Passengers : customers, actors and sensors of the air transportation system

Philippe Monmousseau

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *02/10/2020* **par :**

PHILIPPE MONMOUSSEAU

Passengers: customers, actors and sensors of the air transportation system

Les passagers: clients, acteurs et capteurs du transport aérien

---

JURY

ANDREW COOK	Professeur d'Université	Rapporteur
VOJIN TOSIC	Professeur Émérite	Rapporteur
MICHAEL SCHULTZ	Professeur Associé	Examinateur
ÉRIC FÉRON	Professeur d'Université	Président de Jury
AUDE MARZUOLI	Chercheur	Directrice de thèse
DANIEL DELAHAYE	Professeur d'Université	Directeur de thèse
RICARDO HERRANZ	Chef d'Entreprise	Invité

École doctorale et spécialité :
AA : Mathématiques Appliquées et Informatique
Unité de Recherche :
Laboratoire de Recherche ENAC
Directeur(s) de Thèse :
Daniel DELAHAYE, Aude MARZUOLI
Rapporteurs :
Andrew COOK et *Vojin TOSIC*

Abstract

Air transportation uses planes to transport passengers efficiently between two airports, and its development has been driven by the continuous improvement of planes as a safe and efficient means of transportation. However, if the COVID-19 pandemic has taught the air transportation system one lesson, it's that a problem affecting passengers can be far more detrimental to the air transportation system than a problem affecting planes. Acknowledging the fact that passengers are omnipresent and necessary to the air transportation system, this study proposes to consider passengers as sensors of the air transportation system and harness data generated by passengers to evaluate in near real time the flight-centric metrics traditionally used to evaluate the air transportation system performance. Data generated by passengers have the additional benefit of offering a means of evaluating the interactions between passengers and the other stakeholders of the air transportation system, such as airlines and airports. The journey of a passenger starting and ending beyond the boundaries of airport facilities, the data generated by passengers throughout their journey can also be used to evaluate the full door-to-door journey of a passenger of the air transportation system.

Résumé

Le transport aérien est fondé sur l'utilisation de l'avion pour transporter des passagers entre deux aéroports, et son développement est allé de pair avec l'amélioration continue de l'efficacité et de la sécurité des avions comme moyens de transport. Cependant, si la pandémie liée au COVID-19 nous a appris une leçon, c'est qu'un problème qui touche les passagers du transport aérien peut avoir bien plus de conséquences sur le système dans son ensemble qu'un problème qui concerne les avions. Partant du principe que les passagers sont omniprésents et nécessaires au transport aérien, cette thèse propose de considérer les passagers comme des capteurs du transport aérien, et d'utiliser les données générées par les passagers pour évaluer la performance du transport aérien en quasi temps réel. Ces données générées par les passagers ont également l'avantage d'offrir un moyen d'évaluer les interactions entre les passagers et les autres acteurs du transport aérien, en particulier les aéroports et les compagnies aériennes. Comme le parcours d'un passager commence et se termine au delà des limites d'un aéroport, les données générées par les passagers tout au long de ce parcours peuvent également être utilisées pour évaluer le trajet porte-à-porte complet d'un passager du transport aérien.

Acknowledgements

First and foremost, I would like to thank my advisors Eric Feron, Daniel Delahaye and Aude Marzuoli for their constant support and for trusting me almost blindly from the start of my PhD. I will always remember my very first discussion with Eric and Daniel at ENAC around a coffee machine that helped me decide to kickstart this PhD adventure. They promised me three years of fun and travels, and they did not disappoint!

I am happy to credit Eric for his unwavering enthusiasm regarding the unorthodox approach of my thesis, approach that was initiated and supported by Aude all along my explorations. And I am grateful to Daniel for offering me a stable and healthy environment for my research and for supporting my addiction to sports.

I was lucky to take part to a project proposed by the Verizon Big Data research team with Aude and another by NASA-Ames with Nikunj Oza, that both helped me understand the benefits of considering datasets generated by passengers for analyzing the air transportation system.

I would like to thank Laurent Lapasset at ENAC for his initiatives regarding the calculation clusters that greatly accelerated most of the calculations undertaken for this thesis and for his technical support regarding database structures.

I would also like to thank Marcel Mongeau and Andrija Vidosavljevic for helping me with my swimming addiction and for their regularity in their sports agenda throughout the years.

I am also grateful to my fellow PhD students and friends on both side of the Atlantic for all the good times, with special thoughts to my partners in crime: my roommate Nordine Sebkhi in Atlanta for putting up with me outside of work and showing me the city even once I moved to Toulouse, and my labmate Gabriel Jarry in Toulouse for all the distractions offered during my thesis. And I have to give credit to Sanaa Ikli, who managed to put up with Gabriel and me in the same office for two years...

Finally I would like to thank my family for their continuous support and trust since the very beginning.

Contents

List of Figures

List of Tables

Chapter 1

The need for a complementary passenger-centric approach to the evaluation of the air transportation performance

The air transportation system is an important means of transportation for passengers worldwide, with a steady increase from 2013 to 2019 leading to an all-time high number of passengers in 2019 for U.S. airlines according to the numbers reported by the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) [2] and presented in Figure 1.1. In Europe, Eurostat [3] reported a record number of air passengers traveling in the European Union with more than 1.1 billion air passengers in 2018 [4].

Flight delays remain a major issue both in the United States and in Europe. In 2017, 38.5% of flights in Europe arrived with a delay greater than 5 minutes [5] and 27.8% of U.S. domestic flights arrived with a delay greater than 5 minutes [2].

Flight delays and how these delays propagate with the concerned aircraft have been thoroughly studied in the literature. The majority of these studies are based on the on-time performance reports published by the BTS, which provide flight-level information for each flight, indicating for all scheduled flights whether it was canceled or delayed, and whether the scheduled departure and arrival times are the same as the actual departure and arrival times. A study showed that a 2 hour root delay could result in a multiplied delay of more than 4 hours [6], prompting the research community to better understand how flight delays propagated. Delay is usually propagated either by the aircraft, the crew or the passengers, but the study of the effect of delay propagation on these three categories was often limited to the first two

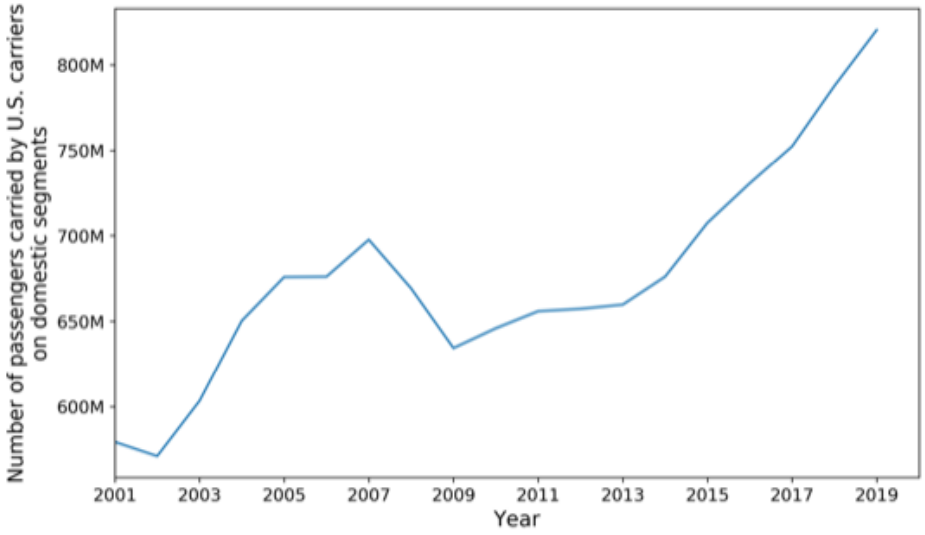


Figure 1.1: Evolution of the yearly number of passengers transported by U.S. airlines on domestic segments and reported to the Bureau of Transportation Statistics.

categories.

Passengers are thus affected by flight delays, but they are also actively part of the air transportation system. This makes them at the core of this system, which explains why both airports and airlines have to consider the passenger experience and their evaluation of the system. Many surveys have been conducted to try and assess passenger satisfaction of airline or airport quality of service, but they all share the same limitations, such as a narrowed scope and a tedious planning difficult to repeat frequently. Furthermore, these studies usually focus only on their experience within one segment of the travel (from airport to airport), whereas passengers are more interested in their full door-to-door travel experience. More recent studies have highlighted the disproportionate impact of airside disruptions on passenger door-to-door journeys, indicating that flight delays do not accurately reflect the delays imposed upon passengers' full multi-modal itinerary.

This led NextGen [7] in the United States and ACARE Flightpath 2050 [8] to advocate a shift from flight-centric metrics to passenger-centric metrics to evaluate the performance of the Air Transportation System. The failures and inefficiencies of the air transportation system not only have a significant economic impact but they also stress the importance of putting the passenger at the core of the system. This advocated shift still has yet to be implemented

by the governing agencies. In a report published in 2016, EUROCONTROL and the FAA presented metrics regarding punctuality that combines airline and passenger views into a single view [9].

It is to be noted that, even though passengers are at the core of the air transportation system, limited quantitative information about passenger movements is publicly shared, especially in airports, which can be considered as the main bottleneck of passenger flow. The management of different airport processes is shared between various stakeholders, from airlines to government, airport authorities and third parties, who do not necessarily rely on each other to make decisions that may affect others. Passengers' satisfaction is largely driven by their experience at the airport, and this experience is the result of the combined control exerted by many stakeholders.

Larger scale studies with a focus on air transportation was recently possible thanks to the increasing use of mobile phone devices as datasources since most individuals now carry a cell phone, and heavily use it through out the day. Though these studies give a full door-to-door view of trips making use of air transportation, mobile phone data are proprietary data and are not often publicly available. In order to operate in real-time, it is thus necessary to also look into other sources of passenger data available on a national scale.

Data gathered from passengers mobile phone in the aforementioned studies can be considered as data gathered by reading signals passively emitted by passengers during their travel, in the sense where the passenger is not actively trying to communicate with the air transportation system via their mobile phone. On the other hand, the ubiquity of mobile phones allows passengers to actively share their experience via social media.

The aim of this thesis is to explore the possibilities offered by these passenger-generated data sources in order to gain additional insight on the state of the air transportation system. The chosen databases are not specific to any country and could be gathered in most regions of the world. Furthermore, they can be easily updated in real time or close-to real time, enabling a regularly updated evaluation of the air transportation system from a passenger perspective. The exploration of these databases leads to the implementation of methods that yield information relevant to passengers but that should also be used by air transportation stakeholders in order to better understand where they stand with respect to other stakeholders and how they could improve.

The work presented in this thesis started in 2017 and focused on the first severe perturbation of the air transportation system that happened shortly thereafter, a major winter storm that shut down three US airports in January 2018. The initial study of this perturbation [10] that initiated this thesis is presented in Appendix A: *Passenger-centric metrics for Air Transportation*

leveraging mobile phone and Twitter data by Marzuoli, A., Monmousseau, P., Feron, E. and presented at the Data-Driven Intelligent Transportation Workshop - IEEE International Conference on Data Mining 2018.

This study validates the need for a passenger-centric approach in order to monitor the state of the air transportation system in close-to real time during severe perturbations. It prompted the implementation of an estimator of the national number of delays in the United States based on Twitter data generated by passengers [11] (*Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources* by Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D. and presented at the Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria) later improved to estimate the hourly number of delays at an airport level [12] (*Passengers on social media: A real-time estimator of the state of the US air transportation system* by Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D. and presented at the ENRI Int. Workshop on ATM/CNS (EIWAC 2019), Tokyo, Japan). This second work is presented in Appendix B, and the latest estimator model, which also enables the real-time estimation of the hourly number of cancellation per airport, is detailed in Chapter 3.

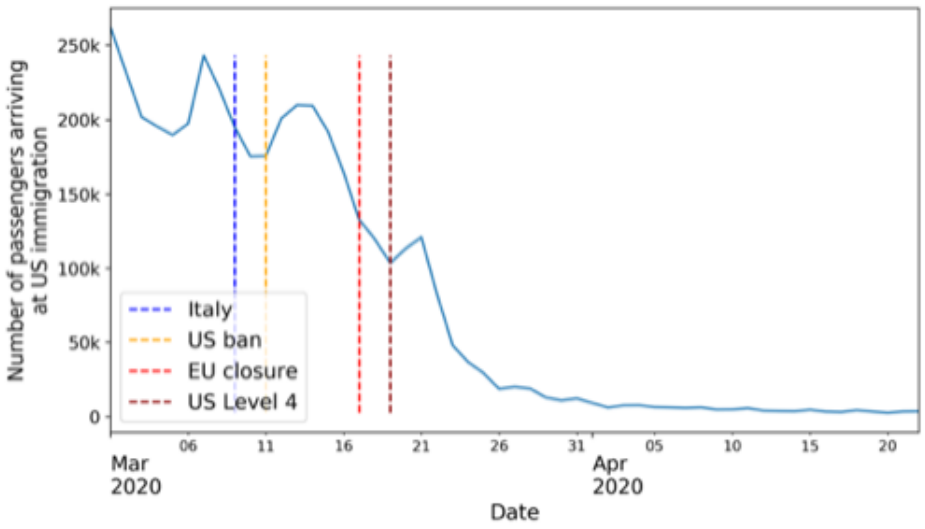


Figure 1.2: Evolution of the daily number of passengers arriving at all US airports of entry from the US Customs and Border Protection data.

The COVID-19 pandemic, and the associated travel restrictions on passengers, have generated an unprecedented drop in international air transportation (see Figure 1.2 and Section 4.1.1) and stressed the necessity of tak-

ing a passenger-centric approach for an up-to-date assessment of the situation. The tools initially developed for analyzing severe but short-scaled perturbations, such as the winter storm of January 2018, have been adapted to offer a real-time analysis of the effects of long-term perturbations such as the COVID-19 pandemic. A first study of the interactions between passengers and airlines and between passengers and airports during the pandemic has been conducted and published online before the release of official flight data [13] (*Putting the Air Transportation System to sleep: a passenger perspective measured by passenger-generated data* by Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D.). This study is improved and described in greater detail in Chapter 4.

Air transportation is a multi-modal transportation system meaning that passengers have to consider their full door-to-door journey when planning to take a plane. A first model to estimate the full door-to-door travel times in Europe based on data available online was created and presented in [14] (*Door-to-door travel time analysis from Paris to London and Amsterdam using Uber data* by Monmousseau, P., Delahaye, D., Marzuoli, A., Feron, E. and presented at the Ninth SESAR Innovation Days (2019), Athens, Greece). This model was adapted to the US, taking into account additional available databases in [15] (*Door-to-door Air Travel Time Analysis in the United States using Uber Data* by Monmousseau, P., Delahaye, D., Marzuoli, A., Feron, E. and presented at the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), IEEE, Singapore, Singapore). The combined model of the full door-to-door travel time valid for both the US and Europe is presented in Chapter 5.

This model of door-to-door travel time has highlighted the disproportionate amount of time passengers can spend at airports, therefore several works aimed at improving the passengers wait time and experience at airports were also conducted during this thesis. The wait time at immigration is explored in [16] (*Doorway to the United States: An Exploration of Customs and Border Protection Data* by Monmousseau, P., Marzuoli, A., Bosson, C., Feron, E., Delahaye, D. and presented at the 38$^{\text{th}}$ Digital Avionics Systems Conference (DASC2019), San Diego, California, USA). A tool to predict the passenger flow at security checkpoints is presented in [17] (*Predicting Passenger Flow at Charles De Gaulle Airport Security Checkpoints* by Monmousseau, P., Jarry, G., Bertosio, F., Delahaye, D., Houalla, M. and presented at the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), IEEE, Singapore, Singapore). A novel approach to sentiment analysis with a direct application to passengers and airlines is implemented in [18] (*Towards a more complete view of air transportation performance combining on-time performance and passenger sen-*

timent. by Monmousseau, P., Puechmorel, S., Delahaye, D., Marzuoli, A., Feron, E. and presented at at the 9th International Conference on Research in Air Transportation (ICRAT '20), Tampa, Florida, USA). These exploratory works were conducted in parallel to the main work of this thesis and are presented in Appendix D.

Acknowledging the fact that passengers are omnipresent and necessary to the air transportation system (Chapter 2), this thesis proposes to consider passengers as sensors of the air transportation system and harness data generated by passengers to evaluate in real time the flight-centric metrics traditionally used to evaluate the air transportation system performance (Chapter 3). Data generated by passengers have the additional benefit of offering a means of evaluating the interactions between passengers and the other stakeholders of the air transportation system, such as airlines and airports, most useful when no flight-centric data are readily available (Chapter 4). The journey of a passenger starting and ending beyond the boundaries of airport facilities, the data generated by passengers throughout their journey can also be used to evaluate the full door-to-door journey of a passenger of the air transportation system (Chapter 5).

Finally, Chapter 6 summarizes the main contributions of this thesis and discusses some potential research directions.

Chapter 2

Background

2.1 Delays within the Air Transportation system

The U.S. Department of Transportation's Bureau of Transportation Statistics (BTS) reported that U.S. airlines carried an all-time high number of passengers in 2019 - 928.9 million systemwide, 813.36 million domestic and 115.55 million international [2]. In Europe, a record number of air passengers traveled in the European Union in 2018 with more than 1.1 billion air passengers reported in 2018 by Eurostat [4]. Flight delays are however a major issue both in the United States and in Europe. In 2017, 44.4% of flights in Europe departed with a delay greater than 5 minutes and 38.5% arrived with a delay greater than 5 minutes [5]. In the US, it represents 27.0% of departing flights and 27.8% of arriving flights [2].

Flight delays and how these delays propagate with the concerned aircrafts have been thoroughly studied in the literature, see Sternberg et al. [19] for a survey and taxonomy analysis of flight delay prediction. The majority of these studies were based on the on-time performance measures of the BTS, which provide flight-level information for each day, indicating for all scheduled flights, whether a flight was canceled or delayed, and comparing scheduled versus actual departure and arrival times. Since 2013, it also displays the reason given by the concerned airline for the delay if there is any (e.g. weather or mechanical).

Beatty et al. [6] first proposed the concept of a Delay Multiplier as a measure of the propagation of delays based on the aircraft and crew schedules of an airline. This study highlighted the need of considering delay propagation, since a 2 hour initial delay could result in a multiplied delay of more than 4 hours. Schaeffer and Millner [20] analyzed propagation of weather

based delays through their Detailed Policy Assessment Tool. The air traffic system is modeled as a network of queues and they showed that significant delay can propagate to the first leg of travel when the capacity-to-demand ratio is too low. Mueller and Chatterji [21] created probabilistic models of departure and arrival delays by fitting Poisson and Normal distributions to the historic delay data from 10 airports. Wang et al. [22] created a recursive flight delay propagation model tuned using historic flight data that separates controllable factors and random factors in order to better understand how an airport configuration can impact delay propagation.

Xu et al. [23] proposed a Bayesian network model of the airports system based on human expertise and validated on historical data of the US national airspace in order to estimate delay interaction between airports. Later Liu and Yang [24] also implemented an improved Bayesian Network model for the Chinese airspace in order to estimate flight delays. Liu and Ma [25] used a similar Bayesian Network model to analyze delay propagation in China, concluding that in some cases cancellations were beneficial to halt major delay propagation.

AhmadBeygi et al. [26] studied the propagation of delays as propagation trees using the schedule of two type of airlines - hub-and-spoke and low-cost. They showed that though around 40% of flights do not propagate their initial delay, for half of the remaining flights the propagated delay more than doubles the initial delay. Tu et al. [27] decomposed push-back delays into seasonal and daily variations in order to implement a flexible continuous probability model able to estimate delays and tested it on the historic data from a specific airport and a specific airline. Sridhar and Chen [28] proposed to predict short-term delays based on the weather impacted traffic index (WITI) [29] and a predicted weather index along with air traffic demand. Klein et al. [30] also proposed a model for short-term delay prediction based on a more granular version of WITI, separating it into twelve different components per airport. Sridhar et al. [31] compared the performance of different neural networks for predicting various aircraft delays using various metrics, showing that models should be season-based and that weather related metrics are a good proxy of flight delays.

Churchill et al. [32] proposed two models for analyzing delay propagation, a microscopic model taking each aircraft as a different unit and a macroscopic considering the arrival and departure flows at each airport, each model giving insights at their own level. The microscopic model showed that propagated delay account for 20 to 30% of the total reported flight delays and the macroscopic model highlights the dependencies between airports with respect of delay propagation. Rebollo and Balakrishnan [33, 34] implemented a network model to classify and predict future delays on specific links or specific

airports using two years of flight-centric and weather-related data. Pyrgiotis et al. [35] later created a stochastic and dynamic queuing model designed to quickly compute approximate delays at 34 US airports, treating these airports as a set of interconnected individual queuing systems. Fleurquin et al. [36] introduce a notion of airport congestion used to measure the level of system-wide delays by considering the size of clusters of congested airports. They also created a historic data-driven model taking into account aircraft and crew connections as well as passenger connections. They conclude that crew and passenger connections are the most effective in introducing delays.

Aljubairy et al. [37] took a different approach from previous studies by considering Internet of Things rather than historical data. They propose a framework to scrape and clean data from both weather and flight-related sensors available in real-time enabling them to classify the delay of an upcoming flight. The derived model can then be used to visualize the performance of seven Chinese airports and their associated airlines based on the estimated flight delay.

Gopalakrishnan and Balakrishnan [38] compared several methods in predicting delays at US airports using a network approach to air traffic delays. Using date related and flight delay related features, they evaluated the performance of various machine learning models along with a delay network dynamic model in predicting various delay measures from two hours up to twenty-four hours in advance. Roy et al. [39] implemented three theoretical vulnerability metrics of the air transportation system based on a Laplacian graph-view approach to flight delays and evaluated them based on simulated situations of severe weather conditions as well as cyber-attack impacting the air traffic management system. Li et al. [40] proposed a graph signal processing approach to delays in the US air transportation system. Based on ten years of flight-centric data, they extracted the spatial delay trends of the network and used it to analyze the effect of several severe weather perturbations such as hurricanes and winter storms. Li et al. further improved this graph signal processing approach in [41] proposing an outlier analysis framework and applying it to compare the US and China airspaces with respect to their spatial delay specificities.

To the best of the author's knowledge, these previous works to predict or classify flight delays were all centered on flight-centric information coming from a variety of sources with different levels of public availability, yet using only very little passenger-centric data.

2.2 Passengers are the core of the system

Already in 1980, Conner [42] illustrated the need to consider the balance between passenger comfort and its associated cost in decision making both for public and profit-making services. Later in 1992, Lemer [43] advocated for the need of unified airport performance measures that would balance the expectations of passengers, airlines and airports along with the expectations of other actors (such as shops or governments). Delay time and crowding was already a measure of airport performance though no systematic way of measuring it was available. Matthews [44] presented an airport performance measure based on hourly passenger flows, which considers that an airport should be able to cope for all hourly passenger flows with the possible exception of the top 5% peak flows. The importance of airport experience in customer, i.e. passenger, satisfaction towards both airline and airport services is highlighted in the study of Pruyn and Smidts [45], where they show that customer satisfaction is largely affected by their experience at waiting areas, both in terms of wait times and wait environment.

Understanding the passenger experience, or at least the passenger perception of airport and airline quality has since been the focus of many studies. Robertson et al. [46] took a reversed engineering process approach and proposed a model for estimating passenger arrival at airports with a 30 minute window using publicly available airline data in order for airlines and airports to better engineer the full passenger experience within the airport. Later Brown and Madhavan [47, 48] created a simulated model of passenger flow through airline check-in and airport security checkpoints from data gathered at Norfolk airport confirming these two areas as main chokepoints for airport passenger flow.

Tsaur et al. [49] first proposed to introduce surveys based on fuzzy set theory in order to analyze airline service quality. They applied their survey to evaluate the performance of three Taiwanese airlines and concluded that the most important attributes are courtesy, safety and comfort. Chang and Yeh [50] proposed a method of evaluating airline service quality using a multi-criteria analysis survey enabling airlines to better understand what were their internal and external advantages with respect to their local competition. Aksoy et al. [51] conducted a survey of customer satisfaction on four different city pair trips comparing a domestic airline with the destination city associated foreign airline and concluded that there were significant differences between the two passenger groups, indicated by different profile, behavior and expectations with respect to the airline customer service. Magri Junior and Alves [52] conducted a performance analysis of six Brazilian airports using passenger centered quality indicators developed by the Air-

ports Council International (ACI) [53]. These thirty-six quality indicators concern all areas and functions within the airport, from facility cleanliness to availability of service and presence of flight information display systems. They were assessed only over three days per airport due to the difficulty and the wide spatial range of the measurement process. Gkritza et al. [54] analyzed a eight month long phone survey between 2002 and 2003 on passenger satisfaction at security screening points and concluded that passenger satisfaction was not solely determined by wait times, though wait times were a significant factor, and that factors influencing this satisfaction could vary over time.

Hunter [55] performed a thorough survey of airline perception related studies from 1995 to 2006, pointing out the decrease in customer service throughout the airline industries. She also analyzed the relation between passenger expectation of service and passenger perception of service with the air rage phenomenon, finding that when passenger perception or expectation decreased, passengers were more understanding toward air rage outbreaks even though they were not more inclined to behave in such an extreme fashion. Pakdil and Aydin [56] proposed a new survey structure to encompass more dimensions of airline service quality and tested it on three different flight segments for one airline. They concluded that passenger's past experience was the most important factor in selecting the airline even if there was always a gap between passenger expectation of service and passenger perception of service, indicating that airlines could use more incentives to improve their customer service. Chou [57] proposed a survey model for the evaluation of airport service quality and used it to compare the performances of two major Taiwanese airports. Their model indicated that staff courtesy (from airlines, customs and immigrations) were the most important service criteria for the surveyed passengers. Chou et al. [58] later applied the same method to evaluate airline service quality for a Taiwanese airline and concluded that safety, customer complaint handling and courtesy were the top three service dimensions for passengers.

Popovic et al. [59] conducted a video study at Brisbane Airport in order to analyze the interactions of passengers with airport staff and infrastructure. They observed that staff were more focused on helping the technology and the information displays rather than the passengers and that the activities (both necessary and discretionary) undertaken by passengers were impacted by the hand luggage they had to carry. Chiou and Chen [60] decomposed airline service quality into a chain of seven services (from seat reservation to complaint response) and analyzed both the overall service framework and the service quality chain based on surveys distributed to passengers from a Chinese low cost carrier. They concluded that low satisfaction trickled down

the service quality chain, meaning that airlines should improve the different element of this chain from beginning to end. They also showed that service quality had the second biggest effect on behavioral intentions, behind service value. For more informations on the various survey-based methods used, de Oña and de Oña [61] conducted a survey of survey based analysis of public transportation system. They concluded that even though researchers keep trying to improve the complexity of the models to better model passenger satisfaction of a public transportation system, managers and practitioners use simpler models in order to reach their goal of improving passenger perceived service quality for an increase of income.

These passenger surveys conducted at airports for airports or airlines, while very detailed, remain limited to small samples of passengers and short time periods, and may not be representative. They are also expensive and time consuming to implement, making their use for measuring the performance of the full air transportation system cumbersome and difficult to update.

A passenger approach to analyzing flight delays was first introduced by Bratu and Barnhart [62] who developed a Passenger Delay Calculator to show that flight-centric metrics do not accurately reflect passenger delays, especially due to flight cancellations. Later in [63] they calculated passenger delay using monthly data from a major airline operating a hub-and-spoke network. They show that disrupted passengers, whose journey was interrupted by a capacity reduction, are only 3% of the total passengers, but suffer 39% of the total passenger delay. Wang et al. in [64, 65] showed that high passenger trip delays are disproportionately generated by canceled flights and missed connections. Nine of the busiest 35 airports cause 50% of total passenger trip delays. Congestion, flight delay, load factor, flight cancellation time and airline cooperation policy are the most significant factors affecting total passenger trip delay. These studies have highlighted the disproportionate impact of airside disruptions on passenger door-to-door journeys. Flight delays do not accurately reflect the delays imposed upon passengers' full multi-modal itinerary.

This led NextGen [7] in the United States and ACARE Flightpath 2050 [8] to advocate a shift from flight-centric metrics to passenger-centric metrics to evaluate the performance of the Air Transportation System. The failures and inefficiencies of the air transportation system not only have a significant economic impact but they also stress the importance of putting the passenger at the core of the system [66, 67]. Both the USA and Europe aim to take a more passenger-centric approach, with ACARE Flightpath 2050 setting some ambitious goals, including some that are not measurable yet due to lack of available data. In the US, the Joint Planning and Development

Office has proposed and tested metrics regarding NextGen's goals, but there are still metrics missing from the passenger's viewpoint, especially regarding door-to-door travel times [68]. Following this new international impulse, the shift from flight-centric information to passenger-centric metrics was first explored by Cook et al. [69] within the project POEM - Passenger Oriented Enhanced Metrics, where they designed propagation-centric and passenger-centric performance metrics, and compared them with existing flight-centric metrics. Several years later, the advocated shift from flight-centric metrics to passenger-centric metrics still has to be actually implemented by the governing agencies. In a report published in 2016, EUROCONTROL and the FAA presented metrics regarding punctuality that combines airline and passenger views into a single view [9].

Passengers are at the core of this system and, yet, limited quantitative information about passenger movements is publicly shared. Each aviation stakeholder only has access to a partial view of the passenger-side of air transportation operations. Airline passenger information - such as: Tickets, boarding passes, boarding time - is airline proprietary. Each airline therefore has a partial view of passenger movements on board aircraft and on the ground (from check-in kiosks and counters to boarding the aircraft). In the USA, the BTS provides aggregated passenger data per market but no granular information. Airports gather customs or security records, shuttle traffic, parking occupancy, sometimes measure queue lengths, while third-parties collect online traces through WiFi hotspots and Bluetooth beacons [70]. These real-time information, combined with historical data, were used to analyze and predict passenger flow to an Australian immigration booth [71] or within several Dutch train stations [72] as well as for the analysis and prediction of passenger occupancy in a Chinese airport [73]. The studies are limited to a fraction of the full system (one or two airport terminals) indicating the difficulty of gathering a system-wide data-driven picture of passenger behavior.

Sun et al. [74] proposed a passenger-centric analysis of the robustness of the worldwide airport network by introducing a measure based on passengers not affected by rerouting when an airport or group of airport fails and testing it against twelve different attacks on the airport network. Sun and Wandelt [75] later considered the robustness of the airline network using this same passenger-centric measure, noting that traditional airlines with a limited number of hubs could break down entirely with a smaller number of affected airports while other airlines can withstand failures to more than five targeted airports without being entirely disintegrated.

2.3 A passenger-centric shift: towards a multi-modal approach to air transportation

2.3.1 Data sharing for a multi-modal approach

In 2003, Pels et al. [76] conducted a study that showed the importance of the access to airports in the choice of the airport for both business and leisure travelers, already indicating the importance of considering the full door-to-door trip and the multi-modal integration of airports with cities to increase airport attractiveness. Grotenhuis et al. [77] studied the different need for information for multi-modal trips, decomposing the trip into three stages: a pre-trip planning stage, a wayside stage while waiting or transferring from a mode to another, and an on-board stage. They concluded that various information sources are needed and that the information needed was different depending on the stage considered and the passenger profile (e.g. first-time or frequent traveler).

Seamless door-to-door travel and data sharing was later deemed as needed by the European Commission's 2011 White Paper [8] and was reconfirmed by the Federal Aviation Administration (FAA) in 2017 [78]. Data sharing was already a main focus in the early 2000s and led at an air system level to the creation of the architecture SWIM - System Wide Information Management [79] - by Europe and later adopted by the FAA. Sipe and Moore [80] suggests that digital data sharing can improve operational efficiency if air traffic management functions are reallocated between the various elements of the air transportation system.

Klock et al. [81] showed the importance of simplifying and broadening the access to intermodal information in order to make public transportation more competitive against private cars. Focusing on trips between New York and Washington D.C. using a mix of car, rail, bus or plane, they showed that intercity travels could have their time and environmental impact improved by 10% and 25% if the proper information were gathered and used for trip planning.

The concept of Multimodal, Efficient Transportation in Airports and Collaborative Decision Making (META-CDM) was later introduced by Laplace et al. [82] and proposed to link both airside CDM and landside CDM, thus taking into account the passenger perspective. In this perspective, Kim et al. [83] proposed an airport gate scheduling model leading to improved efficiency with a balance between aircraft, operator and passenger objectives. Dray et al. [84] illustrated the importance of multimodality by considering ground transportation as well during major disturbances of the air trans-

portation system in order to offer better solutions to passengers.

Marzuoli et al. [85] later applied the concepts of multi-modal collaborative decision making as a post-analysis of the Asiana crash at San Francisco International airport in 2013 and concluded that not only considering ground transportation for diverted passengers would have reduced the average passenger delay by one hour, but with sufficient information sharing between airlines and airports all the concerned flights could have been diverted towards the three airports within the Bay Area rather than across the state or to another state. Marzuoli et al. also conducted the first analysis of multimodal perturbation propagation [86], showing that the Asiana crash had significant repercussions in flight traffic but also in the road and public transit systems surrounding the airport. Dray et al. [87] proposed a framework for considering ground transportation to reduce airline costs and passenger delays when airports suffer disturbances leading to cancelled flights over a period of one to ten hours. Their study looked into the network of the top fifty European airports and concluded that using ground transportation for a small portion of stranded passengers could reduce the airline cost by 20% and the mean passenger delay up to 70%.

Both NextGen and ACARE Flightpath 2050 intend to not only improve the predictability and resilience of the Air Transportation System, but also to reduce door-to-door travel time for passengers. Regarding door-to-door travel times, ACARE FlightPath 2050 aims at having 90% of travelers within Europe being able to complete their door-to-door journey within 4 hours [8].

Door-to-door travel time estimation with a multi-modal approach has been previously studied but for travels contained within the same metropolitan area. Peer et al. [88] studied door-to-door travel times and schedule delays for daily commuters in a Dutch city, showing the importance of considering the correlation of travel times across different road links when estimating the overall travel time. Salonen and Toivonen [89] investigated the need of comparable models and measures for trips by car or public transport within Helsinki, introducing a multi-modal approach when considering the walking and waiting necessary to reach a station or a parking spot. Duran-Hormazabal and Tirachini [90] focused on travel time variability for multimodal trips within Santiago, Chile, using both GPS data and surveyors to estimate the time spent in the different considered modes (walking, car, bus and metro). These studies emphasized the importance of considering all relevant modes when estimating door-to-door travel times, but were limited in scope by the area considered and the data available. Wandelt et al. [91] proposed a method to extract the worldwide railroad network from open source data, which can then be used to improve the estimation of door-to-door multi-modal travel times for trips having a rail component.

Grimme and Martens [92] proposed a model to analyze the feasibility of the 4 hour goal within FlightPath 2050 based on airport to airport flight times and a uniform model of access and egress to airports. Sun et al. [93] implemented a door-to-door minimum travel time estimation based on open source maps and datasets in order to study the possible competitiveness of air taxis. Cook et al. [94] proposed an event-driven model of the door-to-door travel based on sample data within the project Dataset2050[1].

2.3.2 Passengers as a signal

Larger scale studies with a focus on air transportation was later possible thanks to the increasing use of mobile phone devices as datasources since most individuals now carry a cell phone, and heavily use it through out the day. Phone carriers collect Call Detail Records (CDR), indicating when an individual makes a phone call, texts, or browses online, as well as their approximate location when doing so. Please note that such records belong to the carriers and are generally not publicly available. Only in a few instances have partial data sets been anonymized and released for research applications.

As early as 2008, Work and Bayen demonstrated the use of smartphones to monitoring highway traffic in the Bay Area [95]. Gonzalez et al. showed how large scale studies of CDRs can help understand individual mobility patterns [96]. Blondel et al. provided a thorough survey [97] of applications of mobile phone data from mobility, to urban planning and help towards development in Africa for instance [98, 99]. De Montjoye et al. [100] built a Python toolbox to help researchers analyze, visualize and build robust features from mobile phone data. Douglass et al. [101] provided high resolution population estimates from mobile phone data. Alexander et al. [102] showed that CDRs can be used to identify home and work locations reliably and allow the extraction of additional frequent locations, activity travel diary validated comparing them to household surveys. Picornell et al. [103] leveraged CDRs to study the relationship between travel behavior and social networks, highlighting the role of social networks in the presence of individuals at locations other than home and work. Toole et al. [104] focused on using CDRs for urban planning, and in particular travel-demand estimation to provide validated origin-destination matrices on the ground and road usage patterns. More recently Bachir et al. [105] used CDRs along with four other data-sources to study origin-destination flows for the Greater Paris region.

In the field of analyzing air transportation, precursor work was made by

[1] `www.dataset2050.eu`

Marzuoli et al. in [106] using mobile phone data in order to analyze the performances of airports from the passengers' perspective. This study validated the use of this passenger-centric data to better assess the overall health of the Air Transportation System. In Europe, within the BigData4ATM project[2], Garcia-Albertos et al. [107] presented a methodology for measuring the door-to-door travel time using mobile phone data and applied it between two Spanish cities, Madrid and Barcelona. Burrieza et al. [108] later used this same data to showcase a model enabling to better characterize passengers going through Madrid Barajas airport than traditional surveys. Garcia-Albertos et al. [109] also used this method and dataset to analyze some of ACARE Flightpath 2050 goals and showed that full door-to-door trips going through Madrid Barajas airport were far from the four hour ambition.

Though these studies give a full door-to-door view of trips making use of air transportation, mobile phone data are proprietary data and are not often publicly available. In order to operate in real-time, it is thus necessary to also look into other sources of passenger data available on a national scale.

2.3.3 Non-traditional data sources for air transportation

Data gathered from passengers mobile phone in the aforementioned studies can be considered as data gathered by reading signals passively emitted by passengers during their travel, in the sense where the passenger is not actively trying to communicate with the air transportation system via their mobile phone. On the other hand, the ubiquity of mobile phones allows passengers to actively share their experience via social media.

And indeed with more than 200 millions active mobile social media users in Europe [110], social media is another popular source of data previously used for studying large-scale behaviors, in particular Twitter. Twitter is a popular social microblogging service, in which users post messages, called *tweets*, containing no more than 280 characters (with an initial upper limit of 140 characters until November 2017). With more than 64.2 millions active users in the United States in April 2020 [111], Twitter is in effect an important pool of user-created data.

Twitter has already been the main focus of many studies, including studies on its network topology by Java et al. [112], Krishnamurthy et al. [113] and Huberman et al. [114], as well as studies on the different categories of tweets using various text-mining and machine learning techniques. These studies have to address the double difficulty of the important amount of

[2] `www.bigdata4atm.eu`

posted tweets along with the small size of each tweet. Read [115] used the explicit meaning of emoticons in order to efficiently extract tweets easy to label for sentiment analysis, while Coletta et al. [116] combined classification and clustering techniques to overcome the shortness of tweets for sentiment analysis. *Hashtags* - user defined tags sometimes present in tweets - were used to efficiently cluster tweets into six coarse-level topics (e.g. news, sports and entertainment) by Rosa et al. [117] and into nine general domains (e.g. music, sports and political) by Tsur et al. [118]. Lehmann et al. [119] studied popularity peaks of hashtags, which indicates the occurrence of an event, focusing on the social propagation differences between the four prototypical class of temporal peaks - namely on the day of the event, on the days leading to the event, before and after the event, and on the days after the event.

Regarding large-scale events, the use of Twitter during natural disasters has been the focus of many post event studies. Kireyev et al. [120] analyzed topics contained within tweets written following two earthquakes in 2008, Vieweg et al. [121] and Palen et al. [122] studied how Twitter was being used throughout foreseeable natural disasters (e.g. for pre-warning, warning and evacuation), with the example of the Red River flood in 2009. Terpstra et al. studied how a real time Twitter analysis could have provided valuable information for the operational response of a natural disaster crisis management with the case of the storm hitting a festival in Belgium [123].

Sakaki et al. [124] used the fact that some tweets are geolocalized to consider Twitter users as sensors for a faster detection and information propagation during earthquakes. The use of Twitter for real-time surveillance of disease propagation has also been analyzed and implemented in some cases with for example the case of the 2009 H1N1 pandemic by Chew and Eysenbach [125]) and the case of the 2012-2013 influenza epidemic by Bronitaski et al. [126]. Houston et al. [127] conducted a thorough survey of the use of social media during disasters and narrowed down fifteen categories of uses for social media before, during and after a disaster. Takahashi et al. [128] then analyzed these categories during a typhoon in the Philippines to better understand which kind of users would participate to the different uses.

More recently, Priya et al. [129] proposed a framework to retrieve tweets relevant to earthquakes in order to assess infrastructure damage following the earthquake and applied it to earthquakes in Italy and Nepal. Srivastava and Sankar [130] combined weather data and Twitter data to extract critical data relevant to extreme weather perturbations in real-time with a focus on hurricanes making landfall in the US.

Another popular use of Twitter as a user generated textual data is sentiment analysis and many studies have focused on improving sentiment analysis since Pang et al. [131] thanks to the increase of available online reviews.

However, most works on Twitter sentiment analysis focus on analyzing and improving the performance of classifiers such as Pak and Paroubek in [132] or Da Silva et al. in [133] and lack an application of the classifiers output. A thorough survey and classification of sentiment analysis methods was undertaken by Pang and Lee in [134].

Passenger sentiment analysis on Twitter seems a promising approach to the creation of a passenger-centric metric, and most works mining Twitter data for the air transportation field actually focus on how airlines are perceived by passengers by means of sentiment analysis [135] or sentiment classification [136]. Misopoulos et al. [137] analyzed airline customer service experiences both by manually labelling tweets related to airlines containing one of three keywords ("good", "fail" and "lounge") into six categories (personal, positive, negative, promotion, question or news) and then by applying sentiment analysis to the gathered tweets.

Very few works actually propose an application of the classifiers output. Wang et al. [138] presented a framework to visualize real-time sentiment during political events in the United States using a crowd-sourced labeling method. Siau [139] used sentiment and topic analysis to extract from around a thousand tweets the information needed to calculate a proxy of the Airline Quality Rating, a flight centric metric including a measure of customer complaints introduced by Bowen et al. [140]. Samonte et al. [141] proposed a sentiment analysis pipeline with some simple post analysis of the classification results and applied it to local airlines in the Philippines.

A more recent work from Gitto and Mancuso [142] focused on a different category of actors of the air transportation system by analyzing the brand perception of 118 airports worldwide. Khandpur et al. [143] took a security approach and proposed a framework to determine real-time relative airport threat levels by analyzing tweets containing expert-determined keywords along with any news article referenced in these tweets. Gunarathne et al. [144] looked into the interaction between passengers and airlines and shows that airlines are more likely to respond to customers with greater popularity, and have a tendency to respond more to complaints than to compliments.

Though these works give some insight on how passengers perceive the state of specific actors within the air transportation system, they do so for the benefit of the airlines and airports, not of the passengers.

2.4 Conclusion

The number of flights have been steadily increasing in the last ten years, along with the number of carried passengers, and flight delays remain a major

concern for the regulating agencies and for passengers. Flight delays and how they propagate via aircraft, airports, passengers and crew have therefore been thoroughly investigated in the literature, mostly thanks to the availability of flight-centric data. Studies have however shown that flight delay is not a good representation of passenger delay, and that the passenger experience can be disproportionately impacted by flight delays and cancellations. The need for passenger-centered metrics to complement the measures of the air transportation system performance is being advocated by federal and supranational agencies, and the shift from a flight-centric view to a passenger-centric view is still a work in progress.

Passenger experience at airports or with airlines has been traditionally measured via thorough survey-based studies, yet with a usually small passenger sample and over a limited time period. A broader approach is therefore necessary, especially given the fact that the journey of a passenger is not limited to the airport to airport segment. A recent promising approach is to consider passengers as a signal throughout their journey, thanks to data emitted by their smartphones. Though this approach does give a door-to-door view of the passenger journey, restrictions on data property and data privacy add limitations to the use of data for public research. The ubiquity of smartphones has also enabled the increase use of social media in real time, enabling researchers to study the effects of large-scale events on people, especially via the social media Twitter. Twitter has started to be used to study some aspects of the air transportation system, with the majority of studies focusing on sentiment analysis applied to tweets related to airlines. This thesis proposes to explore further how data generated by passengers can be used to offer a new perspective of the air transportation system, with a focus on data available in real-time. In Chapter 3, a method to transform the Twitter stream into a reliable real-time estimator of the number of delayed and cancelled flights in the United States is presented.

