Flight prioritisation tends to increase causality in general, even though it tends to destroy feedback loops.

5.3.2 Effect of E-AMAN scope on arrival manager

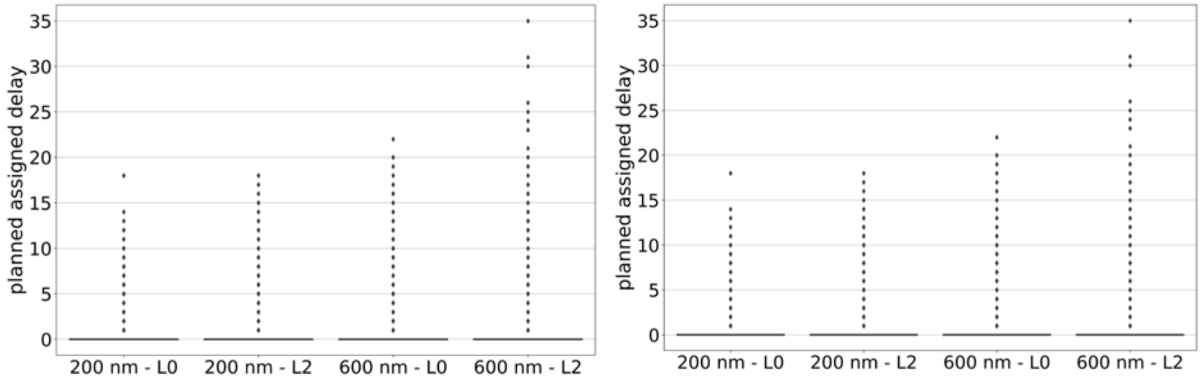
5.3.2.1 E-AMAN mechanism behaviour

There are two moments when flights are issued delay when approaching an airport issued with an E-AMAN, when the flight enters the E-AMAN:

1. planning horizon;
2. tactical horizon.

In the planning horizon, the slot assigned to the flight considers the slots which are not assigned yet and an optimisation criteria (arrival delay for E-AMAN implementation at Level 0, total expected cost for implementation at Level 2). Part of this delay can be absorbed by reducing the cruise speed producing some fuel saving.

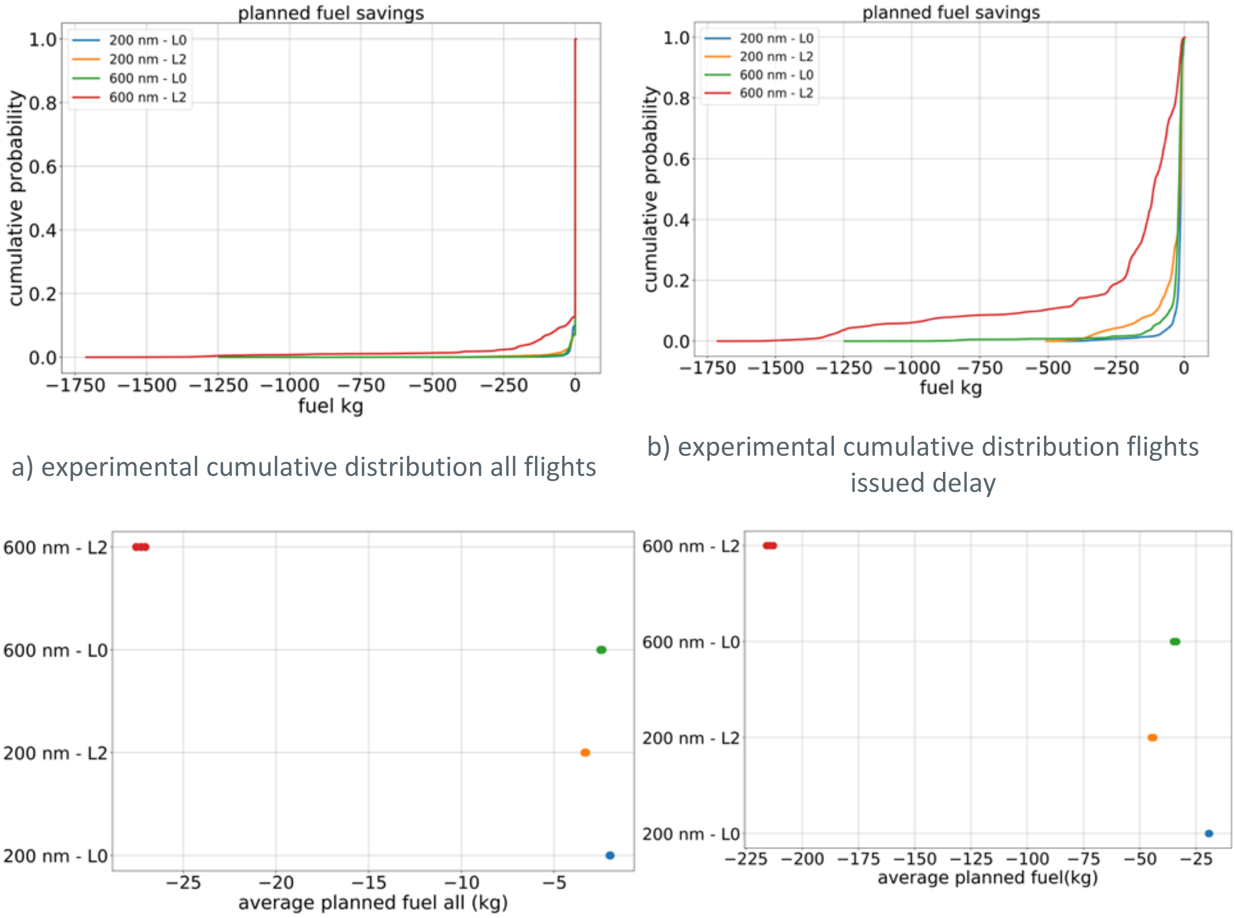
When a flight enters the tactical horizon, the final slot is assigned to the flight and the issued delay will be performed as holding.



a) experimental cumulative distribution

b) boxplot

Figure 32: Delay assigned at planning horizon.



a) experimental cumulative distribution all flights

b) experimental cumulative distribution flights issued delay

c) 95% confidence interval average fuel all flights

d) 95% confidence interval average fuel flights issued delay

Figure 33: Planned extra-fuel expected at planning horizon for flights issued with delay (fuel savings).

As shown in Figure 32, the delay issued at the planning horizon tends to be small and similar on all options. Larger E-AMAN radius and Level 2 implementation assign, in some cases, large delays (e.g., delays of over 20 minutes). This is, as shown in Figure 33 due partially because the E-AMAN in Level 2 is trying to minimise the expected total cost, and in some cases, some delay might represent fuel savings as the speed is reduced. In the baseline implementation (Level 0) as the radius increases from 200 to 600 NM, the fuel that is potentially saved increases, as there is more distance to absorb delay: total fuel save by slowing down at 200 NM Level 0 in average is 19.6 t (2 kg per flight), at 600 NM in Level 0 in average the total fuel is 24.2 t (2.5 kg per flight). In Level 2, as the focus is on the total expected cost, fuel plays a more relevant role and even with a radius of 200 NM the expected fuel savings are larger than for the baseline implementation at 600 NM: at 200 NM Level 2, the total fuel saved is 32.7 t (average of 3 kg per flight), at 600 NM Level 2 the total fuel is 269.0 t (average of 27kg per flight). Here we can see the importance of the cost function as it prioritises the total cost. This means that earlier slots might not be assigned originally so that fuel can be saved.

Figure 34 represents the cost functions for a given flight when entering the planning horizon of the E-AMAN. In this case it can be observed the importance of cost of fuel which is higher than the other expected costs of delay.

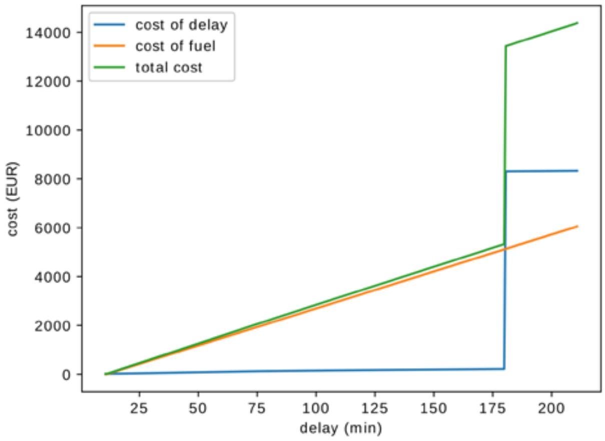
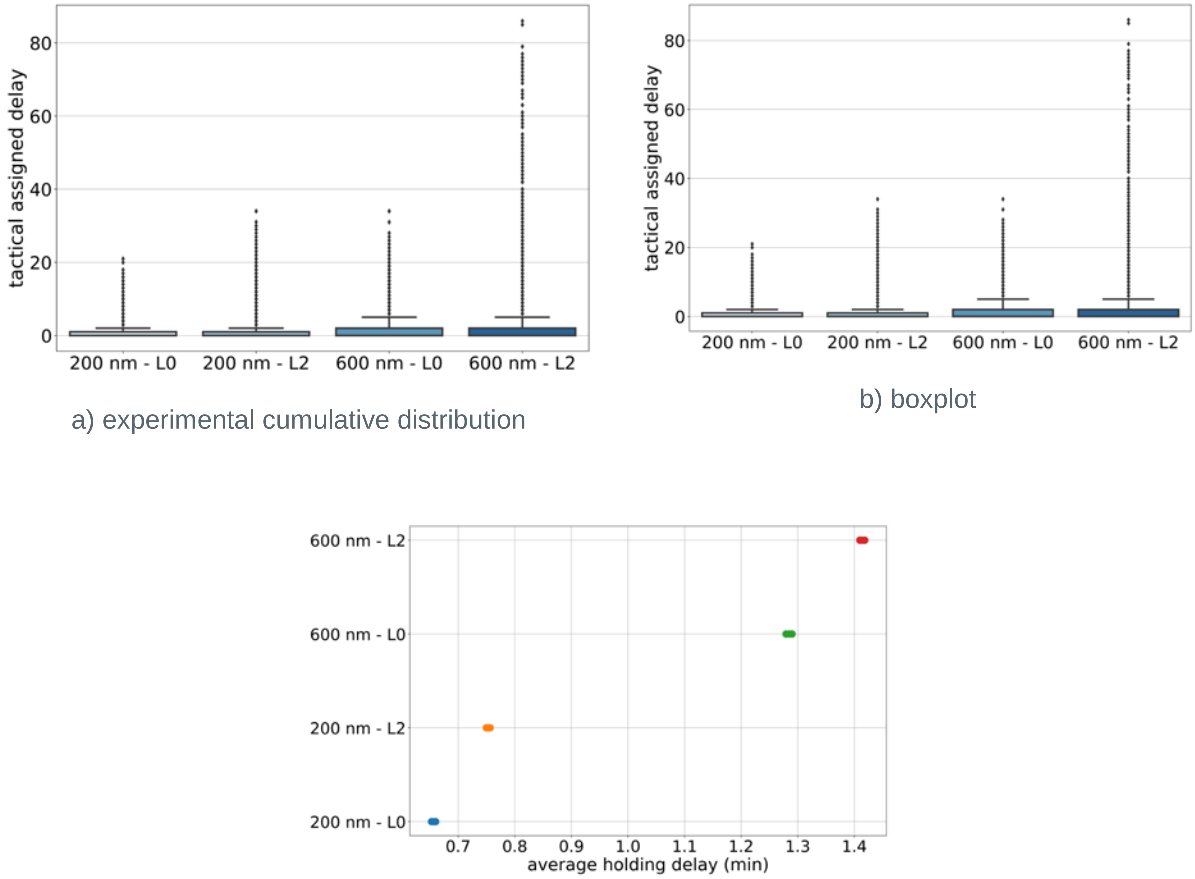


Figure 34: Cost of delay for a given flight on entering E-AMAN planning horizon, considering cost of fuel and other cost of delay (pax and non-pax related).