Chapter 3

Passengers on social media: A real-time estimator of delays and cancellations in the US air transportation system

This chapter presents a pipeline that transforms the activity of passengers on the social media Twitter as a real-time estimator of the state of the US Air Transportation system. A new feature extraction process is implemented on this passenger-generated dataset that enables an accurate estimation of the hourly number of abnormal flights in the United States. These estimation models based on passenger-generated data have a higher performance than time-series forecasting models trained on historic flight-centric data. Analyzing the importance of the features extracted from the Twitter stream in the estimation process highlights the importance of taking a passenger perspective when analyzing the performance of the air transportation system.

Our first work [11] uses publicly available Twitter data created by passengers to accurately estimate and predict the hourly status of the US air transportation system aggregated at a national level. This method was further improved in [12] to reliably estimate the hourly delays at departure and at arrival per airport. The derived model as well as the results and their analysis are presented in Appendix B for an easier reference.

This chapter builds upon these previous works in order to present a novel passenger-centric tool to estimate the state of the air transportation system by estimating the hourly number of abnormal flights of eight major airlines at each of the 34 major airports within the United States. The regressor models used for this estimation are based on three different levels of content-related features created from the flow of social media posts.

The rest of the chapter is structured as follows: Section 3.1 presents the estimation problem considered and the filtering process enabling the extraction of features from the Twitter stream. Section 3.2 then compares the created estimation models with prediction models based on the historic values of the number of abnormal flights and analyzes their respective performances. Section 3.3 discusses the data and method used and concludes with potential future research directions.

3.1 Extracting features from the Twitter stream for a real-time estimation of flight-centric values

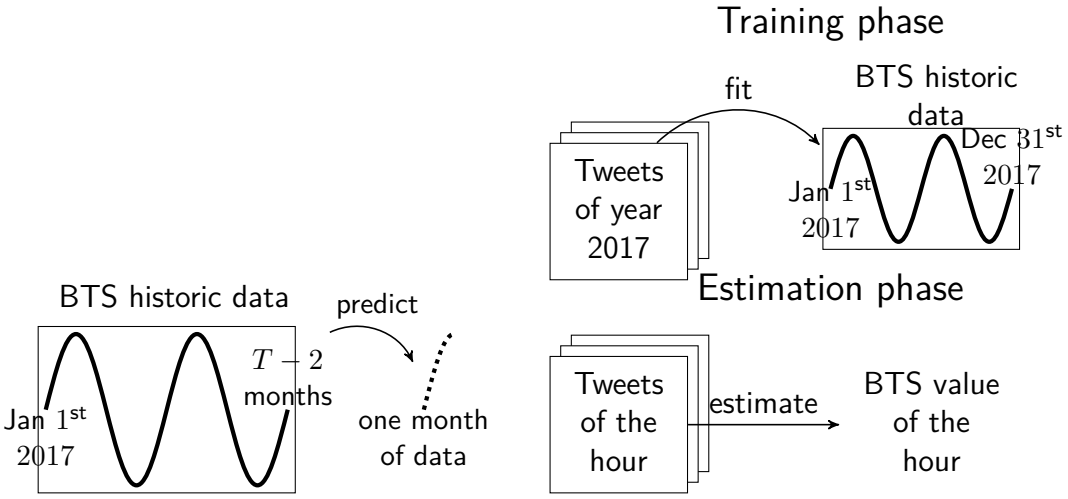
This section presents the filtering process implemented in order to create features from the Twitter stream, which are then used to estimate in real-time the hourly number of delays and cancellations across 34 US airports in Section 3.2.

3.1.1 The advantages of using social media

Estimation vs. prediction of BTS values

The Bureau of Transportation Statistics (BTS) [2] centralizes flight information such as on-time departure for domestic flights, and publishes monthly reports two to three months later. Therefore, any tool aimed at estimating today's National Air Space performance using BTS data only must do so by using data that is at least two months old. A real-time estimator of the number of abnormal flights per airport based on data available online and in real-time, such as tweets, could be of use for all stakeholders of the air transportation system, including passengers. Abnormal flights are here defined as flights departing with a delay greater than 15 minutes, flights arriving with a delay greater than 15 minutes and cancelled flights.

Figure 3.1 presents the different approaches considered in this study, i.e. predicting BTS values based on historical BTS data (Figure 3.1(a)) versus estimating BTS values using real-time available passenger-centric data (Figure 3.1(b)).



(a) Prediction of BTS values using historical BTS data

(b) Estimation of BTS values using passenger-generated Twitter data.

Figure 3.1: Prediction based on historical values vs. estimation based on real-time data.

When predicting using historical BTS data, Facebook's time-series forecasting tool Prophet [145] is used. The Prophet tool is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality [146]. It is described as robust to outliers and missing data with no

parameter tuning necessary, therefore the default parameters of the Prophet tool are used for this forecasting benchmark. As illustrated in Figure 3.1(a) the Prophet tool uses all previous BTS data from January 2017 up to two months before the data to be predicted.

The estimation process based on Twitter data is illustrated in Figure 3.1(b): The models are trained once based on 2017 data and then used to estimate the hourly BTS values from 2018 using only the tweets gathered from the considered hour. For this study, a random forest regressor [147] implemented in the scikit-learn python library [148] is used with the following hyper parameters: a maximum depth of 10, a maximum number of 30 estimators and a minimum sample split of 2.

Overview of passenger Twitter activity

Following the initial work performed in [11] and [12], the goal of this study is to use the social media activity of passengers, airlines and airports - in particular their Twitter activity - in order to build an estimator of the flight-centric health of the US air-transportation system at an airport level. In this study, the flight-centric health of an airport is described by delay and cancellation related information contained within the BTS data. This data is publicly available usually with a two to three month delay and this study limits itself with the BTS data from January 2017 to December 2019.

The period of Twitter activity considered in this study also spans from January 2017 to December 2019. The Twitter stream is first filtered by searching and extracting all the tweets related to one of the handles of 8 major US airlines or to one of the handles of 34 major US airports. The full list of handles can be found in Table 3.1. A tweet is related to a handle if it is written by the owner of the handle, if it is a direct reply to the owner of the handle or if it contains the handle within its text. All tweets written by these airlines or airports Twitter accounts are categorized as "customer service tweets". All the other tweets related to these airlines and airports Twitter handles that were not written from the corresponding airline or airport Twitter account are categorized as "passenger tweets".

Figure 3.2 shows the total number of tweets related to each airline and airport over the year 2017 against the total number of flights flown by each airline or from each airport. As can be seen in Figure 3.2(a), airlines tend to be associated to more tweets than airports, with the three main airlines gathering more than 800,000 tweets over the year 2017 each. The number of tweets related to each airline is not necessarily correlated to the number of flights flown per airline. Delta generated the most tweets over 2017 even though Southwest Airlines carried out the most flights in 2017. Zooming

Table 3.1: Twitter handles used for gathering tweets.

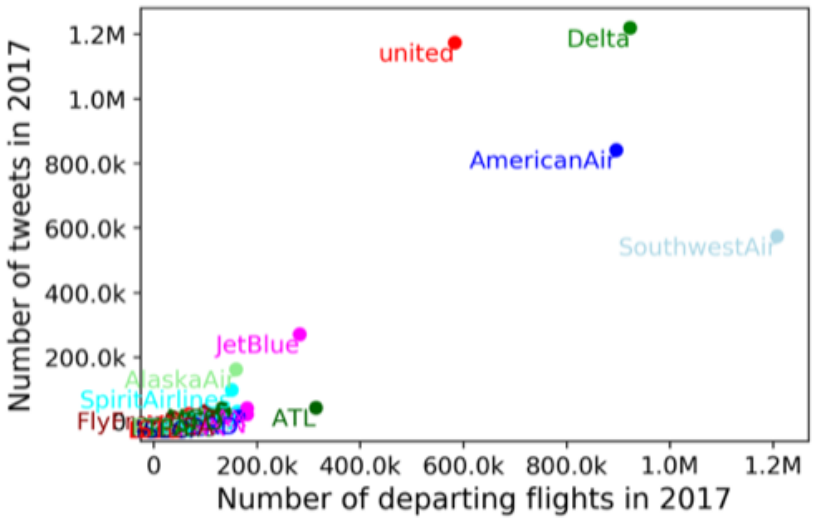
Category	**Twitter handles**
Airlines	@united, @Delta, @AmericanAir, @AlaskaAir, @SouthwestAir, @SpiritAirlines, @JetBlue, @FlyFrontier, @FrontierCare
Airports	@JFKairport, @ATLairport, @flyLAXairport, @fly2ohare, @DFWAirport, @DENAirport, @CLTAirport, @LASairport, @PHXSkyHarbor, @iflyMIA, @iah, @EWRairport, @MCOAirport, @MCO, @SeaTacAirport, @mspairport, @DTWeetin, @BostonLogan, @PHLAirport, @LGAairport, @FLLFlyer, @BWI_Airport, @Dulles_Airport, @MidwayAirport, @Reagan_Airport, @slcairport, @SanDiegoAirport, @flyTPA, @flypdx, @flystl, @flySFO, @HobbyAirport, @flynashville, @Fly_Nashville, @AUStinAirport, @KCIAirport

in from the airport perspective in Figure 3.2(b) indicates that most airports generated less than 30,000 tweets in 2017, Orlando International Airport (MCO), Los Angeles International airport (LAX) and Hartsfield-Jackson Atlanta International airport (ATL) are outliers with around 40,000 tweets over the year. ATL is also an exception from a flight volume perspective, since it is the only airport with over 300,000 departing flights in 2017, the other airports having all less than 200,000 departing flights over the year 2017.

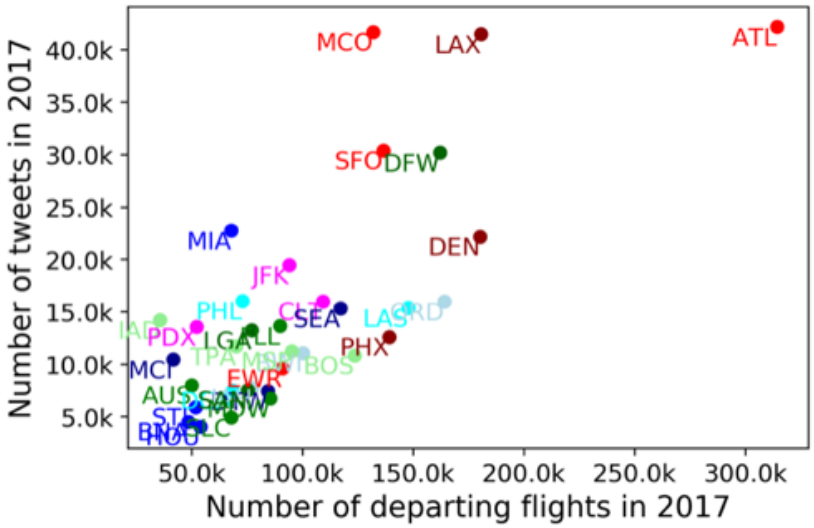
Estimating the number of abnormal flights of each considered airport requires to first extract this information from the BTS dataset for each airport. Three types of abnormal flights are considered here from a passenger's perspective: Flights departing with a delay greater than 15 minutes, flights arriving with a delay greater than 15 minutes, and the cancelled flights. Once all the flights departing or initially scheduled to depart from an airport and all the flights arriving at the same airport are selected, the following values are aggregated per hour:

- NumDepDelay15: Number of flights departing with a delay greater than 15 minutes
- NumArrDelay15: Number of flights arriving with a delay greater than 15 minutes
- NumCancelled: Number of flights cancelled

These values are calculated at each airport considered in Table 3.1 and based only on flights flown by the eight airlines of that same table.



(a) Both airlines and airports



(b) Only airports

Figure 3.2: Number of tweets vs. the number of flights during the year 2017 for the airlines and airports under consideration.

3.1.2 Feature extraction from airline and airport related tweets

This section presents the feature extraction process that takes place on the filtered Twitter dataset described in Section 3.1.1. For airlines and airports with several Twitter handles, their Twitter handles are considered jointly. For

example, tweets related to "@FlyFrontier" and "@FrontierCare" are merged and considered as tweets related to Frontier Airlines. For simplicity they were labelled as "@FlyFrontier" related tweets.

Volume features

Volume related features are extracted identically for all airlines and airports whose Twitter handles are presented in Table 3.1. For all the volume related features, a distinction between passenger tweets and customer service tweets is made. The first volume related features considered are the hourly number of passenger tweets and the hourly number of customer service tweets for each airline and each airport.

In addition to the hourly volume of tweets, the hourly volume of tweets containing a certain specific keyword is also extracted from the filtered Twitter dataset. Six keywords are chosen for this study related to cancellations and delays: 'delay', 'wait', 'hours', 'cancel', 'refund' and 'voucher'. These keywords were chosen since they relate closely to the aim of this study, i.e. estimating the number of delayed and cancelled flights, even when taken out of context. In order to consider all the relevant tweets without having to exhaust all the possible forms of the chosen keywords (e.g. "delay" can be written within the words "delayed", "delays", etc.), regular expression filters are created for each keyword: Any tweet containing at least one word starting with the considered keyword is kept and the number of resulting tweets is then aggregated per hour. As for the hourly volume of tweets, the hourly number of tweets containing a keyword is calculated separately for passenger tweets and customer service tweets.

Sentiment features

The next group of features extracted from the gathered tweets are features based on the sentiment analysis of these tweets. For these features, only tweets written in English or in Spanish are considered. The language of each tweet is initially taken as the one indicated by Twitter's API. The tweets labelled as "unknown" are then processed through a language recognition algorithm and their language label are updated accordingly. Using the Natural Language Toolkit NLTK [149] and based on the work of [150], the number of common stop-words contained in a tweet is extracted for each available language in NLTK and the language with the highest count is selected. Due to the limited length of each tweet, a bias towards English has been introduced as well in the count ordering, i.e. if English and another language have the same count of common stop-words, English takes precedence.

Twitter sentiment analysis usually consists in labelling whether a tweet conveys a *positive* or a *negative* mood. For this labelling process to be the more accurate possible, good training sets containing pre-labelled tweets have to be created, with a similar quantity of tweets conveying a *positive* mood and of tweets conveying a *negative* mood. For each language, the labelled dataset created is based on the works of [115, 151]. 49,030 English tweets and 1,998 Spanish tweets were extracted from the total dataset of tweets written in 2017 by airline customer services and by passengers using emoji filters. The emojis used are associated with a positive or negative sentiment, indicated in Table 3.2, which enables to assign automatically a positive or negative sentiment label to every tweet.

Table 3.2: Emoji sentiment association.

Category	**Emojis**
Positive	":)", "=)", ":-)", ";)", ";-)", ":-D", ":D", "=D"
Negative	":(", ":-(", "=(", ":-@", ":'(", ":-|"

Each tweet goes through the following processing pipeline in order to transform its text into a vector of tokens that will be fed to the sentiment classifiers. A token can be either a single word, a generic keyword, a bigram or a trigram. A bigram is a combination of two consecutive words commonly used together within the full considered dataset, and a trigram is a combination of three consecutive words commonly used together. For example, "record locator" is a bigram commonly used by American Airlines customer service. Generic keywords are used to reduce the sparsity of the considered vocabulary. For example, generic keywords replace mentions to the considered airlines and airports (e.g. "@united" becomes "AIRLINE") and mentions to other Twitter users ("@someone" becomes "MENTION"). Generic keywords also substitute association of date related words, e.g. "January 12th 2018" is replaced by the keyword "DATE" and "2pm" by the keyword "TIME". Additional generic keywords indicate if a picture is embedded in the tweet or if the tweet contains a link to a website. Furthermore, since every tweet in the training set contains an emoji, the generic keyword "EMOJI" replaces each emoji found using a regular expression filter in order to remove any potential bias on the sentiment learning process.

Words in a tweet can be loosely written, with for example repeated letters indicating an emphasis on a specific word, such as "loooove" or "looooooooooove", which has the potential of greatly increasing the sparsity of the considered vocabulary. In order to limit this increase in sparsity, the number of duplicate letters within a word is limited to two: both "loooove" and "looooooooooove"

are simplified to "loove". Similarly to the work of Read [115], negative bigrams are created by merging some negation words - "no" for English and Spanish, as well as "not" and "never" for English - with the word that follows it. A last step to reduce the size of the vocabulary without removing any important token is to remove the tokens occurring in fewer than twenty tweets or in more than 75% of the tweets within the training dataset.

Five classifiers for each language are trained on the datasets extracted from the emojis of Table 3.2: a naive Bayesian classifier [152], an AdaBoost classifier [153], a random forest classifier [147], a gradient boosting classifier [154] and a logistic regressor [155] using the scikit-learn python library [148] and tested on the labeled dataset provided for a Kaggle competition [156] containing airline related tweets from February 1st 2015 for the classifiers based on the English tweets.

Once the classifiers are trained, they yield a score of 1 if they consider that a positive sentiment is conveyed within the tweet and a score of 0 if they consider that a negative sentiment is conveyed within the tweet. This score is based on a predicted probability for a tweet of conveying a positive sentiment that each classifier was trained to estimate and that is then rounded to the closest integer (0 or 1). The five trained classifiers are transformed into regressors by removing the rounding step and considering directly the probability for a tweet of being classified as conveying a positive sentiment. The output of the five obtained regressors is then averaged into one single sentiment score. A sentiment score of 0 indicates that the tweet conveys a negative sentiment and a sentiment score of 1 indicates that the tweet conveys a positive sentiment. The sentiment scores for English and Spanish tweets are finally aggregated per hour, per airport/airline and per user category, similarly to the volume feature extraction presented in Section 3.1.2.

This tokenization process introduces two additional keywords that can be added to the volume features presented in Section 3.1.2: counting the number of tweets containing a picture and the number of tweets containing a website link. Thus, eight keywords are actually considered for the extraction of volume related features: 'delay', 'wait', 'cancel', 'hours', 'refund', 'voucher', 'PICTURE', 'WEBSITE'.

Topic features

The last group of features extracted from the filtered Twitter database is based on topic analysis using Latent Dirichlet Allocation [157] (LDA). In LDA, each document of the considered corpus is modeled as a finite mixture of topics. A topic is defined as a distribution over the words composing the full corpus of documents. The topic distribution of each document and the

word distribution of each topic can be determined using variational Bayes approximations and was implemented in Python by Rehurek and Sojka [158] within the Gensim library.

In order to consider only topics relevant to the goal at hand, i.e. estimating the number of delays and the number of cancellations per airport, a pipeline to calculate the topic distributions related to a specific keyword is implemented. The keywords considered for these features are 'delay', 'cancel', 'refund' and 'voucher'. The first step is to extract all tweets written in 2017 containing the keyword. Then, the tweets written during the same hour, e.g. from 1:00 pm to 2:00 pm, are merged into a single document, since LDA does not work very well with short documents and tweets are limited to 280 characters. These documents are then transformed into vectors of tokens, similarly as for the sentiment analysis presented in Section 3.1.2. LDA is then used to extract the five main topics within this corpus of hourly documents. The pipeline for creating the 20 topics related to the four chosen keywords is presented in Figure 3.3.

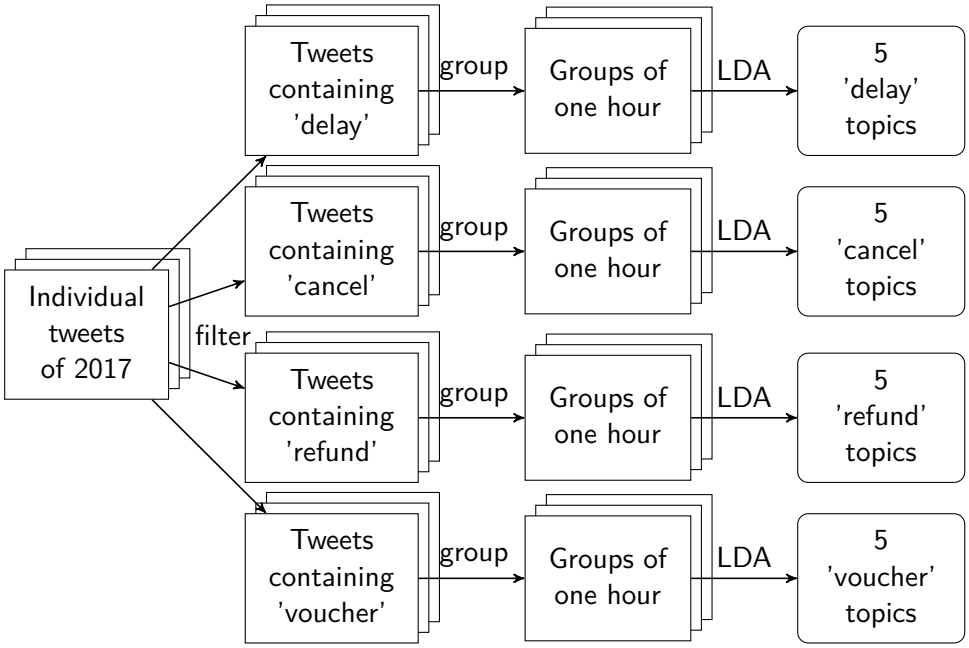


Figure 3.3: Creation process of the keyword-related topics.

A representation of the five topics related to the keyword "delay" extracted from the 2017 database is presented in Table 3.3. The word distribution of each topic is then applied on each individual tweet and then used to calculate the topic distribution contained within each tweet. Figure 3.4 presents an example of how a tweet goes through the pipeline that calculates its dis-

tribution of delay related topics. The twenty topic distributions are then individually averaged per hour and per airline/airport. The hourly standard deviations of the twenty topic distributions are also extracted.

Table 3.3: Representation of the five topics related to the keyword "delay".

Topic	**Word distribution (top 10 words)**
Topic 0	$0.068 \cdot$ "delay" $+ 0.038 \cdot$ "sorry" $+ 0.033 \cdot$ "get" $+0.031 \cdot$ "SIGNATURE" $+ 0.029 \cdot$ "way" $+0.028 \cdot$ "flight" $+ 0.024 \cdot$ "due" $+ 0.023 \cdot$ "soon" $+0.023 \cdot$ "delayed" $+ 0.023 \cdot$ "delays"
Topic 1	$0.072 \cdot$ "AIRLINE" $+ 0.067 \cdot$ "flight" $+ 0.055 \cdot$ "delayed" $+0.035 \cdot$ "delay" $+ 0.026 \cdot$ "MENTION" $+ 0.018 \cdot$ "hour" $+0.016 \cdot$ "hours" $+ 0.014 \cdot$ "plane" $+0.012 \cdot$ "TIME" $+ 0.010 \cdot$ "get"
Topic 2	$0.083 \cdot$ "AIRLINE" $+ 0.059 \cdot$ "delayed" $+ 0.053 \cdot$ "flight" $+0.024 \cdot$ "delay" $+ 0.016 \cdot$ "flights" $+ 0.015 \cdot$ "time" $+0.013 \cdot$ "hours" $+ 0.012 \cdot$ "MENTION" $+0.009 \cdot$ "hour" $+ 0.007 \cdot$ "PICTURE"
Topic 3	$0.140 \cdot$ "delays" $+ 0.080 \cdot$ "MENTION" $+0.027 \cdot$ "WEBSITE" $+ 0.022 \cdot$ "weather" $+ 0.020 \cdot$ "flights" $+0.019 \cdot$ "check" $+ 0.016 \cdot$ "due" $+ 0.013 \cdot$ "status" $+0.012 \cdot$ "normal" $+ 0.010 \cdot$ "PICTURE"
Topic 4	$0.074 \cdot$ "delay" $+ 0.056 \cdot$ "sorry" $+ 0.052 \cdot$ "SIGNATURE" $+0.031 \cdot$ "flight" $+ 0.022 \cdot$ "delayed" $+ 0.020 \cdot$ "hear" $+0.019 \cdot$ "know" $+ 0.018 \cdot$ "us" $+0.017 \cdot$ "apologize" $+ 0.016 \cdot$ "delays"

Summary

Given the temporal nature of the data analyzed, the following features are chosen to keep track of the date: month of the year, day of the month, day of the week and hour in the day. A simplified diagram of the extraction process is presented in Figure 3.5. The following 2,608 features are considered:

- Hourly volume of passenger tweets for each airport/airline (8 airlines and 34 airports giving 42 features)
- Hourly volume of customer service tweets for each airport/airline (42 features)
- Hourly volume of passenger keyword-related tweets for each airport/airline (42x8 features)

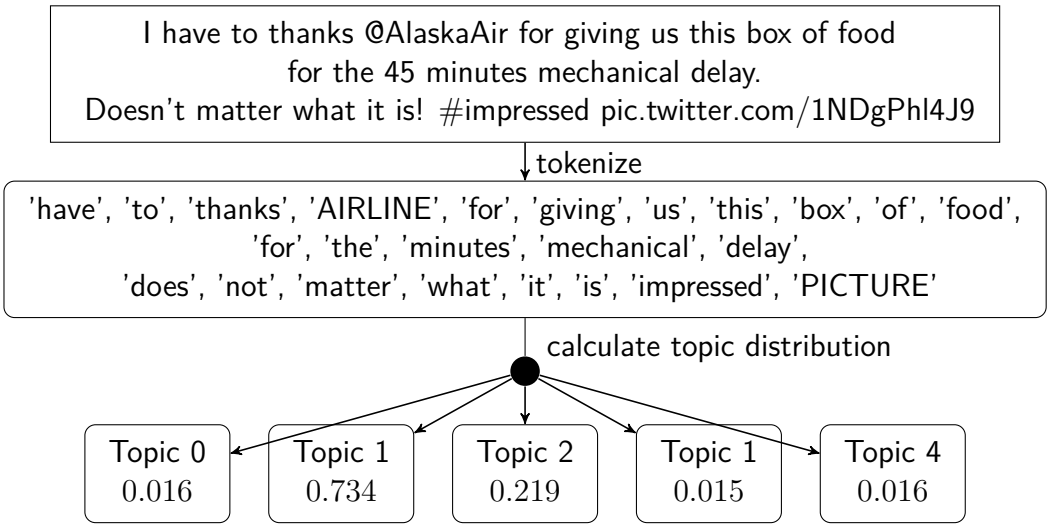


Figure 3.4: Example tweet going through the pipeline that calculates its distribution of delay related topics.

- Hourly volume of customer service keyword-related tweets for each airport/airline (42x8 features)
- Hourly average of passenger tweet sentiment for each airport/airline (42 features)
- Hourly average of customer service tweet sentiment for each airport/airline (42 features)
- Hourly standard deviation of passenger tweet sentiment for each airport/airline (42 features)
- Hourly standard deviation of airline/airport tweet sentiment for each airport/airline (42 features)
- Hourly average of topic distributions for each keyword for each airport/airline (42x20 features)
- Hourly standard deviation of topic distributions for each keyword for each airport/airline (42x20 features)
- Month of the year, Day of the month, Day of the week and Hour in the day (4 features)

3.2 Results

This section presents the output of both models (estimation based on passenger data and prediction based on BTS historic data) for four airports on several periods from 2018 to 2019. A performance comparison of these

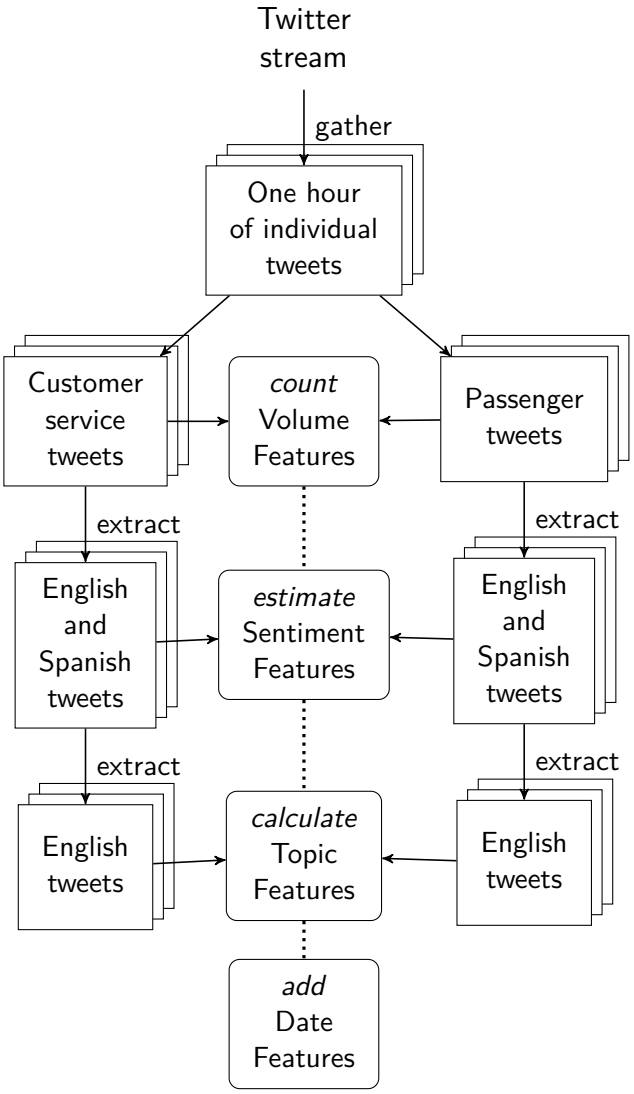


Figure 3.5: Diagram of the full feature extraction process.

two models for the 34 major US airports considered and based on the mean absolute errors described in Appendix C.1. The four chosen airports are Hartsfield–Jackson Atlanta International Airport (ATL), Boston Logan International Airport (BOS), Newark Liberty International Airport (EWR) and John F. Kennedy International Airport (JFK). ATL is the airport with the highest variability in the hourly number of delayed flights and the hourly number of cancelled flights and will illustrate the difficulty of the real-time estimation of the number of delayed flights and of the number of cancelled flights. BOS, EWR and JFK have also a high variability in the hourly num-

ber of delayed flights and the hourly number of cancelled flights, and they were the airports the most affected by the January 2018 bomb cyclone, which is the focus of this section.

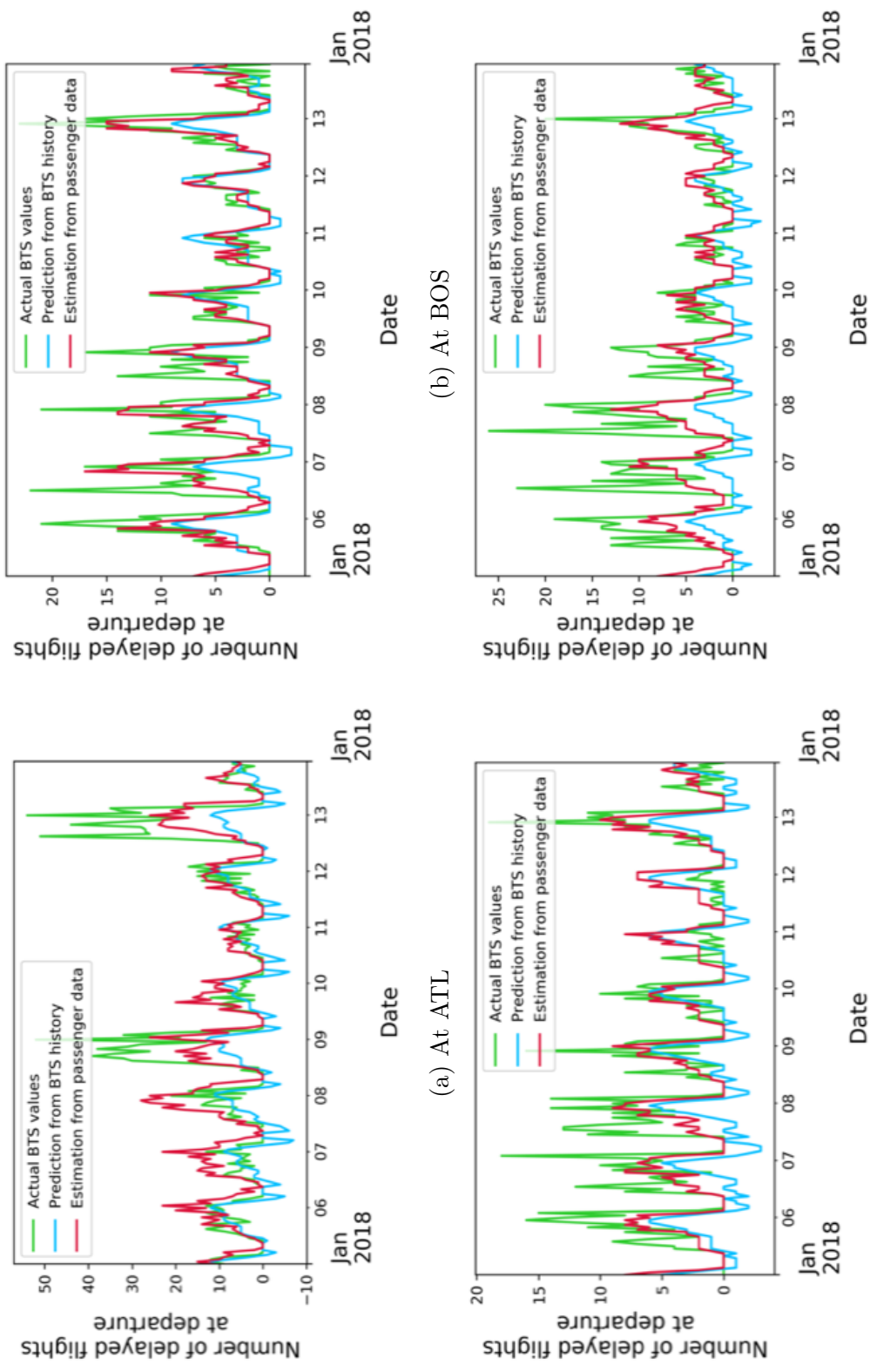
From January 2nd 2018 to January 6th 2018, a massive blizzard nicknamed "historic bomb cyclone" disrupted the Eastern Coast of the United States with a peak of violence on January 4th 2018 as it exploded in the area of the Mid-Atlantic states. The airports JFK and LaGuardia (LGA) in New York were closed for safety measures due to the weather conditions [159]. More than 70% of EWR flights and 20% of JFK flights were announced to be cancelled on January 4th 2018 [160]. Since this blizzard is an exceptional event, its effects on the air transportation system are not expected to be captured by the prediction model based only on historical BTS values. The prediction model serves as a baseline to highlight the difficulty of predicting or estimating the number of abnormal flights per hour and per airport.

3.2.1 Estimation of the number of flights with a delay greater than 15 minutes following the January 2018 bomb cyclone

At departure

Figure 3.6 shows the actual number of flights departing with a delay greater than 15 minutes, the predicted number of flights departing with a delay greater than 15 minutes based on historic BTS values and the estimation of the number of flights departing with a delay greater than 15 minutes based on the Twitter data at ATL, BOS, EWR and JFK for each hour over the ten days following the January 2018 bomb cyclone: January 5th-14th 2018. In these figures, the output of both models was rounded to the closest integer.

The high increase in the number of flights departing with a delay greater than 15 minutes from BOS (Figure 3.6(b)) and JFK (Figure 3.6(d)) following the bomb cyclone landfall between January 5th 2018 and January 9th 2018 followed by two "normal" days is best captured by the real-time estimation based on passenger-generated data than by the prediction based on historic BTS values. The difference between estimation and prediction is less visible at EWR (Figure 3.6(c)), though still with an advantage for the estimation based on passenger-generated data. The estimation of the number of flights departing with a delay greater than 15 minutes from ATL (Figure 3.6(a)) follows better the actual variations of the number of flights departing with a delay greater than 15 minutes than the prediction based on the historic BTS values. The important increases in the number of delayed flights of January 8th and 12th 2018 are not fully captured by the estimation model, though it



(a) At ATL (b) At BOS

(c) At EWR (d) At JFK

Figure 3.6: Comparison of the estimation of the number of flights departing with a delay greater than 15 minutes from ATL, BOS, EWR and JFK using the features extracted from Twitter with the prediction based on historical BTS values of delays over the period January 5th, 2018 to January 14th, 2018. The actual number of delayed flights is indicated in green.

still outperforms the prediction model on these two days.

By construction, the Prophet tool captures the daily, weekly and seasonal variations present in the training dataset (i.e. the year 2017), which explains why it predicts for each day a similar daily variation with the same number of peaks during the day yet with amplitudes varying depending on the month and the day of the week. Since it also extrapolates the underlying trends, it predicts negative values, usually at night when there are no flights, which underlines some limitations of the Prophet tool in this case.

From the random forest regressor used in the estimation model, the importance of each features in order to obtain a good prediction is calculated using the Mean Decrease Impurity measure defined by Breiman in [161]. The obtained feature importance scores are then normalized to make their sum equal to one. The 2,608 features created can be categorized into twenty-five types of features in order to obtain a better understanding of their associated feature importance:

- Date related features
- Features related to the raw number of passenger tweets (num_pax)
- Features related to the raw number of customer service tweets (num_cie)
- Features related to the number of passenger tweets containing a keyword (8 *keyword*_pax)
- Features related to the number of customer service tweets containing a keyword (8 *keyword*_cie)
- Features related to the sentiment expressed in passenger tweets (sent_pax)
- Features related to the sentiment expressed in customer service tweets (sent_cie)
- Features related to the topics of a keyword (4 *keyword*_topics)

Table 3.4 presents the top ten types of features and their aggregated importance for the estimation of the number of flights departing with a delay greater than 15 minutes at ATL, BOS, EWR and JFK.

The low importance of date related features for ATL (3.1% at the 7^{th} position) indicates that the number of flights departing ATL with a delay greater than 15 minutes does not have important daily, weekly or monthly trends, which is also indicated by the bad performance of the prediction based on historic BTS values at ATL presented in Appendix C.1. On the opposite, the high importance of date related features for EWR (54.49% in 1^{st} position) indicates that the number of flights departing EWR with a delay greater than 15 minutes have important daily, weekly or monthly trends, which explains why the estimation model and the prediction model have a similar behavior in Figure 3.6(c) in their estimation and prediction in the afternoon of the number of flights departing EWR with a delay greater

Table 3.4: Top ten feature types (and their aggregated feature importance) for estimating the number of flights departing with a delay greater than 15 minutes at four airports

#	ATL	JFK	BOS	EWR
0	delay_pax (39.46%)	delay_pax (36.38%)	date (36.75%)	date (54.49%)
1	num_pax (18.95%)	date (15.65%)	delay_pax (19.40%)	delay_pax (11.05%)
2	cancel_topics (9.07%)	num_pax (9.23%)	delay_topics (7.96%)	delay_topics (7.53%)
3	delay_topics (8.56%)	voucher_topics (8.45%)	num_pax (7.93%)	voucher_topics (7.08%)
4	voucher_topics (8.14%)	cancel_topics (8.14%)	voucher_topics (7.91%)	cancel_topics (6.88%)
5	refund_topics (7.14%)	delay_topics (7.95%)	cancel_topics (7.33%)	refund_topics (6.81%)
6	date (3.10%)	refund_topics (7.67%)	refund_topics (7.17%)	num_pax (1.85%)
7	sent_pax (1.68%)	num_cie (2.90%)	sent_pax (1.68%)	sent_pax (1.83%)
8	num_cie (1.22%)	sent_pax (1.70%)	num_cie (0.85%)	num_cie (0.47%)
9	delay_cie (0.49%)	sent_cie (0.50%)	WEBSITE_pax (0.57%)	sent_cie (0.43%)

Table 3.5: Top ten feature types for estimating the number of flights arriving with a delay greater than 15 minutes at four airports

#	ATL	JFK	BOS	EWR
0	delay_pax (40.01%)	delay_pax (28.50%)	delay_pax (36.30%)	date (47.87%)
1	num_pax (11.17%)	date (26.41%)	date (14.51%)	delay_pax (10.06%)
2	delay_topics (9.72%)	delay_topics (9.74%)	refund_topics (9.72%)	cancel_topics (8.71%)
3	refund_topics (9.52%)	refund_topics (9.56%)	voucher_topics (9.68%)	refund_topics (8.67%)
4	cancel_topics (9.46%)	voucher_topics (9.36%)	delay_topics (9.67%)	delay_topics (8.33%)
5	voucher_topics (9.34%)	cancel_topics (9.01%)	cancel_topics (8.93%)	voucher_topics (8.07%)
6	date (5.92%)	sent_pax (1.61%)	num_pax (4.52%)	num_pax (3.46%)
7	sent_pax (1.65%)	delay_cie (1.07%)	sent_pax (2.25%)	sent_pax (1.81%)
8	num_cie (0.62%)	num_pax (1.05%)	num_cie (1.11%)	num_cie (0.51%)
9	sent_cie (0.42%)	sent_cie (0.89%)	WEBSITE_pax (0.89%)	PICTURE_pax (0.49%)

than 15 minutes. For the four airports, the importance of delay related features vindicates the choice of keywords within the feature creation process presented in Section 3.1.

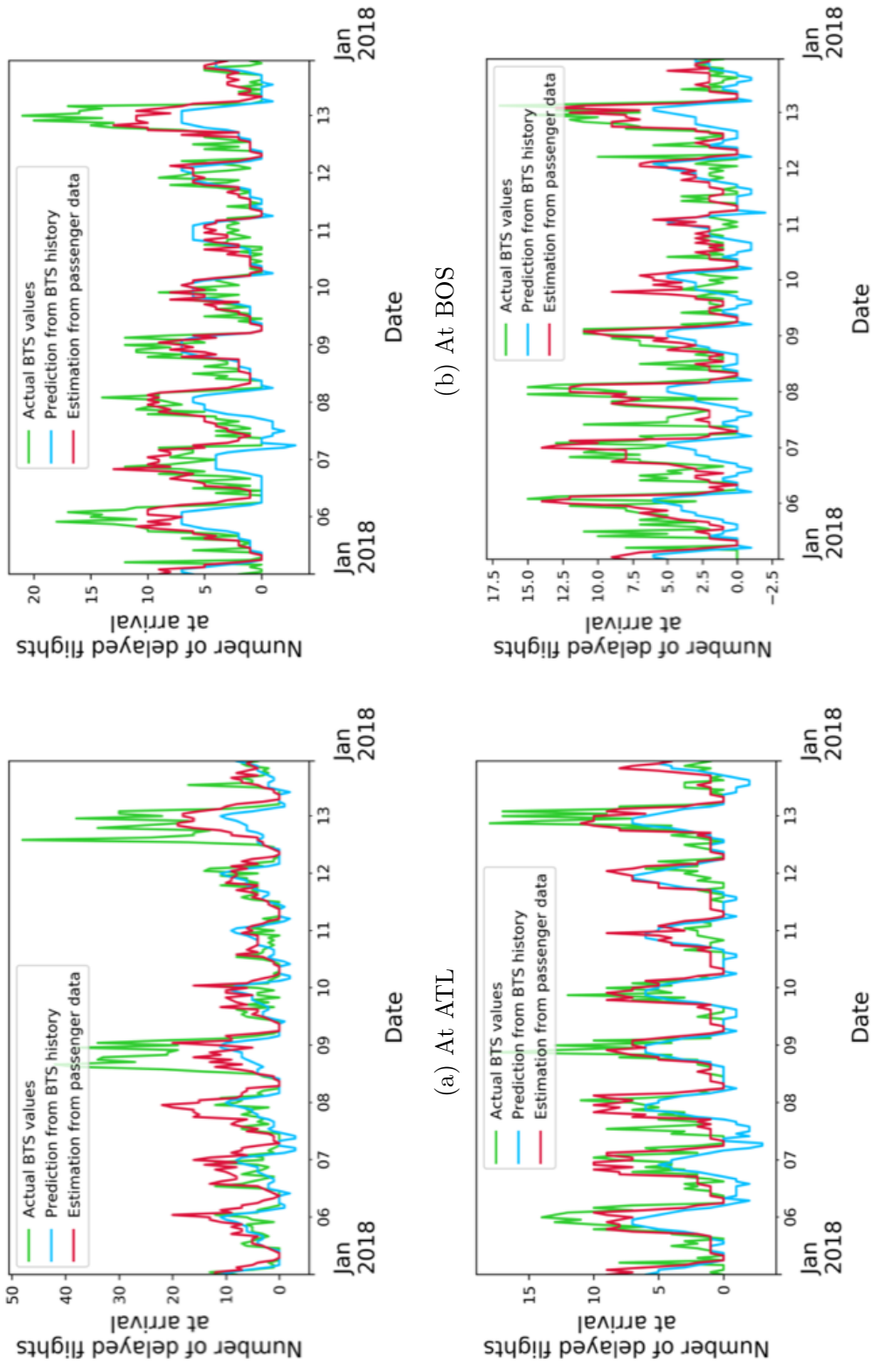
At arrival

Figure 3.7 shows the actual number of flights arriving with a delay greater than 15 minutes, the predicted number of flights arriving with a delay greater than 15 minutes based on historic BTS values and the estimation of the number of flights arriving with a delay greater than 15 minutes based on the Twitter data at ATL, BOS, EWR and JFK for each hour over the ten days following the January 2018 bomb cyclone: January 5^{th}-14^{th} 2018. In these figures, the output of both models was rounded to the closest integer.

Similar conclusions as in Section 3.2.1 can be drawn from these figures. The high increase in the number of flights arriving with a delay greater than 15 minutes at BOS (Figure 3.7(b)) and JFK (Figure 3.7(d)) following the bomb cyclone landfall between January 5^{th} 2018 and January 9^{th} 2018 followed by three "normal" days is best captured by the real-time estimation based on passenger-generated data than by the prediction based on historic BTS values. The estimation of the number of flights arriving with a delay greater than 15 minutes at ATL (Figure 3.6(a)) captures better the important increases in the number of delayed flights of January 8^{th} and 12^{th} 2018 than the prediction model based on the historic BTS values, though the increases are not totally captured in volume.

As for the estimation of the number of flights departing with a delay greater than 15 minutes, the importance of each feature for the estimation of the number of flights arriving with a delay greater than 15 minutes is calculated and then aggregated using the same feature groups as in Section 3.2.1. Table 3.5 (page 37) presents the top ten types of features and their aggregated importance for the estimation of the number of flights arriving with a delay greater than 15 minutes at ATL, BOS, EWR and JFK.

Similarly to the estimation of the number of flights departing with a delay greater than 15 minutes, the importance of date related features for estimating the number of flights arriving with a delay greater than 15 minutes is low for flights arriving at ATL (5.92%) and high for flights arriving at EWR (47.87%). This indicates that there are no important daily, weekly or monthly trends for both the number of delayed departing flight and the number of delayed arriving flights at ATL but that these trends are important for both the number of delayed departing flight and the number of delayed arriving flights at EWR. The features counting the number of passenger tweets containing the keyword delays are predominant for ATL, JFK and



(a) At ATL (b) At BOS

(c) At EWR (d) At JFK

Figure 3.7: Comparison of the estimation of the number of flights arriving with a delay greater than 15 minutes at ATL, BOS, EWR and JFK using the features extracted from Twitter with the prediction based on historical BTS values of delays over the period January 5$^{\text{th}}$, 2018 to January 14$^{\text{th}}$, 2018. The actual number of delayed flights is indicated in green.

BOS and comes second for EWR, emphasizing the importance of this single keyword for the estimation of the number of delayed flights.

3.2.2 Estimation of the number of cancelled flights

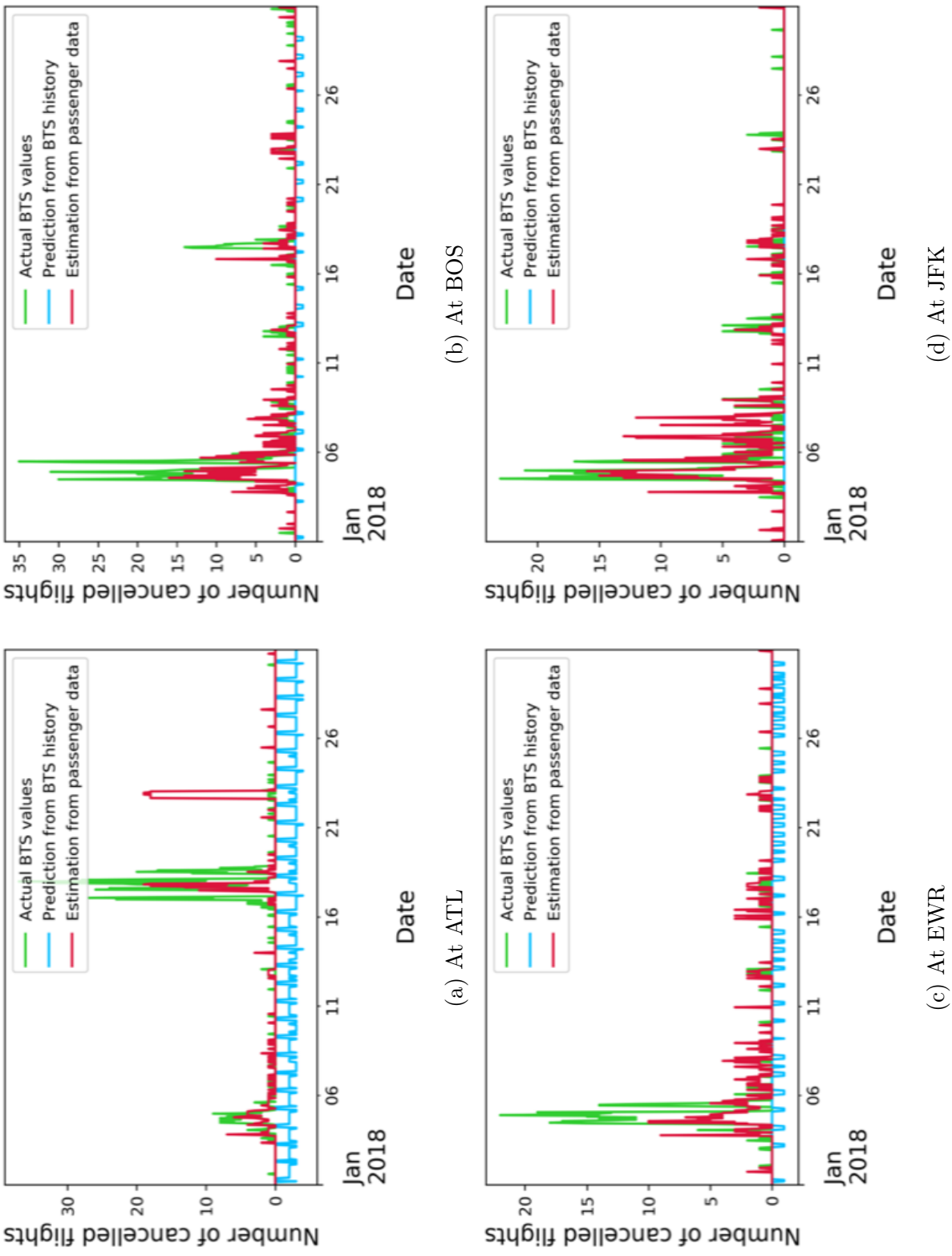
January 2018

Figure 3.8 shows the actual number of cancelled flights, the predicted number of cancelled flights based on historic BTS values and the estimation of the number of cancelled flights based on the Twitter data at ATL, BOS, EWR and JFK for each hour over the full month of January 2018. In these figures, the output of both models was rounded to the closest integer.

A first clear takeaway from these four plots is that predicting the number of cancelled flights based only on BTS historic values is totally ineffective for the month of January 2018 for these four airports. Figure 3.8(a) shows that this method predicts constantly a negative number of cancelled flights at ATL except for a couple of hours every day in the early mornings when it predicts that there are zero cancelled flights. This indicates that the prediction model captured a slowly decreasing trend for cancelled flights at ATL in the historic BTS data over first ten months of the year 2017, which leads the model to predict a negative number of cancelled flights in 2018 even though there were no negative values in the training set. At BOS (Figure 3.8(b)) and at EWR (Figure 3.8(c)), the predicted number of cancelled flights oscillates between -1 flight cancelled and 0 flight cancelled per hour. And at JFK (Figure 3.8(d)), the prediction model predicts that there are absolutely no cancelled flights over the whole month of January 2018.

On the other hand, the estimation model based on passenger-generated data captures better the periods where cancellations occurs, though not always the exact volume cancellations. For example, the increase in the number of cancelled flights due to the bomb cyclone in early January 2018 is clearly captured in the estimated number of cancelled flights at the four airports under consideration. The other periods in January with an increase in the number of cancelled flights is also well captured by the estimation model for BOS (Figure 3.8(b)) and the estimation model for JFK (Figure 3.8(d)). At ATL (Figure 3.8(a)), the period of high cancellations from January 16$^{\text{th}}$ 2018 to January 18$^{\text{th}}$ 2018 is present in the estimation of the number of cancelled flights but with some important underestimations on January 16$^{\text{th}}$ 2018 and January 18$^{\text{th}}$ 2018. On the opposite, the estimated number of cancellations of January 22$^{\text{nd}}$ 2018 is highly overestimating the actual number of cancellations.

As for the estimation of the number of flights departing or arriving with



(a) At ATL (b) At BOS

(c) At EWR (d) At JFK

Figure 3.8: Comparison of the estimation of the number of cancelled flights from ATL, BOS, EWR and JFK using the features extracted from Twitter with the prediction based on historical BTS values of cancellations over the period January 1st, 2018 to January 31st, 2018. The actual number of cancelled flights is indicated in green.

a delay greater than 15 minutes, the importance of each feature for the estimation of the number of cancelled flights is calculated and then aggregated using the same feature groups as in Section 3.2.1. Table 3.6 presents the top ten types of features and their aggregated importance for the estimation of the number of cancelled flights at ATL, BOS, EWR and JFK.

As a confirmation of the difficulty of estimating or predicting the number of cancelled flights based on historical data alone, there are no date related features in the top ten feature types. Date related features actually account for between 0.10% and 0.21% of the feature importance for estimating the number of cancelled flights at these four airports. The features accounting for the hourly number of passenger tweets containing the keyword "cancel" are the most important by far for estimating the number of cancelled flights at these airports. The features accounting for the number of passenger tweets containing the keyword "cancel" and the features related to the cancellation topics are also the most important features for estimating the number of cancelled flights at 24 airports out of 34. The features related to the topics related to the chosen keywords have a greater importance for the estimation of the number of cancelled flights than for the estimation of the number of delayed flights (Section 3.2.1) for JFK, BOS and EWR.

July 2019

In order to see how the estimation model based on passenger-generated data fares through time, another month where many flights were cancelled over several short periods is considered here, the month of July 2019. The estimation models are the same as in Section 3.2.2, i.e. they were only trained once on data from 2017, while the prediction models have access to the BTS history of cancellations from January 1st 2017 to April 30th 2019.

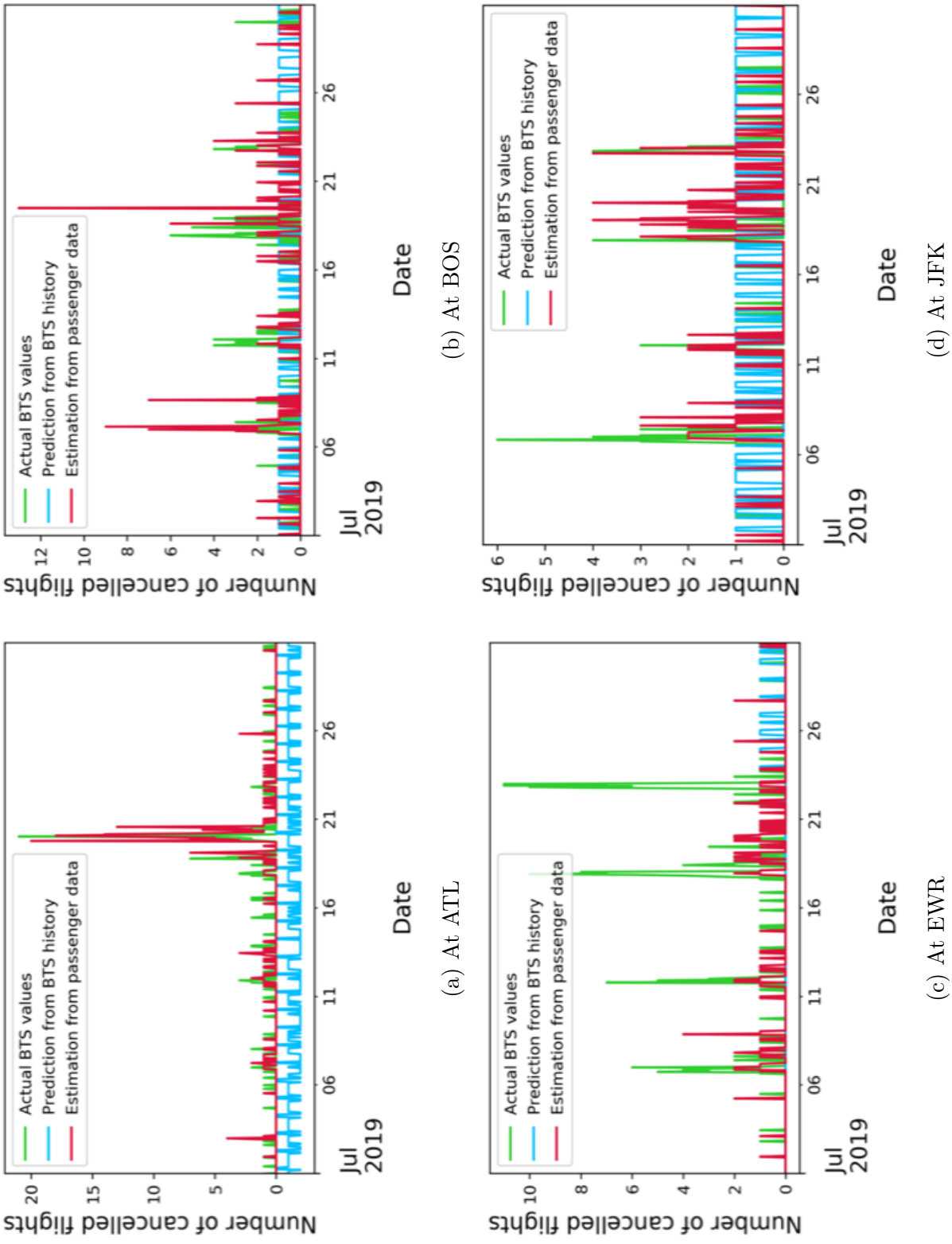
Figure 3.9 shows the actual number of cancelled flights, the predicted number of cancelled flights based on historic BTS values and the estimation of the number of cancelled flights based on the Twitter data at ATL, BOS, EWR and JFK for each hour over the full month of July 2019. In these figures, the output of both models was rounded to the closest integer.

Though the prediction models have access to more than two years of cancellation data, they are still unable to capture the actual evolution of the number of cancellations for the month of July 2019. The prediction model for ATL (Figure 3.9(a)) still predicts a negative number of cancelled flights except for one hour per day in the early morning where it predicts zero cancelled flights. The prediction models for BOS (Figure 3.9(b)), EWR (Figure 3.9(c)) and (Figure 3.9(d)) predicts either 0 or 1 cancelled flight.

The estimation model based on passenger-generated data for ATL (Fig-

Table 3.6: Top ten feature types for estimating the number of cancelled flights at four airports

#	**ATL**	**JFK**	**BOS**	**EWR**
0	cancel_pax (25.04%)	cancel_pax (34.33%)	cancel_pax (25.09%)	cancel_pax (31.77%)
1	hours_pax (16.19%)	voucher_topics (13.60%)	cancel_topics (17.21%)	cancel_topics (16.11%)
2	delay_topics (8.98%)	delay_topics (12.12%)	delay_topics (15.47%)	voucher_topics (13.47%)
3	num_pax (8.95%)	cancel_topics (11.40%)	voucher_topics (14.32%)	delay_topics (12.72%)
4	refund_topics (8.75%)	refund_topics (9.15%)	num_pax (9.59%)	refund_topics (11.69%)
5	voucher_topics (7.58%)	sent_pax (4.10%)	refund_topics (7.70%)	sent_pax (3.66%)
6	cancel_topics (6.48%)	num_pax (3.66%)	sent_pax (2.94%)	num_pax (2.58%)
7	wait_pax (5.51%)	delay_pax (2.64%)	WEBSITE_pax (2.49%)	WEBSITE_pax (1.63%)
8	delay_pax (4.00%)	WEBSITE_pax (2.29%)	PICTURE_pax (0.83%)	PICTURE_pax (1.49%)
9	sent_cie (2.97%)	PICTURE_pax (1.87%)	cancel_cie (0.67%)	voucher_pax (0.81%)



(a) At ATL (b) At BOS

(c) At EWR (d) At JFK

Figure 3.9: Comparison of the estimation of the number of cancelled flights from ATL, BOS, EWR and JFK using the features extracted from Twitter with the prediction based on historical BTS values of cancellations over the period July 1st, 2019 to July 31st, 2019. The actual number of cancelled flights is indicated in green.

ure 3.9(a)) captures correctly the different cancellation periods of July 2019 both in range and in volume. The estimation models for BOS (Figure 3.9(b)) and JFK (Figure 3.9(d)) capture correctly the range of the cancellation periods and the volume to a lesser extent. The estimation model for EWR (Figure 3.9(c)) is not as effective as the models of the other three airports, but still outperforms the associated prediction model.

3.3 Discussion & Conclusion

3.3.1 Conclusion

The proposed feature extraction process transforms the Twitter stream into a real-time estimator of the hourly number of abnormal flights of the US air transportation system. The abnormal flights considered here are flights departing with a delay greater than 15 minutes, flights arriving with a delay greater than 15 minutes and cancelled flights. The estimation models built on the features extracted from the Twitter stream estimate better the actual number of abnormal flights than the prediction models based on the historic BTS data available.

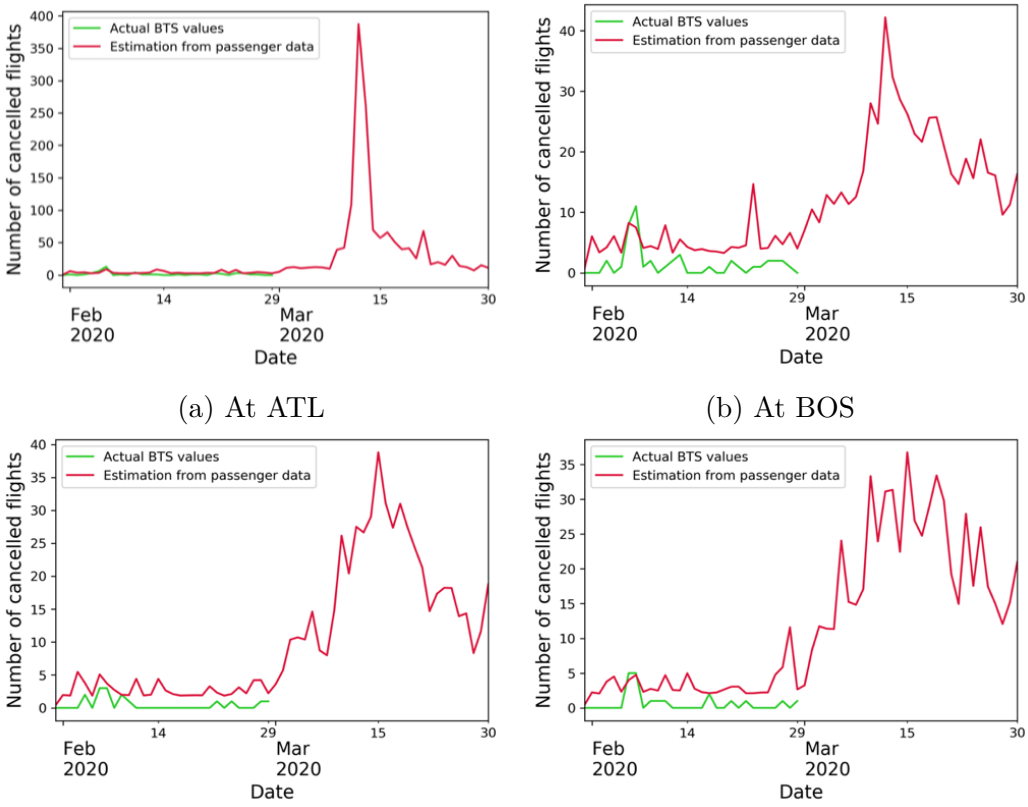
This new estimation model based on passenger-generated data is the result of a continuous improvement of the previous estimation models proposed in [12], since both approaches exploit raw volume information as well as different levels of content information within the Twitter stream. Separating passenger tweets from company tweets and focusing on specific topics lead however to more human-understandable features that help better understand the major differences between predictions based on historic BTS values and estimation based on passenger generated data. Though the feature extraction process presented in Section 3.1 also considers tweets written by the customer services of airlines and airports, Tables 3.4-3.6 indicate that features related to passenger tweets are more important than features related to customer service tweets for the estimation of the hourly number of abnormal flights, emphasizing the importance of considering passenger-generated data.

Future studies should look into the impact of incorporating available flight-centric information to the estimation model (e.g. the number of scheduled flights) on estimation performances as well as the importance given to these flight-centric features compared to the presented passenger-centric features. Retraining the estimation models on each new monthly BTS report should be investigated. Analyzing the evolution of the feature importance scores in such a scenario could lead to a monthly analysis of the passenger perception of the system. These analysis could complement the flight-centric

reports by adding some passenger-related context. Furthermore, estimating

3.3.2 Cancellations following the COVID-19 public health crisis

In the specific case of the COVID-19 pandemic, further discussed in Chapter 4, the model presented can be used to notice an important situation change affecting passengers of the air transportation system. Though no



(a) At ATL

(b) At BOS

(c) At EWR

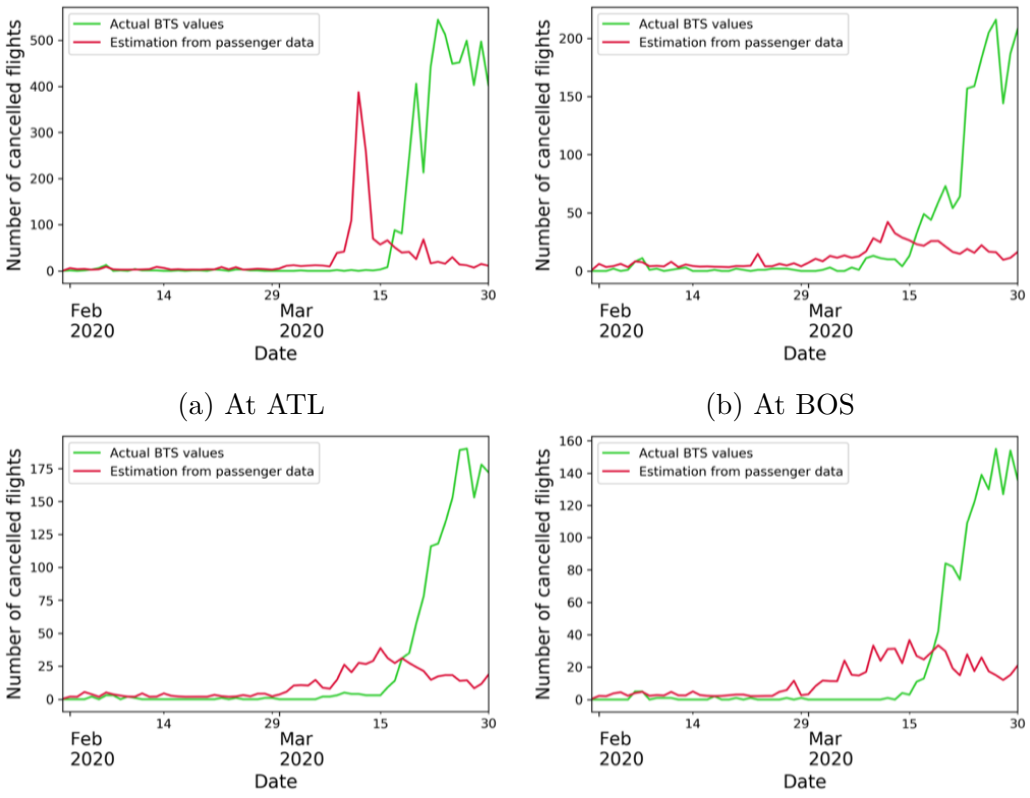
(d) At JFK

Figure 3.10: Estimation of the number of cancelled flights from ATL, BOS, EWR and JFK using the features extracted from Twitter and aggregated per day over the period February 1st, 2020 to March 31st, 2020. The actual number of cancelled flights is indicated in green when available on May 10th 2020.

BTS data was available for March 2020 until mid May 2020 for the period of Spring 2020 (February and March 2020), the data generated by the pas-

sengers on Twitter gives a vivid picture of the situation as they experience it.

Figure 3.10 shows the estimation of the number of cancelled flights using the models based on data generated by passengers aggregated by day and the available corresponding BTS values on May 10th 2020 from February 1st 2020 to March 31st 2020 for ATL, BOS, EWR and JFK. For all four airports, there is an important increase in the estimated number of cancelled flights at the beginning of March 2020. The increase is most important for the estimated daily number of cancelled flights at ATL (Figure 3.10(a)) with a spike on March 12th-13th 2020. Using these estimations in March 2020 could



(a) At ATL

(b) At BOS

(c) At EWR

(d) At JFK

Figure 3.11: Estimation of the number of cancelled flights from ATL, BOS, EWR and JFK using the features extracted from Twitter and aggregated per day over the period February 1st, 2020 to March 31st, 2020. The actual number of cancelled flights is indicated in green when available on May 28th 2020.

have helped airports and passengers better understand which regions were

the most affected by cancellations following the start of the COVID-19 public health crisis.

Waiting for the release of the actual BTS values in order to assess this situation does give an accurate picture of the scale of the cancellations resulting from the COVID-19 public health crisis, but it was necessary to wait until the second half of May 2020 in order to obtain the processed figures. Figure 3.11 shows the estimation of the number of cancelled flights using the models based on data generated by passengers aggregated by day and the available corresponding BTS values on May 28$^{\text{th}}$ 2020 from February 1$^{\text{st}}$ 2020 to March 31$^{\text{st}}$ 2020 for ATL, BOS, EWR and JFK.

The BTS data tells us that the actual increase in cancellations started later than the estimated increase in cancellations at these four airports. This is probably due to the fact that both airlines and passengers realized in advance that flights were to be cancelled, prompting some reaction on Twitter. This reaction, and the interaction between airlines and their passengers, is the focus of Chapter 4, which proposes new metrics to measure in real-time the impact of long-term perturbations on passengers and applies them to the COVID-19 pandemic during Spring 2020.

Chapter 4

Introducing passenger-generated metrics to assess the impact of COVID-19 on the air transportation system

The COVID-19 pandemic has had a significant impact on the air transportation system worldwide. This chapter aims at analyzing the effect of the travel restriction measures implemented during the COVID-19 pandemic from a passenger perspective on the US air transportation system. Four metrics based on data generated by passengers and airlines on social media are proposed to measure how the travel restriction measures impacted the relation between passengers and airlines in close to real-time. Three metrics based on data generated by passengers and visitors at airports are proposed to measure how the public health crisis has impacted the wait times at airports from a passenger perspective. The first reports presenting these metrics came ahead of official data related to the same sequence of events, thereby showing the value of passenger-borne data in an industry where corporate priorities, institutional prudence, and passenger satisfaction come close together.

4.1 Motivation

4.1.1 The COVID-19 pandemic and the resulting travel restrictions from a US perspective

In response to the pandemic situation resulting from the outbreak of the corona disease 2019 (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), travel restrictions measures were implemented by various countries, impacting both domestic travel and international travel [162].

Italy was the first country to enforce a national lockdown [163] on March 9$^{\text{th}}$ 2020, after introducing on February 21$^{\text{st}}$ 2020 an initial measure confining only the northern region of Lodi. Two days after Italy's lockdown announcement, on March 11$^{\text{th}}$ 2020, the United States banned non-US travelers who had been to China, Iran and 26 member states of the European Union (EU) to enter the US, and later extended the ban to non-US travelers who had visited the United Kingdom and Ireland on March 16$^{\text{th}}$ 2020 [162]. The EU officially closed the external borders of 26 of its member states to nearly all non-EU residents on March 17$^{\text{th}}$ 2020 [162]. On March 19$^{\text{th}}$ 2020, the US Department of State issued a Level 4 Global Health Travel Advisory, which cautions all US citizens against international travel, still in place as of May 6$^{\text{th}}$ 2020 [164].

This dramatic sequence of events forms the thread against which the air transportation system has had to progressively put itself to a semi-comatose state to address fast-growing public health and economic concerns. For these

reasons, the following dates are indicated with dotted lines in every graph throughout this chapter in order to better visualize the timeline of each figure.

1. The **Lodi** region lockdown in Italy: February 21st, 2020
2. **Italy**'s lockdown: March 9th, 2020
3. **US ban** of non-US travelers from the EU, China and Iran: March 11th, 2020
4. **EU** external border **closure**: March 17th, 2020
5. **US Level 4** Global Health Travel Advisory: March 19th, 2020

Figure 4.1 presents the number of passengers arriving at US immigration across all airports of entry using the "Airport Wait Times" data from the Customs and Border Protection (CBP) website [165]. This plot illustrates clearly the effect of these travel restriction measures on the international traffic coming to the US. For a more detailed presentation of the available CBP

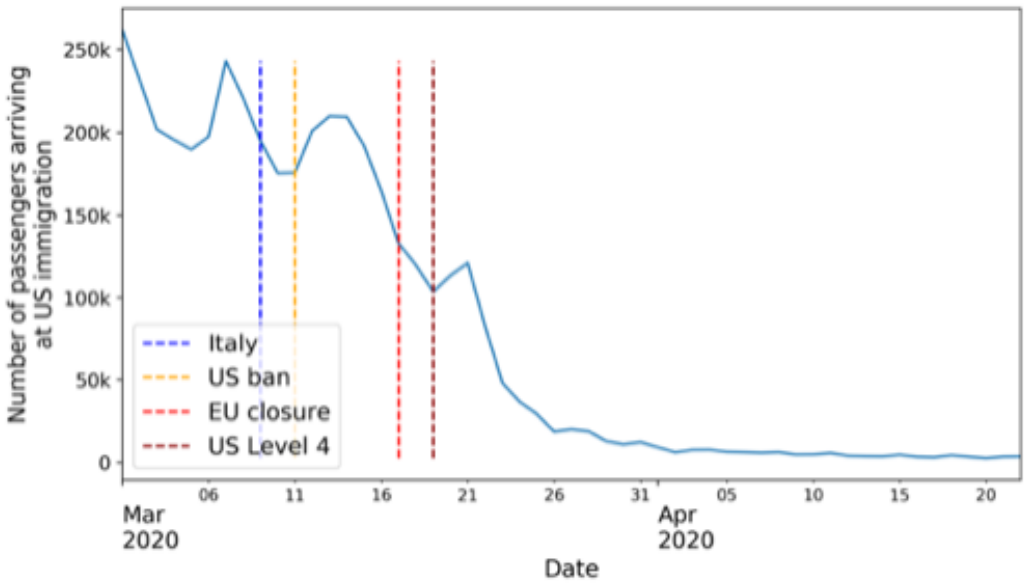


Figure 4.1: Evolution of the daily number of passengers arriving at all US airports of entry from CBP data.

dataset, the authors recommend reading [16] (reproduced in Appendix D.2), which also presents an analysis of the wait times at US airport immigration services from January 2013 to January 2019.

4.1.2 The limitations of traditional approaches to assess the impact of COVID-19 on the air transportation system

The travel restrictions, and the other measures taken by a majority of countries worldwide, are having an unprecedented impact on the air transportation system. Until official flight data are released in the United States regarding international and domestic air transportation, there are no means of measuring this impact on the US air transportation system, except by relying on non-traditional data sources.

Traditionally, the metrics used to measure the state of the US air transportation system are focused on flight performances, such as the amount of delay per flight, the number of delayed flights, the number of cancelled flights and the number of carried passengers. The data considered for these metrics are gathered by the US Department of Transportation Bureau of Transportation Statistics (BTS) [2]. The data are first processed by airlines and airports and then provided to the BTS, which then publishes the data as a monthly report. The BTS reports pertaining to on-time flight data are usually published with a latency of two months. This latency is not well adapted for monitoring and analyzing the effects of situations such as the COVID-19 pandemic on the US air transportation system.

This chapter proposes an alternative approach to analyzing the air transportation system by focusing on airline performances with respect to their passengers using data generated by airlines and by passengers. The importance for airlines of improving the waiting environment at airports in order to improve passenger satisfaction is already highlighted in [45] and is generalized for riders at transit stations in [166]. In the specific case of US air transportation, Twitter is an important medium for direct communication between passengers and airlines. For example, over the month of January 2020, more than 300 tweets were written on average every day by the customer services of four major US carriers (Southwest Airlines, Delta Airlines, American Airlines and United Airlines) and more than 800 tweets were written on average every day by their customers.

This chapter proposes several passenger-centric metrics constructed from passenger-generated data in order to offer a passenger-centric perspective of the air transportation system, with a focus on the relation between airlines and passengers and on the waiting experience of passengers and visitors at airports.

The rest of this chapter is structured as follows: Section 4.2 describes the first two metrics based on a Twitter sentiment analysis and how they can

be used in light of the COVID-19 situation. Section 4.3 then describes two additional metrics based on selected keywords and how they can be used to assess the performance of airline communication during the COVID-19 pandemic. Section 4.4 focuses on the evolution of wait times at US airports for passengers and visitors and proposes three additional metrics. Section 4.5 concludes this chapter and discusses future research directions.

4.2 Impact of the COVID-19 on airline and passenger mood

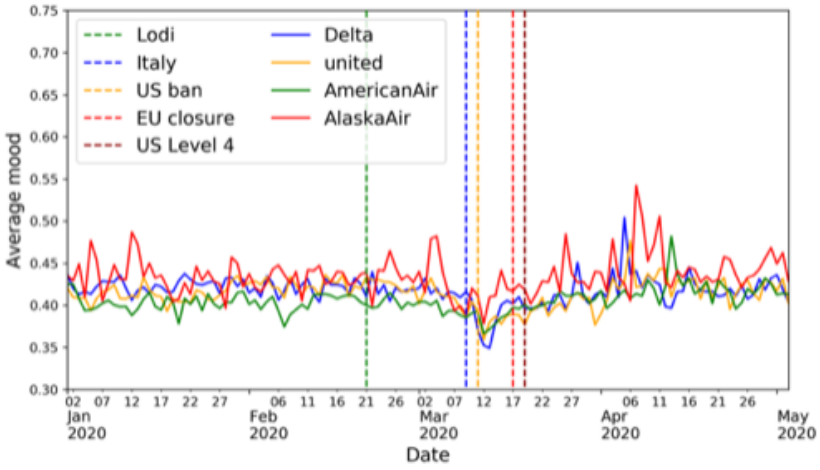
Eight airlines, and their associated Twitter handles, are considered in the analysis below: American Airlines (@AmericanAir), Delta Air Lines (@Delta), United Airlines (@united), Alaska Airlines (@AlaskaAir), Southwest Airlines (@SouthwestAir), JetBlue Airways (@JetBlue), Spirit Airlines (@SpiritAirlines) and Frontier Airlines (@FlyFrontier and @FrontierCare). The first four are legacy airlines, and the last four are low-cost carriers. All tweets written from these airlines Twitter accounts were scraped from January 1st 2020 to May 3rd 2020 and are categorized as "customer service tweets". All tweets written over that same period and mentioning at least one of the airline handles that was not written from the corresponding airline Twitter account were also scraped and categorized as "passenger tweets".

4.2.1 Daily mood evolution

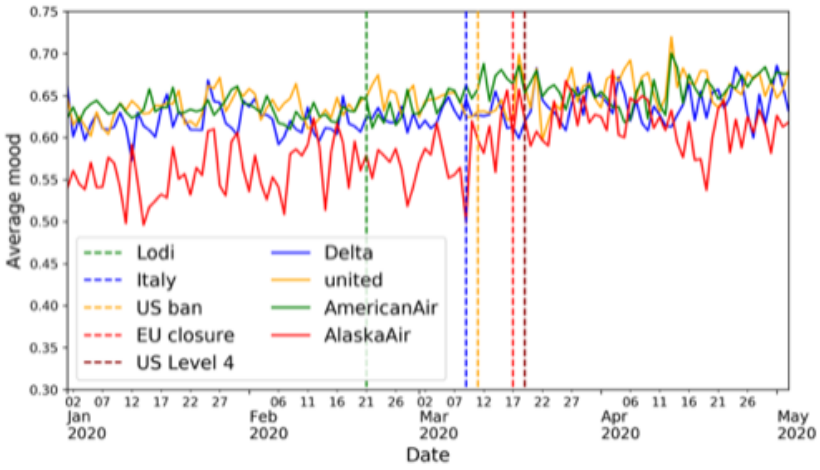
The sentiment extraction process presented in Section 3.1.2 is used to extract the sentiment expressed within each tweet. This expressed sentiment is then averaged on a daily level in order to compare the effect of the travel restriction measures on the expressed passenger mood with their effects on the expressed airline mood. Legacy airlines are usually considered as offering a higher quality service to customers than low-cost carriers, with an average of close to 296 tweets written a day by the customer service of the four considered US legacy airlines versus an average of 112 tweets written a day by the customer service a day for the four considered low-cost carriers. The evolution of the mood expressed by passengers and airline customer services is presented in the following subsections, first for the legacy airlines and then for the low-cost carriers.

Case of legacy airlines

Figure 4.2 shows the evolution of the mood expressed by the four legacy airlines considered and by their passengers from January 1st 2020 to May 3rd 2020.



(a) From passengers of major airlines



(b) From customer service

Figure 4.2: Daily average mood expressed in tweets containing airline Twitter handles for four legacy airlines between January 1st 2020 and May 3rd 2020. The expressed mood score can vary between 0, indicating a negative mood, and 1, indicating a positive mood.

From Figure 4.2(a), a drop in the mood expressed by passengers can be observed starting right after the Lodi lockdown with a steep decrease right

after the US travel ban for the three major airlines (Delta Air Lines, United Airlines and American Airlines). The sentiment extracted from the tweets from Delta's passengers has the steepest descent but also the sharpest recovery. The case of Alaska Airlines exhibits special characteristics: a #AlaskaHappyHour campaign, giving Twitter users the opportunity of winning a free flight to Alaska, was taking place early March 2020. This campaign could explain why the expressed mood in passenger tweets increased between March 1st 2020 and March 5th 2020 and could have compensated a potential decrease in the passenger expressed mood linked to the travel ban announcement.

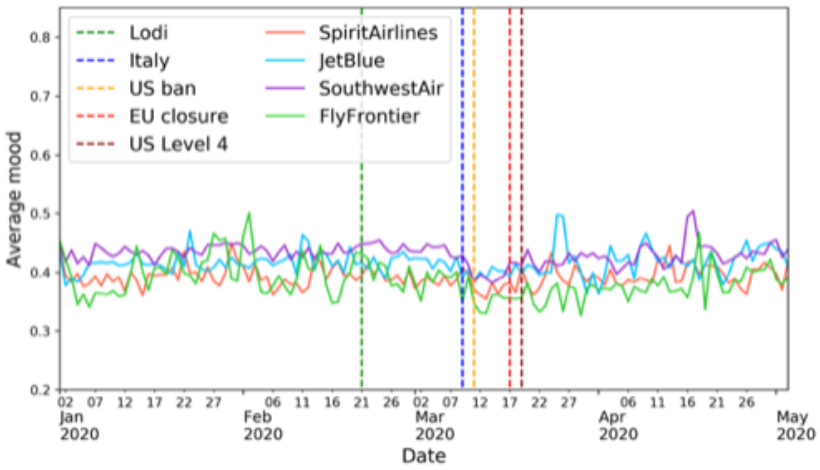
Regarding the mood expressed in tweets written by the airline customer services, shown in Figure 4.2(b), it only decreases for Delta Air Lines and United Airlines starting at the announcement of Italy's lockdown. An opposite reaction is seen with the mood expressed by American Airlines customer service, which increases over that same period. Comparing Figure 4.2(a) and Figure 4.2(b) shows that Delta Air Lines and Alaska Airlines have the highest expressed mood on average within their passenger tweets over the considered period, but the lowest expressed mood within their customer service tweets of the four legacy airlines. An explanation of the better mood expressed by their passengers could be that these airlines expressed a mood closer to their passengers' actual mood. A gap between the mood extracted from passenger tweets and the mood extracted from airline customer service tweets is visible from one figure to another, with airline customer service tweets expressing a mood about 0.2 points higher than passenger tweets.

Case of low-cost carriers

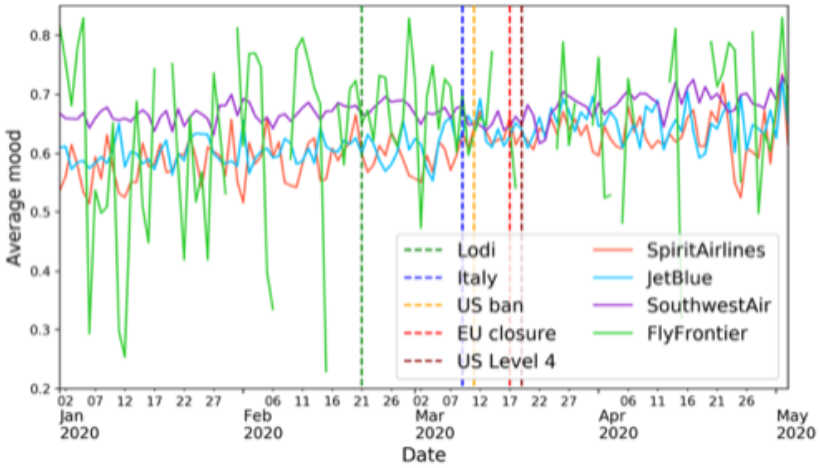
Similar conclusions can be drawn when analyzing the mood associated to tweets from passengers and customer services of low-cost carriers. Figure 4.3 shows the evolution of the expressed mood from January 1st 2020 and May 3rd 2020 in the passenger and customer service tweets of the four low-cost carriers considered.

Figure 4.3(a) indicates that the mood expressed by passengers of Spirit Airlines and Frontier Airlines is significantly lower on average than the mood expressed by passengers of JetBlue Airways and Southwest Airlines over the months of February and March 2020. There is a spike in the mood extracted from tweets written by JetBlue passengers around March 26th 2020. This date is the day when the governor of New York thanked JetBlue for offering free flights to health care workers in order to help the state handle the spread of COVID-19[1]. It also corresponds to the period when an update of their mobile

[1] `https://twitter.com/NYGovCuomo/status/1242941085535608835`



(a) From passengers



(b) From customer service

Figure 4.3: Daily average mood expressed in tweets containing airline Twitter handles for four low-cost airlines between January 1st 2020 and May 3rd 2020. The expressed mood score can vary between 0, indicating a negative mood, and 1, indicating a positive mood.

application contained the message "Now, go wash your hands", prompting an amused reaction of their passengers. The drop in the mood expressed in the tweets written by legacy airline passengers after Italy's lockdown is less visible in the tweets written by passengers of low-cost carriers, with the exception of the mood expressed by passengers of Southwest Airlines.

Looking at the mood expressed by low-cost carrier customer services presented in Figure 4.3(b), the mood expressed by the customer service of Fron-

tier Airlines displays a highly varying behavior, oscillating between 0.23 and 0.83 with discontinuities since on certain days no tweets were written by their customer service. For the other three low-cost carriers, the gap between the mood extracted from the tweets written by Southwest Airlines customer service and the mood extracted from the tweets written by the customer services of the other two carriers reduces significantly the day after Italy's lockdown. Similarly as for legacy airlines, a gap of about 0.2 points is visible between the mood expressed within passenger tweets and airline customer service tweets by comparing Figure 4.3(a) and Figure 4.3(b).

4.2.2 Passenger-centric metrics

Based on the observations presented in Section 4.2.1, two passenger-centric metrics are proposed to measure the relation between airline customer services and their passengers. The first proposed metric aims at measuring the evolution of the airline mood relative to the mood of their passengers. Diverging mood evolutions are given a low score: if the average mood expressed by passengers is decreasing, the average mood expressed in the tweets written by the airline customer service should not be increasing.