The delay assigned at planning stage is, however, not fixed as there is uncertainty on the trajectory of the flights within the E-AMAN, but specially on the forthcoming demand. The planned landing sequence is broken due to lack of managing uncertainty (within flights and with new flights appearing). As the system is implemented, each time a flight enters one of the radii, the sequence of landing is optimised. This means that there is no capacity reserved for new arrivals which are not already in the E-AMAN scope. As shown in Figure 35 this leads to significant amount of delay that is assigned tactically (and performed as holding).



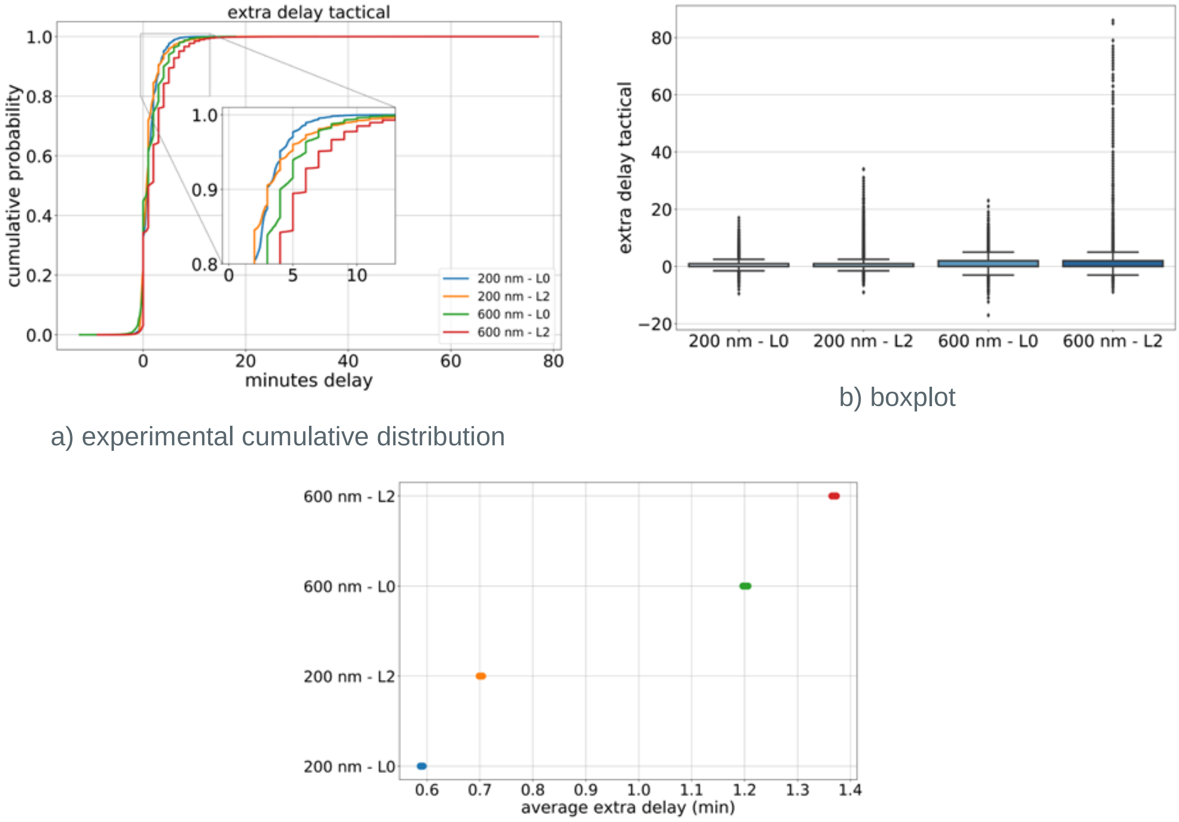
a) experimental cumulative distribution

b) boxplot

c) 95% confidence interval average holding delay

Figure 35: Delay issued at tactical horizon (as holding delay).

Figure 35 also presents how the length of the radius is the major contributor to the amount of delay issued at the tactical horizon. This is due to the fact that the larger the radius the higher the uncertainty and therefore the more changes will be produced between the planned assigned slot and the final assigned one. Note also how implementation at Level 2 tends to issue higher delays than the same radius implementation at Level 0. This is due to the fact that Level 2 is considering the total expected cost of all flights within the E-AMAN and might consider that a swap might be beneficial while at Level 0, as the objective is the arrival delay, any two flights are equivalent for the system.



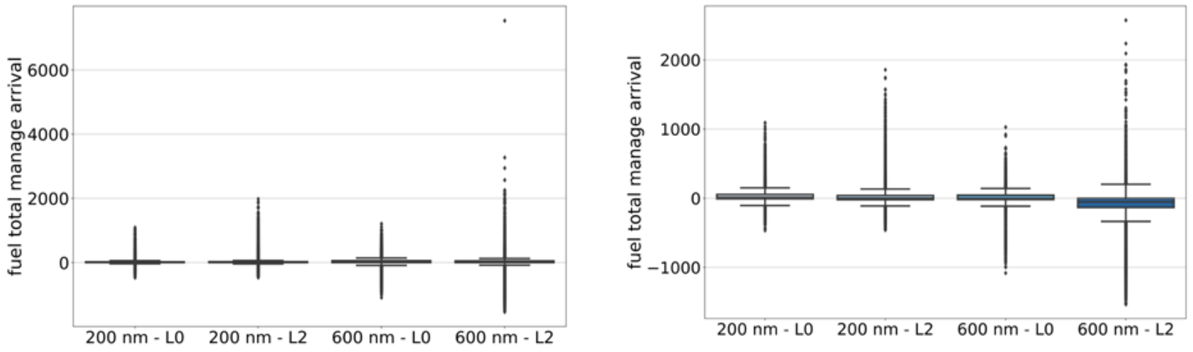
a) experimental cumulative distribution

b) boxplot

c) 95% confidence interval average extra delay

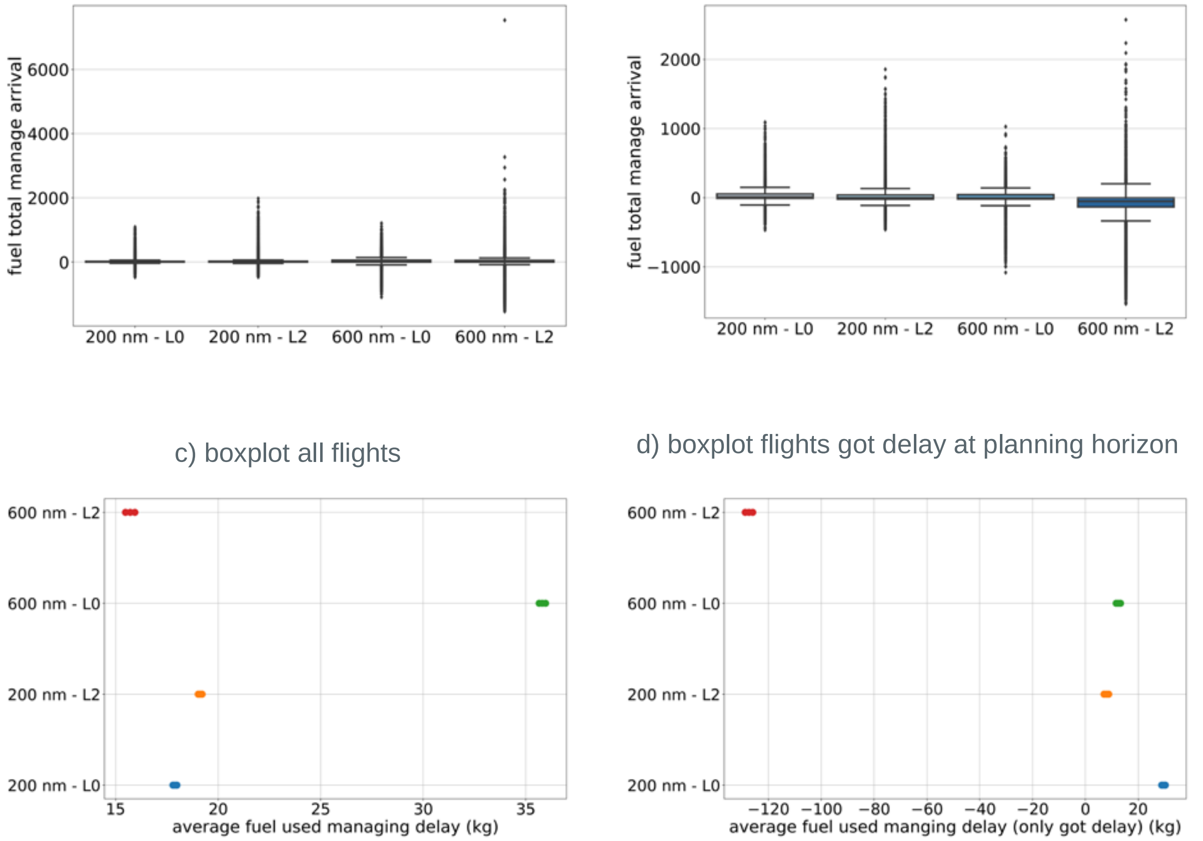
Figure 36: Extra delay required between delay assigned at planning and tactically.