Proposed passenger-centric metric 1 *The airline* ***empathy score*** *is defined as the Pearson correlation between the evolution of the average mood expressed by passengers in their tweets and the evolution of the average mood expressed by the airline customer service in their tweets.*

The empathy score Ξ is calculated using the following formula:

$$\Xi = \frac{\sum_i (p_i - \bar{p})(c_i - \bar{c})}{\sqrt{\sum_i (p_i - \bar{p})^2 \sum_i (c_i - \bar{c})^2}} \tag{4.1}$$

where the set $\{p_i\}_i$ (resp. $\{c_i\}_i$) is the ordered set of the daily expressed mood in passenger tweets (resp. in airline customer service tweets), and $\bar{p}$ (resp. $\bar{c}$) is the average daily expressed mood over the considered period in passenger tweets (resp. in airline customer service tweets).

The empathy score Ξ goes from -1 to 1, with a score of 1 meaning that the airline customer service expressed mood is in agreement with the mood expressed by their passengers. On the opposite, a score of -1 indicates that the mood expressed by the airline customer service is in complete opposition of phase with the mood expressed by their passengers. Such a score would indicate that the mood expressed by the airline customer service increases when the mood expressed in passenger tweets decreases, and *vice-versa*. A

Table 4.1: Airline ranking based on the proposed **empathy score** Ξ and the **sentiment gap** Δ applied to the period of March 1st 2020 to March 31st 2020.

Rank	**Airline**	Ξ	**Rank**	**Airline**	Δ
1	Alaska Airlines	0.476	1	Frontier Airlines	0.104
2	Southwest Airlines	0.456	2	Alaska Airlines	0.179
3	Frontier Airlines	0.374	3	Delta Air Lines	0.228
4	Spirit Airlines	0.146	4	JetBlue Airways	0.228
5	United Airlines	0.129	5	Spirit Airlines	0.237
6	JetBlue Airways	0.066	6	Southwest Airlines	0.244
7	Delta Air Lines	0.029	7	United Airlines	0.246
8	American Airlines	-0.393	8	American Airlines	0.260

score of 0 indicates that the mood expressed by the airline customer service and the mood expressed by their passengers are uncorrelated.

The second proposed metric aims at measuring the gap observed between the mood expressed by passengers in their tweets and the mood expressed in the tweets written by airline customer services.

Proposed passenger-centric metric 2 *The airline* ***sentiment gap*** *is the average difference between the mood expressed by passengers and the mood expressed by airlines.*

The airline sentiment gap Δ is calculated using the following formula:

$$\Delta = \frac{1}{N} \sum_{i} (p_i - c_i) \tag{4.2}$$

where N is the number of days considered and the set $\{p_i\}_i$ (resp. $\{c_i\}_i$) is the ordered set of the daily expressed mood in passenger tweets (resp. in airline customer service tweets), as for the airline empathy score Ξ presented in equation (4.1).

The airline sentiment gap Δ goes from -1 to 1 with a gap of 0 indicating that airline customer services and passengers express the same average mood in their tweets. A gap of 1 indicates a mood expressed by an airline customer service equal to 1 (i.e. the highest possible mood) and a mood expressed by the airline passengers equal to 0 (i.e. the lowest possible mood) on every day of the considered period. A gap of -1 indicates the opposite scenario.

Table 4.1 shows the ranks and scores of the seven airlines associated with each of the two passenger-centric metrics proposed in this section. Both the empathy score Ξ and the sentiment gap Δ were calculated over the period from March 1st 2020 to March 31st 2020.

4.3 Keyword-based metrics

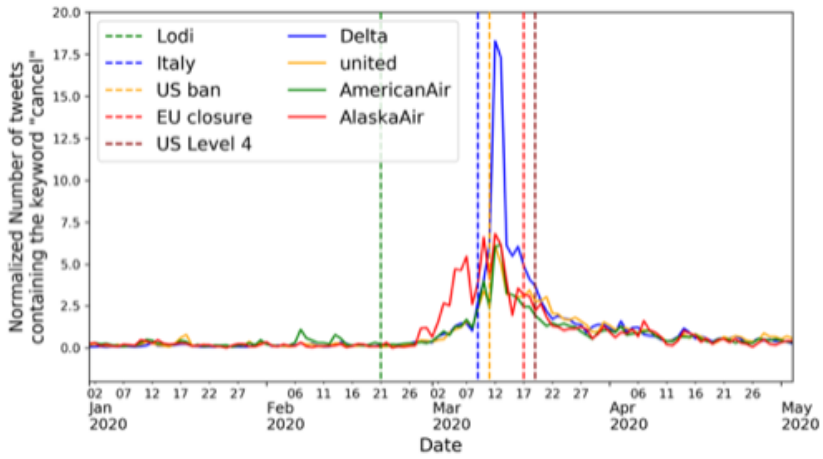
4.3.1 Cancellations

When some exceptional situation occurs, an important increase in the use of specific keywords within the stream of tweets written by the affected users can take place. For example, if many cancellations occur, many passengers will connect to Twitter and write tweets containing the keyword "cancel" to express their concerns directly to the airline they have bought tickets from. In this analysis, any word starting with the keyword "cancel", such as "cancellation" or "cancelled", is considered as a keyword "cancel".

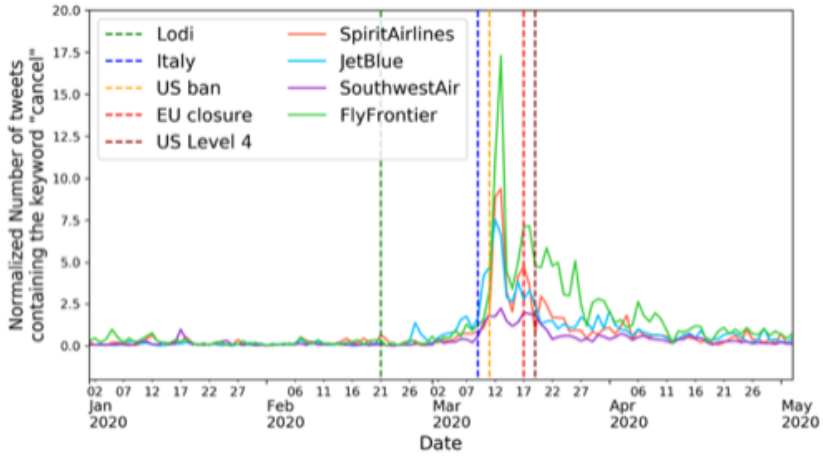
Figure 4.4 shows the evolution of the normalized number of tweets written by passengers and containing the keyword "cancel" between January 1st 2020 and May 3rd 2020 for four US legacy airlines and four US low-cost carriers. The normalization is based on the total number of passengers carried by each airline in 2018 and available in the yearly BTS reports [1].

Figure 4.4(a) indicates that the passengers of the four legacy airlines react as early as Italy's lockdown announcement with an important increase in the number of tweets containing the keyword "cancel". A second spike in the number of passenger tweets containing the keyword "cancel" then occurs once the US announces that it bans all travelers from the EU, China and Iran. Figure 4.4(a) shows that Delta Air Line passengers were, in proportion, about three times more vocal about cancellations on Twitter than the other legacy airlines at this period. This could be an indication that Delta Air Line had a greater proportion of passengers traveling within or through the EU at that time. The number of tweets from Alaska Airlines passengers containing the keyword "cancel" had an early spike compared to the tweets written by passengers from the other legacy airlines. That early spike could be linked to the fact that most of the early US cases of COVID-19 were discovered on the US West Coast first, which is where the main hub of Alaska Airlines is located.

Figure 4.4(b) shows the evolution of the number of tweets containing the keyword "cancel" written by passengers of the four low-cost carriers. Southwest Airlines passengers were, in proportion, less vocal on Twitter on the matter of cancellation than passengers of the other low-cost carriers, with a slight increase in the number of tweets containing the keyword "cancel" that is almost entirely contained within the period between the announcement of Italy's lockdown and the start of the US Level 4 Global Health Travel Advisory. JetBlue Airways passengers display a behavior similar to passengers of legacy airlines in this case. Passengers of Spirit Airlines and Frontier Airlines waited until the US travel ban announcement to communicate mas-



(a) From passengers of legacy airlines



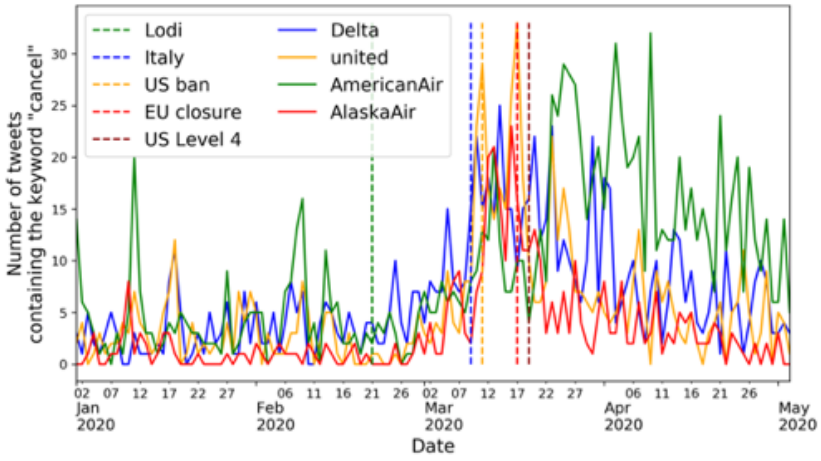
(b) From passengers of low-cost airlines

Figure 4.4: Number of tweets containing the keyword "cancel" and written by passengers normalized by the number of transported passengers per carrier over the year 2018 using BTS data [1]

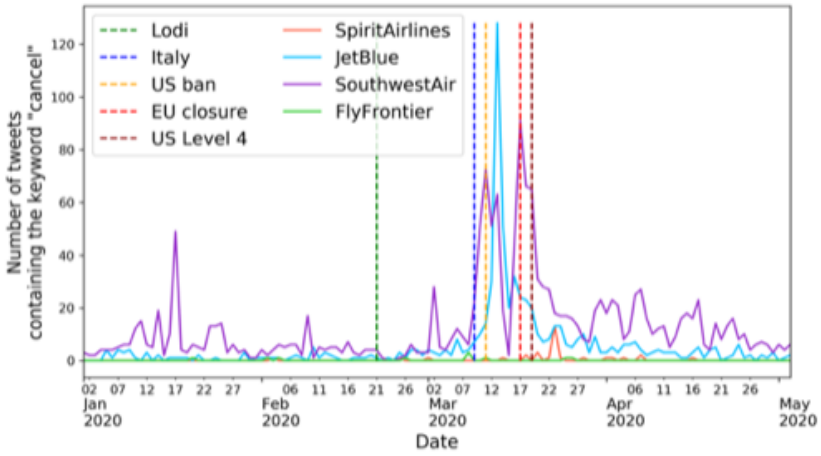
sively on Twitter their concerns using the word "cancel". The second spike in the number of tweets containing the keyword "cancel" starting at the announcement of the EU border closure is more important and lasts longer for tweets written by passengers of Frontier Airlines.

Figure 4.5 shows the evolution of the number of tweets containing the keyword "cancel" and written by airline customer services between January 1st 2020 and May 3rd 2020 for the same four US legacy airlines and three US low-cost carriers. Please note that the y-axis scale is different in Figure 4.5(a)

and Figure 4.5(b).



(a) From customer service of legacy airlines



(b) From customer service of low-cost airlines

Figure 4.5: Number of tweets containing the keyword "cancel" in tweets written by airline customer services

Regarding tweets written by legacy airline customer services, the evolution of the number of tweets containing the keyword "cancel" shown in Figure 4.5(a) presents similarities for three of the four airlines. There is a significant increase in the number of customer service tweets containing the keyword "cancel" starting the day Italy announced its lockdown and then a slow decrease. For tweets written by American Airlines customer service, the number of tweets containing the keyword "cancel" increases as for the other three airlines, but it does not decrease afterwards but fluctuates at a level

more important than during the period before the travel restriction measures where announced.

Regarding low-cost carriers, Figure 4.5(b) shows that each carrier use the keyword "cancel" on different occasions. The number of occurrences of the keyword "cancel" within tweets written by Southwest Airlines passengers has two important spikes around each of the US announcements referenced in the plot. JetBlue has a single massive spike on March 13$^{\text{th}}$ 2020. Both carriers then spent more than two weeks with a higher level of occurrences of the keyword "cancel" than in February 2020. Spirit Airlines customer service never wrote more than three tweets containing the keyword "cancel" in a day except on March 23$^{\text{rd}}$ 2020. Frontier Airlines customer service used the keyword "cancel" only in six tweets over the full month of March 2020.

Based on the observations from the plots in Figure 4.4, an important increase in the normalized number of passenger tweets containing the keyword "cancel" can be treated as an unwanted situation that airlines have to deal with.

Definition 1 *A keyword-related **Twitter situation** is defined as an increase over a predefined threshold of the normalized number of passenger-written tweets containing the keyword.*

Two metrics to measure the airline reaction to such a situation are proposed here. The aim of the first metric is to measure the effectiveness of the airline response to these keyword-related situations.

Proposed passenger-centric metric 3 *The keyword-related Twitter situation **quality response** score of an airline is the time needed for the airline to bring the normalized number of passenger tweets containing the keyword below a predefined threshold.*

The Twitter situation quality response score associated to the keyword "cancel" with a threshold of q normalized tweets κ^q_{cancel} is calculated using the following formula:

$$\kappa^q_{\text{cancel}} = d^q_{f,\text{cancel}} - d^q_{0,\text{cancel}} \tag{4.3}$$

where $d^q_{0,\text{cancel}}$ is defined as the first day of the considered period where the normalized number of passenger tweets containing the keyword "cancel" is greater than q, and $d^q_{f,\text{cancel}}$ is defined as the last day of the considered period where the normalized number of passenger tweets containing the keyword "cancel" is greater than q.

This proposed quality metric measures the time needed for the airline to bring the number of passenger tweets containing the keyword back to a

normal state. When measuring the response of long term perturbations, such as the COVID pandemic, this time is measured in days.

The number of passenger tweets containing the keyword is normalized by the total number of passengers carried by the airline over the year 2018 in this case, similarly to the data presented in Figure 4.4, and this normalization should be updated with the most recent numbers once they are available.

The aim of the second metric is to measure the communication effort produced by the airline in order to handle the situation linked to the increase of number of tweets containing the keyword under consideration.

Proposed passenger-centric metric 4 *The keyword-related Twitter situation* ***quantity response*** *score of an airline is calculated by integrating the number of tweets containing the keyword and written by the airline customer service over the number of days associated to the keyword-related Twitter situation.*

The formula used to calculate the Twitter situation quantity response score associated to the keyword "cancel" with a threshold of q normalized tweets γ^q_{cancel} is the following:

$$\gamma^q_{\text{cancel}} = \int_{d^q_{0,\text{cancel}}}^{d^q_{f,\text{cancel}}} n_{\text{cancel}}(t)\ dt \tag{4.4}$$

where $d^q_{0,\text{cancel}}$ and $d^q_{0,\text{cancel}}$ are the same as for the quality response score κ^q_{cancel} in equation (4.3), and $n_{\text{cancel}}(t)$ is the number of tweets written by the airline customer service containing the keyword "cancel" on day t.

Table 4.2 presents these two proposed metrics in the case of the keyword "cancel" considering that the predefined threshold indicating when a situation starts and ends is 1. Table 4.2 illustrates the necessity of considering both the quality response score and the quantity response score hand in hand. Southwest Airlines has the best scores from both perspective but Spirit Airlines has the second best quality response score but the second worst quantity response score. This would indicate that passengers from Spirit Airlines are more resilient to cancellation situations than passengers of the other airlines; they go back to a close-to normal Twitter chatter about cancellation with almost no cancellation related communication efforts on Twitter of Spirit Airlines.

4.3.2 Refund

Figure 4.6 shows the evolution of the normalized number of tweets containing the keyword "refund" and written by passengers from January 1$^{\text{st}}$ 2020 to May

Table 4.2: Airline ranking based on the "cancel"-related Twitter situation **quality and quantity response scores** $\kappa^1_{\textbf{cancel}}$ (in days) **and** $\gamma^1_{\textbf{cancel}}$ applied to the period of March 1st 2020 to April 30th 2020.

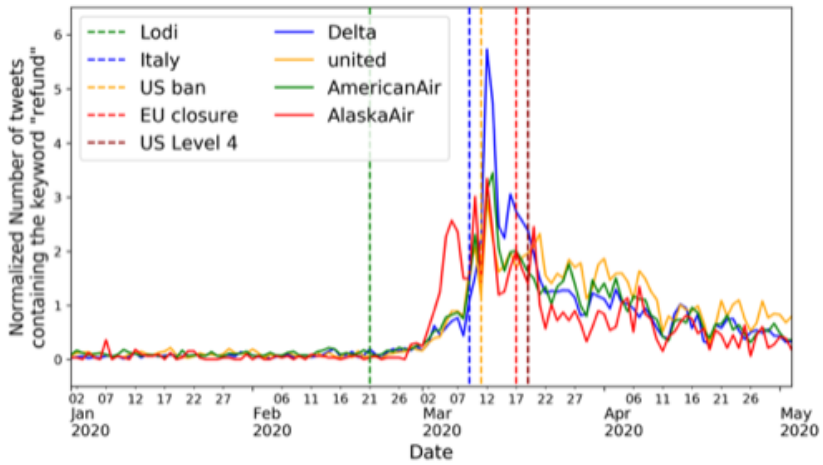
Rank	**Airline**	$\kappa^1_{\textbf{cancel}}$	**Rank**	**Airline**	$\gamma^1_{\textbf{cancel}}$
1	Southwest Airlines	11	1	Southwest Airlines	50.64
2	Spirit Airlines	26	2	American Airlines	15.34
3	United Airlines	34	3	Delta Air Lines	11.98
4	American Airlines	35	4	United Airlines	11.62
5	Delta Air Lines	41	5	JetBlue Airways	10.28
6	Alaska Airlines	44	6	Alaska Airlines	6.82
7	Frontier Airlines	53	7	Spirit Airlines	0.96
8	JetBlue Airways	54	8	Frontier Airlines	0.11

3rd 2020 for the same eight US airlines using the same normalization process as for the keyword "cancel".

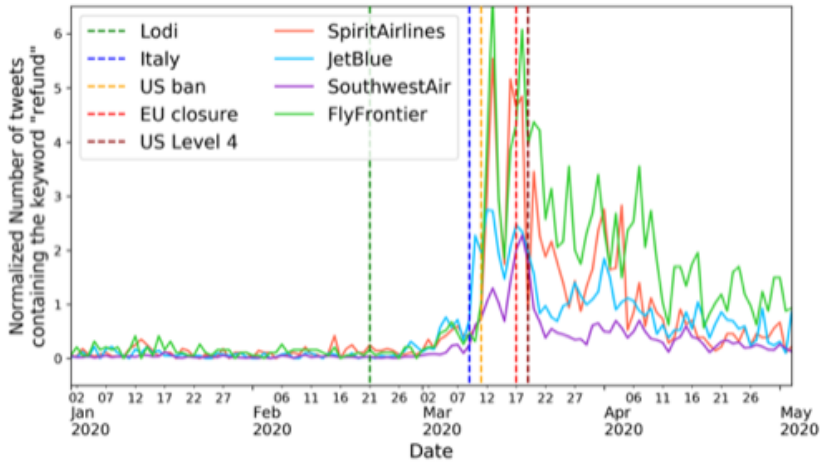
The evolution of the number of passenger tweets containing the keyword "refund" is similar to the evolution of the number of occurrences of the keyword "cancel" but at a lower proportion. Figure 4.6(a) shows that the number of occurrences of the keyword "refund" in tweets written by passengers of all four legacy airlines steeply increases at the announcement of Italy's lockdown and then very slowly decreases. Passengers of Alaska Airlines have an anticipated spike in the number of tweets containing the keyword "refund" at the beginning of March 2020. Figure 4.6(b) shows that the increase in the number of tweets containing the keyword "refund" and written by Southwest Airlines passengers is still lower than the number of tweets containing the keyword "refund" and written by the passengers of the other low-cost carriers. The number of tweets containing the keyword "refund" and written by Southwest Airlines passengers gets back to a normal level faster than for the passengers of the other low-cost carriers. The spike in the number of tweets containing the keyword "refund" and written by Spirit Airlines and Frontier Airlines passengers starts only at the announcement of the US travel ban.

Figure 4.7 shows the evolution of the number of tweets containing the keyword "refund" and written by airline customer services from January 1st 2020 to May 3rd 2020 for the same eight US airlines.

Figure 4.7(a) shows the evolution of the number of tweets containing the keyword "refund" and written by the customer services of the four considered legacy airlines. The initial increase is similar than for the keyword "cancel" (Figure 4.5(a)), however there is then a second increase towards the end of March 2020, this increase being most visible within the tweets written by American Airlines customer service. From a low-cost carrier perspective,



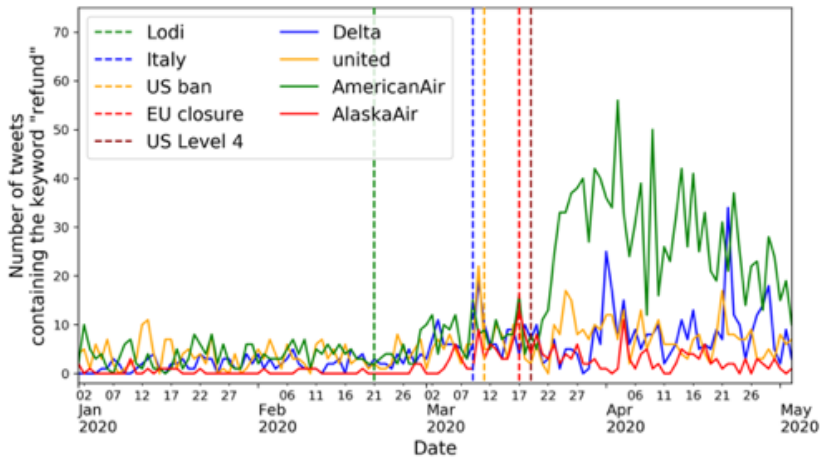
(a) From passengers of legacy airlines



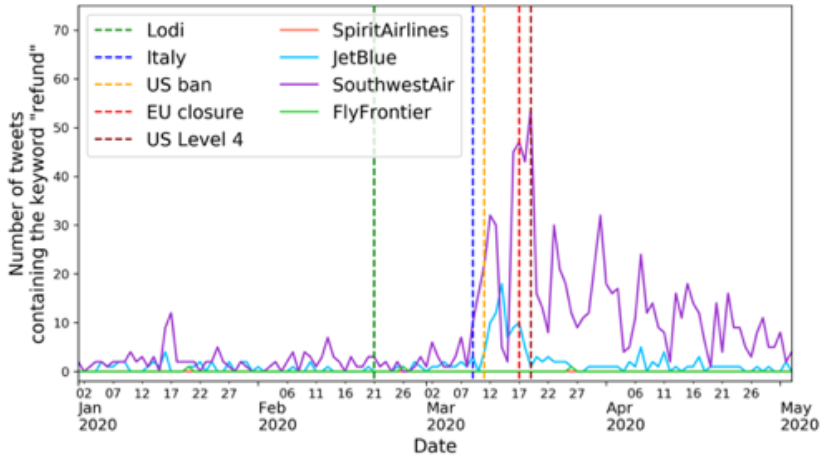
(b) From passengers of low-cost airlines

Figure 4.6: Number of tweets containing the keyword "refund" and written by passengers normalized by the number of transported passengers per carrier over the year 2018 using BTS data [1]

Figure 4.7(b) illustrates the same characteristics as in Figure 4.5(b): There are two spikes around the US announcements for the number of tweets containing the keyword "refund" in tweets written by Southwest Airlines customer service, this time with higher fluctuations afterwards, and one major spike on March 13th 2020 for the number of tweets containing the keyword "refund" and written by JetBlue Airways customer service. Only one tweet containing the keyword "refund" was written by Frontier Airlines customer service over the month of March 2020 and none written by Spirit Airlines



(a) From customer service of legacy airlines



(b) From customer service of low-cost airlines

Figure 4.7: Number of tweets containing the keyword "refund" and written by airline customer services

customer service since January 1st 2020.

The same two metrics associated to the "cancel"-related Twitter situation presented in Section 4.3.1, i.e. the quality response score and the quantity response score, can be used for this "refund"-related Twitter situation. Table 4.3 presents these two proposed metrics in the case of the keyword "refund" using the same predefined threshold of 1 for delimiting a Twitter situation.

As for the handling of the "cancel"-related Twitter situation, Southwest Airlines had the most effective (best quality response score) and most proactive (best quantity response score) of the eight airlines. The same resilience

Table 4.3: Airline ranking based on the "refund"-related Twitter situation **quality and quantity response scores** $\kappa^1_{\textbf{refund}}$ (in days) **and** $\gamma^1_{\textbf{refund}}$ applied to the period of March 1st 2020 to April 30th 2020.

Rank	**Airline**	$\kappa^1_{\textbf{refund}}$	**Rank**	**Airline**	$\gamma^1_{\textbf{refund}}$
1	Southwest Airlines	8	1	Southwest Airlines	32.25
2	Spirit Airlines	29	2	American Airlines	22.81
3	American Airlines	31	3	United Airlines	7.67
4	Alaska Airlines	35	4	Delta Air Lines	7.46
5	Delta Air Lines	37	5	Alaska Airlines	4.03
6	JetBlue Airways	39	6	JetBlue Airways	3.03
7	United Airlines	51	7	Frontier Airlines	0.02
7	Frontier Airlines	51	8	Spirit Airlines	0.00

is shown by passengers of Spirit Airlines during this "Refund"-related Twitter situation as for the "cancel"-related Twitter situation.

4.4 Impact the COVID-19 travel restriction measures on airports

Using datasets provided by the US Department of Homeland Security Customs and Border Protection [165] (CBP) and SafeGraph [167], this section presents an analysis of the impact of COVID-19-induced travel restrictions on the US airports, as they transpire through passenger-generated data gathered until April 22nd 2020.

4.4.1 Overall impact on the number of passengers/visitors at airports

International passengers

Travel restrictions do not ban entirely international travel, and there were still passengers arriving at most US airports of entry after the implementation of these travel restrictions. However, starting March 13th 2020, US citizens who have been in high risk areas and are returning to the United States must arrive at one the thirteen following airports of entry: [162]

- ATL: Hartsfield-Jackson Atlanta International Airport
- BOS: Boston-Logan International Airport
- DFW: Dallas Fort Worth International Airport
- DTW: Detroit Metropolitan Airport

- EWR: Newark Liberty International Airport
- HNL: Daniel K. Inouye International Airport
- IAD: Washington-Dulles International Airport
- JFK: John F. Kennedy International Airport
- LAX: Los Angeles International Airport
- MIA: Miami International Airport
- ORD: Chicago O'Hare International Airport
- SEA: Seattle-Tacoma International Airport
- SFO: San Francisco International Airport

The effect of these travel restrictions on international travel coming to the US can be analyzed thanks to the "Airport Wait Times" data from the Customs and Border Protection (CBP) website [165]. The data provided are hourly aggregates and are usually available on the following day they are generated. The immediate availability of the data is due to the fact that CBP measures directly the signal generated by passengers, which is generated through their passports once passengers clear the immigration process, and does not have to wait for an airline or airport to process and provide the data.

Among other information, the dataset contains the number of passengers arriving at immigration per hour, the average wait time at immigration per hour, and the number of open immigration booths per hour. For a more detailed presentation of the available dataset, the authors recommend the reading of [16], which also proposes an analysis of these wait times from January 2013 to January 2019. The data considered in the current report ranges from January 1st 2020 to April 22nd 2020.

Looking first at the evolution of the total number of passengers arriving at US immigration booths per day across all airports, the total number of passengers arriving at US immigration booths drops from an average of 218,700 per day between February 23rd 2020 and March 15th 2020 to an average of only 5,000 passengers per day between April 1st 2020 and April 22nd 2020. This represents a drop of 97.7% of international passenger inflow within two weeks. The day by day evolution of the total number of passengers arriving at US immigration from March 1st to April 22nd is shown in Figure 4.8. This figure also indicates the last date where immigration data are available for each airport with no CBP immigration data available on April 22nd 2020. This corresponds to 22 airports. Only Raleigh–Durham International Airport (RDU) closed its immigration service between the US ban of EU travelers and before the US entered a Level 4 travel advisory. In addition John Wayne Airport (SNA) has no immigration data since January 5th 2020. Beginning March 22nd 2020, the number of airports not generating

any immigration data steadily increases with nine airports shutting down their immigration services within ten days. Another nine airports then stop generating immigration data within ten days beginning April 12th 2020.

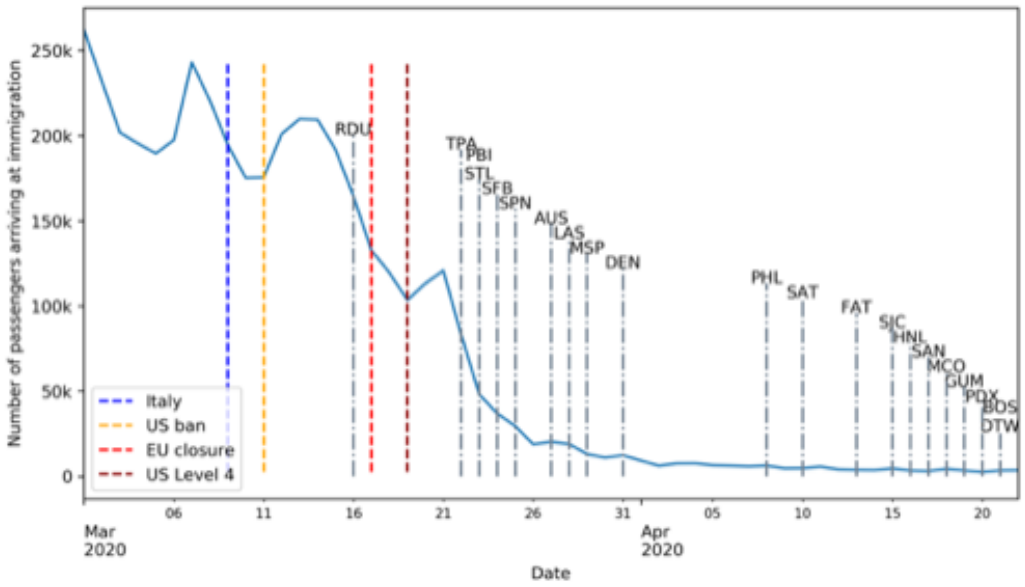


Figure 4.8: Evolution of the daily number of passengers arriving at all US airports of entry. The dates of last recorded CBP data for airports with no immigration data on the date of April 22nd 2020 are indicated as dotted lines.

Among the airports with no immigration data on April 22nd 2020 indicated on Figure 4.8 is BOS, which last day of recorded immigration data was on April 21st 2020, even though it is one of the selected airport of entry for US citizens coming from high-risk areas. This illustrates the fact that the influx of international passengers is so low that BOS had no arriving international passenger to their immigration services for at least one day.

Impact on the number of airport visitors

From a domestic perspective, weekly patterns at specific points of interest (POI) are available within the data provided by SafeGraph[2]. From these patterns, it is possible to have an estimate of the number of airport visitors per hour by considering all available POI associated with an airport. Airport visitors are a broader category than air passengers, since this category also encompasses airport staff and people dropping off or picking up passengers.

[2] `https://docs.safegraph.com/docs/weekly-patterns`

The data available for this study ranges from February 27th 2020 to April 18th 2020.

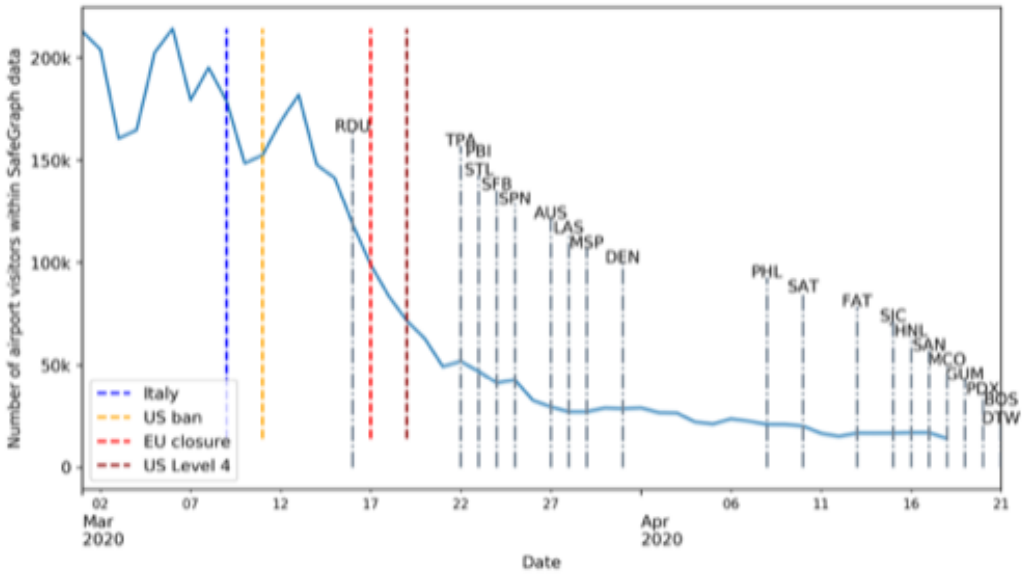


Figure 4.9: Evolution of the total number of daily airport visitors using SafeGraph data. The dates of last recorded CBP data for airports with no immigration data on the date of April 22nd 2020 are indicated as dotted lines.

Looking first at the evolution of the total number of airport visitors per day across all airports, the number of airport visitors captured within the SafeGraph data drops from an average of 176,800 visitors per day between February 27th 2020 and March 15th 2020 to an average of only 20,200 visitors per day between April 1st 2020 and April 18th 2020. This represents a drop of 88.6% airport visitors within two weeks. The day by day evolution of the total number of airport visitors from March 1st to April 18th is shown in Figure 4.9. Similarly to Figure 4.8, this figure also indicates for each airport with no CBP immigration data available on April 22nd the last date where immigration data are available.

Figure 4.9 shows that US domestic travel was already impacted before the rise to a Level 4 travel advisory: The number of airport visitors contained within the SafeGraph data drops from 152,400 on March 11th 2020 down to 71,600 on March 19th 2020, which represents a 53% drop.

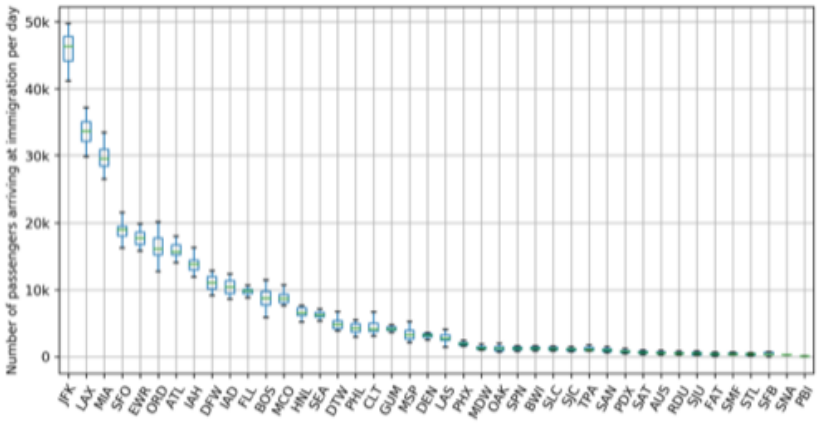
4.4.2 Distribution of the impact across airports

At immigration

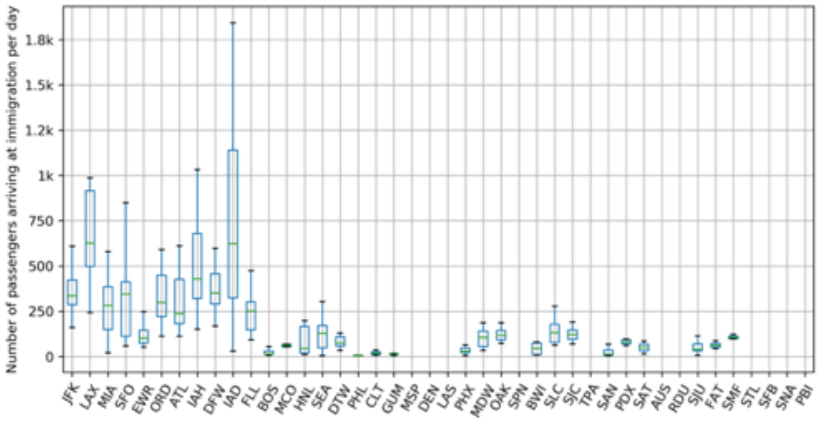
Following the global impact of the COVID-19 travel restrictions, their impact on the individual airports is analyzed in the following section. Figure 4.10 compares the individual airport situation of the first two weeks of April 2020 with the first two weeks of April 2019. Figure 4.10(a) shows the boxplots of the number of passengers arriving at immigration per day for each airport over the period of April 1st-22nd 2019. The median number of arriving passengers is indicated in green and each box lower and upper bounds represent respectively the 1st and 3rd quartile. The whiskers above and below each box give a visualization of the full range of the considered data even though extreme values are not drawn. The airports are ordered by their median daily number of passengers arriving at immigration over that period. Figure 4.10(b) shows the boxplots of the number of passengers arriving at immigration per day for each airport over the period of April 1st-22nd 2020. The airports in this figure are in the same order as for Figure 4.10(a). Please note that the y-axis are not the same between Figure 4.10(a) and Figure 4.10(b) due to the important drop in the number of passengers arriving at US airports of entries after the implementation of the travel restriction measures.

Figure 4.10(a) is a snapshot of the "normal" situation regarding the number of passengers arriving at US immigration over the first three weeks of April, while Figure 4.10(b) is a snapshot of a pandemic situation regarding the number of passengers arriving at US immigration. Please note that the y-axis are different between Figure 4.10(a) and Figure 4.10(b) due to the important drop in international travel following the COVID-19 travel restrictions. The thirteen airports chosen for handling the return of US citizens from high-risk areas are all in the top 16 airports with the highest median daily number of passengers arriving at immigration, along with George Bush Intercontinental Airport (IAH), Fort Lauderdale–Hollywood International Airport (FLL) and Orlando International Airport (MCO).

The drop in the number of passengers arriving at immigration per day is clearly visible between the years 2019 (Figure 4.10(a)) and 2020 (Figure 4.10(b)) for the airports with the most arriving passengers. JFK is the airport with the highest number of passengers arriving at immigration per year since 2013 [165] and has the most important drop in volume going from a median number of passengers arriving at immigration of 45,900 between April 1st-22nd 2019 down to a median of 360 April 1st-22nd 2020. For JFK, this drop represents a drop of 99.3% between the median number of passengers arriving at immigration of these two periods. For all the considered airports,



(a) April 1st-22nd 2019



(b) April 1st-22nd 2020

Figure 4.10: Boxplots of the number of arriving passengers per day for each airport of entry to the US over the first three weeks of April for the years 2019 and 2020.

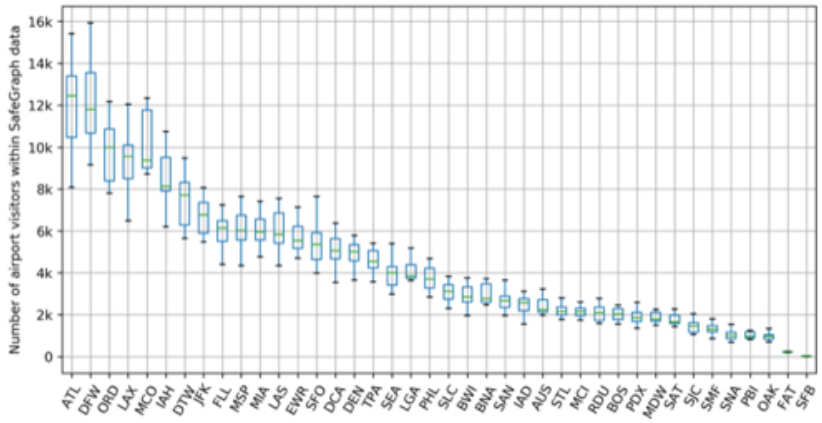
the corresponding drop is between 70.7% for Sacramento International Airport (SMF) and 100% for the eleven airports without no immigration data between April 1st 2020 and April 22nd 2020.

Looking at the airport ranking based on the median number of passengers arriving at immigration per day over the period of April 1st-22nd, Figure 4.10(b) shows that it has been reshuffled from year 2019 to year 2020: JFK dropped to the sixth place and IAD climbed to the second place right behind LAX. IAD has however the highest average number of passengers arriving at immigration per day over the period of April 1st-22nd 2020 with 726 passengers a day on average, LAX being second with 658 passengers a

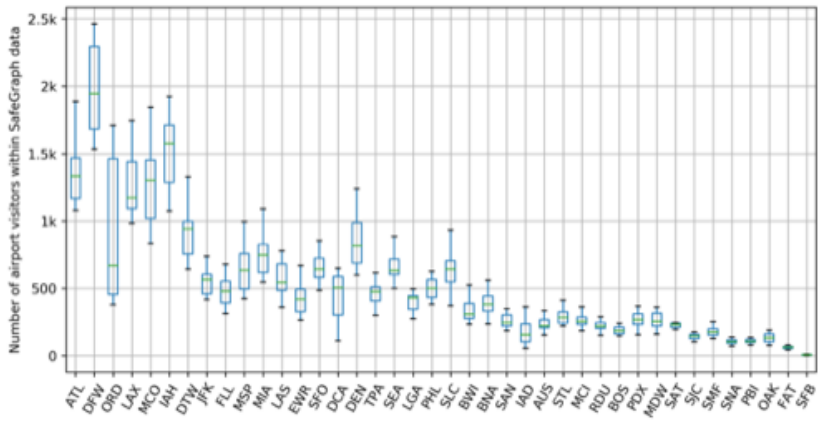
day on average.

At airports

A similar comparison of the number of airport visitors before and after the travel restriction measures can be conducted based on the SafeGraph data. Due to data availability, this comparison has to take place between March 2020 and April 2020. Figure 4.11 shows the boxplots of the number of airport



(a) March 1st-15th2020



(b) April 1st-15th2020

Figure 4.11: Boxplots of the number of airport visitors per day for 44 US airport with available SafeGraph data over the first two weeks of March and April 2020.

visitors per day for 40 US airport with available SafeGraph data over the first two weeks of March 2020 (Figure 4.11(a)) and April 2020 (Figure 4.11(a)).

The airports on these two plots are sorted by their median daily number of airport visitors over the period of March 1st-15th 2020. Please note that the y-axis are not the same between Figure 4.11(a) and Figure 4.11(b) due to the important drop in the number of passengers arriving at US airports of entries after the implementation of the travel restriction measures.

As for the number of passengers arriving at immigration, where JFK has both the highest median daily number of passengers and the most important drop in volume, ATL has both the highest median daily number of airport visitors and the most important drop in volume. ATL has a drop of 11,100 airport visitors in the SafeGraph data between these two weeks, which represents a 89.3% drop. While some airports stopped generating immigration data following the travel restrictions, no airport stops completely of receiving visitors, though the drop is important for all 40 considered airports, ranging from 72.5% for Fresno Yosemite International Airport (FAT) to 93.8% for IAD.

4.4.3 Proposed passenger-centric metrics

For passengers arriving at immigration

With 35 airports experiencing a drop in the median number of passengers arriving at immigration greater than 90%, all airports are severely impacted by the COVID-19 measures from a passenger-volume perspective. A few questions related to operations are examined here from a passenger perspective. Since there are far fewer passengers arriving at immigration, does the immigration process go faster? The number of agents operating immigration booths has also decreased due to the corona virus, but it is possible to consider an immigration load factor.

Definition 2 *The **immigration load factor** is defined as the ratio of the number of passengers arriving at immigration per hour with the number of open immigration booths per hour.*

The immigration load factor ρ is calculated using the following formula:

$$\rho = \frac{n_{\text{PAX}}}{n_{\text{booth}}} \tag{4.5}$$

where n_{PAX} is the total number of passengers arriving at immigration during the considered hour and n_{booth} is the number of immigration booths operating during that same hour.

The immigration load factor ρ indicates the load in terms of passengers for each immigration booth per hour. A lower load indicates that each immigration booth has fewer passengers to process per hour. From a passenger

perspective, a lower load for a given number of passengers, indicates that there are more immigration booths open, so the average processing time should be lower and thus a passenger at immigration would have to wait less to be processed.