These differences between planned and executed are captured in Figure 36. The difference between the delay that has been issued at the planning horizon and the delay required at the tactical horizon added to the delay that has been absorbed by slowing down are presented. As described, the larger the radius, the larger this inefficiency which is higher for Level 2.



a) experimental cumulative distribution all flights

b) experimental cumulative distribution flights got delay at planning horizon



c) boxplot all flights

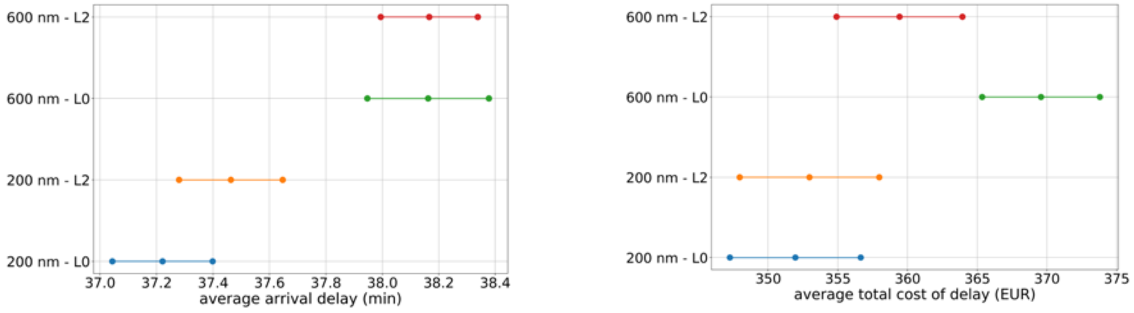
d) boxplot flights got delay at planning horizon

e) 95% confidence interval average fuel all flights

f) 95% confidence interval average fuel flights got delay at planning horizon

Figure 37: Fuel required to manage arrival sequence (savings as slowing down and holding).

Finally, these delays are traduced on fuel required to perform those delays as presented in Figure 37. It can be observed that the implementation at Level 2 leads to higher variation on the fuel used per individual flight, but that at 600 NM with Level 2, some fuel reductions are observed in average with respect to the baseline at 200 NM.



a) average arrival delay

b) average cost of delay

Figure 38: 95% confidence interval: average delay and cost of delay per flight

Besides usage (and cost) of fuel, the delay and cost of delay due to passenger and non-passenger cost increase in average when the radius of the E-AMAN is increased but not significantly as observed in Figure 38.

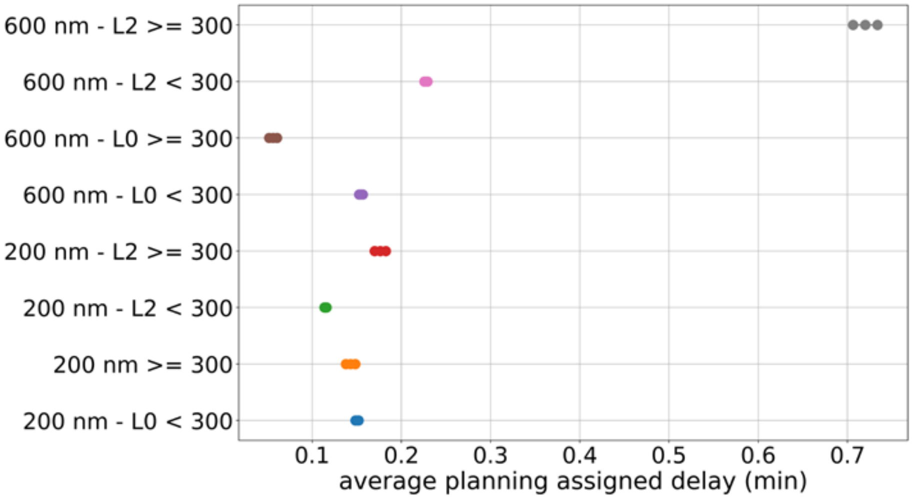


Figure 39: 95% confidence interval: assigned delay at planning radius per aircraft size

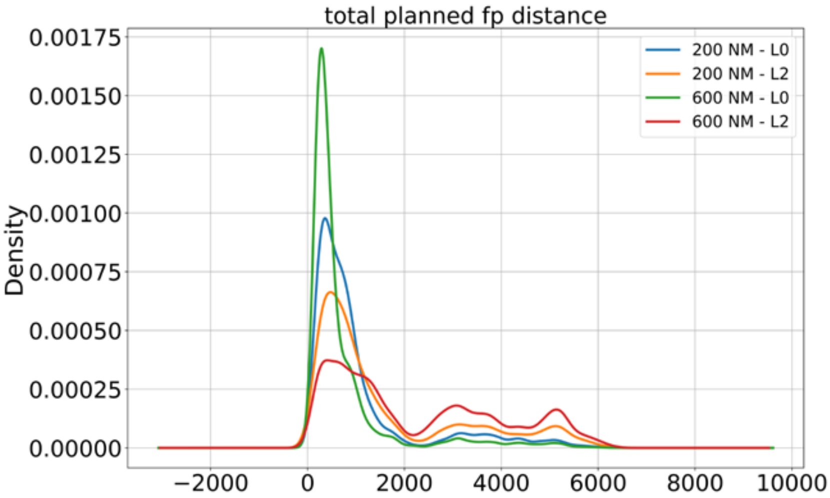


Figure 40: Flight plan distance (NM) of flights with delay issued at planning horizon

An interesting phenomenon is that in the baseline implementation (Level 0), as the radius increases, larger aircraft (with more than 300 pax) tend to get smaller delay assigned at planning horizon. This is consistent with the fact that larger planes tend to cover larger distances, therefore, when the radius is increased, they enter the scope of the E-AMAN before and more slots (earlier) are available. However, in Level 2, this is not the case, as those flights seem to be assigned larger delays as they can absorb delay and smaller flights might have connecting passengers with higher cost of delay and are hence prioritised. This is shown in Figure 39. As shown in Figure 40, the implementation of the mechanism at Level 2 tends to assign delay to flights which have larger flight plans.

5.3.2.2 Classical metrics: delay and costs

The percentage variations in flight delay metrics and in cost-related metrics in the advanced scenarios (Level 2) with respect to the baseline (200 NM with Level 0) are shown in Figure 41 and Figure 42. The value of the metrics and their interquartile range are reported in Table 16 and Table 19. From Figure 41 it emerges clearly that average flight delays increase when the FAC mechanism in Level 2 is active, and the increase is larger when the radius is larger. This remains true when we restrict the analysis to all flights landing in the airports with the FAC mechanism implemented (see Figure 43), in which case the increase in average delays is even larger (note the different vertical scale in the plot). The large percentage increase of the gate-to-gate delay (~18% in the restricted case) signals an increase of the flight time for flight landing at E-AMAN airports. The average excess cost also slightly increases, both on the entire system and on the restricted sample (see Figure 44). Interestingly, as pointed out before, the holding fuel cost does not decrease (note that the saving due to slowing down are not considered here) nor do the passenger-related costs.

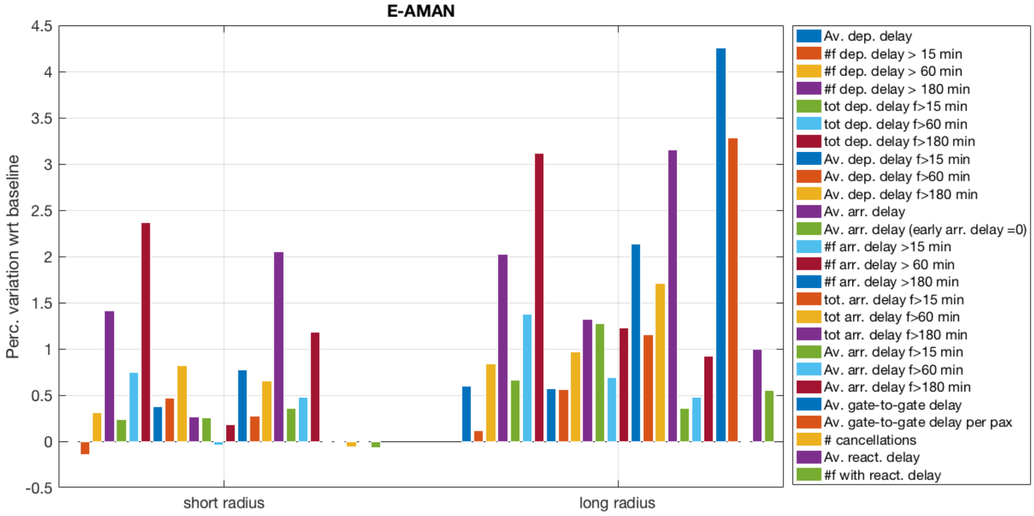


Figure 41: Percentage change in classical delay metrics in E-AMAN (i) - advanced scenarios with respect to the baseline.

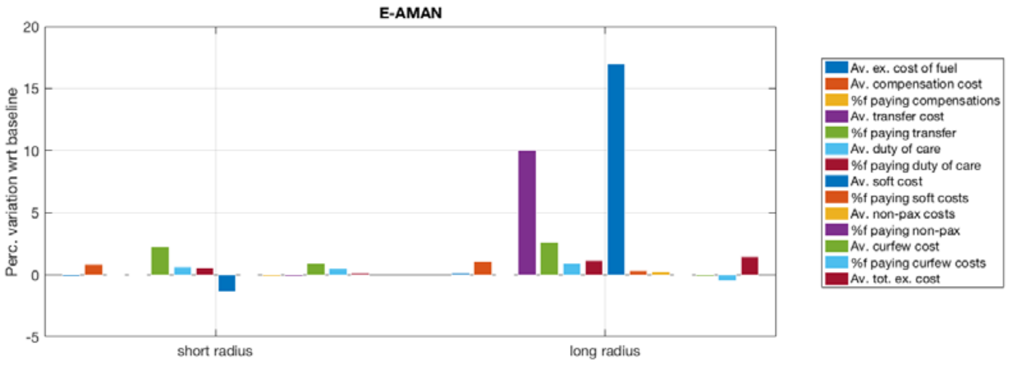


Figure 42: Percentage change in classical cost metrics in E-AMAN (i) - advanced scenarios with respect to the baseline.