The daily immigration load factor ϱ is calculated as the daily average immigration load factor using the following formula:

$$\varrho = \frac{1}{\delta t} \sum_t \rho_t \tag{4.6}$$

where $\{\rho_t\}_t$ is the set of immigration load factors for every operating hour t of the day under consideration and δt is the number of operating hours of the day.

Assumption 1 *If the daily immigration load factor decreases, then the daily average wait time for passengers at immigration should decrease as well.*

Based on this reasoning, an **immigration quality score** is proposed: It measures how well Assumption 1 is verified for an airport immigration service over a selected period of days.

Proposed passenger-centric metric 5 *The **immigration quality score** for an airport of entry is defined as the correlation between the daily average wait time for passengers at its immigration service and the daily average immigration load factor of the airport over a given period.*

The formula used to calculate the immigration quality score χ is the following:

$$\chi = \frac{\sum_i (\varrho_i - \bar{\varrho})(\tau_i - \bar{\tau})}{\sqrt{\sum_i (\varrho_i - \bar{\varrho})^2 \sum_i (\tau_i - \bar{\tau})^2}} \tag{4.7}$$

where the set $\{\varrho_i\}_i$ (resp. $\{\tau_i\}_i$) is the ordered set of the daily immigration load factors (resp. the daily average wait times at immigration), and $\bar{\varrho}$ (resp. $\bar{\tau}$) is the average of the set $\{\varrho_i\}_i$ (resp. $\{\tau_i\}_i$).

This immigration quality score is equal to 1 if Assumption 1 is perfectly verified, to 0 if the daily average wait time for passengers at immigration is uncorrelated with the daily average immigration load factor and to -1 if the opposite of Assumption 1 occurs over the considered period, i.e. a decrease in the daily average immigration load factor implies an increase in the daily average wait time for passengers at immigration.

This proposed passenger-centric metric is applied to the period pre-COVID (January 1$^{\text{st}}$ 2020 to February 29$^{\text{th}}$ 2020) and to the period post-COVID

Table 4.4: Airport partial ranking based on the proposed immigration quality score χ applied to the period of pre-COVID of January 1st 2020 to February 29th 2020 and to the period post-COVID of March 1st 2020 to April 22nd 2020 for the 40 considered US airports of entry.

	Top ten best airports					**Top ten worst airports**			
	Pre-COVID		**Post-COVID**			**Pre-COVID**		**Post-COVID**	
Rank	**Airport**	**Score**	**Airport**	**Score**	**Rank**	**Airport**	**Score**	**Airport**	**Score**
1	SFB	0.92	SFB	0.98	40	JFK	-0.32	SLC	-0.1
2	SPN	0.91	LAS	0.93	39	ATL	-0.17	AUS	0.17
3	PBI	0.72	SFO	0.89	38	MDW	-0.17	PDX	0.17
4	RDU	0.72	MSP	0.85	37	PDX	-0.09	HNL	0.17
5	GUM	0.68	PHX	0.84	36	OAK	-0.05	IAD	0.19
6	CLT	0.66	MIA	0.83	35	IAH	-0.01	BWI	0.21
7	SAT	0.63	DEN	0.82	34	LAX	0.02	MDW	0.25
8	MCO	0.62	FAT	0.82	33	MSP	0.02	SJU	0.29
9	TPA	0.61	CLT	0.81	32	SFO	0.12	OAK	0.34
10	PHL	0.6	JFK	0.81	31	FAT	0.14	DFW	0.4

(March 1st 2020 to April 22nd 2020) for 40 US airports of entry. Table 4.4 shows the associated partial ranking (top ten best airports and top 10 worst airports) for these two periods.

The airport still generating immigration data on April 24th 2020 with the worst drop between the period pre-COVID and the period post-COVID is IAD, going from 17th down to 36th, and the airport still generating immigration data on April 24th 2020 with the best increase in rank is JFK, with 30 places gained and with an increase in score from the negative value of -0.32 to the positive value of +0.81.

For airport visitors in general

With 38 airports having a drop in the median number of airport visitors greater than 80%, all airports are also severely impacted by the COVID-19 travel restrictions measures from a visitor-volume perspective. Visitors in general avoid airports, but some are still going to the airports after the travel restriction measures. The same question as for the immigration process can be asked: Are these visitors processed faster since there are less visitors?

The data for visitors available for this is different than the data available for passengers arriving at immigration, therefore a different approach has to be considered here. The SafeGraph data contains weekly bucketed dwell times for each considered location. The dwell time is the time spent at that location, be it waiting, shopping, walking, etc. The buckets are: less than 5 minutes, between 5 and 20 minutes, between 21 and 60 minutes, between 61 minutes and 240 minutes and more than 240 minutes. From these weekly bucketed dwell times, two complementary passenger-metrics are proposed to measure an airport efficiency to process visitors.

Proposed passenger-centric metric 6 *The weekly* ***airport visitor efficiency score*** *for an airport is defined as the weekly proportion of airport visitors spending less than 60 minutes at an airport.*

The airport visitor efficiency score η for a given week is calculated using the following formula:

$$\eta = \frac{n_{<60}}{N} \tag{4.8}$$

where $n_{<60}$ is the number of airport visitors that spend less than 60 minutes at the airport during the week under consideration and N is the total number of airport visitors during that same week.

Proposed passenger-centric metric 7 *The weekly* ***airport visitor sluggishness score*** *for an airport is defined as the weekly proportion of airport visitors spending more than 240 minutes at an airport.*

The airport visitor sluggishness score ζ for a given week is calculated using the following formula:

$$\zeta = \frac{n_{>240}}{N} \tag{4.9}$$

where $n_{>240}$ is the number of airport visitors that spend more than 240 minutes at the airport during the week under consideration and N is the total number of airport visitors during that same week.

The time thresholds within these two metrics are also chosen due to the format of the data, and could be adjusted to less aggregated data. Airport visitors staying less than 60 minutes are essentially visitors dropping off or picking up a passenger, and potentially some passengers on domestic flights, where the overall security screening process is faster than for international flights. Therefore, the idea behind the airport visitor efficiency score is to measure how efficiently airports keep the flow of people coming in and out of their facilities. Regarding the 240 minutes threshold, most airlines and airports recommend their passengers on international flights to arrive two to three hours ahead of their flight's scheduled departure time, therefore the idea behind the airport visitor sluggishness score is to measure the validity of this recommendation.

Airport staff can be counted as airport visitors using this dataset and they are likely to stay more than 240 minutes at the airport, increasing the number of airport visitors staying longer than this threshold. Therefore, an airport with a high airport visitor sluggishness score could either be an airport with many passengers taking more than four hours to clear their entire airport process, or an airport with a disproportionate number of airport staff compared to the number of airport visitors.

Regarding the actual implementation of these scores, since there are several locations per airport within the SafeGraph data, e.g. "LAX Terminal 4" and "LAX Terminal South" for LAX, an estimation of the proposed airport visitor efficiency score is calculated by taking the minimum weekly proportion of airport visitors spending less than 60 minutes at a location within an airport over all considered airport locations. Similarly, an estimation of the proposed airport visitor sluggishness score is calculated by taking the maximum weekly proportion of airport visitors spending more than 240 minutes at a location within an airport over all considered airport locations.

These proposed passenger-centric metrics are applied to the period pre-COVID (March 1$^{\text{st}}$ 2020 to March 15$^{\text{th}}$ 2020) and to the period post-COVID (April 5$^{\text{th}}$ 2020 to April 19$^{\text{th}}$ 2020) for 44 US airports. These periods contain 2 weeks each and therefore 2 points of data each. The scores are calculated for each week and then averaged over the period. Table 4.5 shows the partial ranking (top ten best airports and top 10 worst airports) associated to the

Table 4.5: Airport partial ranking using the proposed airport visitor efficiency score η applied to the period of pre-COVID of March 1st 2020 to March 15th 2020 and to the period post-COVID of April 5th 2020 to April 19th 2020 for the 44 considered US airports based on SafeGraph data.

	Top ten best airports					**Top ten worst airports**			
	Pre-COVID		**Post-COVID**			**Pre-COVID**		**Post-COVID**	
Rank	**Airport**	**Score**	**Airport**	**Score**	**Rank**	**Airport**	**Score**	**Airport**	**Score**
1	SJC	0.69	SJC	0.6	44	LGA	0.0	LAX	0.0
2	GUM	0.69	SMF	0.56	43	LAX	0.0	SLC	0.0
3	MCI	0.66	RDU	0.55	42	SFB	0.1	MIA	0.12
4	RDU	0.65	SAT	0.55	41	DEN	0.13	DEN	0.14
5	SMF	0.64	AUS	0.53	40	MIA	0.21	SFO	0.16
6	OAK	0.63	OAK	0.53	39	ATL	0.24	BNA	0.18
7	SAT	0.62	MCI	0.53	38	EWR	0.24	ATL	0.19
8	AUS	0.62	GUM	0.5	37	SLC	0.25	DTW	0.19
9	TPA	0.61	PHX	0.49	36	DCA	0.31	IAD	0.2
10	FAT	0.6	STL	0.48	35	PDX	0.32	EWR	0.2

Table 4.6: Airport partial ranking using the proposed airport visitor sluggishness score ζ applied to the period of pre-COVID of March 1st 2020 to March 15th 2020 and to the period post-COVID of April 5th 2020 to April 19th 2020 for the 44 considered US airports based on SafeGraph data.

	Top ten best airports					**Top ten worst airports**			
	Pre-COVID		**Post-COVID**			**Pre-COVID**		**Post-COVID**	
Rank	**Airport**	**Score**	**Airport**	**Score**	**Rank**	**Airport**	**Score**	**Airport**	**Score**
1	GUM	0.04	GUM	0.0	44	LGA	1.0	LAX	1.0
2	MCI	0.05	SMF	0.16	43	SFB	0.73	SFB	0.71
3	SMF	0.05	SJC	0.17	42	DEN	0.62	DEN	0.67
4	SJC	0.05	HNL	0.19	41	SLC	0.52	SFO	0.55
5	AUS	0.05	RDU	0.2	40	EWR	0.49	LGA	0.55
6	OAK	0.05	PHX	0.2	39	MSP	0.48	ATL	0.53
7	STL	0.05	OAK	0.2	38	DTW	0.35	SLC	0.5
8	SAT	0.06	AUS	0.21	37	LAX	0.27	EWR	0.49
9	PHX	0.06	CLT	0.21	36	DFW	0.27	SNA	0.48
10	PBI	0.06	SAT	0.21	35	ATL	0.27	MDW	0.47

proposed airport visitor efficiency score η for these two periods.

In Table 4.5, a score of 1 indicates that all airport visitors within the SafeGraph data spend less than one hour at the same location within the airport, while a score of 0 indicates that all airport visitors within the SafeGraph data spend more than one hour at the same location within the airport. Some airports have a score of 0 due to locations receiving very few visitors (fewer than 5) over the considered week that were captured within the SafeGraph data, and all those visitors stayed more than one hour at that same airport location.

Table 4.6 shows the partial ranking (top ten best airports and top 10 worst airports) associated to the proposed airport visitor sluggishness score ζ for the same two considered periods. In Table 4.6, a score of 0 indicates that all airport visitors within the SafeGraph data spend less than four hours at the same location within the airport, while a score of 1 indicates that all airport visitors within the SafeGraph data spend more than four hours at the same location within the airport. Similarly as with the visitor airport efficiency score, some airports have a score of 1 due to locations receiving very few visitors (less than 5) over the considered week that were captured within the SafeGraph data, and all those visitors stayed more than four hours at that same airport location.

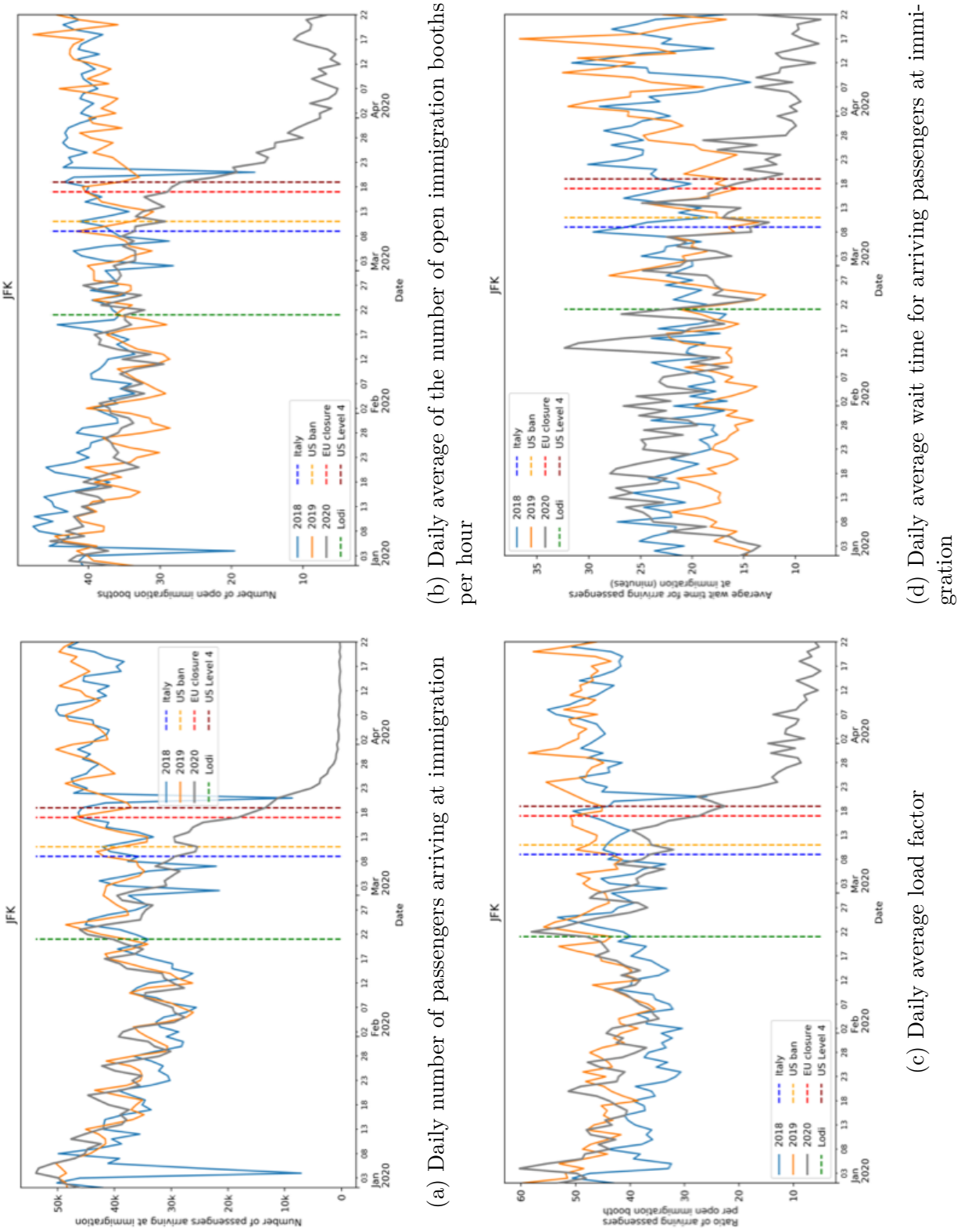
4.4.4 Cases of JFK and IAD immigration process

Following the ranking resulting from the metric proposed in Section 4.4.3, this section focuses on JFK and IAD, which undergo the largest changes in behavior linked to the COVID-19 travel restriction measures.

JFK

JFK had the best increase in rank using the proposed immigration quality score presented in Table 4.4, and this section aims at analyzing the available CBP immigration data. The effect of the travel ban measures presented in Section 4.1.1 on passengers arriving at JFK's immigration is presented in Figure 4.12 through four different views by comparing data from 2020 with CBP data from the years 2018 and 2019 between January 1st and April 22nd.

Figure 4.12(a) shows the daily evolution of the number of passengers arriving at JFK's immigration and confirms the important drop in the number of arriving international passengers to the US from an average 35,600 thousand passengers arriving at immigration per day down to barely 360 passengers per day. The drop in the number of passengers arriving at JFK immigration is more important when comparing over the same period of the years



(a) Daily number of passengers arriving at immigration

(b) Daily average of the number of open immigration booths per hour

(c) Daily average load factor

(d) Daily average wait time for arriving passengers at immigration

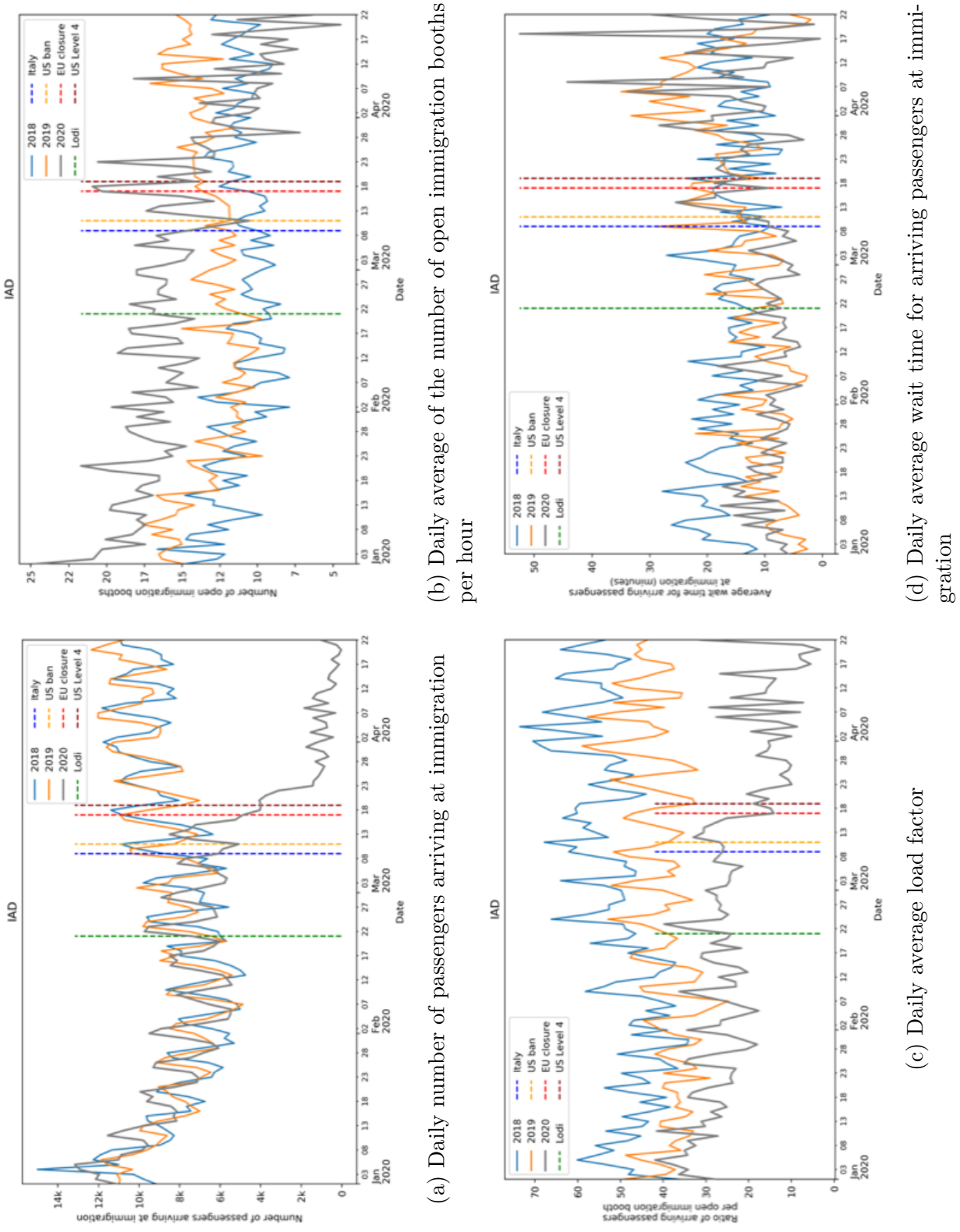
Figure 4.12: JFK airport: Comparison of CBP data from January 1st to April 13th for the years 2018 to 2020.

2018 and 2019, with an average of 45,900 thousand passengers per day in April 2019. Figure 4.12(b) shows the daily evolution of the average number of open immigration booths per hour at JFK and presents a similar drop than in Figure 4.12(a), the number of open immigration booths dropping from around an average of 35.4 per operating hour down to an average of 7.5 per hour. This drop is however less important in proportion compared to the drop in the number of passengers arriving at immigration. Figure 4.12(c) shows the evolution of the daily average load factor (Definition 2 and equations (4.5)-(4.6)). After the lockdown and travel ban measures, the daily load factor drops significantly from an average of 42.5 before the measures down to around 8.5, which represents a 80% drop. This indicates that after the measures, an immigration booth has about five times fewer passengers to process per hour. This has a direct positive impact to the average wait time at immigration for passengers. Figure 4.12(d) shows the daily evolution of the average wait time for passengers at JFK's immigration. It was reduced by half after the lockdown and travel ban measures, from around 21.5 minutes to around 10.5 minutes, compared to the usual April levels of 26 minutes in 2019 and 23 minutes in 2018.

IAD

IAD's rank experiences the worst drop using the proposed immigration quality score presented in Table 4.4, and is the focus of this section. Figure 4.13 shows the impact of the travel restriction measures for passengers arriving at IAD's immigration through the four same perspectives as the analysis of JFK.

Figure 4.13(a) shows the daily evolution of the number of passengers arriving at IAD immigration. It confirms that, even though in 2020 that number has dropped from an average of 7,200 thousand in February 2020 to an average of 726 in April 2020 after the implementation of the travel ban measures, the drop is less important than for JFK (Figure 4.12(a)). Though this is still a 93% drop for the number of passengers arriving at immigration in April between the years 2019 and 2020, with a daily average of 10,400 thousand passengers in 2019, the number of open immigration booths did not decrease as much as for JFK. Figure 4.13(b) shows the daily evolution of the average number of open immigration booths per hour at IAD. The daily average of open booths per hour over the month of April 2020, with an average of 10.1 per hour, is similar to the daily average over the month of April 2018, with an average of 11 per hour, and only slightly lower than the number of open booths over the month of April 2019, with an average of 14.6 per hour. Over the period of January to March, the daily average of the number of open



(a) Daily number of passengers arriving at immigration

(b) Daily average of the number of open immigration booths per hour

(c) Daily average load factor

(d) Daily average wait time for arriving passengers at immigration

Figure 4.13: IAD airport: Comparison of CBP data from January 1st to April 13th for the years 2018 to 2020.

booths per hour is significantly higher in 2020 than in 2018 or 2019, with an average of 10.2 for 2018 and of 11.7 for 2019. The daily load factor evolution at IAD is similar to JFK's. Figure 4.13(c) shows the daily evolution of the immigration load factor at IAD. The decrease in passengers after the travel ban measures led to a load factor that oscillates around an average of 26.7, which is three to four times lower than the usual load factor of this period. The drop is of 67% with the year 2019 and of 75% with the year 2018. Even though passengers arriving at immigration starting March 20$^{\text{th}}$ 2020 have at least three times more available open booths than in the previous year, the wait time for passengers at immigration did not improve, unlike for passengers arriving at JFK immigration. Figure 4.13(d) shows the daily evolution of the average wait time for passengers arriving at IAD immigration. The average wait time has increased throughout the travel ban measures and even reached the same level as during the previous years. It went from an average of 8.1 minutes in February 2020 to an average of 17 minutes in April 2020, compared to an average of 14.6 minutes in 2018 and of 26.3 minutes in 2019.

4.5 Discussion & Conclusion

4.5.1 Airline score summary

Figure 4.14 presents a radar plot for each of the eight considered airlines indicating their normalized scores.

The normalizations were conducted using the following formulas:

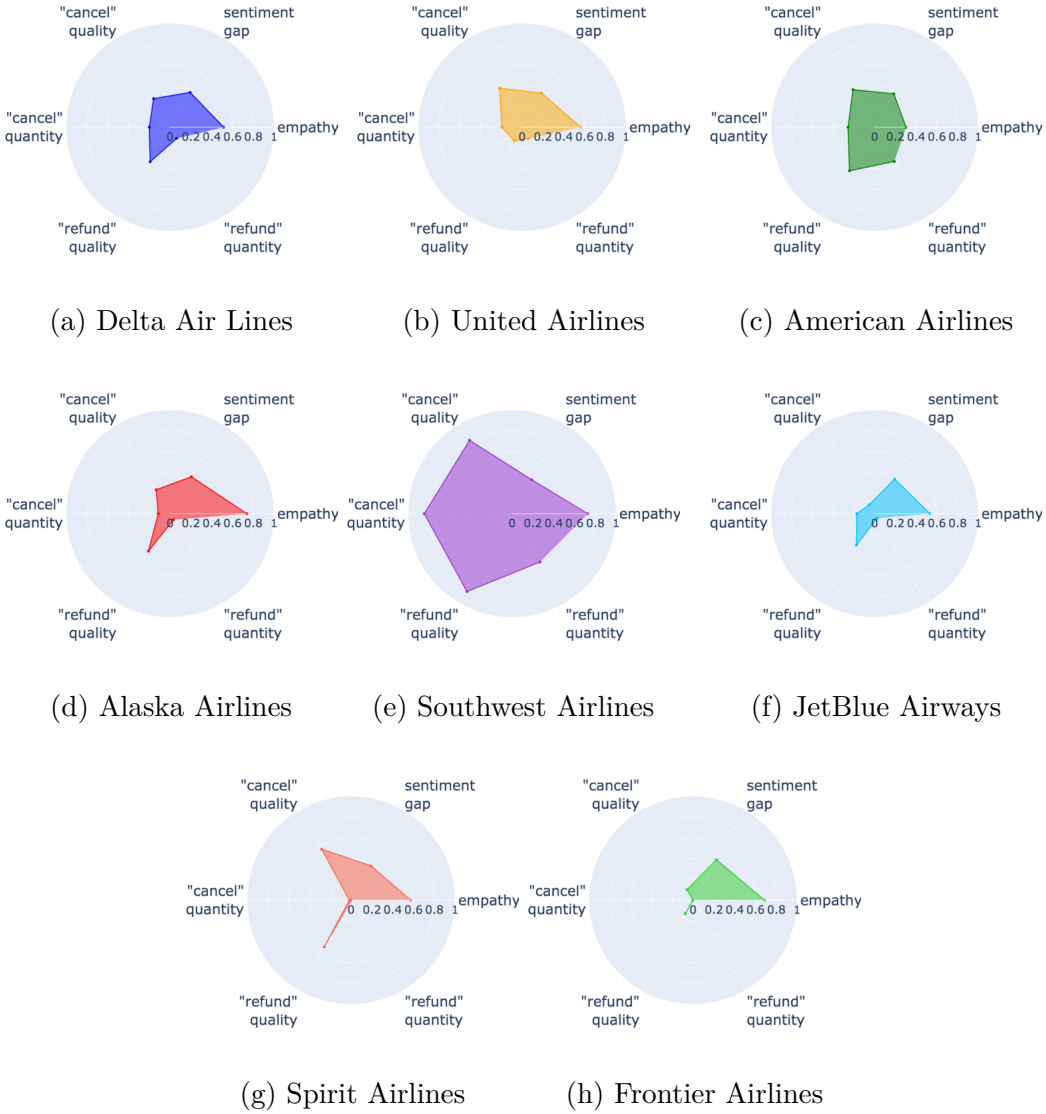
$$\hat{\Xi} = \frac{1 + \Xi}{2} \tag{4.10}$$

$$\hat{\Delta} = \frac{1 - \Delta}{2} \tag{4.11}$$

$$\hat{\kappa}^q_{\text{keyword}} = \frac{1 - \kappa^q_{\text{keyword}}}{\delta T} \tag{4.12}$$

$$\hat{\gamma}^q_{\text{keyword}} = \frac{\gamma^q_{\text{keyword}}}{\delta T} \tag{4.13}$$

where δT is the number of days of the full period over which the keyword-related Twitter situation response scores are calculated. All-but-one of the normalized scores go from the worst score of 0 to a good score of 1. The score can be greater than 1 in the case of a keyword-related Twitter situation response quantity score, but that scenario did not occur here. Regarding the normalized sentiment gap, a score of 0.5 indicates a normal score of 0, a normalized score of 0 indicates a score of 1 and a normalized score of 1 indicates a score of -1.



(a) Delta Air Lines (b) United Airlines (c) American Airlines

(d) Alaska Airlines (e) Southwest Airlines (f) JetBlue Airways

(g) Spirit Airlines (h) Frontier Airlines

Figure 4.14: Radar plots of the normalized scores associated to the proposed passenger-centric metrics for the eight airlines under consideration.

As can be seen in Figure 4.14, each airline has its own "Twitter profile". Passengers are then free to integrate these different profiles in their decision process for choosing the airline that corresponds the most to their travel needs and wants. Traditionally, the airline and airport choices are shown to be based on fare, access time and journey time [76, 168]. These studies do not take the airlines reputation among passengers as a decision parameter, and the proposed metrics could provide an additional decision layer for passengers.

For example, some risk-averse passengers could decide to opt for an airline that has better "refund"-related scores if they prefer a refund when flights are cancelled, rather than choosing an airline with a lower fare. Similarly, some passengers can consider that the flight experience is important in their airline decision and use the empathy and sentiment gap scores to help them decide which airline choose.

On the other hand, airlines can also compare their Twitter profiles provided in Figure 4.14 in order to improve their interactions with their passengers. For example, an airline with a clear description of their cancellation procedures on their website could use the "cancel" and "refund" related scores to verify if this information is actually easily accessible to passengers and if adequate communication is made on its availability. For example, a low "cancel" quality score would indicate that passengers already have access to the cancellation information.

4.5.2 Discussion

Several limitations and possible improvements should be noted here for a better understanding of the proposed metrics. The data used to estimate the number of visitors at airports and the proportion of time spent per airport location was graciously provided by SafeGraph in order to better understand the COVID-19 situation, and is not usually as easily available. In order to implement the associated metrics, agreements should be held between the different data providers and the group in charge of such metrics. Furthermore, an analysis of the categories of person most likely to be within the gathered data should be undertaken to better tune the final score.

The proposed passenger-centric metrics for airlines were built using Twitter data, which have the major advantage of being available in real-time, and can therefore be easily updated on an hourly basis if needed. Discussion between federal agencies, airlines and passengers should be undertaken in order to further tune the proposed metrics in order to meet the expectations of all concerned parties.

The proposed metrics based on data from Twitter have the added ben-

efit of enabling each passenger and airline to actively influence the scores. It should however be emphasized here that the metrics measure essentially the communication quality and quantity between airlines and passengers via Twitter, and should therefore still be complemented with traditional flight-centric measures for completeness.

This study focused on the effects of the travel restriction measures linked to a major disruption taking its course over an important number of days and tailored the proposed metrics for this timespan. Future studies could also investigate into the adaptation of some of these proposed passenger-centric metrics to measure effects on a smaller scale, e.g. over a single day or a few hours.

The metrics proposed in this chapter can be used to monitor the experience of passengers at airports or with airlines in order to help passengers be better informed in their choice of airline and airport when planning a trip by plane. However, these trips do not start at airports and a multi-modal approach is necessary during the planning phase. Hopefully passengers generate data throughout their multi-modal trip, and Chapter 5 presents how that data can be used to estimate full door-to-door travel times, which can then be used for a better trip planning.

Chapter 5

Estimating door-to-door travel times with the help of data generated by passengers

5.1 Introduction

The journey of a passenger of the air transportation system is not limited to the segment between two airports. Improving the passenger travel experience using door-to-door travel times as a possible metric is one of the ambitious goals set forward by NextGen and ACARE Flightpath 2050. Grimme and Martens [92] proposed a model to analyze the feasibility of the 4 hour goal within FlightPath 2050 based on airport to airport flight times and a simplified model of access and egress to airports. Sun et al. [93] implemented a door-to-door minimum travel time estimation based on open source maps and datasets in order to study the possible competitiveness of air taxis. The model and analysis presented in this chapter are also based on already available online data but with a post operation approach. Data generated by passengers throughout their door-to-door journey are essential to this model, and can be used in an aggregated format in order to respect passenger privacy. The aim here is to create a method based to measure the actual average door-to-door travel time once the trips are over enabling an analysis and comparison of the different possible transportation modes.

A first version of this method was presented in [14] and applied to two intra-European multi-modal trips comparing air to rail. It was then adapted and improved in [15] by leveraging four different data sources (road data, flight data, phone data and census data) in order to compare air trips between five different cities in the United States, three on the West Coast and two on the East Coast. Using recently released Uber data along with other online databases, a reliable estimation of door-to-door travel times is possible, which then enables a comparison of cities performance regarding the good integration of their airports as well as a per segment analysis of the full trip. This model can also be used to compare the reach and performance of different access modes to a city. It also enables a better evaluation of where progress should and can be made with respect to air passenger travel experience.

This chapter is organized as follows: Section 5.2 presents the model and data used to evaluate the full door-to-door journey time. Section 5.3 showcases several applications and comparisons enabled by this model for trips between Amsterdam and Paris, while Section 5.4 focuses on applications within the United States. Finally Section 5.5 concludes this chapter and discusses further research directions.

5.2 The full door-to-door data-driven model

In the specific cases of air and rail travels with no transfers and similarly to [94], [107] and [93], the full door-to-door travel time T can be decomposed into five different trip phases represented in Figure 5.1 and summarized in the following equation:

$$T = t_{\text{to}} + t_{\text{dep}} + t_{\text{in}} + t_{\text{arr}} + t_{\text{from}} \tag{5.1}$$

where

- t_{to} is the time spent traveling from the start of the journey to the departure station (e.g. train station or airport)
- t_{dep} is the time spent waiting and going through security processes (if any) at the departure station
- t_{in} is the time actually spent in flight or on rails
- t_{arr} is the time spent at the arrival station (e.g. going through security processes)
- t_{from} is the time spent traveling from the arrival station to the final destination

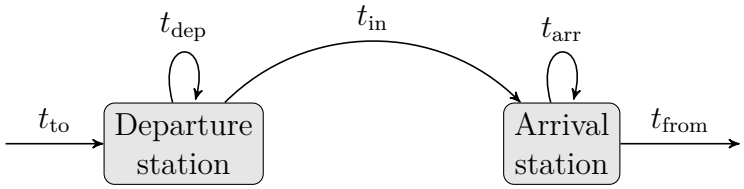


Figure 5.1: Model of the full door-to-door travel time.

Once the full door-to-door travel time model is defined as such, it is still necessary to be able to measure or at least estimate the values of these different times. This chapter proposes a framework aiming at their measurement and estimation based to a large extent on data generated by passengers throughout their journey. This study limits its scope to the two following main transportation modes: by air and by rail. Due to data availability, the case studies presented here only consider six major US cities (Atlanta, Boston, Los Angeles, Seattle, San Francisco and Washington D.C.) and two European capitals (Amsterdam and Paris).

5.2.1 Travel time from the origin location to the departure station and from the arrival station to the final destination

Uber [169] is a ride-sharing service launched in 2010 and implanted in major urban areas on six continents that has recently released anonymized and aggregated travel time data for certain of the urban areas it services. The available data consist of the average travel time, the minimum travel time and the maximum travel time between different zones (e.g. census tracts in the case of US cities) within the serviced area from all Uber rides aggregated over five different periods for each considered day. The five considered periods, which are used throughout this study, are defined as follows:

- Early Morning: from midnight to 7am
- AM: from 7am to 10am
- Midday: from 10am to 4pm
- PM: from 4pm to 7pm
- Late Evening: from 7pm to midnight

Depending on the availability of data, there are days when the travel times between some zones are only aggregated at a daily level. This data are generated by the mobile phones of their riders (via their ride-sharing application) and aggregated in a way that is respectful of their privacy.

Since Uber was initially introduced in the US, the impact of Uber in US urban transit has already been the focus of several studies prior to this data release. Li et al. [170] concluded that at an aggregated level Uber tends to decrease congestion in the US urban areas where it was introduced. Later Erhardt et al. [171] built a model showing that ride sharing companies did increase congestion using the example of San Francisco. Hall et al. [172] focused on whether Uber complemented or substituted public transit by studying the use of public transit system before and after Uber's entry date in different US cities. Wang and Mu [173] studied Uber's accessibility in Atlanta, US by using the average wait time for a ride as a proxy and concluded that the use of Uber was not associated to a specific social category. Since this data release, Pearson et al. [174] proposed a traffic flow model based on this aggregated Uber data and used it to analyze traffic patterns for seven cities world-wide. As Uber rides are part of the road traffic flow, this study considers that Uber's travel times are an accurate proxy of the actual travel times by road. In the case where there are no specific road lanes for bus routes, these travel times are a valid proxy for both car and bus trips. This chapter limits its scope to road access and egress to and from the considered stations.

The analysis of subway alternatives, by using schedules and real time data, is not considered in this chapter.

The data used for this study was gathered using Uber's Movement API[1]. Each US city was divided into their census tracts, Paris was divided into the IRIS zones used by INSEE [175] for census and Amsterdam into its official districts called *wijk*.

5.2.2 Dwell time at stations

The dwell time at a station, either t_{dep} or t_{arr} is defined as the time spent at the station, whether going through security processes, walking through the station or waiting. The time spent at each station depends on the mode considered, the specific trip and whether the passenger is boarding or unboarding. This dwell time at departure can be split into two components: a processing time t_{sec} necessary to get through security (if any) and through the station to the desired gate or track and an extra wait time t_{wait} due to unanticipated delays.

Processing times at US airports are based on the median wait times at airports presented in the study of [106] that are extracted from data generated by the mobile phone of passengers. The US airports considered in this study are the six following airports: Hartsfield-Jackson Atlanta International Airport (ATL), Boston's Logan International Airport (BOS) and Ronald Reagan Washington National Airport (DCA) for the East Coast, Los Angeles International Airport (LAX), Seattle-Tacoma International Airport (SEA) and San Francisco International Airport (SFO) for the West Coast. For the three considered European airports, i.e. Paris Charles de Gaulle Airport (CDG), Paris Orly Airport (ORY) and Amsterdam Airport Schiphol (AMS), this processing time is constant over the airports and determined using most airline's recommendation.

The average dwell times at these airports are summarized in Table 5.1 for US airports and in Table 5.2 for European airports:

Regarding processing times at train stations, based on the recommendation of the train station websites, the departure dwell time is fixed at 15 minutes and the arrival dwell time is fixed at 10 minutes for all train stations. These estimates could be improved by gathering data from GPS or mobile phone sources as well as WiFi beacons within airports and train stations and by using a method similar to the passenger flow study at Sydney International Airport by Nikoue et al. [71].

[1] `movement.uber.com`

Table 5.1: Average dwell time spent at US airports in minutes.

	ATL	**BOS**	**DCA**	**LAX**	**SEA**	**SFO**
Time at departure	110	105	100	125	105	105
Time at arrival	60	40	35	65	50	45

Table 5.2: Average dwell time spent at European airports in minutes.

	AMS	**CDG**	**ORY**
Time at departure	90	90	90
Time at arrival	45	45	45

The extra wait times can be computed when the scheduled and real departure or arrival times are available. For US airports, these wait times are calculated only for departure using the publicly available data from the Bureau of Transportation Statistics (BTS) [2]. They were obtained by subtracting the scheduled departure time from the actual flight departure time. This study assumed that there was no extra wait time at arrival.

5.2.3 Time in flight or on rail

US flights

The actual flight time was calculated based on the data from BTS using the actual departure/arrival times of all direct flights between each city pairs from January 1st 2018 to March 31st 2018. Cancelled flights are not considered in this study and were discarded. Future studies should take into account airline policies with respect to cancellations in order to estimate the impact of a cancelled flight on a passenger's full door-to-door travel time.

European trips

Since there are no centralized flight schedule data in Europe, it is assumed that flights and trains are on time and follow a weekly schedule. Future studies should consider scraping actual flight and train times, in order to take into account delays and perturbations.

5.2.4 Full door-to-door time

This model assumes that travelers plan their departure time to arrive at the departure station exactly t_{sec} minutes before the scheduled departure time of their flight or train. This assumption is used in the determination of t_{to}

since it defines uniquely the period of the day to consider when extracting the Uber average time from the origin location to the departure station. The value of t_{from} is extracted using the actual arrival time of the flight or train. When only daily aggregated times are available in the Uber data, these times are used for each period of the day as a proxy.

5.3 Flights versus trains: a comparison of different access modes to Paris

This section considers the case study of a traveler leaving from Amsterdam city center and willing to reach the Paris area. Three possible means of transportation are under study in this case: travelers can either travel by plane either via Paris Charles De Gaulle airport (CDG) or via Paris Orly (ORY), or they can travel by train via Paris Gare du Nord (GDN).

5.3.1 Flight and train schedules

As assumed previously in Section 5.2.3, the flight and train schedules considered in this study were extrapolated based on a simulated weekly schedule. For the flights between Amsterdam Airport Schiphol (AMS) and CDG or ORY, the weekly schedules are based on the actual flight schedules during the two months of December 2019 and January 2020. These schedules are summarized in Table 5.3 for flights between AMS and ORY, and in Table 5.4 for flights between AMS and CDG. The weekly train schedule between Amsterdam Centraal station and Paris Gare du Nord is similarly based on the actual train schedule of the year 2019 and is summarized in Table 5.5. Night trains were not considered for this study.

Table 5.3: Simulated weekly schedule from Amsterdam to Paris via ORY.

Mo	**Tu**	**We**	**Th**	**Fr**	**Sa**	**Su**	**Ams.**	**Paris**
x	x	x	x				10:25	11:45
				x			14:45	16:05
x	x	x	x	x			18:50	20:10
						x	19:40	21:00

These schedules already contains the major differences between the three considered modes: Flying through ORY is the option with the fewest possibilities with at most two daily flights through ORY compared to an almost hourly schedule for the other two modes. Another notable difference can

Table 5.4: Simulated weekly schedule from Amsterdam to Paris via CDG.

Mo	**Tu**	**We**	**Th**	**Fr**	**Sa**	**Su**	**Ams.**	**Paris**
x	x	x	x	x	x	x	06:50	08:10
x	x	x	x	x	x	x	07:20	08:45
x	x	x	x	x	x	x	08:10	09:35
x	x	x	x	x	x	x	09:30	10:55
x	x	x	x	x	x	x	10:20	11:45
x	x	x	x	x	x	x	12:25	13:40
x	x	x	x	x	x	x	13:55	15:15
x	x	x	x	x	x	x	14:50	16:10
x	x	x	x	x	x	x	16:35	17:50
x	x	x	x	x	x	x	17:45	19:00
x	x	x	x	x		x	19:10	20:30
x	x	x	x	x	x	x	20:25	21:45

be seen with respect to the station-to-station travel times: flights between Amsterdam and Paris (both CDG and ORY) take 1h20 (± 5 minutes) while train rides between Amsterdam and Paris GDN take 3h20 (± 3 minutes).

These weekly schedules were used to generate flight and train schedules from January 1st 2018 to September 30th 2019. For each flight and each train of these expanded schedules, the full door-to-door travel time is estimated assuming a departure from Amsterdam city center. All available zones within the Paris area are considered as potential final arrival zones.

5.3.2 Average total travel time mode comparison

A first use of this door-to-door model is to give a means of evaluating and comparing the range of each considered mode, helping to better understand the urban structure and behavior from a transportation point of view. The same daily periods as those used in the Uber data (see Section 5.2.1) are considered here to regroup the trips into five groups depending on the time of arrival at the final destination. For each day and each period, the mean per arrival zone of the average door-to-door travel time was calculated for each mode and the mode with the best mean was kept. It is then possible to count over the twenty-one month period how many times a mode has been the best during each daily period for each zone. This distribution of modes over the different zones can help travelers to choose which mode to favor depending on the desired arrival zone and on the desired time of arrival. It can also help urban planners to better understand the road network linking

Table 5.5: Simulated weekly schedule from Amsterdam to Paris via GDN.

Mo	**Tu**	**We**	**Th**	**Fr**	**Sa**	**Su**	**Ams.**	**Paris**
x	x	x	x	x		x	06:15	09:35
x	x	x	x	x	x	x	07:15	10:38
x	x	x	x	x	x	x	08:15	11:35
x	x	x	x	x	x		09:15	12:35
					x	x	10:15	13:38
x	x	x	x	x	x	x	11:15	14:35
x	x	x	x	x	x	x	13:15	16:38
x	x	x	x	x		x	14:15	17:35
x	x	x	x	x	x	x	15:15	18:35
x	x	x	x	x		x	16:15	19:35
x	x	x	x	x		x	17:15	20:35
x	x	x	x	x	x	x	18:15	21:38
x	x	x	x	x		x	19:15	22:35
						x	20:15	23:38

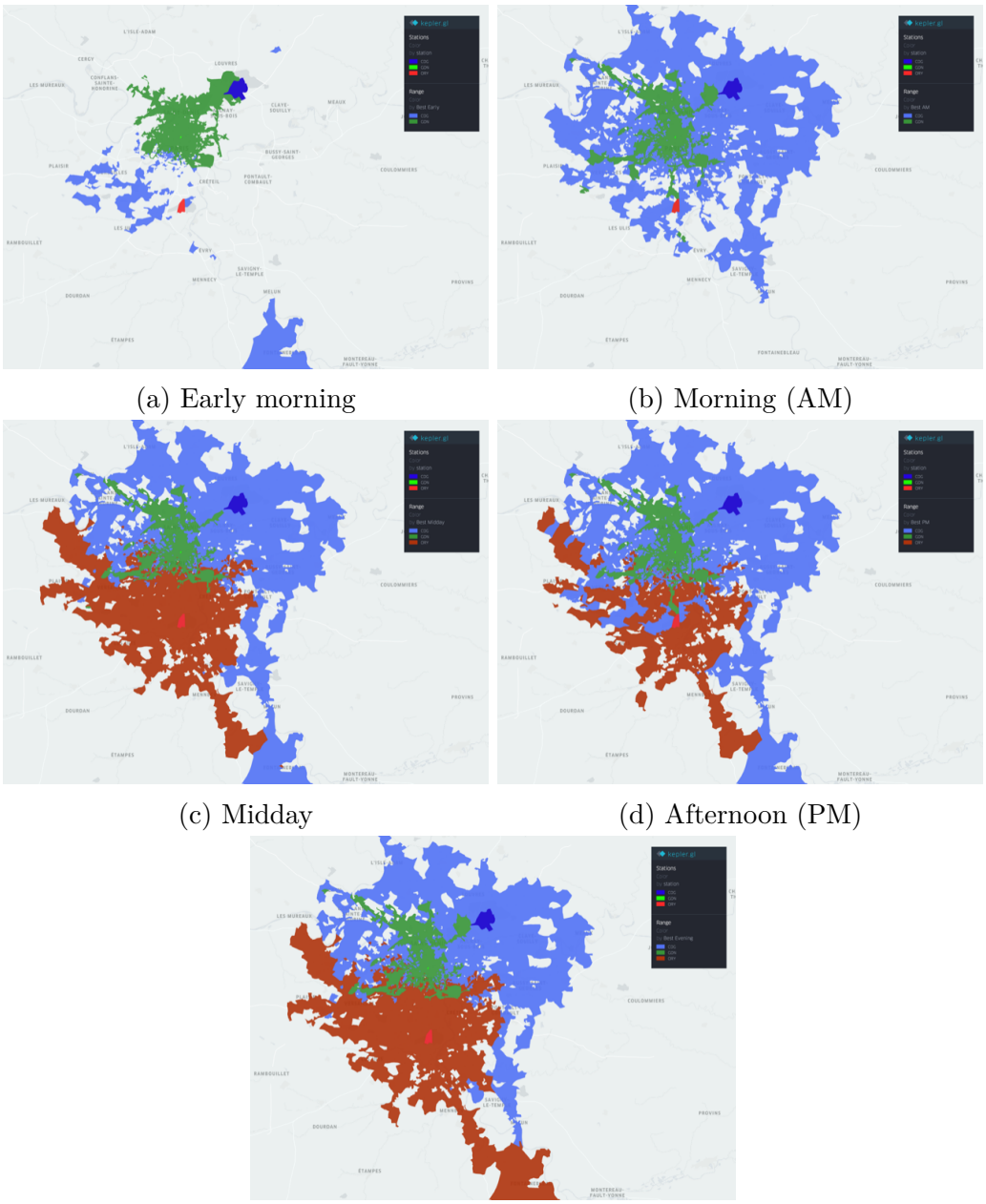
the different stations to the city.

Figure 5.2 shows the fastest mode to reach the different zones in the Paris dataset for the five different periods of the days used by the Uber dataset. For each zone and each period, the fastest mode associated is the mode having the highest number of days with the lowest average total travel time over the considered date range. The zones best reached through CDG are indicated in blue, ORY in red and GDN in green.

Several conclusions can be drawn from these maps. Looking at all five maps, the absence of zones reached through ORY (in red) is particularly noticeable in the morning periods (both early morning and AM) with an important chunk of South-West Paris not being reached by Uber rides neither from GDN nor from CDG. These maps would advocate for an increase in frequency for the AMS-ORY flights from a traveler perspective.

Focusing on the early morning map (Figure 5.2(a)), a non intuitive fact appearing is that it is on average faster to reach zones closer to CDG by taking the train through GDN. As a matter of fact, the only zones where it is not better to take the train in the early morning are zones situated on the opposite side of Paris from CDG. However, the associated flights are the ones landing at 21:45 the previous day, whereas the zones reached by train are associated with early morning trains.

From a structural perspective, the highway linking Paris to CDG is visible on all five maps since it enables travelers through GDN to reach zones



(a) Early morning

(b) Morning (AM)

(c) Midday

(d) Afternoon (PM)

(e) Evening

Figure 5.2: Comparison of the average total travel times to the Paris area between the three considered arrival stations (CDG: blue, ORY: red, GDN: green) for a trip starting from Amsterdam city center for different trip initiation periods.

close to CDG faster than if they flew to CDG directly. The *Boulevard Périphérique* circling Paris is also a major aid to GDN and is visible on the maps where the competition between GDN and CDG is fierce. The section of the *Boulevard Périphérique* farthest from GDN (i.e. in the south-west between *Porte de Versailles* and *Porte de Gentilly*) is however overtaken by either airport depending on the period of the day. The rest of GDN influence zone is fairly invariant from a period to another.

The range of ORY is limited during the afternoon (Figure 5.2(d)), with CDG taking over some zones close to ORY. This is essentially due to the limited number of flights landing in the afternoon (one per week, every Friday) compared to the daily arrival of CDG flights.

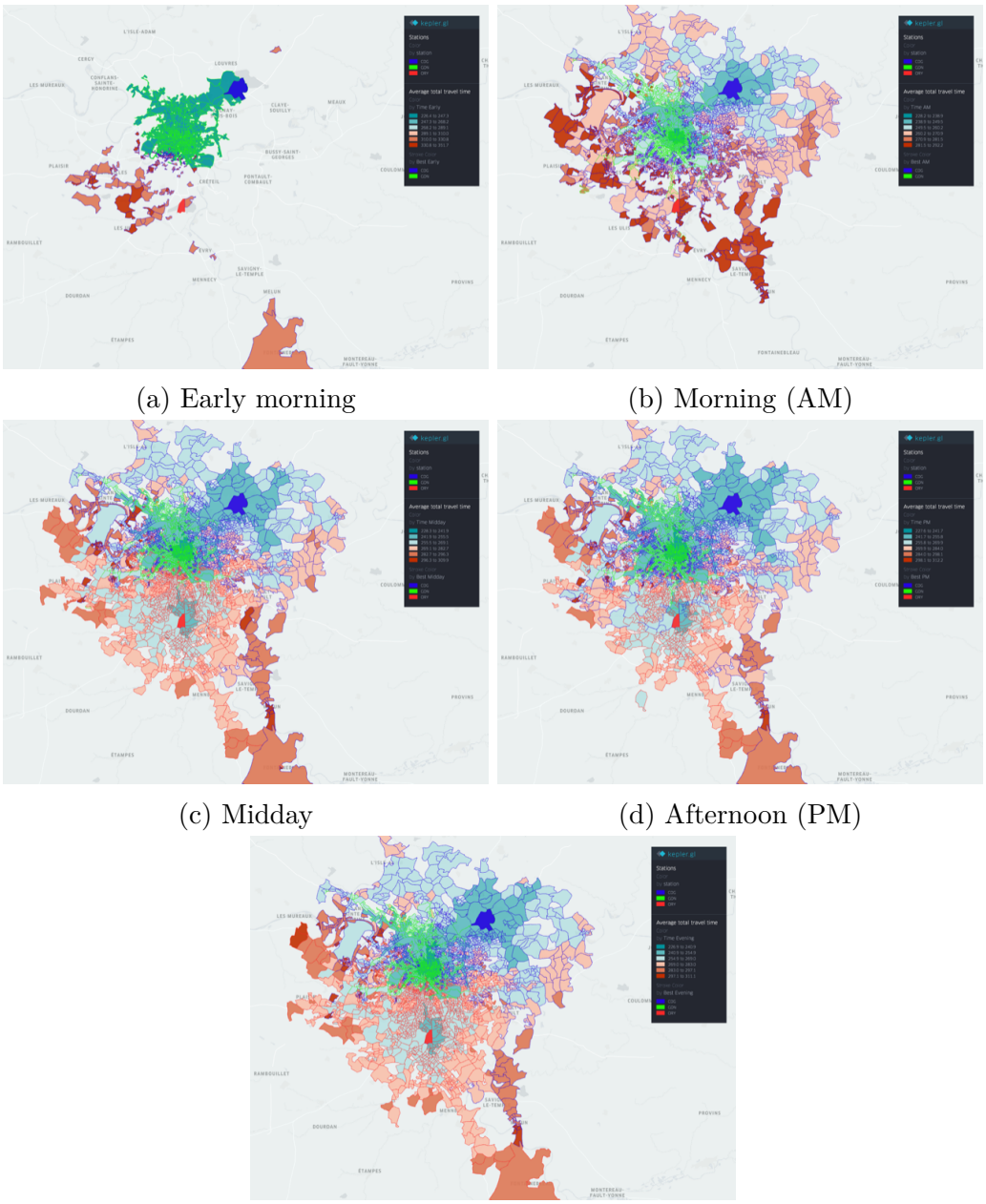
5.3.3 Average total travel time distribution analysis

Once the best mode to reach each zone is determined, it is possible to analyze their associated full door-to-door travel times. This approach gives an overview of the level of integration of airports, train stations, and road structure and can indicate zones which are less reachable than others and would thus require more attention from urban planners.

Figure 5.3 displays the average full door-to-door travel time to reach the different zones in the Paris dataset for the five different period of the days used by the Uber dataset based on the analysis presented in Section 5.3.2. The period is again determined using the arrival time of the full door-to-door trip. The contour of each zone indicates the mode used to reach it using the same color code, i.e. the zones best reached through CDG are indicated in blue, ORY in red and GDN in green. Though the time color scales are different from one map to another, the first green scale represents trips lasting less than four hours ($\pm$ two minutes) with the exception of trips ending in early mornings (Figure 5.3(a)) where the first green scale represents trips lasting less than 4 hours and 7 minutes. For a better comparison, the distribution of the number of zones per period reached within four time intervals (less than 4h00, between 4h00 and 4h30, between 4h30 and 5h00, and more than 5h00) is presented in Table 5.7.

From both Figure 5.3 and Table 5.7 several differences can be noted concerning the reach of each mode. Zones for which taking the train is the best option are undertaken in less than 4h30 for 99% of them, while they represent only 66% of the zones best reached through CDG and 69% of those through ORY. Zones reachable in less than 4 hours are best reached by train through GDN for 98% of them, the last 2% being reached only via CDG, a trip through ORY always requiring more than 4 hours under this model.

Focusing now on the relative integration of the two airports within the



(a) Early morning

(b) Morning (AM)

(c) Midday

(d) Afternoon (PM)

(e) Evening

Figure 5.3: Comparison of the average total travel times to the Paris area between the three considered arrival stations (CDG, ORY, GDN) for a trip starting from Amsterdam city center for different trip initiation periods. The contour color of each zone indicates the best mode to reach it.

Table 5.6: Color code per period of the day for the average full door-to-door travel times presented in Figure 5.3.

Early	AM	Midday	PM	Late
3h46-4h07	3h48-3h58	3h48-4h01	3h47-4h01	3h46-4h00
4h07-4h28	3h58-4h09	4h01-4h15	4h01-4h15	4h00-4h14
4h28-4h49	4h09-4h20	4h15-4h29	4h15-4h29	4h14-4h29
4h49-5h10	4h20-4h30	4h29-4h42	4h29-4h44	4h29-4h43
5h10-5h30	4h30-4h41	4h42-4h56	4h44-4h58	4h43-4h57
5h30-5h51	4h41-4h52	4h56-5h09	4h58-5h12	4h57-5h11

Table 5.7: Number of zones per mode and period of the day grouped by full door-to-door travel time intervals. The original dataset is the same as that used to generate Figure 5.3.

Mode	Time interval	Early	AM	Midday	PM	Late
CDG	$t \leq$ 4h	0	4	6	5	11
	4h $< t \leq$ 4h30	22	1306	866	1189	845
	4h30 $< t \leq$ 5h	0	653	433	498	384
	$t >$ 5h	187	0	15	13	11
GDN	$t \leq$ 4h	398	247	247	290	314
	4h $< t \leq$ 4h30	775	818	731	719	641
	4h30 $< t \leq$ 5h	0	14	8	6	8
	$t >$ 5h	0	0	0	0	0
ORY	$t \leq$ 4h	0	0	0	0	0
	4h $< t \leq$ 4h30	0	0	906	563	997
	4h30 $< t \leq$ 5h	0	0	397	259	425
	$t >$ 5h	0	0	0	0	1

Parisian road structure, one can notice that the range of the first two scales of green surrounding the airports is larger for CDG than for ORY for all three periods where they are both active, both in surface and in number of zones. This indicates that CDG has a better road egress structure than ORY, and that improvements should be considered for ORY to be more competitive with respect to CDG.

Looking at the combination of all three modes, one is forced to notice that there is a major dissymmetry between the north and the south of Paris in terms of access times from Amsterdam. It is possible to find a path between Paris city center and CDG going only through zones from the first two green scales, while there is a discontinuity between Paris city center and ORY. Despite being at proximity of an airport (ORY), most zones south-east of

Paris are not as easily reached as other zones further away from airports and from the train station.

5.3.4 Safest total travel time

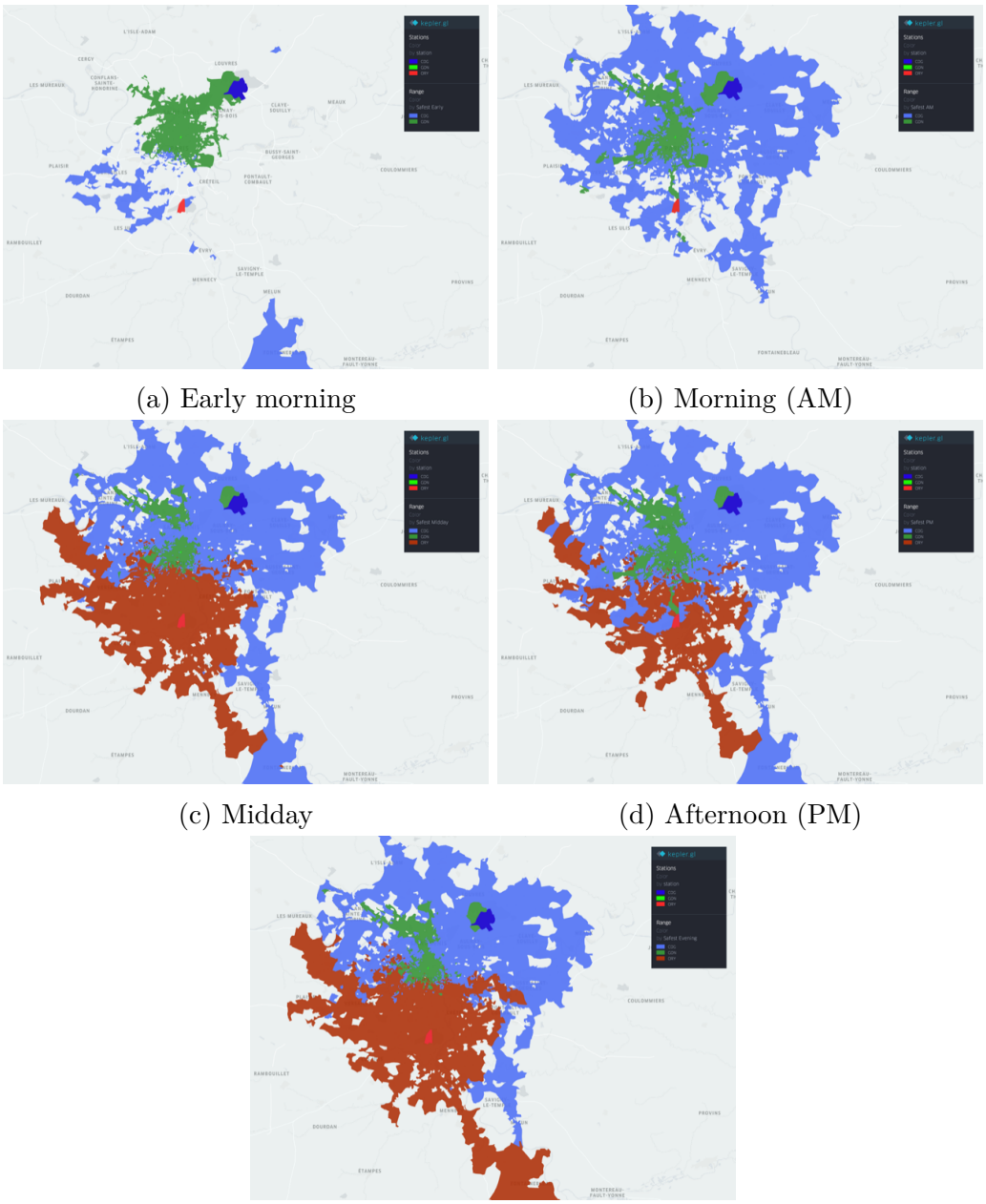
This full door-to-door travel time model assumes that passengers choose their departure time in order to arrive exactly t_{sec} before the scheduled departure of their plane or train and that they also know how long it takes to reach the departure station. However, in reality, there is an uncertainty in the time the traveler will spend reaching the airport and in the airport processing times. This uncertainty often leads to an additional buffer time implying an earlier departure time for the traveler. Using the presented model with the available data, it is possible to find which is the most reliable mode to use per arrival zone. The most reliable mode for a given arrival zone is defined as the mode with the lowest variability in travel time, i.e. the mode where the difference between the maximum travel time and the minimum travel time to reach that zone is the lowest. This comparison is useful for passengers or trips that require an accurate arrival time rather than a minimum travel time.

Figure 5.4 shows the most reliable mode on average to reach the different zones in the Paris dataset for the five different period of the days used by the Uber dataset. As for the previous analysis, the period was determined using the departure time of the full door-to-door trip and uses the same color code, i.e. the zones reached most reliably through CDG are indicated in blue, ORY in red and GDN in green. For each zone and each period of the day, the most reliable mode associated is the mode having the highest number of days with the lowest average variability travel time over the considered date range.

Though Figure 5.4 and Figure 5.2 are similar, there are some major differences between average efficiency and average reliability noticeable in these maps. For example, though it is on average faster to reach by train the zones close to the highway leading to CDG, after 10:00 it is safer from a time variability perspective to reach them via CDG. From a reliability perspective, CDG has claimed the quasi totality of the zones surrounding it, except in the early morning where trips through GDN are still better. When comparing all three modes, it appears that GDN is the most adversely affected by this metric, with its range smaller than when considering rapidity.

5.3.5 Impact of faster processing times

One of the major difference between air and rail travel is the necessary processing time both at departure and at arrival. In this particular study, with



(a) Early morning

(b) Morning (AM)

(c) Midday

(d) Afternoon (PM)

(e) Evening

Figure 5.4: Comparison of the average variability of travel times to the Paris area between the three considered arrival stations (CDG: blue, ORY: red, GDN: green) for a trip starting from Amsterdam city center for different trip initiation periods.

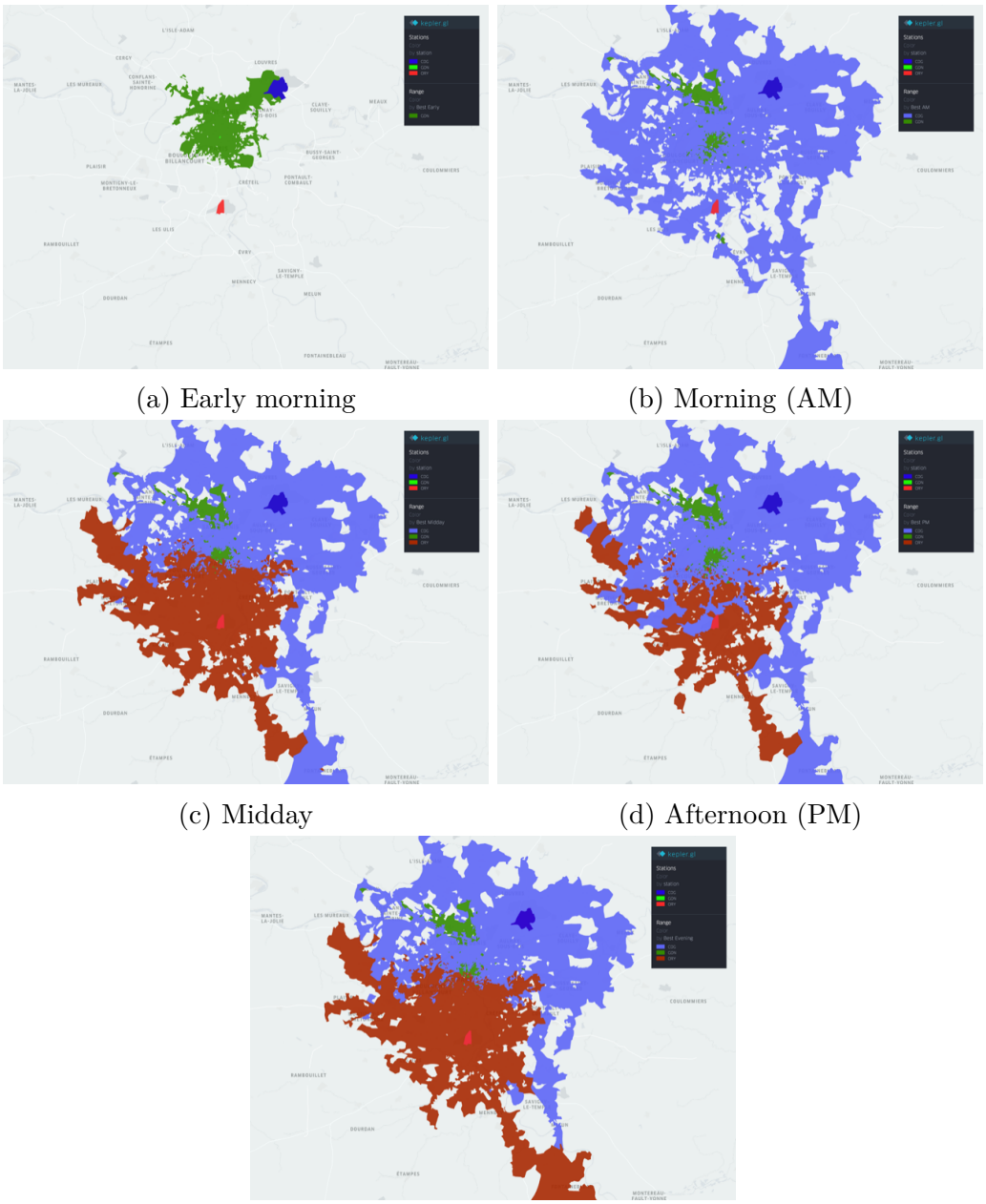
a flight time of about 80 minutes, the current assumption of a departure processing time of 90 minutes implies that travelers spend more time at their departure airport than in flight, which greatly impacts the rapidity of air travel. The presented model allows to modify these assumed processing times in order to study the impact of improving these times both from an airport perspective and a passenger perspective. Let's assume in the following analysis that the processing time at airports is improved from 90 to 60 minutes at departure and from 45 to 30 minutes at arrival. These modifications could be achieved in reality considering that this is an intra-Schengen trip and that there isn't any border control.

Figure 5.5 shows which is the fastest mode on average to reach the different zones in the Paris dataset for the five different period of the days used by the Uber dataset. As for the previous analysis, the period was determined using the arrival time of the full door-to-door trip. The range of each mode is indicated with the contour of each zone using the same color code, i.e. the zones reached most reliably through CDG are indicated in blue, ORY in red and GDN in green. For each zone and each period, the fastest time associated is the average travel time using the fastest mode determined in Section 5.3.2.

Table 5.8: Number of zones per mode and period of the day grouped by full door-to-door travel time intervals in the case of faster airport processing times. The original dataset is the same as that used to generate Figure 5.5.

Mode	**Time interval**	**Early**	**AM**	**Midday**	**PM**	**Late**
CDG	$t \leq$ 4h	0	1921	1492	2146	1525
	4h $< t \leq$ 4h30	0	762	180	133	113
	4h30 $< t \leq$ 5h	0	3	1	0	0
	$t >$ 5h	0	0	0	0	0
GDN	$t \leq$ 4h	398	289	318	384	263
	4h $< t \leq$ 4h30	797	132	49	41	33
	4h30 $< t \leq$ 5h	0	7	6	6	7
	$t >$ 5h	0	0	0	0	0
ORY	$t \leq$ 4h	0	0	1514	819	1656
	4h $< t \leq$ 4h30	0	0	49	13	40
	4h30 $< t \leq$ 5h	0	0	0	0	0
	$t >$ 5h	0	0	0	0	0

The first major difference with this processing time improvement can be seen for trips arriving in the early morning (Figure 5.5(a)): all zones previously reached through CDG are no longer accessed at this period since



(a) Early morning

(b) Morning (AM)

(c) Midday

(d) Afternoon (PM)

(e) Evening

Figure 5.5: Comparison of the average total travel times to the Paris area assuming faster airport processing times between the three considered arrival stations (CDG: blue, ORY: red, GDN: green) for a trip starting from Amsterdam city center for different trip initiation periods.

they were associated to the 21:45 flight of the previous day. This indicates that all trips start and end on the same day, with no trips finishing after midnight. Another conclusion is that if a traveler from Amsterdam needs to reach Paris in the early morning, the fastest way on average is to take the train through GDN.

Looking at trips arriving later in the morning (Figure 5.5(b)), trips through CDG are greatly advantaged by this time improvement, with CDG taking over more than half of GDN previous influence zone. This range increase from CDG can be explained both by faster door-to-door travel times and by the increase of trips through CDG arriving in the morning (rather than at midday). As a matter of fact, besides in the early morning, GDN loses its competitiveness against both airports, with its range greatly shrinking in size. The competition between CDG and ORY is however unchanged, which is understandable since they both received the same processing time improvement.

From a time perspective, Table 5.8 summarizes the distribution of the number of zones per period reached within the same four time intervals as previously (i.e. less than 4h, between 4h and 4h30, between 4h30 and 5h, and more than 5h). This table shows that all trips are now conducted in less than five hours and that 99.8% of the zones reachable are reached in less than 4h30. ORY sees some major improvements with now 97.5% of the zones best reached through it reached in less than four hours (compared to no trips in less than 4h in the initial model), while increasing the number of zones it reaches the fastest.

Using a map representation similar to Section 5.3.3, but not presented here due to space considerations, it is possible to notice a 20-30 minutes shift in the time distribution for every period except for early morning trips since train processing times were unchanged. The upper bound travel time is also unchanged for trips arriving in the morning, which would indicate that for some zones, the processing time improvement resulted in no improvement or even a worsening of the full trip travel time. Besides that exception, it is to be noted that in this case a 45 minutes improvement in airport processing time leads only to a maximum of 30 minutes of average total travel time improvement due to the influence of train trips through GDN.

5.4 A multi-modal analysis of the US air transportation system

This section presents additional insights that can be gained from this full door-to-door travel thanks to the availability of complementary data. The United States is a federal state the size of a continent, therefore various aggregated and centralized datasets are more easily available to all. Several of these datasets are used in this section to add applications to the presented full door-to-door model. This US study limits itself to the period from January 1st 2018 to March 31st 2018.

5.4.1 Flight schedule

As presented in the model definition in Section 5.2.3, both the scheduled flight times and the actual flight schedules of most domestic flights can be obtained via the Bureau of Transportation Statistics (BTS) [2]. This study considers only the six US airports presented in Section 5.2.2, three East-coast airports - Hartsfield-Jackson Atlanta International Airport (ATL), Boston's Logan International Airport (BOS) and Ronald Reagan Washington National Airport (DCA) - and three West-coast airports - Los Angeles International Airport (LAX), Seattle-Tacoma International Airport (SEA) and San Francisco International Airport (SFO).

The average number of direct flights for the five considered day periods are presented in Table 5.9. This table only counts flights that were not cancelled from January 1st 2018 to March 31st 2018. During this three-month period, 38,826 flights were considered, which corresponds to 3,523 early flights, 8,170 morning flights, 13,451 midday flights, 6.695 afternoon flights and 6,987 evening flights.

The full door-to-door travel times were then calculated for each scheduled flight from January 1st 2018 to March 31st 2018 using the model presented in Section 5.2.

5.4.2 Leg analysis

The full door-to-door travel times are initially calculated for every census tract pair with sufficient Uber data between the census tract and the corresponding airport. This yields an important number of travel times for each considered flight. A method to aggregate these travel times into one travel time per city pair would be to weigh the travel time associated to each census tract with the proportion of passengers initiating their trips from there, or

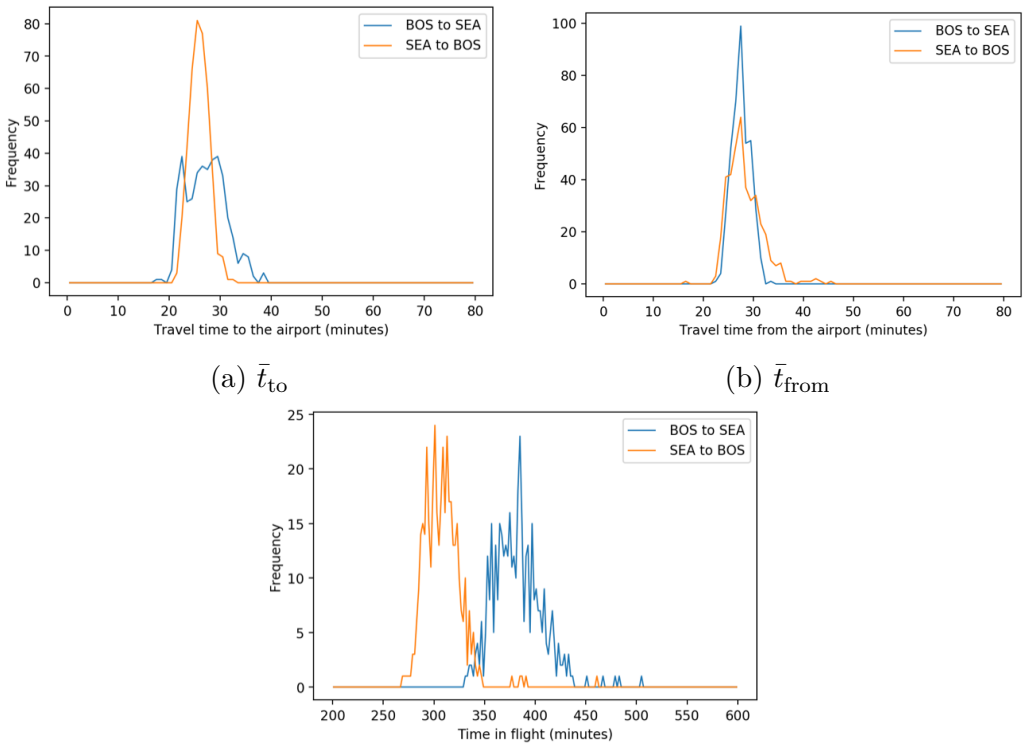
Table 5.9: Average number of flights for each period of the day between the considered US city pairs between January 1st 2018 and March 31st 2018.

Flight leg	**Early**	**AM**	**Midday**	**PM**	**Late**
ATL - BOS	1.33	10.42	22.74	8.23	11.06
ATL - DCA	2.65	12.84	18.90	8.97	13.00
ATL - LAX	2.74	9.87	24.61	6.10	7.13
ATL - SEA	1.00	3.97	7.90	3.74	2.90
ATL - SFO	0.00	8.35	7.26	2.90	7.13
BOS - ATL	7.94	9.97	19.52	13.06	2.77
BOS - DCA	3.71	13.55	22.19	11.19	8.74
BOS - LAX	2.07	9.87	8.13	12.42	2.35
BOS - SEA	1.08	6.55	0.00	3.16	2.42
BOS - SFO	5.35	11.29	7.94	5.65	5.00
DCA - ATL	9.84	5.48	20.26	11.00	9.77
DCA - BOS	3.55	9.16	24.13	12.03	10.65
DCA - LAX	0.00	5.68	0.00	5.74	0.00
DCA - SEA	0.00	2.84	0.00	2.84	0.00
DCA - SFO	0.00	2.90	0.00	2.84	0.00
LAX - ATL	2.55	16.19	21.32	2.81	7.55
LAX - BOS	2.94	10.68	8.42	2.15	10.74
LAX - DCA	0.00	5.74	5.74	0.00	0.00
LAX - SEA	6.65	10.39	24.55	13.97	14.32
LAX - SFO	7.68	18.65	44.19	23.65	22.61
SEA - ATL	5.81	2.94	6.84	0.00	3.77
SEA - BOS	2.52	1.33	4.48	0.00	4.94
SEA - DCA	0.00	2.84	2.87	0.00	0.00
SEA - LAX	12.74	11.32	20.42	11.77	13.48
SEA - SFO	9.39	11.35	25.00	9.23	14.55
SFO - ATL	2.94	4.06	12.77	0.00	5.87
SFO - BOS	1.04	8.16	13.55	4.81	7.94
SFO - DCA	0.00	2.84	2.87	0.00	0.00
SFO - LAX	10.45	26.61	35.52	23.68	20.87
SFO - SEA	8.94	8.00	21.77	14.32	15.81

finishing their trip there. The distribution of passengers over the different census tracts could be estimated using mobile phone data. This data are not available for this study, therefore the number of passengers originating from or finishing within a census tract is assumed to be proportional to the

population density of the considered census tract. The information relative to the US census tracts are obtained from an online database[2] based on the US government 2010 census. The proposed aggregation method leads to a single value for t_{to} and t_{from} per flight, which will be denoted as $\bar{t}_{\text{to}}$ and $\bar{t}_{\text{from}}$, and thus a single full door-to-door travel time $\bar{T}$ per flight.

Once aggregated the full door-to-door travel time model enables a more condensed leg-by-leg comparison of trips between two cities. As an example, the city pairs (Boston, Seattle) and (Seattle, San Francisco) are compared using this approach.



(a) $\bar{t}_{\text{to}}$ (b) $\bar{t}_{\text{from}}$

(c) t_{in}

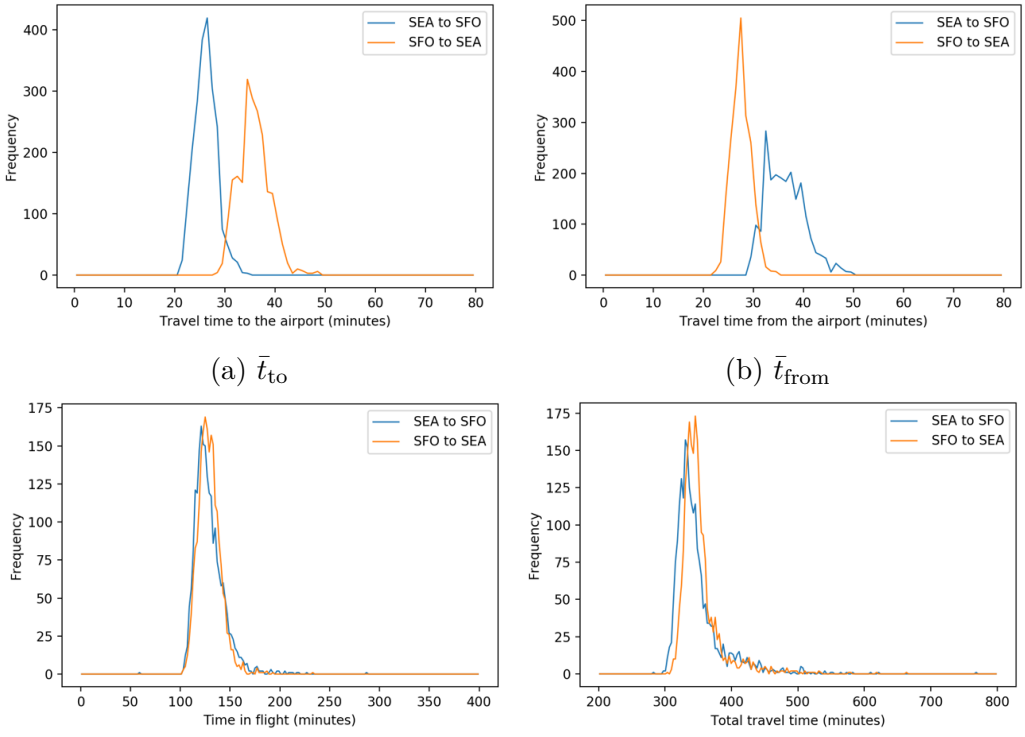
Figure 5.6: Histograms of the aggregated time spent going to ($\bar{t}_{\text{to}}$) and from ($\bar{t}_{\text{from}}$) the airports as well as the time spent in flight for both ways of the journey Boston - Seattle, from January 2018 to March 2018

Starting with the city pair (Boston, Seattle), Figure 5.6 shows the histograms of the aggregated time spent going to ($\bar{t}_{\text{to}}$) and from ($\bar{t}_{\text{from}}$) the airports as well as the time spent in flight for both ways between January 2018 and March 2018. Figures 5.6(a) & 5.6(b) show that the quasi-totality of

[2] `www.usboundary.com`

the weighted egress or access time distributions are under 30 minutes for both airports. This indicates that both cities have integrated their airports in a similar fashion. Assuming the processing times presented in Table 5.1, the flight time is the major difference between travelling from Boston to Seattle or the other way round. This difference is essentially due to an important West-East wind on the chosen flight paths.

The second example city pair (Seattle, San Francisco), with both cities in the same timezone, leads to a different conclusion. Figure 5.7 shows the histograms of the aggregated time spent going to ($\bar{t}_{\text{to}}$) and from ($\bar{t}_{\text{from}}$) the airports as well as the time spent in flight for both ways between January 2018 and March 2018.



(a) $\bar{t}_{\text{to}}$ (b) $\bar{t}_{\text{from}}$

(c) t_{in} (d) $\bar{T}$

Figure 5.7: Histograms of the aggregated time spent going to ($\bar{t}_{\text{to}}$) and from ($\bar{t}_{\text{from}}$) the airports as well as the time spent in flight and the aggregated full door-to-door times $\bar{T}$ for both ways of the journey San Francisco - Seattle, from January 1$^{\text{st}}$ 2018 to March 31$^{\text{st}}$ 2018.

Unlike the flight time distributions of the previous city pair (Figure 5.6(c)), Figure 5.7(c) shows that for these two West-coast cities that the flight time distributions are similar for both flight directions. Figure 5.7(a) shows that

for a majority of the considered flights linking SEA to SFO the weighted time $\bar{t}_{\text{to}}$ to reach SEA from Seattle is 30 minutes or less (with an average of 26 minutes), whereas for a majority of the considered flights linking SFO to SEA the weighted time $\bar{t}_{\text{to}}$ from San Francisco to SFO is greater than 30 minutes (with an average of 36 minutes). The same conclusions apply to the weighted times $\bar{t}_{\text{from}}$ to leave each airport. Figure 5.7(b) shows that reaching San Francisco from SFO takes also more than 30 minutes (with an average of 35 minutes) while reaching Seattle from SEA takes less than 30 minutes (with an average of 27.5 minutes). Figures 5.7(a) & 5.7(b) clearly state that SEA is better integrated to Seattle than SFO is to San Francisco. Figure 5.7(d) shows the histogram of the weighted total times $\bar{T}$ for the city pair (Seattle, San Francisco) over the considered period. The slight shift of four minutes of the distributions in favor of the direction San Francisco - Seattle is essentially due to the fact that the processing time at arrival t_{arr} is five minutes faster at SEA than at SFO (Table 5.1). If SFO were to have a similar processing time at arrival, the other direction would be slightly faster.

These histogram plots are useful to gain insight on each specific leg of the full trip and to more easily compare each leg between different city pair trips. Another representation leads to a better understanding of the time spent in each leg proportionally to the time spent on the overall trip. For each trip, the percentage of time spent at each phase is calculated based on the full door-to-door travel time. The average percentage time spent is then calculated for each phase and for each city pair trip. Figure 5.8 shows the bar plot of these average percentage times for the thirty considered city pairs. The city pairs are sorted according to the percentage of time spent in the actual flight phase. With the proposed full door-to-door model, for all considered trips, passengers spend on average more time at the departure airport than riding to and from the airports. This figure also shows that, with this model, for some short-haul flights, such as between SFO and LAX or between BOS and DCA, passengers spend on average more time at the departure airport than in the plane. Refining the full door-to-door model by considering tailored airport processing times t_{sec} at departure depending on the city pair and not only on the departure airport could lead to a different conclusion. This modification of the model would however require a stronger access to passenger data.

The proposed full door-to-door model combined with the census data available enables a better comparison of the integration of each airport within its metropolitan area. To each census tract is associated an internal point within its boundaries, and this internal point can be used to automatically calculate the distance between airports and each census tract of their metropolitan area. Figure 5.9 shows the scatter plot of the average daily ride

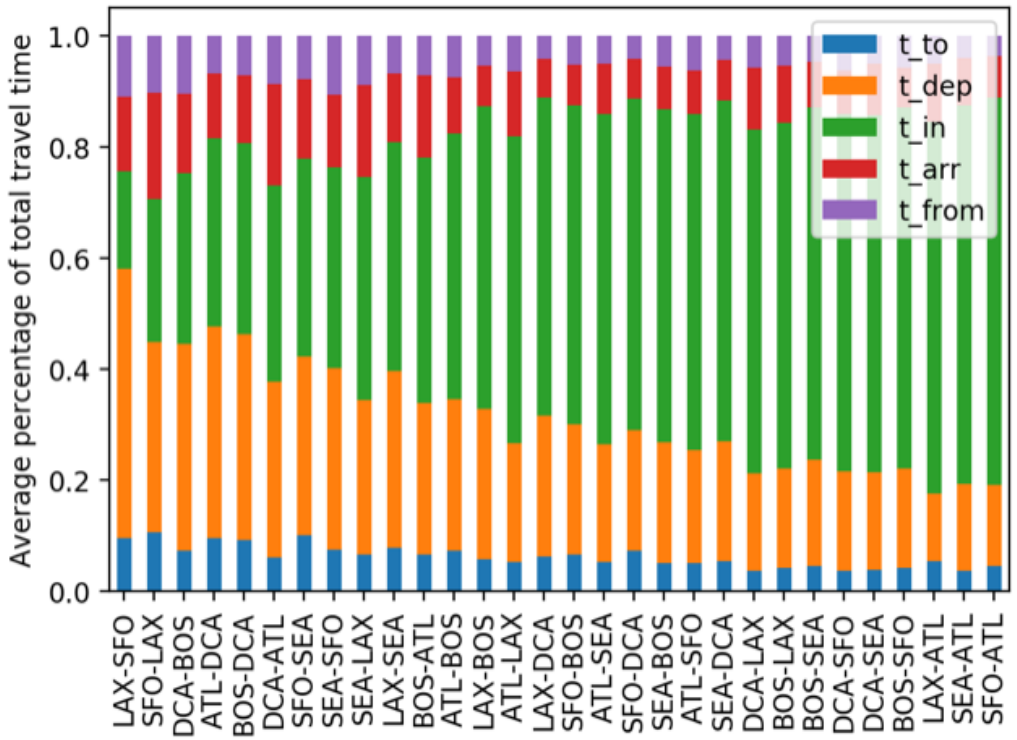


Figure 5.8: Bar plot of the average proportion of the time spent within each phase of the full door-to-door journey for all thirty considered trips.

time to each airport versus the geodesic distance to the airport for the six considered airport. The geodesic distance is the shortest distance between two points along the surface of the Earth. Additionally, the plot also figures a linear regression of these average time with respect to the distance to the airport. A steeper slope for the linear regression indicates that it takes longer to reach the airport from a given distance. Figure 5.9 highlights the disparity between the range of each airport within the available data: DCA has a range limited to 20 km while SFO attracts Uber riders from more than 120 km away. The other four airports have a similar range. The difference in slope of their associated linear regression is however useful to rank their integration within their region of attraction. From this perspective, Seattle has the best integrated airport, i.e. the smallest slope, followed by Atlanta, Boston and then Los Angeles.