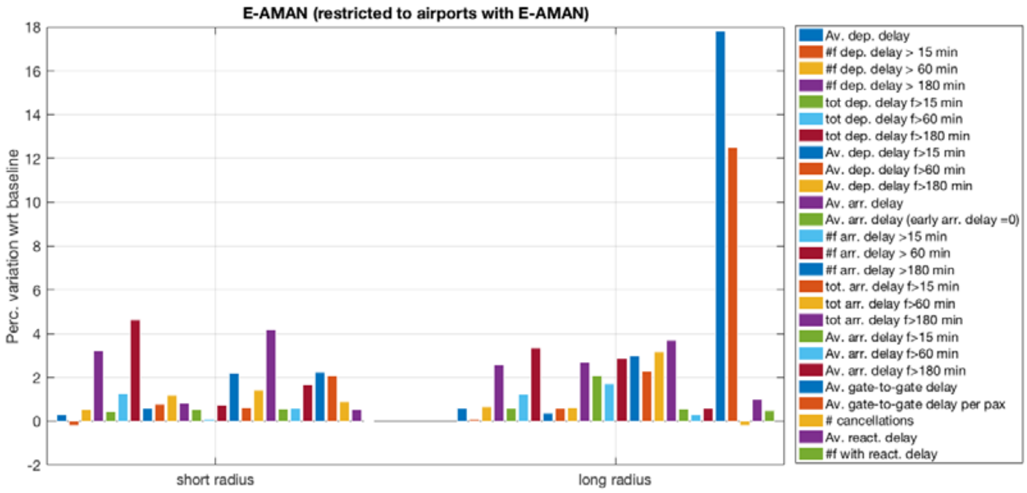


Figure 43: Percentage change in classical delay metrics in E-AMAN (ii) - advanced scenarios, restricted to the airports where the E-AMAN is implemented.

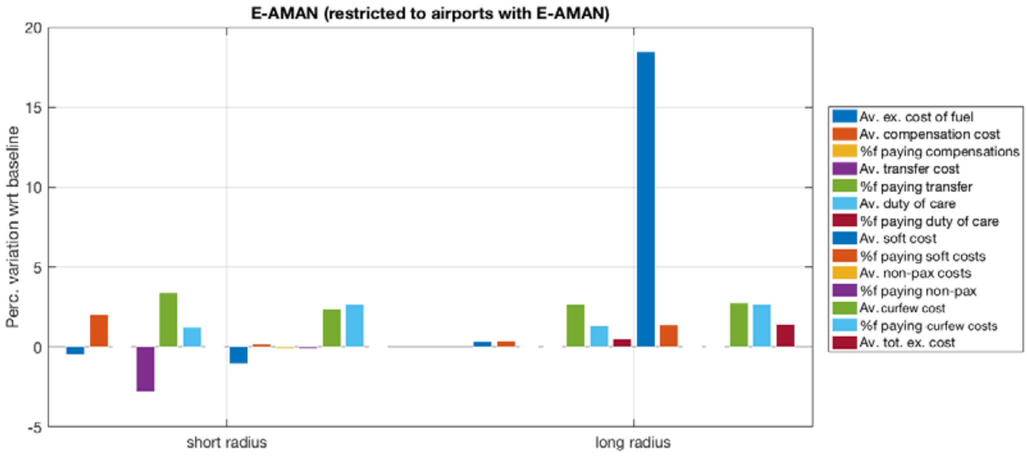


Figure 44: Percentage change in classical cost metrics in E-AMAN (ii) - advanced scenarios, restricted to the airports where E-AMAN is implemented.

5.3.2.3 Passenger-related metrics

The percentage variations in passengers-related metrics in the advanced scenarios with respect to the baseline are shown in Figure 45 and in Figure 46 for the restricted sample of itineraries that leave or pass from or arrive to an airport with E-AMAN (representing ~80% of the itineraries). The metrics average values and their interquantile ranges are reported in Table 17 and Table 18. In the nominal radius case, there is a clear difference between the connecting passengers average delay (overall and positive), which is increasing with respect to the baseline, and the one of non-connecting passengers, which stays roughly the same. In the long range case, instead, both types of passengers are negatively affected. As an effect of the increased delays, more and larger compensations are paid on average, and more passengers receive duty of case (though smaller on average). In both cases there is a decrease in the average delay of passengers having very large delays (>180 min), however the fraction of such passengers increases. In the long radius case the fraction of modified itineraries is also increasing. The situation is qualitatively similar when the analysis is restricted to the itineraries passing from E-AMAN airports, but the worsening is more accentuated.

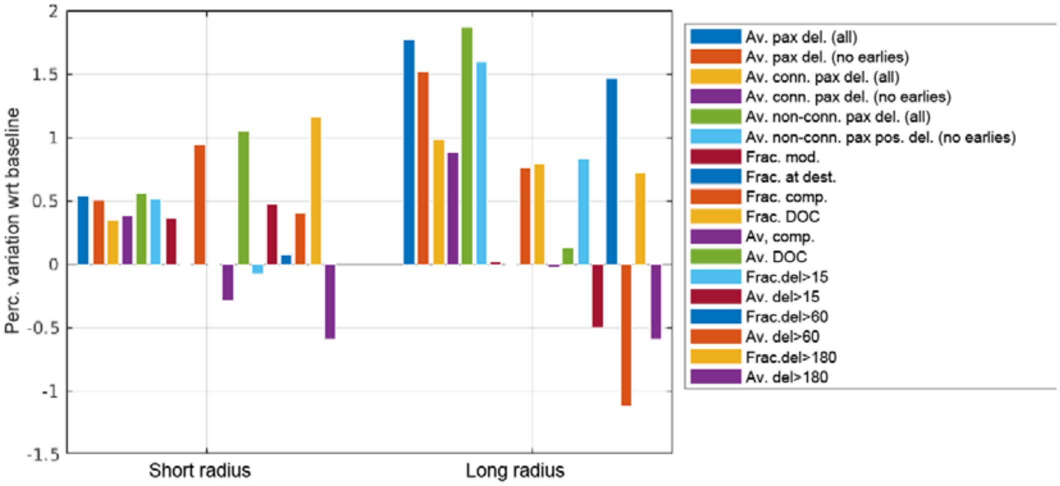


Figure 45: Percentage change in passenger-related metrics in E-AMAN (i) - advanced scenarios with respect to the baseline.

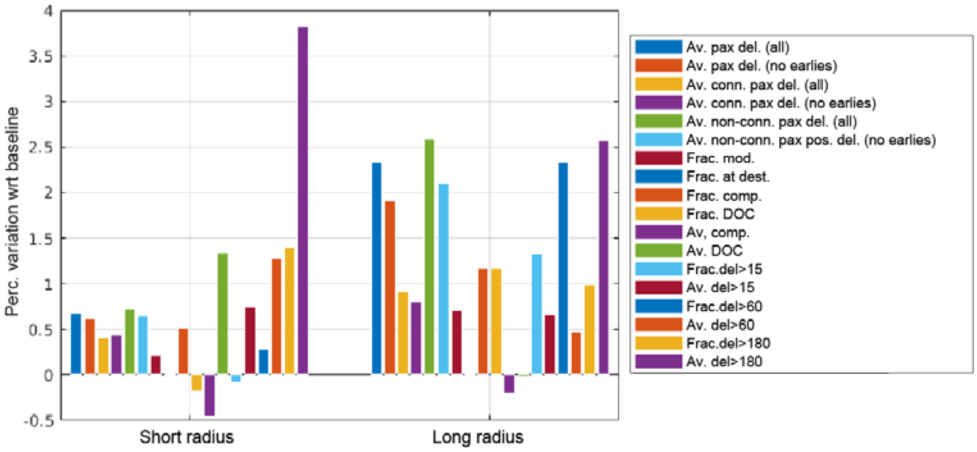


Figure 46: Percentage change in passenger-related metrics in E-AMAN (ii) - advanced scenarios with respect to the baseline, restricting the sample to the itineraries that leave or pass from or arrive to an airport with E-AMAN.

5.3.2.4 Centrality metrics

The percentage variations in centrality metrics in the advanced scenarios with respect to the baseline are shown in Figure 47. All types of centrality losses increase on average (except for a very small decrease of the incoming passenger centrality loss in the long radius scenario), as a consequence of the increased delays on the network.

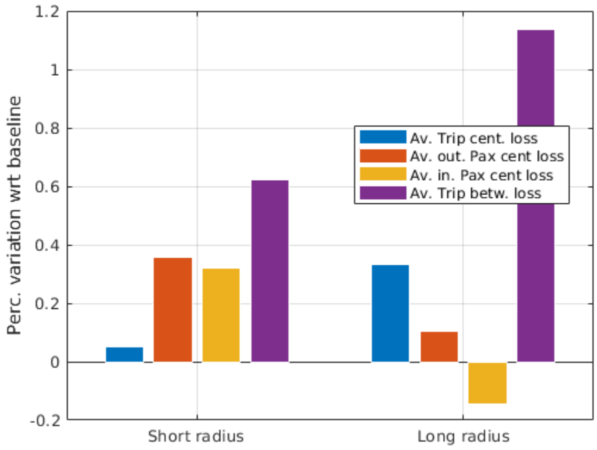


Figure 47: Percentage change in centrality metrics in FACadvanced scenarios with respect to the baseline.

5.3.2.5 Causality metrics

The percentage variations of the causality metrics in the E-AMAN advanced scenarios with respect to the baseline are shown in Figure 48. In the case of nominal radius, the average level of causality measured by link density remain approximately the same with respect to the baseline, for all Granger causality networks, in line with the really small variation of the departure delay of flights. However, it increases significantly in the case of long radius, measured as the increase of link density in the Granger causality in mean network. This is consistent with the average increase of the departure delay in the same scenario. Nevertheless, the opposite pattern is observed when considering the case of Granger causality in tail networks, thus suggesting that extreme delay events become less correlated. Finally, the other network metrics, i.e., reciprocated links and feedback triplets, do not display any clear pattern of variations in the case of short radius. On the contrary, they show a slight percentage decrease in the case of long radius, thus highlighting the partial disruption of the feedback subsystems of delay amplification, mediated by more than one leg effects. This result contradicts apparently the positive variations of all delay metrics in the long radius E-AMAN advanced scenario (Level 2). However, this effect can arise because of the local scope of the E-AMAN mechanism: it induces some delays in the planning horizon and/or the holding phase (explaining the positive variations of all delay metrics), but the same delays could reduce the correlation between the delays of different flights departing at different airports, thus maybe reducing the correlation between the states of congestion at the airports.