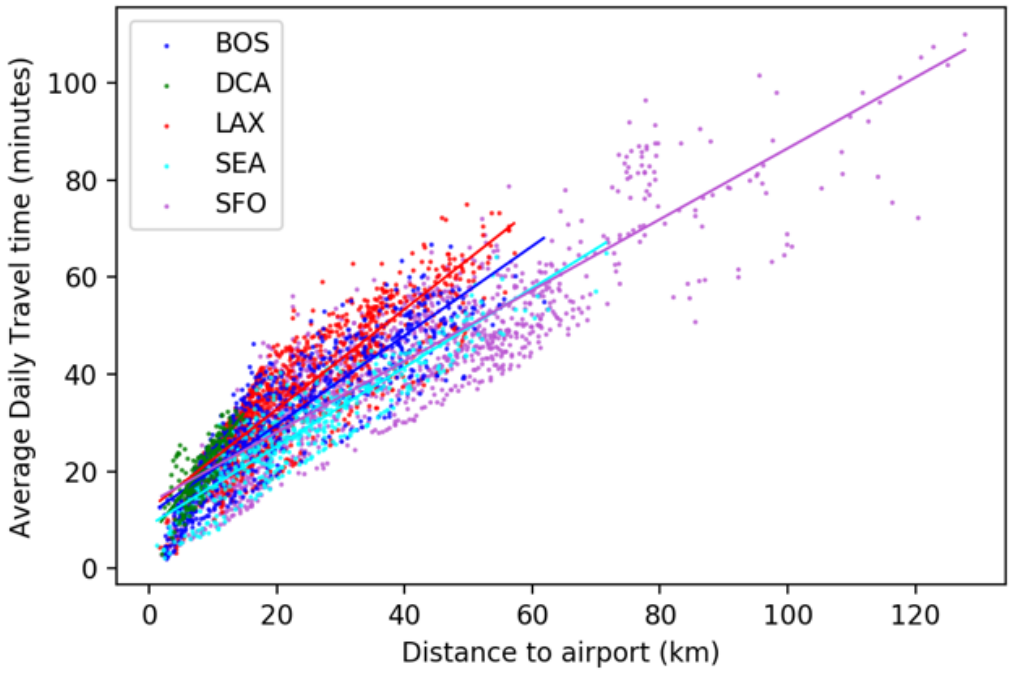


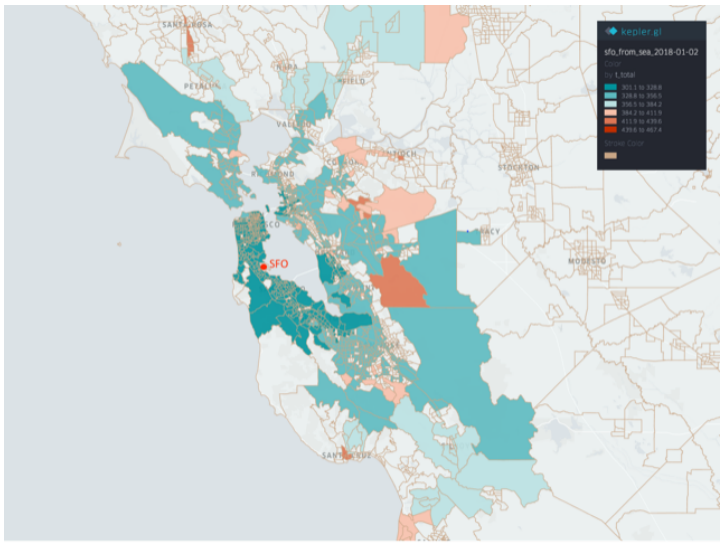
Figure 5.9: Scatter plot of the average ride time to the airport t_{to} versus the distance to the airport from January 1st 2018 to March 31st 2018. Straight lines indicate the linear regression fit for each city.

5.4.3 Reach analysis

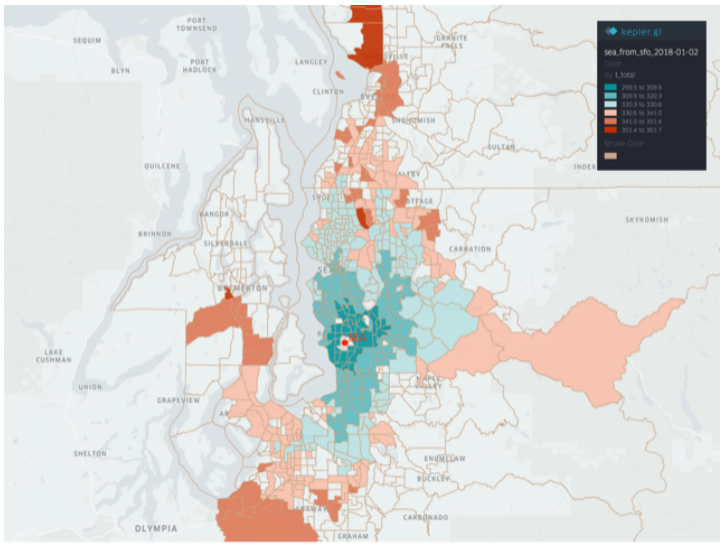
Similarly to the Parisian modal comparison, this full door-to-door model can also be used in the US to visualize the accessible range of a city starting from a specific census tract along with the time necessary to reach each possible census tract in the destination city. The resulting map can help in better understanding the urban structure of the metropolitan area or the impact of severe weather on the full door-to-door trip and not only on the flight segment.

Comparison of airport integration within the metropolitan road structure

Figure 5.10 shows the average full door-to-door travel times associated with trips in both directions between Seattle and San Francisco on January 2$^{\text{nd}}$ 2018. For each direction, the origin census tract is the census tract containing the city hall. Please note that the color scale representing the full door-to-door travel times are different from one map to another. The full color scale



(a) From Seattle to San Francisco (minimum travel time: 299 minutes, maximum travel time: 362 minutes)



(b) From San Francisco to Seattle (minimum travel time: 301 minutes, maximum travel time: 467 minutes)

Figure 5.10: Average door-to-door travel times for trips between the city pair (Seattle, San Francisco) starting from their city halls on January 2nd 2018. The color scale is different from one map to another.

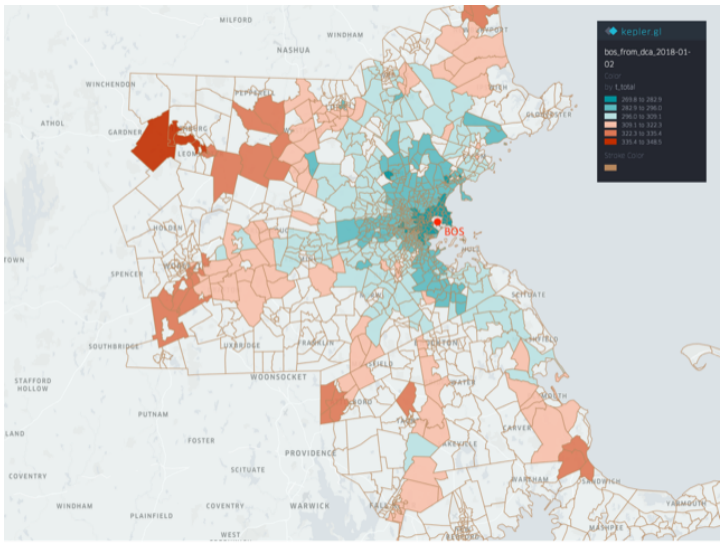
for Figure 5.10(b) ranges from 299 minutes to 362 minutes and is almost completely contained within the first two color levels of Figure 5.10(a), which ranges from 301 minutes to 356 minutes. The full color scale of Figure 5.10(a)

ranges from 301 minutes to 467 minutes. For both cities, there are two main axis of propagation visible thanks to the quasi linear time expansion from the airport. For San Francisco, Figure 5.10(a) shows two axis on each side of the Bay, except for a limited number of zones. These zones are associated with census tracts close to parks and with less housing and fewer roads. The linear propagation in time on both sides of the Bay can be explained by the presence of numerous highways (e.g. I-280 and I-880) on both sides. For Seattle, Figure 5.10(b) shows two perpendicular axis of propagation, one North-South and one East-West. The North-South axis has a longer range and faster propagation than the East-West axis. This can be explained by the presence of the highway I-5 and could suggest the need of an improved East-West road access.

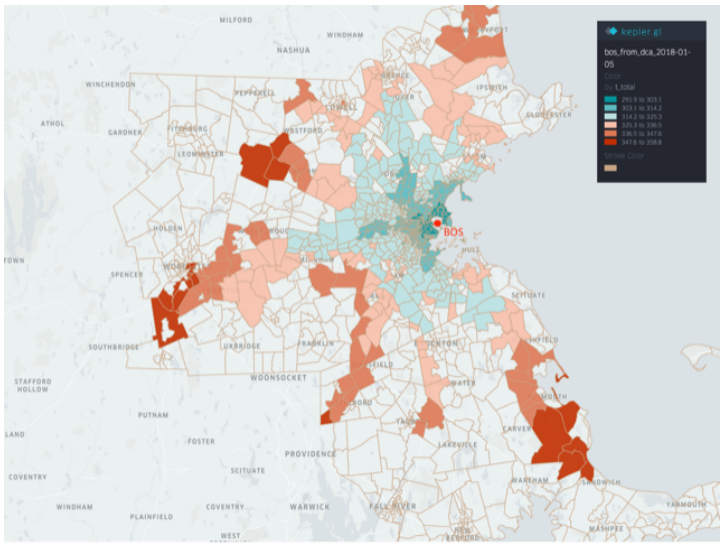
Impact of severe weather analysis

Using the same door-to-door travel time visualization process and applying it to different days can be a tool to better analyze the effects of severe weather perturbations on the full door-to-door journey. As an example, the winter storm previously studied in [10] is analyzed for trips between Washington D.C. and Boston. This winter storm hit the East Coast of the United States on January 4$^{\text{th}}$ 2018, and led to the closure of two airports in New York City, along with the cancellation of the majority of flights flying to or from the North-Eastern US coast. Figure 5.11 shows the map of the average full door-to-door travel times to reach the Boston area starting from Washington D.C. city hall on January 2$^{\text{nd}}$ 2018 - before the landfall of this winter storm - and on January 5$^{\text{th}}$ 2018 - after the landfall of the winter storm.

Please note once more that the color scales representing the full door-to-door travel time are different from one map to another. This difference indicates that on January 5$^{\text{th}}$ 2018 (Figure 5.11(b)), the minimum average full door-to-door travel time to reach any census tract within the Boston area was more than twenty minutes higher than on January 2$^{\text{nd}}$ 2018 (Figure 5.11(a)) from 269 minutes up to 291 minutes. The maximum average full door-to-door travel time was increased by ten minutes, from 348 minutes to 358 minutes. Comparing the two maps, a shift towards the red is noticeable from January 2$^{\text{nd}}$ 2018 to January 5þ 2018, along with some census tracts disappearing from the considered range on January 5$^{\text{th}}$ 2018 due to lack of sufficient Uber ride data. These two observations indicate that the full door-to-door travel times are closer to the maximum average travel time than from the minimum travel time on January 5$^{\text{th}}$ 2018 compared to January 2$^{\text{nd}}$ 2018 and that some zones were might have been sufficiently adversely impacted by the weather to allow rides from the airport to reach them. On



(a) Before landfall, on January 2nd, 2018



(b) After landfall, on January 5th, 2018

Figure 5.11: Average door-to-door travel times from Washington D.C. city hall to Boston over a single day, before and after the Bomb Cyclone of January 2018.

a side note, both maps show that Washington D.C. has road different structure than both Seattle and San Francisco, with a star-shaped propagation of travel times from its airport.

5.4.4 On the importance of a passenger-centric approach to delays

A final application to the full door-to-door model presented in this chapter is to emphasize the difference between flight delay and passenger delay. Since Uber splits the day into five different periods, each with their traffic idiosyncrasies with respect to peak times, it is possible to calculate how much extra travel time is required for a passenger when a flight does not arrive in the scheduled period. For example, a flight supposed to arrive in the early morning that lands after 10:00 AM could result in the passenger getting stranded in traffic when trying to leave the airport. Though airlines are not responsible for road traffic, passengers can choose flights based on their arrival time to avoid peak time traffic.

To calculate this extra travel at aggregated level, the difference of average travel time between the two periods concerned by flights not arriving according to schedule is calculated for each arrival zone and then aggregated using the method presented in Section 5.4.2. Another measure of sensitivity is to consider the maximum difference between the maximum travel times of each zone between the two considered periods. This second measure indicates the worst variation of the travel time upper bound, i.e. the maximum difference a constantly unlucky rider can experience going from the airport to their final destination zone.

Let us consider the flight UA460 from LAX to SFO scheduled to arrive on Thursday February 15, 2018 at 18:02 local time and that landed with a minor delay of 16 minutes. Due to the 45 minutes processing time required to leave the airport, this 16 minutes delay shifts the departure period from the airport from afternoon (PM) to late evening. The aggregated average extra travel time from the airport is of 15 minutes and 40 seconds, i.e. a 16 minutes flight delay triggered an average 31 minutes total delay for the passengers. Looking at the second considered measure, the maximum travel time difference for this flight delay is of 72 minutes, meaning that potentially one passenger could experience a total delay of 88 minutes resulting from this 16 minutes delay experienced by the flight. This first example illustrates that passenger delay and aircraft delay are distinct and cannot be substituted.

Paradoxically, arriving earlier than scheduled for a flight does not necessarily mean that the full door-to-door trip ends earlier. For example, flight VX1929 from LAX to SFO scheduled to arrive on Thursday February 8, 2018 at 15:22 local time actually landed 25 minutes earlier. This implied that the passengers were no longer leaving the airport in the afternoon (PM) period but at midday. The aggregated average extra travel time from the airport is here of 15 minutes and 2 seconds, so on average travelers did arrive ear-

lier than scheduled, but only by about ten minutes and not the twenty-five minutes announced by the airline. However, looking at the second measurement method again, the maximum ride time difference is of 66 minutes and 44 seconds, which means that potentially a passenger could end up arriving forty minutes later than if the flight would have landed on time.

5.5 Conclusion

By leveraging Uber's recently released data of the aggregated travel times of their passengers and integrating them with several other available data sources, this chapter introduces a model of the full door-to-door travel time for multi-modal trips both in Europe and in the United States. Though the model is used for one city pair in Europe and five different cities in the United States, it could however be implemented for any world city pairs with available ride-sharing or taxi data. This model can be adapted depending on the available data regarding the main modes considered, since a weekly schedule with no delay information can lead to some meaningful insights for passengers and city planners alike.

Furthermore, by aggregating the full door-to-door travel times at a city level, the model enables both the pairwise comparison of the different travel times per trip segment between two cities as well as an analysis over time of the time necessary to join two specific cities. It can also be used to evaluate on a national level some passenger-centric objectives within NextGen in the US and ACARE in Europe regarding the good integration of airports within their cities. It can also bring some insights to how multi-modal trips are affected by severe weather perturbations, indicating where improvements can be made. It also brings a valuable measurement of the difference between flight delays and passenger delays, emphasizing the need of passenger-centric metrics for evaluating the performance of the air transportation system, which is not solely constituted of planes.

Further studies should consider using alternative modes to reach the departure station or leave from the arrival station such as the subway. Additionally, knowing the actual daily proportion of travelers using the different approaches (road or rail) would enable a better daily evaluation of the full door-to-door travel time. A possible method to determine this proportion would be by using aggregated information from GPS or mobile phone sources.

Chapter 6

Discussion and conclusion

6.1 Conclusion

Over the past ten years, the number of flights every year has steadily increased to meet the demand of passengers, which are then considered as customers of the air transportation system. The performance of the air transportation system is traditionally measured by flight-centric metrics, such as the number of delayed flights, the amount of delay and the number of cancelled flights. Previous studies have shown that passengers are a vector of flight delay propagation, along with planes and crew. This makes passengers actors of the air transportation system. The work presented in this thesis harnesses the fact that passengers generate data throughout their travel and interaction with the air transportation system to also consider passengers as sensors of the air transportation system.

In Chapter 3 we have shown that data actively generated by passengers on Twitter can be filtered and processed into features that can be used to estimate the number of abnormal flights per airport and per hour in the United States. This estimation pipeline enables any actor of the air transportation system to have a good view of what is currently happening within the system, without having to wait for official flight data to be released. This real time availability is of particular importance when considering severe perturbations, both for short-term perturbations, such as the January 2018 bomb cyclone, and for long-term perturbations, such as the COVID-19 health crisis.

During the COVID-19 health crisis in Spring 2020, passengers were not traveling as much as before, but they remained customers, actors and sensors of the air transportation system. In Chapter 4, we proposed metrics for passengers based on passenger-generated data in order to monitor the

interaction between passengers and airlines and the interaction between passengers and airports. The data used for these metrics are available faster than official flight data, from the real time availability of Twitter data to a week latency for SafeGraph mobile data, versus a two month latency for BTS reports. Therefore, these metrics give a partial but up-to-date view of the air transportation system which can then be completed with flight data once they are released.

Considering passengers as sensors of the air transportation system has the added benefit of offering the possibility to capture the inherent multi-modality of the air transportation system, given that passengers generate data throughout their door-to-door journey. This specificity leads to the full door-to-door travel time model presented in Chapter 5. This model can be used to gain a better understanding of the urban network surrounding airports, which can then be used to improve airport integration with cities for the benefit of passengers, airports and airlines. It can also be used to compare the different possible transportation modes between two cities in order to help passengers choose the best option for them depending on their travel preferences.

In conclusion, passenger-generated data can be used to estimate in real time the flight-centric status of the air transportation system, while also giving a complementary view of the system via the interactions between passengers and the other actors of the system. Combined with other data sources, passenger-generated data can also provide an estimation of the actual door-to-door travel time for multi-modal trips.

6.2 Perspectives

The work presented in this thesis and in the appendix is a first step towards putting the passenger back to the center of the air transportation system. It has therefore covered many facets of the air transportation system in order to highlight the central role that passengers play within the air transportation system. Several research directions presented in this thesis can therefore be further investigated.

The estimator built in Chapter 3 is trained using only data from 2017 and used to estimate cancellations up to May 2020. A possible improvement to the proposed estimator would be to integrate new flight data and perturbations within a continuous training process. This regular update could also potentially lead to the integration within the extracted features of the rapidly changing behavior of the Twitter stream. Furthermore, the estimator could be adapted to estimate the number of flights arriving with a delay greater

than 30 minutes or 60 minutes, since these magnitudes of delay have a far greater impact on the passenger experience.

The passenger-centered metrics presented in Chapter 4 are tailored for long term perturbations in order to answer to the lack of fight information during the COVID-19 health crisis. An interesting research direction would be to adapt those metrics to short term perturbations, which would require to consider the data generated by passengers at an hourly level rather than at a daily or weekly level.

Twitter being used worldwide, an adaptation of the proposed models and metrics based on the data generated by passengers on Twitter to other regions of the world, such as the European Union, could provide more insights on the regional interconnections of the air transportation system.

Finally, the model of the full door-to-door travel time presented in Chapter 5 could be enhanced by integrating additional transportation modes to and from the airports, such as subway and public transportation, and by gaining access to aggregated dwell times by airport and by train station thanks to an increased data-sharing. Furthermore, a better-tuned model of the processing and dwell time spent at airports could also be derived from data gathered by some Airport Operations Centre (APOC) initiatives, e.g. at Heathrow Airport. The model could then be used to better assess which trip phases can be most easily compressed in order to reach the 4 hour aim of ACARE FlightPath 2050. Any optimization model of the full door-to-door travel time should also take into account the fact that passengers generate revenue for airports during their dwell time [176].

The ubiquity of smartphones and the increase in the use of sensors as part of the Internet of Things is leading to massive amounts of data being collected by various entities, both public and private, within the air transportation system. The work in this thesis shows that these data can be shared in such a way that it protects the passengers privacy while benefiting every actor of the air transportation system.

Furthermore, being able to access real time or close-to real time information is essential for the handling of unforeseen events and can benefit all stakeholders of the air transportation system, including passengers. Air transport operations are highly optimized, increasingly dynamic, especially during uncertain times or degraded situations. The system-wide availability of better information about the system's state, which is comprised of airplanes, flights and passengers, is bound to improve the response to unplanned events

6.3 Contributions

This section lists the different papers submitted and/or published within this PhD. There are eight papers accepted at international conferences, including one paper later accepted for a journal special issue. There are two additional papers submitted to journals and one paper published on arXiv.org. This last paper was published online without peer-review in order for it to be published before the official release of flight-related data by BTS.

6.3.1 Conferences

[10] Marzuoli, A., Monmousseau, P., Feron, E., 2018. *Passenger-centric metrics for Air Transportation leveraging mobile phone and Twitter data*, in: Data-Driven Intelligent Transportation Workshop - IEEE International Conference on Data Mining 2018.

Presented at the IEEE International Conference on Data Mining 2018, Singapore.

[11] Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D., 2019. *Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources.*

Presented at the Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019), Vienna, Austria.

[16] Monmousseau, P., Marzuoli, A., Bosson, C., Feron, E., Delahaye, D., 2019. *Doorway to the United States: An Exploration of Customs and Border Protection Data.*

Presented at the 38$^{\text{th}}$ Digital Avionics Systems Conference, San Diego, California, USA.

Best session paper award.

[12] Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D., 2019. *Passengers on social media: A real-time estimator of the state of the US air transportation system.*

Presented at the ENRI Int. Workshop on ATM/CNS (EIWAC 2019), Tokyo, Japan.

Best student award.

[14] Monmousseau, P., Delahaye, D., Marzuoli, A., Feron, E., 2019. *Door-to-door travel time analysis from Paris to London and Amsterdam using Uber data.*

Presented at the Ninth SESAR Innovation Days, Athens, Greece.

[17] Monmousseau, P., Jarry, G., Bertosio, F., Delahaye, D., Houalla, M., 2020. *Predicting Passenger Flow at Charles De Gaulle Airport Security Checkpoints*

Presented at the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), IEEE, Singapore, Singapore.

[15] Monmousseau, P., Delahaye, D., Marzuoli, A., Feron, E., 2020. *Door-to-door Air Travel Time Analysis in the United States using Uber Data.*

Presented at the 2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT), IEEE, Singapore, Singapore.

[18] Monmousseau, P., Puechmorel, S., Delahaye, D., Marzuoli, A., Feron, E., 2020. *Towards a more complete view of air transportation performance combining on-time performance and passenger sentiment.*

Accepted at the 9th International Conference on Research in Air Transportation (ICRAT '20), Tampa, Florida, US. Not presented due to COVID-19.

6.3.2 Papers submitted to journals

- Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D., 2020. *Passengers on social media: A real-time estimator of the state of the US air transportation system.*

  Accepted at the Air Traffic Management and Systems – IV, Lecture Notes in Electrical Engineering, Springer

- Monmousseau, P., Marzuoli, A., Feron, E. and Delahaye, D., 2020. *Impact of Covid-19 on passengers and airlines from passenger measurements: Managing customer satisfaction while putting the US Air Transportation System to sleep.*

  Published in Transportation Research Interdisciplinary Perspectives, Volume 7, September 2020.

- Monmousseau, P., Marzuoli, A., Feron, E. and Delahaye, D., 2020. *Analyzing and comparing door-to-door travel times for air travel using aggregated Uber data.*

Submitted at IEEE Transactions on Intelligent Transportation Systems.

6.3.3 Papers published on arXiv.org

[13] Monmousseau, P., Marzuoli, A., Feron, E., Delahaye, D., 2020. *Putting the Air Transportation System to sleep: a passenger perspective measured by passenger-generated data.* arXiv:2004.14372 [physics].

Thank you for reading so far!

Appendix A

A first case study using passenger-generated data: The January 2018 bomb cyclone viewed from mobile phone and social media data

A.1 Introduction

This thesis is highly inspired by the approach taken by Marzuoli et al. in [106], and the path to the methods and results presented throughout this thesis started in a study of the passenger experience in airports under major perturbations combining mobile phone data and social media data [10].

This appendix aims at presenting this study, which is a detailed analysis of domestic air passengers behavior during a major air-traffic disturbance, from two complementary passenger-centric perspective: a passenger mobility perspective and a passenger social media perspective. By leveraging over 5 billion records of mobile phone location data per day from a major carrier in the United States, passenger mobility can be reliably analyzed, no matter which airline the passengers fly on or which airport they fly to and from. Such information is currently unavailable to the major aviation stakeholders at such scale and can be used to establish performance benchmarks from a passenger's perspective. Combining it with a Twitter analysis provides a more detailed and passenger-focused analysis than the traditional flight-centric measurements used to evaluate the overall system performance. More generally, these two passenger-centric analysis could be implemented in real-time for a daily evaluation of the Air Transportation System, enabling a faster analysis of the impact of major disruptions, whether due to meteorological conditions or system failures.

These tools are here implemented and tested *a posteriori* in the case of the bomb cyclone that hit the Northeast part of the United States in January 2018, causing the closure of Kennedy International Airport (JFK) and severe capacity decreases at Logan International Airport (BOS), Newark Liberty International Airport (EWR) and LaGuardia Airport (LGA).

The appendix is organized as follows. Section A.2 describes the bomb cyclone and its impact on flight operations, leveraging publicly available on-time performance data from the Bureau of Transportation Statistics. Section A.3 offers a passenger-centric perspective in this appendix, focused on passenger mobility, supported by mobile phone cell-tower location data from a major US carrier. Section A.4 provides a second passenger-centric perspective, focused on passenger travel experience, using publicly available Twitter data. Section A.5 draws the conclusions of the study and provides future research perspectives.

A.2 The Bomb Cyclone and its impact on Air Operations

From January 2nd to January 6th 2018, a massive blizzard nicknamed "Bomb Cyclone" disrupted the Eastern Coast of the United States with a peak on January 4th. More than 90 percent of LGA flights, more than 70 percent of Newark Liberty flights and 20 percent of JFK flights were announced to be cancelled on January 4th. Both JFK and LGA airports were closed for safety measures due to the weather conditions [159] [177]. Port Authority closed JFK airport at 10:45 am on January 4th, expecting reopening at 3 pm. At 2 pm, the reopening was pushed to 8 pm. At 6 pm, it was pushed a second time to the next day, January 5th, at 7 am. On January 7th, the record low temperatures led to water pipes breaking at JFK Terminal 4, forcing a partial evacuation and flooding hundreds of luggage.

A.2.1 Overall impact on the United States

In this section, we selected the top 45 airports in terms of traffic volume in the continental United States and extracted all traffic between these airports from the BTS on-time performance measures. Given the hub-and-spoke structure of the airport network, this represents the majority of domestic operations.

The number of flown flights, aggregated by departing airports each day between December 27th, 2017 and January 12th, 2018, is shown in Figure A.1. This initial flight-centric perspective confirms the major impact the bomb cyclone had on four airports in particular: BOS, EWR, LGA and JFK. The volume of flights on January 4th is an extreme outlier for these airports, which are amongst the busiest in the United States, and is still lower than usual on January 5th.

A.2.2 Focus on the North East

Figure A.1 highlighted the impact of the bomb cyclone on four major airports of the North East of the United States, namely JFK, LGA, EWR and BOS. More precisely, it highlights the abnormal flight operations on January 4th (each of these airports had less than 30 flights overall) and emphasizes the two-day recovery period needed to return to a normal volume of operations.

However, the recovery in terms of schedule adherence and delays took longer, as depicted in Figure A.2(a), showing the number of delayed flights at these airports. First, there are almost no delayed flights on January 4th since

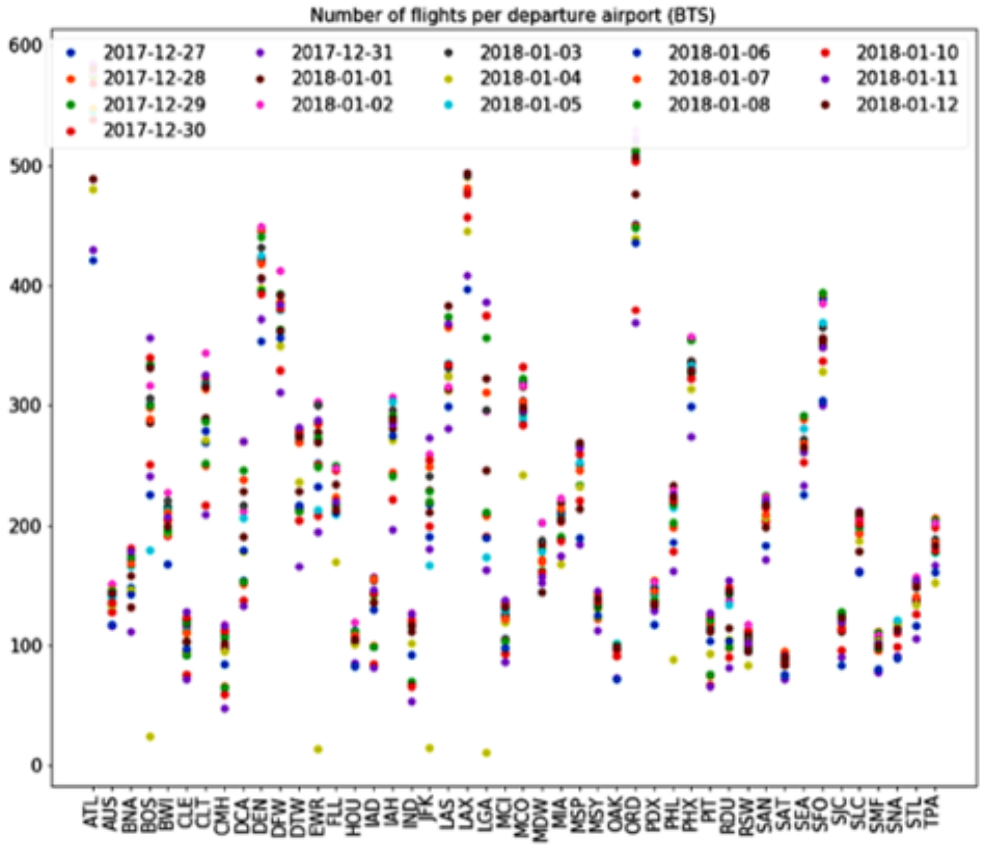
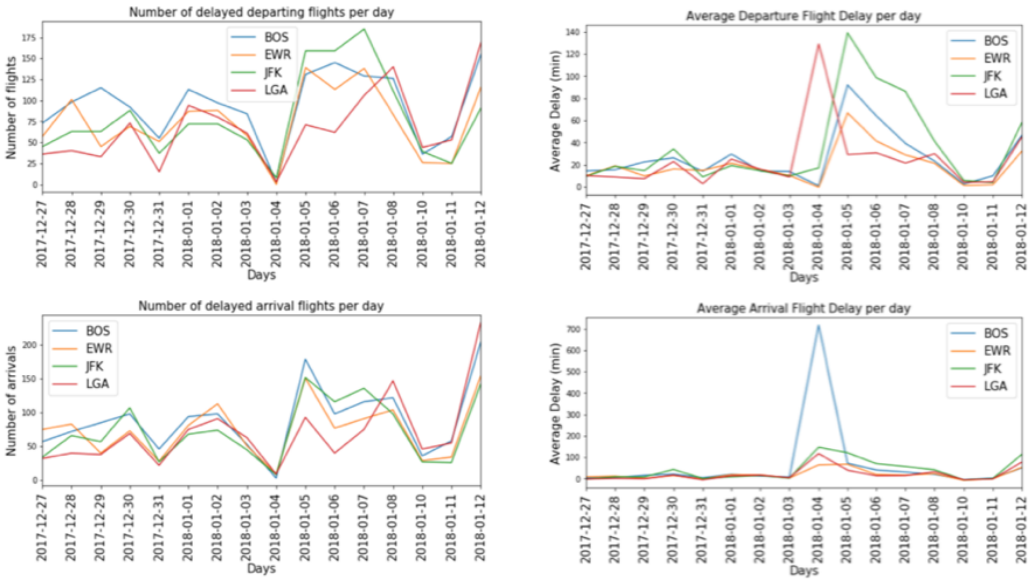


Figure A.1: Number of flights per departure airports (BTS)

the vast majority was cancelled. The recovery period took about five days. Figure A.2(b) presents the average flight delay per day at each airport at departure and arrival. The average departure delay shows different recovery profiles. LGA airport had its peak departure delay (across only 11 flights) on January 4$^{\text{th}}$, while it was on January 5$^{\text{th}}$ for the other airports. Moreover, on the worst day of the Bomb cyclone, January 4$^{\text{th}}$, at BOS, the average flight delay spiked to over 11 hours, for the only 3 flights that landed. For the other three airports, the peak arrival delay is on January 5$^{\text{th}}$.

From the BTS data, it is possible to evaluate the quantitative impact of the bomb cyclone on flight traffic. Yet, it is not enough to fully apprehend the disproportionate impact of the bomb cyclone on passenger experience.



(a) Number of delayed flights per day. (b) Average flight delay per day.

Figure A.2: Number of delayed flights and average flight delay per day

A.3 Bomb Cyclone from mobile location data

In this section, the method of passenger selection validated in [106] is implemented and analyzed for the time period covering the bomb cyclone.

A.3.1 Global view of domestic passengers experience at airports

The top 45 airports in terms of traffic were chosen for this study and latitude/longitude bounding boxes were created for each of them. On a daily basis, 5 billion records are collected by the carrier each time a phone connects to the cellular network and an approximate location is obtained from cell tower triangulation. A record consists in an anonymized user id, a time stamp and the approximate latitude and longitude of the user. Passengers are identified if they have a cell phone record located within the bounding boxes of at least two different airports, provided these airports are not in the same metropolitan area. Once the passengers are detected, only the initial and final time stamps within each bounding box are kept in order to have a reliable estimate of the time spent by the passengers in each airport.

The number of passengers per day for these airports using this method is represented in Figure A.3. From this plot, the same four northeastern

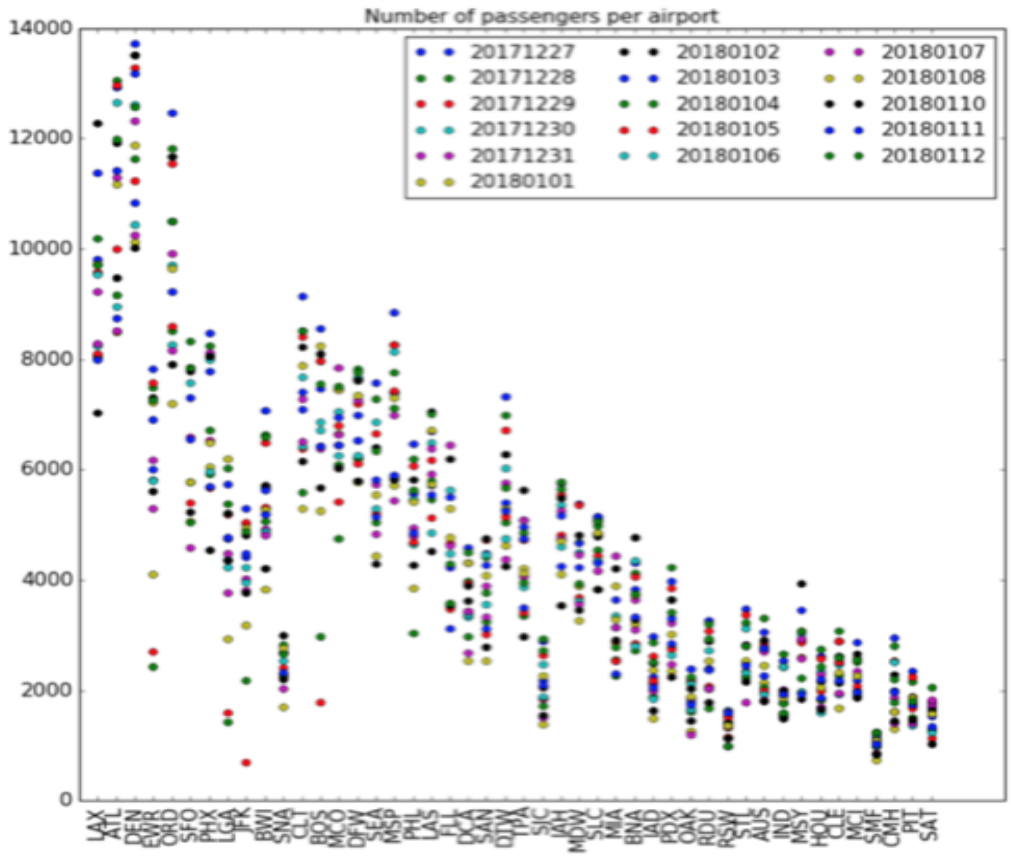


Figure A.3: Number of passengers per airport.

airports are noticeable as outliers on January 4th and 5th, 2018. This simple observation indicates that, from a passenger perspective, the peak of the bomb cyclone's impact was not solely located on January 4th as the BTS data shows.

Making a box plot visualization of the time spent at airports yields a more condensed way of comparing the performance of the airports in terms of passenger time spent within the airport. Figure A.4, which shows the average and quartile distribution of the time spent by passengers at each airport on January 2nd, 2018, at departure or arrival. As expected, passengers typically spend more time at departure than at arrival. January 2nd is selected as a fairly uneventful day, to portray the usual performance of each airport from a passenger's perspective. For example, LAX and MCO have the highest average time spent by passengers at departure, with 130 minutes, but LAS has the highest standard deviation, with 71 minutes. At arrival, the worst performer is DFW with 87 minutes on average.

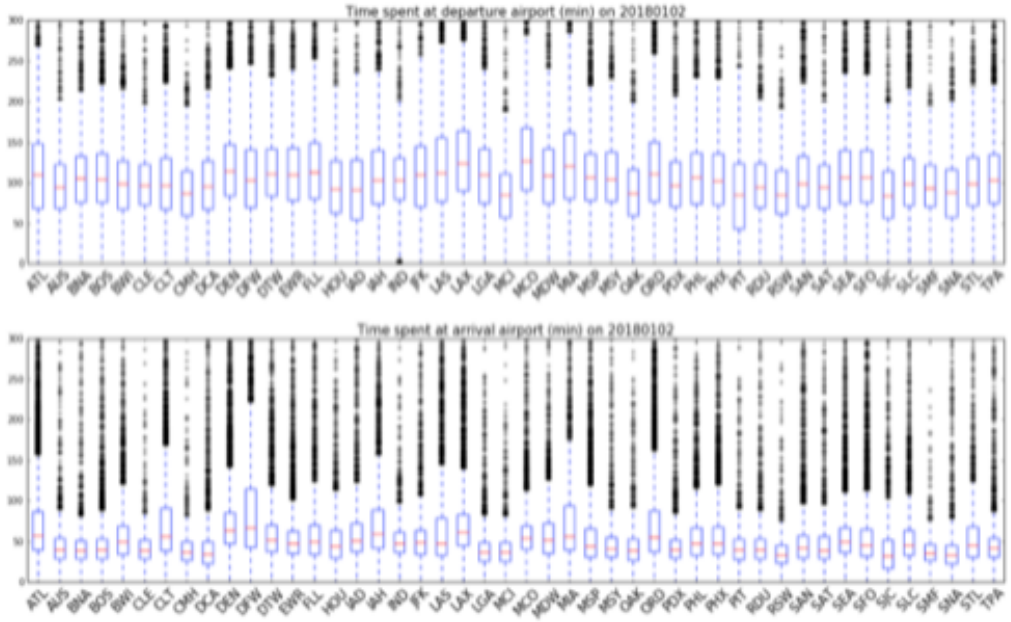


Figure A.4: Time spent at airports by passengers on January 2nd 2018.

Comparing Figure A.4 for January 2nd with Figures A.5 & A.6 - showing January 4th and 5th respectively yield interesting conclusions. On January 4th, the number of passengers at BOS, EWR, JFK and LGA is very small. Between January 2nd and 4th, passengers spent in average less time at departure at the impacted airports (about 10 percent) but with a wider distribution. Visually, this can be illustrated as the box plot sinking and widening compared to its normal state. While the average time spent at departure is similar to that of January 2nd, the standard deviation is about 20 percent higher for these four airports. On January 5th, when there were less cancellations but a peak in flight delay, we observe a peak in time spent at departure. For instance, at JFK, on January 2nd, a departing passenger spends 109 minutes on average, with a standard deviation of 59 minutes. This is a good performance compared to the other airports in the United States. But on January 5th at JFK, the average time at departure for passengers jumps to 194 minutes and the standard deviation to 98 minutes.

These new plots and new methods have confirmed that mobile phone data do pinpoint airports that are going through major disturbances by establishing a reliable benchmark of their performance from the passengers perspective.

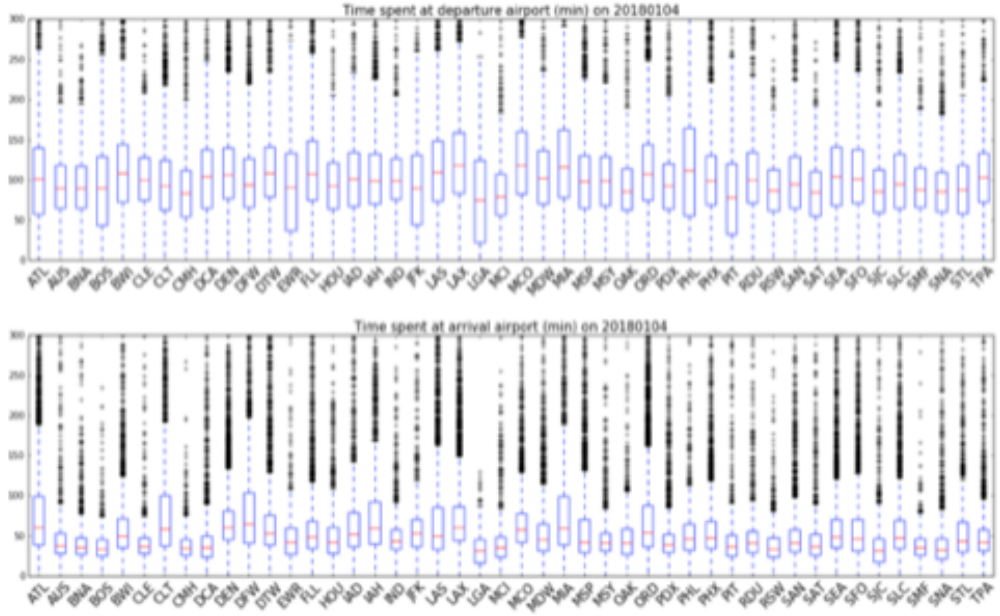


Figure A.5: Time spent at airports by passengers on January 4$^{\text{th}}$ 2018.

A.3.2 Analysis at each airport in the North East

Once the most impacted airports are identified from comparing with the top airports, a more specific analysis can be conducted to better evaluate if this disturbance impacted each of these airports differently.

Visitors

First, we examine the behavior of users visiting the airport, i.e. people who were within the bounding box of an airport. These visitors includes passengers, airport staff, taxi drivers, and anyone driving by as well. This approach is useful to know if the disturbance only affected passengers or a wider group. Figure A.7(a) shows the evolution of the number of visitors per day over two weeks around the bomb cyclone for the four impacted airports.

Each airport considered typically employs between 15,000 and 40,000 people, as airport staff. Thousand of domestic and international passengers transit through each airport. Friends and family drop off and pick up passengers. Several airports are located along major roads or highways, and because location data is noisy, pings might be recorded within bounding boxes around the airport. What matters here is not the absolute number but the relative changes day to day. The signal for passengers is much cleaner thanks to a more elaborate filtering.