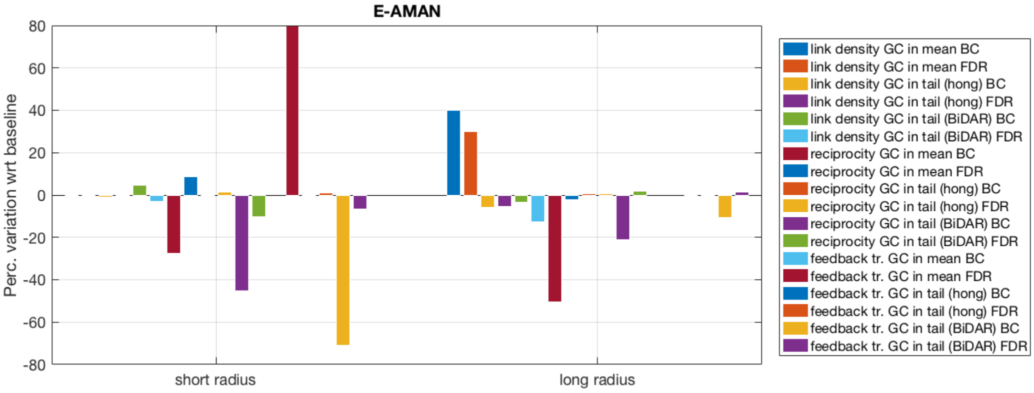


Figure 48: Percentage change of link density, reciprocity, and the number of feedback triplets of the Granger causality networks with E-AMAN implemented for both nominal (short) and long radius, with respect to the baseline. The networks of Granger causality in mean and in tail, for both Hong et al. and BiDAR methods, are built with the states of delay of airports defined as the third quartile of the distribution of departing delays of flights. We consider both corrections for multiple hypotheses testing, namely Bonferroni correction (BC) and false discovery rate (FDR) correction.

5.3.2.6 Conclusions regarding E-AMAN scope

- The larger the E-AMAN radius, the higher the uncertainty associated with the flights therein. The implementation in the model of E-AMAN considers only the flights that are within the E-AMAN radius and does not take into account uncertainties regarding the assignment of arrival slots in the sequence, or future demand. This leads to changes in the landing sequence, which translate into non-optimal behaviour of the system.

- E-AMAN at the more advanced Level 2 tries to minimise the expected cost of the landing sequence. However, the relative cost of fuel means that fuel burn is highly prioritised and large flights might receive relatively large amounts of delay at the planning horizon. This leads to high savings of fuel during the cruise phase as speeds are reduced but, as new flights enter the E-AMAN, and uncertainty manifests itself, the landing sequence is broken. Therefore, higher amounts of delay are assigned at holding. Earlier slots have already been given to other flights, which translate into slightly higher fuel consumption at holding, than under the simple assumptions of Level 0, but also higher arrival delays, and hence costs of delay, also subsequently result.

6 Conclusions and look ahead

6.1 Conclusions

The primary goal of Domino was to improve the state of the art regarding a methodology for analysing the architecture of, and interdependencies within, the air transportation system, by capturing different facets of causality under the impact of a selection of ATM mechanisms. In this regard, the project has made progress on three key topics to reach this goal, as listed below.

- The possibility to model, at a disaggregated level, the full gate-to-gate European air transportation system.
- The introduction of new metrics allowing the capture of some subtle network effects.
- A powerful statistical analysis, combining the capabilities of the model with the power of classical and new metrics to perform a full, network-wide assessment.

The re-implementation of the Mercury simulator, using an agent-based paradigm, is one of the most important achievements of Domino. To our knowledge, it is the only full, ECAC-wide model able to simulate key stakeholders, such as, passengers, airlines and the network manager, in an integrated simulator. With respect to previous versions, it allows us to inject complex behavioural rules for different agents, in particular airlines. Looking forward, Mercury can serve as a test bed for different types of simulations. Different optimisation processes for E-AMAN, rules for flight swapping or trajectory management, levels of congestion, levels of compensation and duty of care for passengers are all examples of modifications which can be tested, relying on a realistic representation for the other components of the model. With respect to other tools, Mercury provides many advantages. For instance, whereas the RNEST tool from EUROCONTROL is more advanced regarding airspace management (including, for example, explicit ATFM regulations and CASA algorithm implementation), Mercury takes into account behavioural (potentially sub-rational) effects from different agents, realistic, stochastic generation of delays, passenger management, and a highly detailed cost of delay model, driving the most important decisions for airlines.

Various metrics have been deployed by Domino. Some are classically used in ATFM (such as average delay), some have been imported from other fields, others have been developed specifically for this project: in particular, the centrality metrics, which take into account the itineraries of passengers and the precise timing of the scheduled flights. At their core, they represent the most relevant metrics in terms of connectivity considering passengers. For their practical usage, Domino has identified some shortcomings, as highlighted in the next section. Causality metrics, on the other hand, have been used before in the air transportation system. Their introduction answered the necessity for decision-makers to understand causal links between subsystems, as opposed to correlations, in order to gain some high-level knowledge. The new methods introduced in Domino, allow us to capture different facets of causality, in particular with an emphasis on how rare events trigger other rare events in the system.

The model allows us to measure a large number of low-level observables, which both poses a problem and raises an opportunity. With so much data, different practical issues arise – such as storage or analysis time – but, more importantly, statistical analyses must be performed with care. Due to the number of observables, the stochastic nature of the simulator, its geographical scope, and the dynamic nature of the system, many different analyses can be performed. In this deliverable, we have focused on the variety of metrics available to the modeller, and on the possibility to restrict the analysis in scope (such as geographically or by stakeholder). Domino has shown that the model can be used to inspect, with a high level of detail, different aspects of the system. In particular, it is able to shed light on the inner functioning of different mechanisms (such as flight swapping or DCI), understanding under which conditions they would, or would not, provide benefits for the different stakeholders. Domino's model sheds light on the role of exogenous and endogenous noise, the behaviours of agents and the initial conditions (passengers, schedules, etc.) on the efficiency of different mechanisms.

6.2 Look ahead

Domino's high-level goal was to provide a tool and a methodology, to analyse the interdependencies existing within the air transportation system. By developing a model able to simulate and produce the right level of information, Domino was able to develop and test several metrics that can be used to analyse the system's architecture from the point of view of systemic, 'domino' effects.

The consortium plans to take this method closer to a real application in the future. Firstly, by generalising the analysis, allowing complex metrics to be used at different levels. For instance, various subsystems can be considered instead of airports, to infer causality links or central nodes. Flights, routes, sectors are all possibilities of subsystems that could be considered, and whose relationships may thus be analysed. This will allow the consideration of adequate measures to avoid the propagation of disruption in the system. Secondly, the partners plan to test more extensively the method, considering other days of operations or specific environments. This will allow us to provide some improvements to the method, and also to improve the extent of the validation.

Validating the model further is a third objective of the partners. This validation will include two processes. The model, Mercury, will need to undergo some systematic comparisons with other available models. These include, for instance, RNEST outputs which can be compared with Mercury, where these are similar in scope and depth. The validation of Mercury lies as much in this comparison itself, as in the understanding of the differences between the models. Expert interviews on some detailed aspects of the models are also needed (for example to calibrate some mechanisms on their baseline implementation, such as adjusting the wait for passenger rules). Finally, using more data to test the model in different environments is paramount to assessing its generalisability.

The metrics presented in this deliverable will need more work in order to assess their usability. This includes some theoretical effort to understand the relationship between these new metrics and established ones (similar to the work presented in D5.1 Metrics and analysis approach), and also a practical effort to render them more intuitive and/or understandable, and thus being candidates to be used as future (key) indicators. It is important to note that these metrics may indeed (as we suggest) be good proxies for others, and/or add further dimensionality and usefulness. This can only be understood via statistical analyses and additional case studies. It is also worth noting that these metrics are independent of the model, and that the model is established, in part, to: (1) to support the increase of the power of the metrics via statistical analyses of the output; (2) to create synthetic data for more or less exploratory scenarios.

This last point is crucial for the partners and will also be developed further in the future. Indeed, for several years, the partners have maintained the ambition to establish a model that could be used as a standard in the field to test different solutions and observe their impact in terms of various KPIs. Hence, the consortium is particularly interested in enhancing the reusability of Mercury, which has been re-implemented in this project with this idea in mind. Harmonious performance assessments (or independent checks) could be achieved through the use of a standard model.

The consortium is eager to develop some of the ideas used during the project regarding the three mechanisms: 4DTA, FAC, and FP. Whilst studies on 4DTA and its impact appeal to airlines regarding their operations, FP is more important to assess for the network manager. The possible introduction of further variations of UDPP (e.g., credits for low-volume users) can be studied with the method presented in Domino. The way in which FAC impacts operations at airports and at the network level, is both important for airports and the network manager. Various optimisation algorithms could be tested with the method presented in Domino.

7 References

[1] Domino Project Consortium, “D4.1 Initial model design,” 2018.

[2] Consortium, Domino, “D5.2 Investigative Case Studies,” 2019.

[3] European Commission, “Commission implementing regulation (EU) No 716/2014 of 27 June 2014 on the establishment of the Pilot Common Project supporting the implementation of the European Air Traffic Management Master Plan,” 2014.

[4] Boeing, “Airports with Noise and Emissions Restrictions,” https://www.boeing.com/commercial/noise/list.page, 2019.

[5] CODA, “IATA Summer Season 2010. Taxi-In (TXI) times in minutes,” 2011.

[6] CODA, “IATA Summer Season 2010. Taxi-Out (TXO) times in minutes,” 2011.

[7] SESAR Joint Undertaking, “E.02.14 - DCI-4HD2D D3.2 Final Technical Report - 00.00.01,” 2016.

[8] EUROCONTROL, “DDR2-Webportal,” 8 June 2018. [Online]. Available: https://www.eurocontrol.int/ddr.

[9] SESAR Joint Undertaking, “E.02.06 - POEM-D6.2 Final Technical Report,” 2013.

[10] SESAR Joint Undertaking, “E.02.06 - POEM-D4.2 Consolidated design review an models refinements,” 2012.

[11] Airbus, “Performance Engineering Programme,” 2014.

[12] CODA, “CODA Digest 2017. All-causes delay and cancellations to air transport in Europe - 2017,” 2018.

[13] Domino Consortium, “D3.3 Adaptive Case Studies description,” 2019.

[14] Domino Project Consortium, “D5.1 Metrics and analysis approach,” 2019.

[15] M. E. J. Newman, Networks: An introduction, doi:10.1007/978-3-319-03518-5-8: Oxford Univ.,

2010.

[16] Habiba, C. Tantipathananandh and T. Berger-wolf, “Betweenness Centrality Measure in Dynamic Networks,” Technical report, DIMACS, 2007.

[17] C. W. Granger, “Investigating causal relations by econometric models and cross-spectral methods,” *Econometrica: Journal of the Econometric Society,* pp. 424-438, 1969.

[18] M. Zanin, S. Belkoura and Y. Zhu, “Network analysis of Chinese air transport delay propagation,” *Chinese Journal of Aeronautics,* vol. 30(2), pp. 491-499, 2017.

[19] P. Mazzarisi, S. Zaoli, F. Lillo, L. Delgado and G. Gurtner, “New centrality and causality metrics assessing air traffic network interactions,” *arXiv preprint arXiv:1911.02487,* 2019.

[20] Y. Hong, Y. Liu and S. Wang, “Granger causality in risk and detection of extreme risk spillover between financial markets,” *Journal of Econometrics,* vol. 150(2), pp. 271-287, 2009.

[21] P. A. Jacobs and P. A. Lewis, “Discrete Time Series Generated by Mixtures. III. Autoregressive Processes (DAR (p)) (No. NPS55-78-022),” *NAVAL POSTGRADUATE SCHOOL MONTEREY CALIF,* 1978.

[22] CODA, “CODA Digest 2014. All-causes delay and cancellations to air transport in Europe - 2014,” 2015.

[23] Domino Project Consortium, “D3.1 Architecture definition,” 2018.

[24] Domino Project Consortium, “D3.2 Investigative case studies description,” 2018.

[25] Domino Project Consortium, “D6.3 Workshop results summary,” 2019.

[26] I. Tsalouchidou, R. Baeza-yates, F. Bonchi, K. Liao and T. Sellis, “Temporal betweenness centrality in dynamic graphs,” *Int. J. Data Sci. Anal.,* 2019.

8 Acronyms

4DTA: 4D Trajectory Adjustment mechanism

ABM: Agent-based model

AIRAC: Aeronautical Information Regulation and Control

AMAN: Arrival Manager

ANSP: Air Navigation Service Provider

AOBT: Actual off-block time

AOC: Airline Operations Centre

ATC: Air Traffic Control

ATFM: Air Traffic Flow Management

ATM: Air traffic management

AU: Airspace user

BADA: Base of Aircraft Data

BC: Bonferroni correction

CASA: Computer Assisted Slot Allocation

CI: Cost Index

COBT: Calculated off-block time

CODA: Central Office for Delay Analysis

DCI: Dynamic cost indexing

DDR2: Demand Data Repository

DMAN: Departure Manager

E-AMAN: Extended Arrival Manager

ECAC: European Civil Aviation Conference

EIBT: Estimated in-block time

EOBT: Estimated off-block time

FAC: Flight Arrival Coordination mechanism

FDR: False discovery rate

FP: Flight Prioritisation mechanism

G2G: Gate to Gate

GC: Granger Causality

HDM: Hub Delay Management scenario

KPI: Key Performance Indicator

MCT: Minimum connecting time

MTT: Minimum turnaround time

NAS: National Airspace

NM: Nautical mile

Pax: Passengers

Q-Q: Quantile-Quantile

RNEST: Research Network Strategic Tool

SIBT: Scheduled in-block time

SJU: SESAR Joint Undertaking

SOBT: Scheduled off-block time

TMA: Terminal Manoeuvring Area

TOC: Top of Climb

TOD: Top of Descend

UDPP: User-Driven Prioritisation Process

9 Annex I – Detailed mechanism tables

9.1 Hub delay management

9.1.1 Delay metrics

9.1.1.1 Flight metrics

Table 12: Values of the classical delay metrics – hub delay management

	Baseline			4DTA Level 2			FP Level 2		
Metric	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile
Mean departure delay of flights	11.7	11.2	12	11.4	11	11.6	11.8	11.2	11.9
Number of flights with departure delay > 15 minutes	7110	7000	7220	6990	6870	7090	7110	6990	7230
Number of flights with departure delay > 60 minutes	556	471	626	531	454	584	568	465	632

Number of flights with departure delay > 180 minutes	38.3	13	33	27.5	12	26	38.7	13	36
Total departure delay of flights with departure delay > 15 minutes (x10^3)	237	223	244	229	217	234	238	222	241
Total departure delay of flights with departure delay > 60 minutes (x10^2)	553	400	606	495	390	537	566	419	589
Total departure delay of flights with departure delay > 180 minutes (x10^2)	120	33.9	102	78.5	34.6	79.4	118	39.5	114
Mean departure delay of flights with departure delay > 15 minutes	33.3	31.6	33.8	32.7	31.5	33.1	33.4	31.7	33.6
Mean departure delay of flights with departure delay > 60 minutes	96.2	86.4	97.4	91.4	85.5	93.3	95.8	86.4	97.8
Mean departure delay of flights with departure delay > 180 minutes	294	259	324	288	255	319	286	245	316
Mean arrival delay of flights	6.9	6.4	7.2	6.8	6.4	7	6.9	6.3	7.1
Mean arrival delay of delayed flights (early arrivals as 0)	11.6	11.1	11.8	11.4	11	11.5	11.6	11.1	11.8
Number of flights with arrival delay > 15 minutes	7380	7280	7480	7330	7210	7430	7380	7240	7490
Number of flights with arrival delay > 60 minutes	635	550	703	621	544	677	646	535	707
Number of flights with arrival delay > 180 minutes	39.3	14	36	28.9	14	28	40	15	39

Total arrival delay of flights with arrival delay > 15 minutes (x10^3)	255	242	262	250	240	255	256	240	260
Total arrival delay of flights with arrival delay > 60 minutes (x10^2)	616	474	665	565	463	614	628	474	643
Total arrival delay of flights with arrival delay > 180 minutes (x10^2)	122	36	108	81	37	79	121	42	116
Mean arrival delay of flights with arrival delay > 15 minutes	34.6	33	35	34.2	33	34.5	34.7	33	34.9
Mean arrival delay of flights with arrival delay > 60 minutes	94.3	85.1	95.8	89.6	84.2	91.4	94	85.5	95.6
Mean arrival delay of flights with arrival delay > 180 minutes	288	254	311	281	249	311	281	243	306
Mean gate-to-gate delay of flights	4.8	4.8	4.9	4.6	4.5	4.6	4.8	4.8	4.9
Mean per-passenger gate-to-gate delay (x10^{-3})	28	27	28	26	26	27	28	27	28
Number of cancelled flights	280	265	291	281	269	290	282	271	293
Mean reactionary delay	3.8	3.4	3.9	3.6	3.4	3.7	3.8	3.4	3.9
Number of flights with reactionary delay	4880	4810	4950	4830	4740	4890	4890	4810	4960

9.1.1.2 Passenger metrics

Table 13: Passenger-related metrics – hub delay management

	Baseline			4DTA Level 2			FP Level 2		
Metric	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile
Average pax delay	4.7	4.17	4.8	4.85	4.47	5.02	4.77	4.23	4.97
Average pax positive delay	10.6	10.1	10.7	10.5	10.1	10.6	10.7	10.1	10.9
Average conn. pax delay	9.28	8.38	9.59	8.54	7.9	9.04	9.5	8.24	9.67
Average conn. pax pos. delay	14.8	13.8	14.9	14	13.5	14.5	15	13.8	15.1
Average non-conn. pax delay	4.27	3.79	4.4	4.51	4.15	4.65	4.33	3.88	4.52
Average non-conn. pax pos. delay	10.2	9.73	10.3	10.1	9.83	10.2	10.3	9.81	10.4
Fraction modified itineraries	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Fraction pax at dest.	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Frac. Pax receiving comp.	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Frac. Pax receiving DOC	0.02	0.02	0.02	0.02	0.01	0.02	0.02	0.02	0.02
Average compensation	39.9	39.2	40.7	40.1	39.4	40.8	40	39.5	40.6
Average DOC	57.8	53.5	62.6	58.5	55.4	62.8	56.6	53.2	62.3
Frac. pax with delay>15	0.25	0.25	0.26	0.26	0.25	0.26	0.26	0.25	0.26
Average delay>15	34.9	33.1	34.7	34.4	33.3	34.4	34.9	33.2	35.1
Frac. pax with delay>60	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Average delay>60	104	95.1	101	97.7	93	98.7	103	94.1	105
Frac. pax with delay>180	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Average delay>180	316	291	331	308	291	321	308	278	328

Table 14: Passenger-related metrics for restricted sample – hub delay management

	Baseline restricted (to passengers departing or landing in an airport during regulations)			4DTA Level 2 restricted (to passengers departing or landing in an airport during regulations)			Baseline restricted (to passengers landing in an airport during regulations)			FP Level 2 restricted (to passengers landing in an airport during regulations)		
Metric	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile
Average pax delay	2.02	1.6	2.41	2.69	2.32	2.95	4.38	4.01	4.76	4.4	3.98	4.77
Average pax positive delay	9.7	9.31	10	9.93	9.56	10.2	10.7	10.4	11.1	10.7	10.3	11
Average conn. pax delay	4.38	3.45	5.06	4.27	3.48	4.85	3.4	2.58	4.2	3.31	2.63	3.9
Average conn. pax pos. delay	11.9	11.1	12.6	11.7	11	12.2	11.2	10.5	11.9	11.1	10.5	11.6
Average non-conn. pax delay	1.46	1.11	1.85	2.32	1.93	2.68	4.77	4.46	5.15	4.82	4.34	5.24
Average non-conn. pax pos. delay	9.18	8.91	9.4	9.52	9.18	9.83	10.5	10.2	10.9	10.6	10.2	10.9
Fraction modified itineraries	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Fraction pax at dest.	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Frac. Pax receiving comp.	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02

Frac. Pax receiving DOC	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Average compensation	49.2	47.7	50.8	49.4	47.5	51.4	51.4	49	54	51.9	50.1	54.8
Average DOC	62.1	57	68	60.4	55.5	65.9	60.2	54.3	67.1	59.3	54.1	65.4
Frac. pax with delay>15	0.24	0.23	0.24	0.24	0.23	0.25	0.26	0.26	0.27	0.26	0.25	0.27
Average delay>15	34.1	32.9	34.9	34.7	33.7	35.4	34.3	33.3	35.1	34.4	33.5	35.1
Frac. pax with delay>60	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Average delay>60	108	99.7	114	108	99	114	112	104	119	112	103	118
Frac. pax with delay>180	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02
Average delay>180	339	320	356	336	320	355	338	313	352	331	310	351

9.1.2 Cost metrics

Table 15: Values of the classical cost metrics – hub delay management

	Baseline			4DTA Level 2			FP Level 2		
Metric	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile	Mean	1st Quartile	3rd Quartile
Average excess cost of fuel	128	127	130	103	102	105	129	128	130
Average cost of compensation	56.6	53.7	59.1	57.1	53.5	59.3	57	54.5	59.5
Fraction of flights paying compensation (x10^{-3})	17	16	18	17	17	18	17	16	18
Average cost of transfer	1.1	0.5	1.2	1	0.6	1.2	1.2	0.6	1.3
Fraction of flights paying transfer (x10^{-5})	83	70	92	86	70	99	87	70	101
Average duty of care cost	122	114	126	121	114	125	124	116	127
Fraction of flights paying duty of care	0.092	0.087	0.096	0.092	0.086	0.097	0.093	0.088	0.098
Average soft costs	8.5	3.6	14.2	9.9	3.7	14.3	10.7	3.8	14.3
Fraction of flights paying soft costs	0.491	0.489	0.494	0.49	0.488	0.492	0.491	0.489	0.493
Average non-pax costs	70.9	68.7	72.3	70.1	68.1	70.9	71.4	68.2	72.4
Fraction of flights paying non-pax costs	0.936	0.935	0.938	0.936	0.935	0.937	0.936	0.935	0.937
Average curfew costs	9.2	3.7	8.6	10	4.9	8.6	9.7	4.9	10.4
Fraction of flights paying curfew costs (x10^{-5})	27	11	26	30	15	26	29	15	31
Average total excess cost	238	227	242	211	203	216	243	231	245

Note: Costs are represented in euros as experienced by the airlines per flight.

9.2 E-AMAN scope

9.2.1 Delay metrics

9.2.1.1 Flight metrics

Table 16: Values of the classical delay metrics – E-AMAN scope

	Baseline - FAC Level 0 - 200 NM			FAC Level 2 - 200 NM			FAC Level 0 - 600 NM			FAC Level 2 - 600 NM		
Metric	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.
Mean departure delay of flights	33.3	32.8	33.8	33.3	32.8	33.9	33.5	32.9	33.9	33.5	33	33.9
Number of flights with departure delay > 15 minutes (x10^2)	158	158	159	158	157	159	158	158	159	158	158	159
Number of flights with departure delay > 60 minutes	4200	4130	4260	4210	4140	4280	4240	4170	4320	4240	4180	4290
Number of flights with departure delay > 180 minutes	213	196	228	216	194	238	220	200	240	217	195	237
Total departure delay of flights with departure delay > 15 minutes (x10^3)	839	826	852	841	826	856	846	829	857	845	831	857

Total departure delay of flights with departure delay > 60 minutes (x10^3)	452	440	464	456	439	472	460	441	469	459	444	472
Total departure delay of flights with departure delay > 180 minutes (x10^3)	101	88	115	104	88	118	106	91.6	119	104	89.8	118
Mean departure delay of flights with departure delay > 15 minutes	53	52.3	53.9	53.2	52.3	54.2	53.4	52.4	54.1	53.3	52.4	54.1
Mean departure delay of flights with departure delay > 60 minutes	108	105	110	108	105	111	109	106	111	108	105	111
Mean departure delay of flights with departure delay > 180 minutes	476	437	512	479	440	518	479	451	511	480	440	517
Mean arrival delay of flights	37.9	37.4	38.4	38	37.4	38.6	38.4	37.7	38.7	38.4	37.9	38.9

Mean arrival delay of delayed flights (early arrivals as 0)	39.2	38.7	39.7	39.3	38.7	39.9	39.6	39	40	39.7	39.1	40.1
Number of flights with arrival delay > 15 minutes (x10^{2})	183	182	183	183	182	183	184	183	184	184	183	185
Number of flights with arrival delay > 60 minutes	5480	5410	5550	5490	5430	5560	5540	5450	5630	5550	5490	5600
Number of flights with arrival delay > 180 minutes	258	240	275	260	240	283	265	247	286	263	241	283
Total arrival delay of flights with arrival delay > 15 minutes (x10^{3})	1010	1000	1030	1020	1000	1030	1020	1010	1030	1030	1010	1040
Total arrival delay of flights with arrival delay > 60 minutes (x10^{3})	576	561	590	579	563	598	586	568	597	586	571	598
Total arrival delay of flights with arrival delay > 180 minutes (x10^{3})	112	98.6	124	114	96.4	129	116	102	128	115	101	130
Mean arrival delay of flights with arrival delay > 15 minutes	55.5	54.8	56.2	55.7	54.8	56.4	55.8	55	56.3	55.7	55	56.5

Mean arrival delay of flights with arrival delay > 60 minutes	105	103	107	106	103	108	106	103	107	106	103	108
Mean arrival delay of flights with arrival delay > 180 minutes	433	399	463	438	404	469	438	411	466	437	403	469
Mean gate-to-gate delay of flights	-4.7	-4.7	-4.6	-4.7	-4.7	-4.7	-4.8	-4.9	-4.8	-4.9	-5	-4.9
Mean per-passenger gate-to-gate delay	-61	-61	-0.06	-61	-61	-61	-62	-62	-62	-63	-63	-62
Number of cancelled flights	327	313	338	327	316	337	324	313	335	327	314	340
Mean reactionary delay	20	19.6	20.3	20	19.6	20.5	20.2	19.8	20.6	20.2	19.8	20.6
Number of flights with reactionary delay (x10^3)	12.0	11.9	12.0	12.0	11.9	12.0	12.1	12.0	12.1	12.1	12.0	12.1

9.2.1.2 Passenger metrics

Table 17: Passenger-related metrics – E-AMAN scope

	Baseline - FAC Level 0 - 200 NM			FAC Level 2 - 200 NM			FAC Level 0 - 600 NM			FAC Level 2 - 600 NM		
Metric	**Mean**	**1st Quart.**	**3rd Quart.**	**Mean**	**1st Quart.**	**3rd Quart.**	**Mean**	**1st Quart.**	**3rd Quart.**	**Mean**	**1st Quart.**	**3rd Quart.**
Average pax delay	34.2	33.9	34.6	34.4	34	34.9	34.7	34.2	35.2	34.8	34.3	35.4
Average pax positive delay	36.1	35.8	36.4	36.3	35.8	36.7	36.5	36	37	36.6	36	37.2
Average conn. pax delay	46.7	45.7	47.6	46.9	45.9	47.8	47.2	46.2	48.1	47.2	46.4	48.2
Average conn. pax pos. delay	48.4	47.4	49.4	48.6	47.7	49.5	48.9	47.9	49.7	48.9	48.1	49.9
Average non-conn. pax delay	33.1	32.7	33.4	33.3	32.9	33.7	33.5	33.1	34	33.7	33.1	34.3
Average non-conn. pax pos. delay	34.9	34.6	35.3	35.1	34.7	35.6	35.4	34.9	35.9	35.5	34.9	36.1
Fraction modified itineraries	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
Fraction pax at dest.	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Frac. Pax receiving comp.	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	0.02
Frac. Pax receiving DOC	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Average compensation	39.2	38.8	39.7	39	38.5	39.6	39.1	38.6	39.6	39.2	38.6	39.9
Average DOC	24.2	23.4	24.9	24.5	23.8	25.1	24.1	23.5	24.7	24.3	23.3	25

Frac. pax with delay>15	0.65	0.64	0.65	0.65	0.64	0.65	0.65	0.65	0.65	0.65	0.65	0.65
Average delay>15	53.7	53.1	54.2	54	53.3	54.8	54	53.3	54.8	54.1	53.3	54.8
Frac. pax with delay>60	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.19	0.2	0.19	0.19	0.2
Average delay>60	105	103	106	106	104	109	106	104	108	106	103	108
Frac. pax with delay>180	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Average delay>180	438	412	462	452	425	488	449	424	474	453	422	485

Table 18: Passenger-related metrics for restricted sample – E-AMAN scope

	Baseline - FAC Level 0 - 200 NM			FAC Level 2 - 200 NM			FAC Level 0 - 600 NM			FAC Level 2 - 600 NM		
Metric	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.
Average pax delay	32.6	32.2	33	32.9	32.3	33.3	33.3	32.8	33.9	33.4	32.8	34.1
Average pax positive delay	34.7	34.2	35	34.9	34.4	35.3	35.3	34.9	35.9	35.3	34.8	36.1
Average conn. pax delay	44.6	43.8	45.6	44.8	43.8	45.7	45.2	44.3	46.2	45	44.2	45.7
Average conn. pax pos. delay	46.4	45.6	47.4	46.6	45.7	47.5	46.9	46	47.9	46.8	45.9	47.5

Average non-conn. pax delay	31.1	30.7	31.4	31.3	30.8	31.8	31.8	31.3	32.4	31.9	31.2	32.6
Average non-conn. pax pos. delay	33.2	32.7	33.4	33.4	32.9	33.9	33.8	33.4	34.4	33.9	33.2	34.6
Fraction modified itineraries	0.01	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.02	0.02	0.01	0.02
Fraction pax at dest.	0.99	0.98	0.99	0.98	0.98	0.99	0.99	0.98	0.99	0.98	0.98	0.99
Frac. Pax receiving comp.	0.02	0.01	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01	0.02
Frac. Pax receiving DOC	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
Average compensation	40.6	40.1	41	40.4	39.7	41	40.5	39.8	41.1	40.5	39.8	41.1
Average DOC	25.4	24.6	26.3	25.8	24.9	26.8	25.4	24.8	26	25.4	24.5	26.3
Frac. pax with delay>15	0.63	0.63	0.63	0.63	0.63	0.63	0.64	0.64	0.64	0.64	0.64	0.64
Average delay>15	52.7	51.9	53.2	53.1	52.2	53.8	53.1	52.5	53.9	53	52	53.9
Frac. pax with delay>60	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.19	0.18	0.18	0.19
Average delay>60	105	103	107	106	103	110	106	104	109	106	102	109
Frac. pax with delay>180	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
Average delay>180	420	388	450	436	394	475	434	406	463	431	387	475

9.2.2 Cost metrics

Table 19: Values of the classical cost metrics – E-AMAN scope

	Baseline - FAC Level 0 - 200 NM			FAC Level 2 - 200 NM			FAC Level 0 - 600 NM			FAC Level 2 - 600 NM		
Metric	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.	Mean	1st Quart.	3rd Quart.
Average excess cost of fuel	187	185	189	187	185	189	191	189	193	188	186	189
Average cost of compensation	71.4	68.4	74.4	72	69.2	74.7	70.5	67	74.3	72.2	69.2	75.4
Fraction of flights paying compensation	35	34	36	35	34	36	35	35	36	35	35	36
Average cost of transfer	2	1.6	2.3	2	1.6	2.3	2.1	1.5	2.1	2.2	1.6	2.6
Fraction of flights paying transfer (x10^{-4})	27	25	29	27	26	30	27	26	28	28	25	30
Average duty of care cost	204	195	211	205	199	211	204	197	215	206	197	213
Fraction of flights paying duty of care	178	174	183	179	175	183	178	174	182	0.18	176	184
Average soft costs	29.5	12.3	48.7	29.1	12.3	48.7	39.6	21.5	49.3	34.5	12.4	49.4
Fraction of flights paying soft costs	0.58	578	581	0.58	579	582	0.58	579	582	582	0.58	583

Average non-pax costs	165	164	166	165	164	167	166	164	168	165	164	167
Fraction of flights paying non-pax costs	967	966	968	966	966	967	967	966	968	967	966	968
Average curfew costs	66.7	63.1	71	67.3	62.4	72.2	68.4	65.2	69.8	66.6	61.8	70.9
Fraction of flights paying curfew costs (x10^{-4})	20	19	21	20	19	22	21	20	21	20	19	21
Average total excess cost	493	477	509	494	473	513	509	488	525	500	480	518

Note: Costs are represented in euros as experienced by the airlines per flight.

10 Annex II – Causality testing methods

In this Appendix, we give more details about the methods to test the presence of a causality relationship between two state variables.

(i) Granger causality in mean

$Y \equiv \{y_t\}_{t=1,\dots,T}$ is said to Granger-cause $X \equiv \{x_t\}_{t=1,\dots,T}$ if we reject the null hypothesis that the past values of Y do not provide statistically significant information about future values of X by assuming VAR(p) as the predictive model. Let us consider X and Y described by

$$\begin{cases} x_t & = \phi_0^1 + \sum_{j=1}^{p} \phi_j^{11} x_{t-j} + \sum_{i=1}^{p} \phi_i^{12} y_{t-i} + \epsilon_t^1 \\ y_t & = \phi_2^1 + \sum_{j=1}^{p} \phi_j^{21} x_{t-j} + \sum_{i=1}^{p} \phi_i^{22} y_{t-i} + \epsilon_t^2 \end{cases}$$

where $\epsilon_t^1, \epsilon_t^2$ are taken to be two uncorrelated white-noise series. The goal of the test [17] is to assess the statistical significance of $\{\phi_i^{12}\}_{i=1,\dots,p}$ by considering as null hypothesis that they are zero, i.e. $H_0\colon \{\phi_i^{12} = 0\}_{i=1,\dots,p}$. The null hypothesis H_0 is equivalent to considering that $\{x_t\}$ evolves according to an AR(p) process. After estimating both VAR(p) and AR(p) models (the order p of the autoregressive processes is selected by means of the Bayesian Information criterion), the Likelihood-Ratio test (or, alternatively, the F-test) can be applied in order to test if VAR(p) outperforms statistically AR(p) in fitting the observations $\{x_t\}$. If it does, H_0 is rejected, meaning that Y 'Granger-causes (in mean)' X.

(ii) Granger causality in tail

The statistical approach introduced by [20] aims to evaluate whether the knowledge of the past extreme events for a random variable Y helps in forecasting the occurrence of future extreme events for another random variable X.

A realisation x_t is defined as *extreme* when it falls in the right (or left) tail of the distribution of X at time t, as claimed before. In particular, assume to know the probability density function of X at time t conditional on past values and let us define $Q_t \equiv Q(x_1, \dots, x_{t-1}, \beta)$ as the $(1-\beta)$-quantile of the conditional probability distribution of X, i.e. $\mathbb{P}(X > \mathrm{Q}_t \,|\, x_1, \dots, x_{t-1}) = 1 - \beta$ almost surely with $\beta \in (0,1)$ defines Q_t implicitly. This define the new binary random variable Z whose binary realisations describe the extreme events of X.

The null hypothesis H_0^{tail} of the test [20] is:

$$\mathbb{P}(X > \mathrm{Q}_t \,|\, \{x_s\}_{s=1}^{t-1}) = \mathbb{P}(X > \mathrm{Q}_t \,|\, \{x_s\}_{s=1}^{t-1}, \{y_s\}_{s=1}^{t-1}) \text{ a.s.} \forall t = 1, \dots, T$$

meaning that the past realisations of Y do not help in predicting the extreme events of X more than the past history of X itself. A rejection of the null hypothesis H_0^{tail} means that Y 'Granger causes in tail' X at level β. For further information on how to make testable this definition, see [20]. In particular, the authors point out that this test is (not rigorously) equivalent to a Granger-type procedure based on the following auxiliary regression for given $M > 0$ (here, M has a role similar to the order p of the autoregressive process in the Granger causality in mean test. However, it is not optimally selected according to some criterion, but arbitrarily chosen.)

$$z_{1,t} \simeq \alpha_0 + \sum_{j=1}^{M} \alpha_{2,j} z_{2,t-j} + u_t$$

where $\{z_{1,t}\}_{t=1,\dots,T}$ and $\{z_{2,t}\}_{t=1,\dots,T}$ are the time series of extreme events of X and Y, respectively, thus checking whether the coefficients $\{\alpha_{2,j}\}_{j=1,\dots,M}$ are jointly zero. Note that this process does not account for the possibility of autocorrelated $\{z_{1,t}\}$, i.e. terms of type $\alpha_{1,j} z_{1,t-j}$ for some $j > 0$. It can be shown (both numerically and analytically) that the statistical testing procedure introduced in [20] displays a significant (i.e. larger than the confidence level of the test) false positive rate of '$Y(Z_2)$ Granger causes in tail $X(Z_1)$' in the case of $\alpha_{2,j} = 0 \forall j$, when two conditions hold: (i) Z_1 is autocorrelated (i.e. $\alpha_{1,j} \neq 0$ for some j) and (ii) '$X(Z_1)$ Granger causes in tail $Y(Z_2)$'.

(iii) A novel method of Granger causality in tail

Let us consider the standard DAR(p) model [21] for the binary random variable Z_1 representing the occurrence of an extreme event for the random variable X (similarly, we associate a binary random variable Z_2 to the random variable Y), i.e.

$$Z_{1,t} = V_t Z_{1,t-\tau_t} + (1 - V_t) U_{1,t}$$

where $Z_{1,t} \in \{0,1\} \forall t$, $V_t \sim \mathcal{B}(\tilde{\nu})$ is a Bernoulli random variable with $\tilde{\nu} \in [0,1]$, $\tau_t \sim \mathcal{M}(\tilde{\gamma}_1, \dots, \tilde{\gamma}_p)$ is a multinomial random variable with $\tilde{\boldsymbol{\gamma}} \equiv \{\tilde{\gamma}_j\}_{j=1,\dots,p}$ such that $\sum_{i=1}^{p} \tilde{\gamma}_i = 1$ and $U_{1,t}$ is a binary random variable sampled according to the Bernoulli marginal distribution $\mathcal{B}(\tilde{\chi})$ with $\tilde{\chi} \in [0,1]$. In other words, at each time V_t determines if copying from the past or sampling according to the marginal. When we copy from the past, then the multinomial random variable τ_t selects the time lag and, accordingly, which past realisation of Z_1 we copy.

Then, let us introduce the generalisation of the DAR(p) model [21] for bivariate binary random variables Z_1 and Z_2 with Markov properties of order p, namely Bi-DAR(p) model, and the following autoregressive process

$$\left\{ \begin{aligned} Z_{1,t} &= V_t((1 - A_t) Z_{1,t-\tau_t^{11}} + A_t Z_{2,t-\tau_t^{12}}) + (1 - V_t) U_{1,t} \\ Z_{2,t} &= S_t(B_t Z_{1,t-\tau_t^{21}} + (1 - B_t) Z_{2,t-\tau_t^{22}}) + (1 - S_t) U_{2,t} \end{aligned} \right.$$

where $Z_{1,t}, Z_{2,t} \in \{0,1\} \forall t$, $V_t \sim \mathcal{B}(\nu)$ with $\nu \in [0,1]$, $S_t \sim \mathcal{B}(\xi)$ with $\xi \in [0,1]$, $A_t \sim \mathcal{B}(\alpha)$ with $\alpha \in [0,1]$, $B_t \sim \mathcal{B}(\beta)$ with $\beta \in [0,1]$, and $\tau_t^{\cdot} \sim \mathcal{M}(\gamma_1^{\cdot}, \dots, \gamma_p^{\cdot})$ with $\sum_{j=1}^{p} \gamma_j^{\cdot} = 1$. The marginals $u_{1,t}$ and $U_{2,t}$ are also Bernoulli random variables with distribution $\mathcal{B}(\chi_1)$ and $\mathcal{B}(\chi_2)$, respectively, with $\chi_1, \chi_2 \in [0,1]$.

The process describes the evolution of binary state Z_1, as: (i) at time t, V_t determines if copying or not from the past; (ii) if yes, A_t determines if copying $Z_{1,t-\tau_t^x}$ (with probability $1-\alpha$) or $Z_{2,t-\tau_t^{12}}$ (with probability α); (iii) how may steps backward is determined by the multinomial random variable $\tau_t^{\cdot}$ which selects the time lag; (iv) otherwise, we toss a coin with success probability χ_1. Equivalently for the state variable Z_2. Hence, the parameter α (or, equivalently, β) controls the level of dependence of Z_1 from Z_2 (and vice versa when considering β): conditional on the probability that a past event affects the current state (i.e. ν), the larger is α, the larger is the probability that a past extreme event for Z_2 triggers an extreme event for Z_1. In that case, taking into account the past information on Z_2 helps in forecasting the current state of Z_1, thus revealing a causal relationship. We can test for Granger causality in tail as follows.

The null hypothesis $\mathbb{H}_0^{tail}$ that the time series $\{Z_{2,t}\}$ *does not* Granger-cause in tail the time series $\{Z_{1,t}\}$ can be stated in terms of the Bi-DAR(p) model as

$$H_0^{tail}\colon \mathbb{P}^{DAR(p)}(\{Z_{1,t}\}|\widetilde{\nu},\widetilde{\boldsymbol{\gamma}},\widetilde{\chi}) = \mathbb{P}^{Bi-DAR(p)}(\{Z_{1,t}\}|\{Z_{2,t}\},\nu,\alpha,\boldsymbol{\gamma}_i,\chi)\text{a.s.}$$

where on the left-hand side we have the likelihood of the DAR(p) process, whereas in the right-hand side it is the likelihood of the Bi-DAR(p) model (the order p of the discrete autoregressive processes is selected by means of the Bayesian Information Criterion).

Note that the two considered models are nested, since the ‘full’ Bi-DAR(p) model contains all the terms of the ‘restricted’ DAR(p) model, but includes also the ‘off-diagonal’ term of interaction. Thus, to make testable the null hypothesis $\mathbb{H}_0^{tail}$, we can apply the likelihood-ratio test to assess the goodness of fit of the two competing nested models by evaluating how much better the full model works than the restricted one: if the likelihoods of the two models are statistically different one from each other, then the null hypothesis is rejected, thus detecting the presence of a non-zero interaction from Z_2 to Z_1, or, equivalently, a relationship of causality in tail from Y to X.

-END OF DOCUMENT-