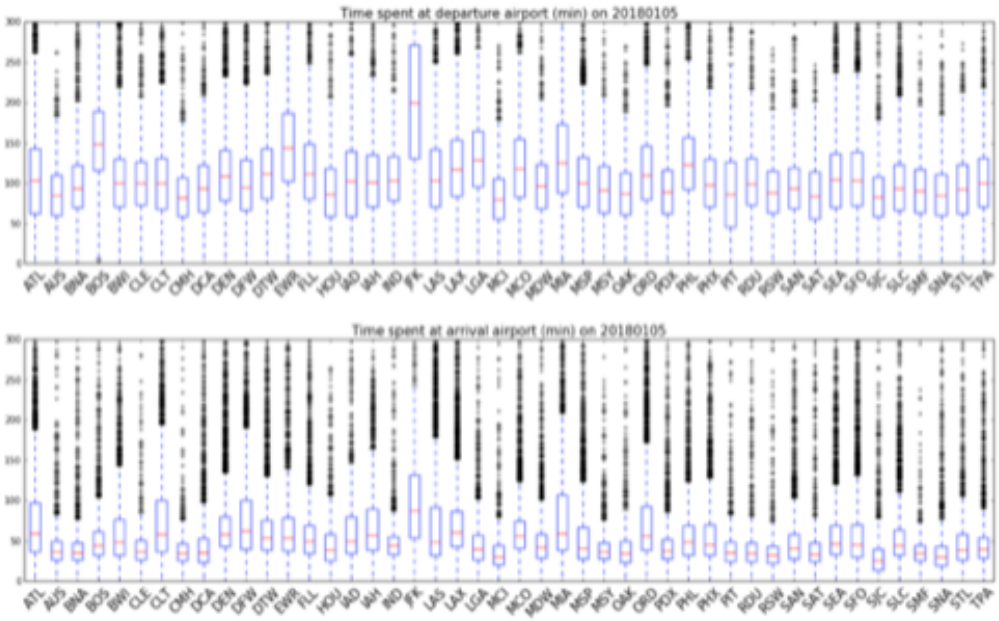
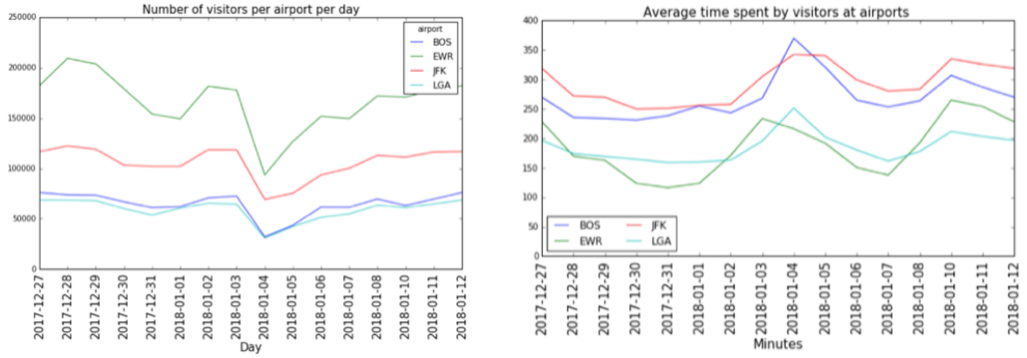


Figure A.6: Time spent at airports by passengers on January 5$^{\text{th}}$ 2018.

The drop on January 4$^{\text{th}}$ is clearly visible for all four impacted airports, with different recovery profiles. EWR had the fastest recovery while JFK's recovery was slower and started a day later.



(a) Number of visitors (b) Average time spent

Figure A.7: Evolution of the number of visitors and of the average time spent by visitors at the most impacted airports by the Bomb Cyclone

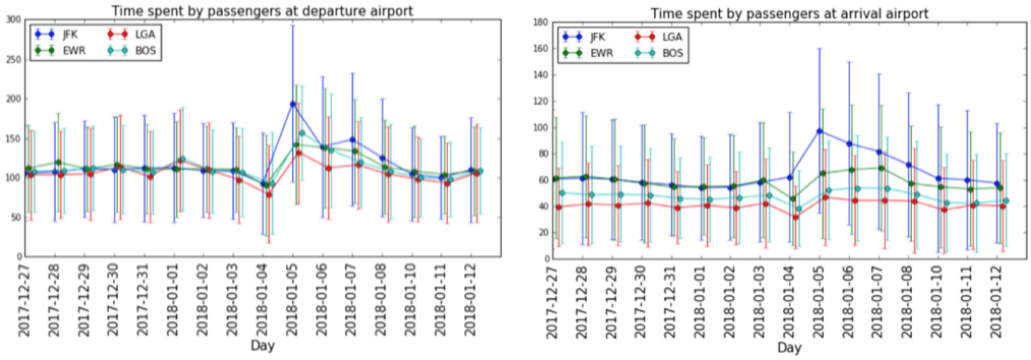
Plotting the distribution of the time spent at these airports by visitors every day yields some striking patterns. From Figure A.7(b), the average time spent by visitors is fairly consistent over the days except on January 4$^{\text{th}}$

and 5th. These plots show that even the number of visitors dropped during the bomb cyclone, most likely because the snow levels made access to these airports difficult. The peak is most visible for BOS.

This visitor-centric view gives us a better insight on the impact of any disturbance for different airports. During the bomb cyclone, both passengers, who flew in or out of the airport, and visitors were impacted. The difference of impact between the different airports for visitors also illustrates how the airports' access routes had an effect or were impacted by the anomaly.

Passengers

The box plots used to create a performance benchmark from a passenger's perspective proposed in section A.3.1 are useful to pinpoint impacted airports, and they can be used in a different configuration to gain additional insight on the differences between normal and disrupted behavior at each airport. Figures A.8(a) and A.8(b) shows the evolution across days of the average and standard deviation of the time spent at departure and arrival for passengers at the airports most impacted by the Bomb cyclone.



(a) At departure airports.

(b) At arrival airports.

Figure A.8: Average and standard deviation of time spent by passengers at departure and arrival airports.

The differences noted previously become obviously visible in terms of averages: on January 4th, at JFK, EWR, LGA and BOS, there is a small decrease in the average time spent at departure with a wider distribution for smaller waiting times followed by an important increase of the average time spent on January 5th with an increased distribution spread as well. Regarding the time spent at arrival, the patterns are similar, although less marked. In terms of standard deviations, JFK has a large increase in the width of the time distribution starting January 5th for the time spent at

departure and a four day recovery period for this parameter while LGA does not have this increase in width as well as a one day recovery period. EWR and BOS experience a lower increase in spread than JFK and they both have a three day recovery period. Overall, the recovery took longer at JFK than at EWR, LGA and BOS.

A.4 Bomb Cyclone on Twitter

A complementary view of passenger experience through this major disturbance can be obtained via passenger activity on social media platforms. This section presents a three step process in analyzing Twitter content in order to better understand the impact of the bomb cyclone from a passenger perspective.

A.4.1 Volume of tweets related to airlines/airports

Twitter's developer API [178] was used to filter and collect relevant tweets from the full Twitter stream, by using specific airlines and airports handles as queries over the same time period as in the previous sections, i.e. from December 27^{th} 2017 to January 12^{th} 2018. A tweet is considered as relevant if it contains the handle of an airline or an airport within its text or if it is published by an airline or an airport Twitter account. The considered Twitter handles for the specific case of the bomb cyclone are presented in Table A.1. This collection of tweets is organized as a database of tweets labeled by airline and by airport. Each entry of this database consists of the tweet ID, the time stamp, the text and the account handle used for the search.

Table A.1: Twitter handles used for gathering tweets relevant to the bomb cyclone perturbation

	Twitter handles
Airlines	@united, @Delta, @AmericanAir, @SouthwestAir, @SpiritAirlines, @VirginAmerica, @JetBlue
Airports	@JFKairport, @EWRairport, @BostonLogan, @LGAairport

A first step in visualizing the content of this temporal text-based database is to count the number of tweets per day and per label, i.e. the airline or airport handle used to retrieve them, and plot their evolution over time.

These volume changes in the social activity of passengers, airlines and airports are a first indicator that can be used to understand the social impact of the bomb cyclone. Figure A.9 shows the daily evolution of the number of tweets related to airlines over the considered period. It illustrates that customers had a different experience depending on which airlines they were flying. The most impacted airlines were Jetblue (B6) and Delta (DL), since they both had an important increase in Twitter volume over a period of four days. American Airlines (AA) and United Airlines (UA) have less important daily variations from January 1st to January 8th, which would tend to indicate a better management of passengers from these two airlines during this perturbation.

Figure A.9: Volume of tweets referring to airlines aggregated by day

Figure A.10 shows the daily evolution of the number of tweets related to airports over the considered period. It highlights how much worse the impact was at JFK airport compared to other North East airports, with a disproportionate amount of tweets associated with JFK Twitter handle from January 4th to January 14th. The peak of tweets is observed on January 7th, when one of its terminal was flooded by a broken water pipe.

To obtain a more fine-grained picture of the situation on social media, tweets are then aggregated on an hourly basis. Figure A.11 shows the hourly evolution of the number of tweets related to airlines over the same two weeks.

This representation brings additional insight concerning the two worst-hit airlines noticed in Figure A.9, Delta and Jetblue. Delta typically has a higher tweet volume over all days than other airlines, even on normal days. On the days following the bomb cyclone, the daily level is similar but there are hourly peaks on January 6th and January 7th. On the other hand, JetBlue shows

Figure A.10: Volume of tweets referring to airports aggregated by day

a large increase in tweets on January 4th in the afternoon, before steadying at a lower level for the following four days, albeit at a higher level than on normal days.

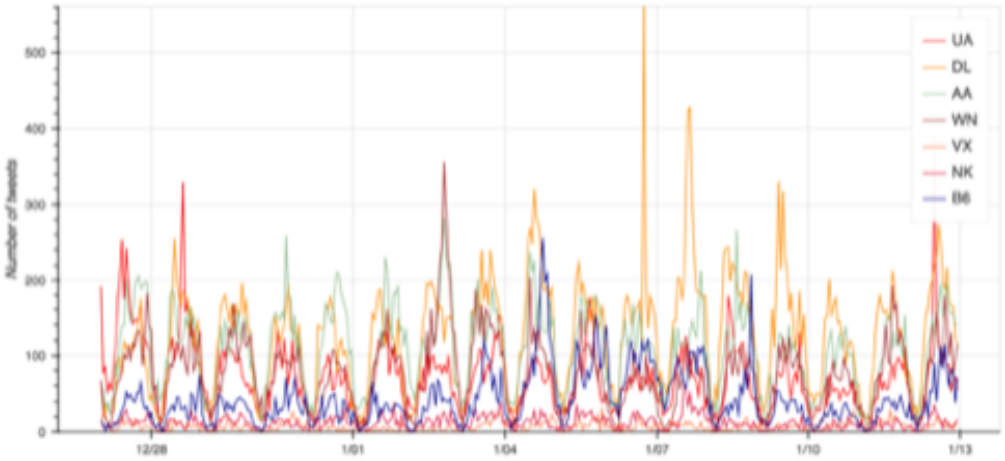


Figure A.11: Volume of tweets referring to airlines aggregated by hour

The visualization of tweet volume is more striking in terms of airports aggregated per hour, with Figure A.12 showing the hourly evolution of the number of tweets related to airports over these two weeks. From the normal small chatter similar to the other airports, JFK becomes a huge source of tweets as soon as 6am on January 4th. And this source takes five full days before slowly disappearing, which is consistent with the phone location anal-

ysis presented previously in Section A.3. This representation has also the advantage of pinpointing the hour when the broken pipe actually started impacting passengers, with a major spike in tweet volume in the early morning of January 7th.

Figure A.12: Volume of tweets referring to airports aggregated by hour

A.4.2 Tweets about delays and cancellations

While monitoring tweet volume provides clues regarding the presence of anomalies and can help pinpoint when they start impacting passengers, Twitter is most useful to obtain contextual information. Using simple filters based on the presence of specific keywords, one can get a better understanding of airline performance and overall passenger satisfaction. In the particular case of air transportation, filters based on cancellation or delay related keywords yield some interesting results. The keywords used for these filters can be found in Table A.2.

Table A.2: Keywords used for filtering tweets

Filter	Keywords
Cancellation	cancellation, cancel, cancelled, postponed
Delay	delay, delayed, wait, waiting, late, postponed, hours

Applying these filters and aggregating all airline-related tweets reveals that cancellations had a greater impact than delays on passengers' social

behavior. Figure A.13 shows the hourly evolution of the number of tweets containing cancellation related keywords from December 27th 2017 to January 12th 2018. The volume of cancellation related tweets increases almost five-fold on January 3rd, the day many cancellations were first announced given the weather forecasts, and keeps increasing on January 4th when the cancellations actually take place. The five day recovery period determined in the phone data analysis from Section A.3 is still visible on Figure A.13.

Figure A.13: Volume of tweets referring to airlines aggregated by hour filtered by cancellation-related keywords

Figure A.14 shows the hourly evolution of the number of tweets containing delay related keywords over the same two week period. Regarding the volume of delay related tweets, the increase is less visible, though still occurs, since the amount of delay actually decreased due to the increase of cancellations. From this perspective also it is clearly visible that the return to normal activity starts only around January 10th.

A.4.3 Topic analysis on tweets

A more elaborate way of exploiting information from tweets is to perform a topic analysis of the tweet database using Latent Dirichlet Allocation (LDA) [157] and comparing "normal" days (January 9th- 11th) with days where the bomb cyclone impacted the East coast (January 4th- 6th). In LDA, each document - here each tweet - is modeled as a finite mixture of topics. A topic is defined as a distribution over the words composing the full set of considered documents. The topic distribution of each document and the word distribution of each topic can be determined using variational Bayes

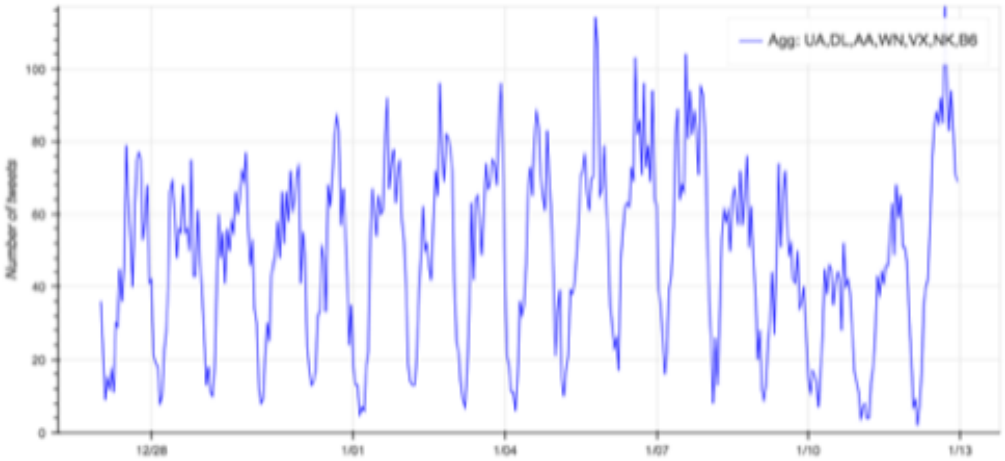


Figure A.14: Volume of tweets referring to airlines aggregated by hour filtered by delay-related keywords

approximations and was implemented in Python by Rehurek and Sojka [158] within the Gensim library. For this initial study, the topic distribution determination algorithm is run five times and the best topic representation is chosen using the coherence measures introduced in [179]. The aim of these coherence measures is to select topics with word distributions the more human understandable possible for a better explainability. Each tweet is thus associated with a finite number of topics, each having a relative importance. When considering a set of tweets, it is possible to add the importance of each topic over all the considered tweets leading to the creation of a topic importance ranking.

A first step in topic analysis is to clean and format the tweets analyzed. For instance, any reference to websites or pictures is replaced by a corresponding keyword. Mentions of other users within a tweet (@someone) and most emojis are similarly replaced. Note that "dm" means "direct message" on Twitter, which is used when a user wants to take a public conversation to a private channel. Since the collected database contains many responses from airlines, the individual signatures of each airline agent is also replaced by a keyword. Dates and times are also generically replaced by keywords. An improved and more detailed cleaning process inspired by these initial considerations, is further presented in Section 3.1.2. The resulting text is then filtered against common stop-words and words occurring only once in the whole month of January 2018 are removed.

Two different methods are used to study the impact of the bomb cyclone using topic analysis. In a first approach, topics are generated using the airline

related tweets across the full month of January 2018, and then their relative importance is ranked for each set of days. Table A.3 shows the top four topics using this method for the chosen set of normal days and Table A.4 shows the four top topics for the set of days impacted by the bomb cyclone. Words written in full capital letters correspond to cleaning keywords as explained in Section 3.1.2.

Table A.3: Top 4 monthly Twitter topics for January 9-11 2018

Rank	Distribution	Top 10 words
1	13.9	MENTION, flight, get, one, know, plane, PICTURE, still, WEBSITE, bags
2	13.0	SIGNATURE, sorry, thanks, flight, get, hear, know, MENTION, time, us
3	12.3	MENTION, PICTURE, WEBSITE, flight, SIGNATURE, help, back, thanks, flying, get
4	11.6	MENTION, flight, get, TIME, thanks, **delayed**, **cancelled**, flying, us, flights

Table A.4: Top 4 monthly Twitter topics for January 4-6 2018

Rank	Distribution	Top 10 words
1	16.1	MENTION, flight, get, TIME, thanks, **delayed**, **cancelled**, flying, us, flights
2	16.0	MENTION, flight, get, one, know, plane, PICTURE, still, WEBSITE, bags
3	10.9	MENTION, PICTURE, WEBSITE, flight, SIGNATURE, help, back, thanks, flying, get
4	10.9	SIGNATURE, sorry, thanks, flight, get, hear, know, MENTION, time, us

Only the ranking of the topics differ from Table A.4 to Table A.3, otherwise they are both composed a topic where passengers ask for information about their flights and bags, two topics where airlines answer to passenger concerns and a topic concerning delays and cancellations. As may be expected, the topic concerning cancellations and delays went from 4^{th} place during the normal days to 1st place during the bomb cyclone.

A second approach provides more specific insight regarding the bomb cyclone. Topics are determined independently for each set of days using only the tweets from the corresponding days. They are then ranked by importance on each set of days, see Table A.5 for the top four topics on January 9th to 11th and Table A.6 for the top four topics on January 4th to 6th.

Table A.5: Top 4 specific Twitter topics for January 9-11 2018

Rank	Distribution	Top 10 words
1	15.7	MENTION, flight, WEBSITE, PICTURE, thanks, time, great, airline, travel, flights
2	15.1	MENTION, flight, PICTURE, WEBSITE, thanks, back, get, thank, bag, SIGNATURE
3	13.8	MENTION, flight, get, PICTURE, one, plane, airport, check, still, help
4	10.3	SIGNATURE, WEBSITE, please_dm, sorry, dm, happy, hi, hear, bag, flight

Table A.6: Top 4 specific Twitter topics for January 4-6 2018

Rank	Distribution	Top 10 words
1	22.4	flight, MENTION, get, TIME, flights, **cancelled**, **jfk**, still, time, **delayed**
2	12.7	SIGNATURE, sorry, MENTION, thanks, us, know, please, flight, team, airport
3	11.2	MENTION, thank, WEBSITE, get, SIGNATURE, PICTURE, time, see, flight, airport
4	11.2	MENTION, PICTURE, flight, thanks, WEBSITE, get, us, help, one, airport

Regarding the set of normal days in Table A.5, the topics are similar to the top three topics found using the first approach in Table A.3 and they encompass usual tweets about vacations, trips, waiting for luggage at airports as well as discussions between passengers and airlines customer services. Regarding the set of days impacted by the bomb cyclone, Table A.6 highlights the large impact that the bomb cyclone had on JFK airport, with the top topic of the corresponding days containing the word "jfk" along with the words "cancelled" and "delayed". It is worth noting that the tweets considered to create these topics are the tweets related to airline handles and not to airport handles, which means that airport related information can be found in airline related tweets.

A.5 Conclusion

To the best of the authors' knowledge, this study constitutes one of the first big data applications of mobile phone data and social media data to the analysis of the impact of large disruptions in air transportation. Leveraging

two weeks of mobile location data in the United States, with more than 5 billion records per day, as well as two weeks of Twitter data, this analysis shows that mobile phones and social media can act as sensors for air traffic passengers, yielding a more complete and richer picture of the situation than traditional flight-centric measurements from the Bureau of Transportation Statistics. Thanks to these independent sources of measurements, various aviation stakeholders, who currently only have access to a partial and private view of passenger behavior, could now reliably measure system-wide passenger-centric metrics. These methods were here implemented in order to provide insights on how the passenger experience was impacted at airports in the North East of the United States during the Bomb Cyclone in January 2018.

Appendix B

Passengers as a real-time estimator of the US air transportation system: a first working model for estimating delays

In this appendix is presented the second version of the process that transforms the flow from the social media Twitter into a real-time estimator of the US Air Transportation system. Two different machine learning regressors have been trained on this 2017 passenger-centric dataset and tested on the first two months of 2018 for the estimation of air traffic delays at departure and arrival at 34 different US airports. Using three different levels of content-related features created from the flow of social media posts led to the extraction of useful information about the current state of the air traffic system. The resulting methods yield higher estimation performances than traditional state-of-the-art and off-the-shelf time-series forecasting techniques performed on flight-centric data for more than 28 airports. Moreover the features extracted can also be used to start a passenger-centric analysis of the Air Transportation system. This appendix is the continuation of previous works focusing on estimating air traffic delays leveraging a real-time publicly available passenger-centered data source [11]. The results of this study suggest a method to use passenger-centric data-sources as an estimator of the current state of the different actors of the air transportation system in real-time.

This appendix proposes to build on this previous work in order to estimate the state of the air transportation system to a finer level. Rather than predicting the number of delays across all the United States, the proposed passenger-centric models are improved and tuned to accurately estimate the state of delays for each of the 35 major airports within the United States. The created models are based on three different levels of content-related features created from the flow of social media posts. First results indicate that these new models can estimate the number of hourly delays with a mean absolute error of less than 3 flights for 26 of the considered airports, and of less than 6 flights for the 9 remaining airports.

The rest of the appendix is structured as follows: Section B.1 describes the datasets and the feature extraction process. The methodology and results of the training process are shown in Section B.2, before being analyzed and exploited in Section B.3. Section B.4 concludes this study and discusses possible future steps.

B.1 Dataset description and feature selection

B.1.1 Dataset description

Following the initial work performed in [11], the goal here is to use passengers behavior on social media - in particular on Twitter - in order to analyze

and estimate the flight-centric health of the US air-transportation system at an airport level. In this study, the flight-centric health of an airport is described by delay related information contained within BTS data. This data is publicly available usually with a two to three month delay and this study limits itself with the BTS data from January 2017 to February 2018.

The Twitter dataset available for this study is the same as in [11] and consists of all the tweets found using a basic search for each handle of 7 major US airlines as well as 34 major US airports (one of them having two Twitter handles). The full list of handles can be found in Table B.1. Each entry consists of a timestamp, a user id, the content of the tweet and the handle used to retrieve the tweet. This dataset spans the entire period from January 1st 2017 to February 28th 2018. The extraction of features from this dataset has been improved since the previous study and is described in Section B.1.2.

Table B.1: Twitter handles used for gathering tweets

Category	**Twitter handles**
Airlines	@united, @Delta, @AmericanAir, @SouthwestAir, @SpiritAirlines, @VirginAmerica, @JetBlue
Airports	@JFKairport, @ATLairport, @flyLAXairport, @fly2ohare, @DFWAirport, @DENAirport, @CLTAirport, @LASairport, @PHXSkyHarbor, @MiamiAirportMIA, @iah, @EWRairport, @MCOAirport, @Official_MCO, @SeaTacAirport, @mspairport, @DTWeetin, @BostonLogan, @PHLAirport, @LGAairport, @FLLFlyer, @BWI_Airport, @Dulles_Airport, @MidwayAirport, @Reagan_Airport, @slcairport, @SanDiegoAirport, @flyTPA, @flypdx, @flystl, @flySFO,@HobbyAirport, @flynashville, @AUStinAirport, @KCIAirport

In order to estimate the flight-centric health of each considered airport, this information first needs to be extracted from the BTS dataset for each airport. Only two types of delayed flights are considered here from a passenger's perspective: Flights departing with any amount of delay, and flights arriving with a delay greater than 15 minutes. Once all the flights departing an airport and all the flights arriving at the same airport are selected, the following values can be aggregated per hour:

- NumDepDelay: Number of flights departing with a delay
- NumArrDelay15: Number of flights arriving with a delay greater than 15 minutes

The aim of this study is to accurately estimate these two values for each airport at every hour using a single passenger-centric dataset.

B.1.2 Feature selection on Twitter data

Volume features

Features were extracted identically for all search handles presented in Table B.1, for the exception of @MiamiAirportMIA, which does not gather enough tweets. In addition to the raw number of tweets per hour per search handle, keyword related information is also extracted from the Twitter dataset. In order to keep all the relevant tweets without having to decline all the possible forms of the chosen keywords (e.g. "delay", "delayed", "delays", etc.), simple regular expression filters were created for each keyword: Any tweet containing a word starting with the related keyword is kept and the resulting tweets are then aggregated per hour. Five keywords were chosen for this study: 'delay', 'wait', 'cancel', 'hours', 'refund'.

Topic features

Another way of exploiting information from the content of these tweets is to perform a topic analysis of the tweet database using Latent Dirichlet Allocation [157] (LDA). In LDA, each document - here each tweet - is modeled as a finite mixture of topics. A topic is defined as a distribution over the words composing the full set of considered documents. The topic distribution of each document and the word distribution of each topic can be determined using variational Bayes approximations and was implemented in Python by Rehurek and Sojka [158] within the Gensim library.

A first step in topic analysis is to clean the documents analyzed, here the tweets. This cleaning process was already performed in [10] and [11] and consists of the following steps: any reference to websites or pictures was replaced by a corresponding keyword. Every mention to another Twitter user within a tweet (@someone) as well as most emojis were similarly replaced. Since this database contains many replies from airlines to their customers, individual signatures of each agent were also replaced by a keyword. Dates and times were also generically replaced by keywords (e.g. "3rd Jan 2017" becomes "DATE" and "4pm" becomes "TIME"). The resulting text was then filtered from common stop-words and from words occurring only once in the whole year of 2017.

For this study, the choice of 100 topics is made and the topic distribution determination algorithm is run five times and the best topic representation is chosen using the coherence measures introduced in [179]. The aim of these coherence measures is to select topics with word distributions the more human understandable possible for a better explainability. As an example, the top five words of a created topic are: "toknowmeistoflywithme",

"nut_allergy", "restrictions_apply", "comfortable_journey" and "mins_secs". The first word represents a *hashtag* for the phrase "To know me is to fly with me" and the other words are actually bigrams. The combination of these five words indicate a topic around passenger well-being aboard a plane.

The topic mixture of each tweet is then calculated based on this choice of 100 topics. Topic related features are then created by averaging the distribution of each topic per hour and per search handle. The hourly standard deviation of each topic distribution is also extracted.

This cleaning process introduces two additional keywords that enables a quick filtering of tweets, and therefore two additional features to add per search handle: tweets containing a picture and those containing a website link. Thus, seven keywords are actually considered for feature extraction: 'delay', 'wait', 'cancel', 'hours', 'refund', 'PICTURE', 'WEBSITE'.

Sentiment features

Sentiment analysis is also used here to enhance the feature set considered. Two different datasets and cleaning method were used to train three different regressors each. The first dataset used was the labelled dataset used in a Kaggle competition [156] and was cleaned using the same process as for the previous LDA learning. The generic keywords from the cleaning process (e.g. 'WEBSITE', 'DATE') were removed before creating the associated dictionary, as well as words appearing in less than 20 tweets or in more than 75% of the full dataset. A second dataset and cleaning process was generated based on the work of Read [115]. Emoji filters were used to extract tweets from the initial dataset and automatically label them with a positive or negative sentiment according to Table 3.2 (page 28). The text cleaning process is improved by merging negation words ("no", "not" and "never") with the word that follows it. The tokens used for the creation of the dictionary are the resulting bigrams, i.e. combinations of two words that follow each other in a tweet, with the same frequency filter as the first method described.

For both methods, three classifiers are trained (a random forest classifier, a naive Bayesian classifier and a logistic regressor) using the scikit-learn library [148]. A sentiment score is then calculated for each tweet by averaging the output of these classifiers, 0 meaning a unanimous negative sentiment and 1 a unanimous positive sentiment. The hourly average of these scores are added to the Twitter feature set.

Summary

Given the temporal nature of the data analyzed, the following features were chosen to keep track of the date: month of the year, day of the month, day of the week and hour in the day. In summary the following 8,484 features are considered:

- Hourly volume of tweets for each search handle (7 airlines and 33 airports giving 40 features): Num_tweets_*handle*
- Hourly volume of keyword-related tweets for each search handle (40x7 features): Num_tweets_*keyword_handle*
- Hourly average of tweets' sentiment (40x2 features): Mean_sent_*method_handle*
- Hourly average of topic distribution for each search handle (40x100 features): Mean_*topic_handle*
- Hourly standard deviation of topic distribution for each search handle (40x100 features): Std_*topic_handle*
- Month of the year, Day of the month, Day of the week and Hour in the day (4 features)

B.2 Estimating delays

The aim of this section is to see how well it is possible to estimate per airport the number of flights departing with a delay and the number of flights arriving with a delay greater than 15 minutes using the features extracted from the Twitter dataset. The dataset was split into a training set consisting of the data from the year 2017, and a testing set with the data from January and February 2018.

B.2.1 Methodology

For each BTS value, two different machine learning regressors were trained on the training data set: a Random Forest regressor and a Gradient Boosting regressor. These regressors were implemented from scikit-learn [148] with identical hyper-parameters. The maximum depth of each regressor was limited to ten, the minimum number of samples for a split was fixed to two and the maximum number of trees was fixed at ten.

As a comparison benchmark, we used Facebook's time-series forecasting tool Prophet [145] on the 2017 BTS data to forecast the full two first months of 2018. The Prophet tool is based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality [146]. It is described

as robust to outliers and missing data with no parameter tuning necessary, therefore the default parameters of the Prophet tool was used for this forecasting benchmark.

Lastly, the standard deviation of the BTS values in the training set were calculated to illustrate the added value of the trained regressors. The performance measures used to compare the different regressors are presented in the upcoming section B.2.2.

B.2.2 Estimation performance measures

In order to measure the performance of the different models, two different indicators were used: the R^2 score and the mean-absolute error (MAE).

The R^2 score, also known as the coefficient of determination, is defined as the unity minus the ratio of the residual sum of squares over the total sum of squares:

$$R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{B.1}$$

where y is the value to be predicted, $\bar{y}$ its mean and f is the predicted value. It ranges from $-\infty$ to 1. A score of 1 indicates a perfect prediction and a score of 0 means that the prediction does as well as constantly predicting the mean value for each occurrence. In the case of a negative R^2, then the model has a worse prediction than if it were predicting the mean value for each occurrence and therefore yields no useful predictions.

Regarding the mean-absolute error, the smaller its value is, the more accurate the prediction is. It is calculated using the following formula:

$$\text{MAE} = \frac{1}{n} \sum_i |f_i - y_i| \tag{B.2}$$

where n is the number of values being predicted.

B.2.3 Estimation results

Figure B.1 shows a comparison per airport of the mean-absolute error of the two trained regressors along with the chosen benchmark for the estimation of the number of flights departing with a delay. The standard deviation of the number of delayed departing flights at each airport during the year 2017 is also included for comparison. The Random Forest models have the best results in this case: they outperform the Gradient Boosting models at all-but-one airports (LAX) and the Facebook Prophet tool on 31 airports out of 34. For 26 airports, the Random Forest models are able to estimate

the hourly number of delayed departing flights with a mean-absolute error of three flights or less, and with an error of less than six flights for the remaining airports.

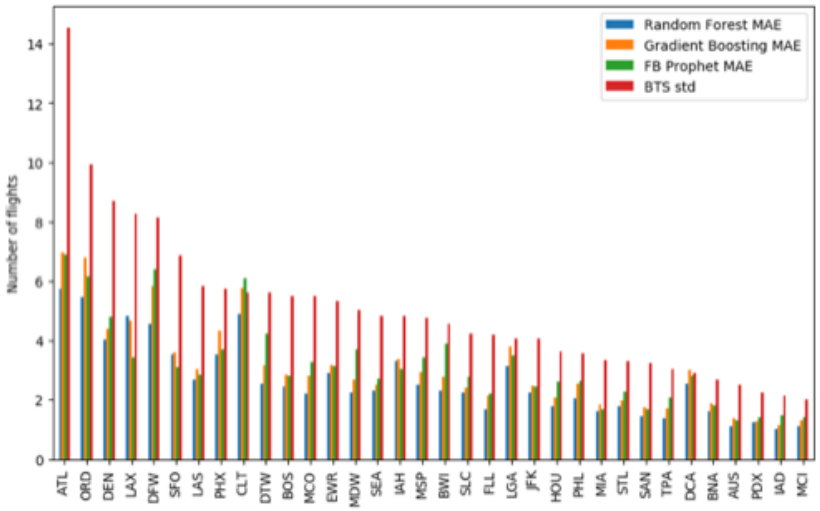


Figure B.1: Comparison of the mean absolute errors per airport for the trained regressors for the estimation of the number of delayed departing flights. The standard deviation of the BTS value on the training set is included for comparison.

Figure B.2 shows a comparison per airport of the mean-absolute error of the two trained regressors along with the chosen benchmark for the estimation of the number of flights arriving with a delay greater than 15 minutes. The standard deviation of the number of delayed arriving flights at each airport during the year 2017 is also included for comparison. The Random Forest models also have the best results in this case though their relative performance are not as important as for delayed departing flights : they outperform the Gradient Boosting models at 27 airports out of 34 and the Facebook Prophet tool on 28 airports out of 34. The absolute performance is however better than for estimating the number of delayed departing flights. For 28 airports, the Random Forest models are able to estimate the hourly number of delayed departing flights with a mean-absolute error of less than three flights, and with an error of less than five flights for the remaining airports.

Figure B.3 shows a comparison per airport of the R^2 score of the two trained regressors along with the chosen benchmark for the estimation of the number of flights departing with a delay. The Random Forest models still

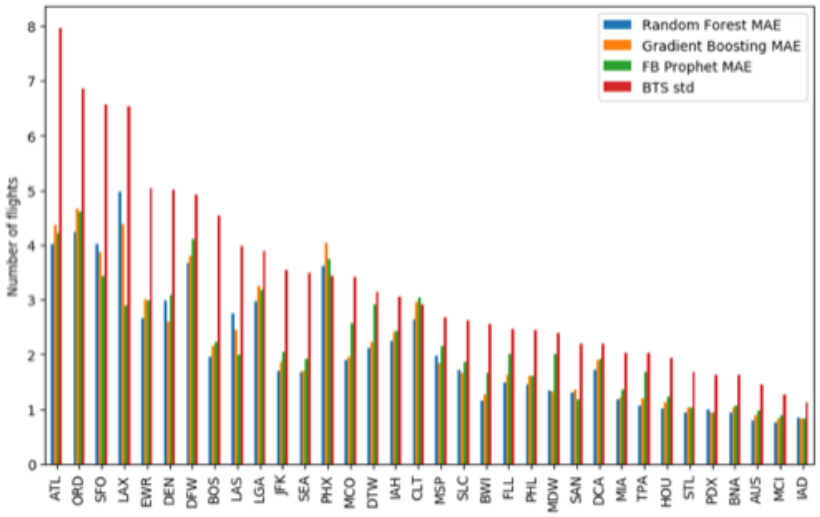


Figure B.2: Comparison of the mean absolute errors per airport for the trained regressors for the estimation of the number of flights arriving with a delay greater than 15 minutes. The standard deviation of the BTS value on the training set is included for comparison.

have the best results in this case, but the model associated with LAX airport also shows the only negative score. They outperform the Gradient Boosting models at 27 airports out of 34 and the Facebook Prophet tool on 28 airports out of 34.

B.3 Analysis and applications

The aim of this section is to analyze the differences between the chosen models as well as to explore possible applications resulting from the extracted features.

B.3.1 Model analysis

Figure B.4 shows the hourly prediction of the number of delayed departing flights at Atlanta airport (ATL) over the period January 12th-16thfor the two trained regressors along with the benchmark and the actual values. This airport was chosen since it has the highest BTS standard deviation for the number of delayed departing and arriving flights, and the period was chosen to illustrate the high variability of the number of delays from a day to another.

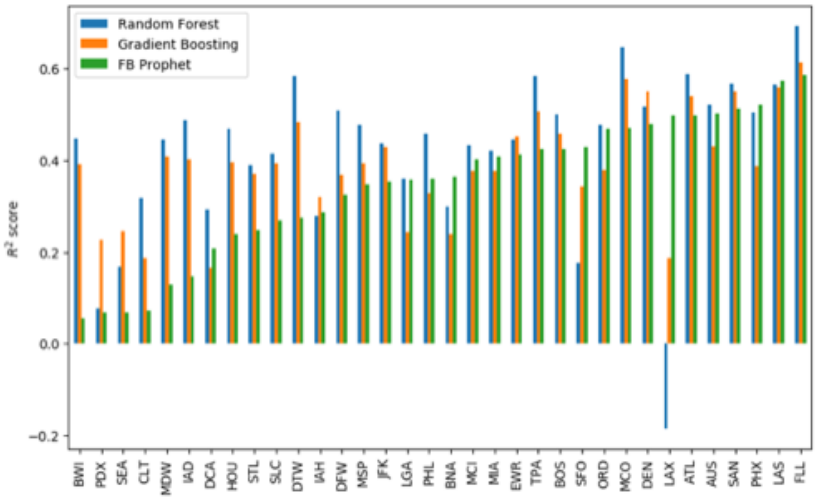


Figure B.3: Comparison of the R^2 scores per airport for the trained regressors for the estimation of the number of delayed departing flights

In this example, January 12 has more than twice as many delayed flights than any other day, as well as important hourly variations.

Figure B.4 illustrates the main differences between the different models. The Prophet tool predicts for each day a similar daily variation with three peaks during the day yet with amplitudes varying depending on the month and the day of the week. It also predicts negative values, which underlines some limitations of the model in this case. The added value from passenger-centric data-sources is better seen on January 12 and 13, where only the Random Forest regressor is able to estimate the higher number of delays on January 12 before correctly estimating the more usual levels of January 13. The Gradient Boosting regressor doesn't estimate outliers as well as the Random Forest regressor due to the difference in their loss functions. That difference is also illustrated by the non-zero minimum of the Gradient Boosting estimation during night time.

B.3.2 Other applications

Real-time sentiment analysis

The extracted features can be fed to the trained models for accurately estimating the number of delayed flights, but they can also be used directly in order to sense the overall passenger mood. Once the sentiment analysis are conducted on the tweets, it is possible to merge them into one score per

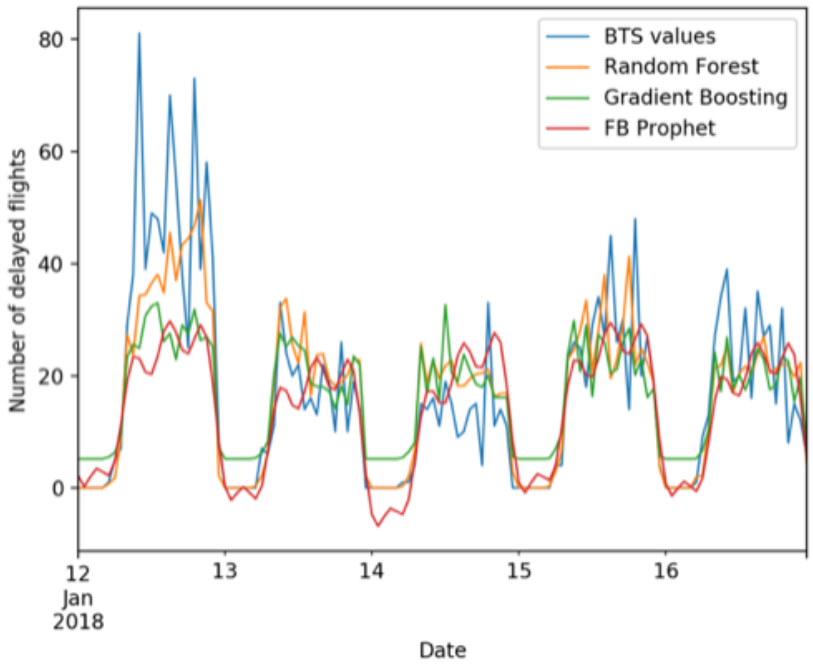


Figure B.4: Predicted number of delayed departing flights at ATL by the trained regressor over the period January 12$^{\text{th}}$, 2018 to January 16$^{\text{th}}$, 2018. The actual number of delayed departing flights is indicated for comparison.

airline and monitor their evolution.

Figure B.5 shows the hourly average mood for three major airlines during the Northeastern bomb cyclone studied in [10]. These three airlines have a similar passenger mood evolution at the beginning and the end of the period, yet United Airlines shows a drop in passenger mood on January 4$^{\text{th}}$, the day when the bomb cyclone actually hit the East coast. Though all three airlines have hubs in New York, United Airlines is the only airline with a hub at Newark International Airport (EWR) and not John F. Kennedy International Airport (JFK) nor LaGuardia Airport (LGA), which were both closed during the bomb cyclone, meaning that United Airlines probably had more dissatisfied passengers to handle on site during these extreme weather conditions.

Airports passenger map

After training the Random Forest models, it is possible to search for the most important features within the 8,484 initial features for each airport. This is achieved by using the Mean Decrease Impurity measure defined by Breiman

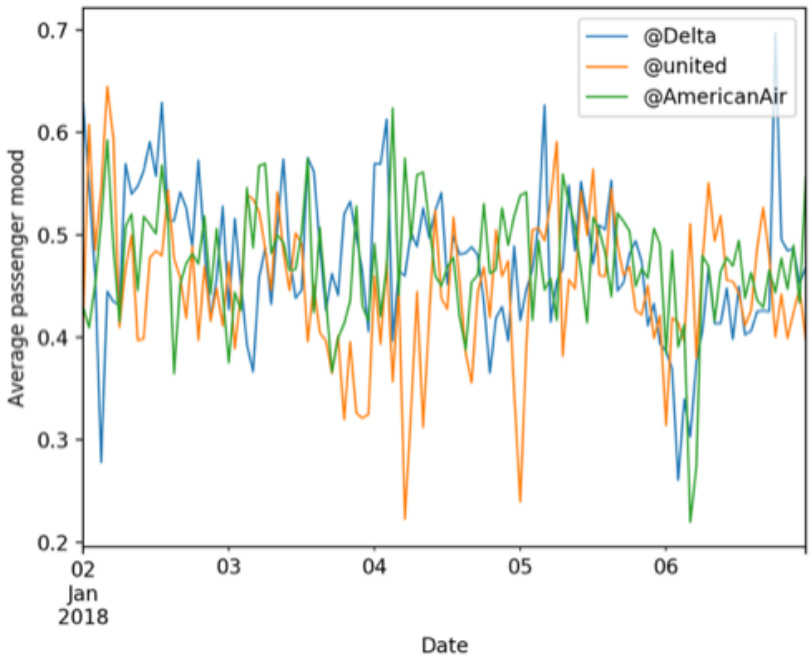


Figure B.5: Average passenger sentiment with respect to three major airlines over the period January 2nd, 2018 to January 6th, 2018, corresponding to a bomb cyclone hitting in the North-East of the US.

in [161] and normalizing the obtained feature importance scores so that the sum of all feature importance scores is equal to one. Table B.2 shows the ten features with the highest feature importance for predicting the number of delayed departing flights in ATL. Besides date related features, four of the top ten features are related to the volume of tweets containing delay keywords.

Table B.2: Top ten features for predicting the number of delayed departing flights at ATL

Rank	**Feature**	**Rank**	**Feature**
1	Hour	6	DayOfMonth
2	Month	7	delay_@SouthwestAir
3	DayOfWeek	8	num_ATL
4	delay_@Delta	9	delay_JFK
5	delay_ATL	10	mean_63_BWI

Once the features gathering 99% of the total importance for estimating the number of delayed flights are extracted, it is possible to group these

features per origin in order to gain some insight on how airports are related from a passenger perspective. For example, once the most important features for estimating the number of delays at ATL are extracted, it is possible to count how many of these features are issued from tweets gathered using the handle of John F. Kennedy International Airport (JFK).

Figure B.6: Map of feature links between Atlanta airport (ATL) and the other airports for estimating the number of delayed departing flights. The larger the link, the more features were kept among the features gathering 99% of the total importance for estimating the number of departing delayed flights at ATL.

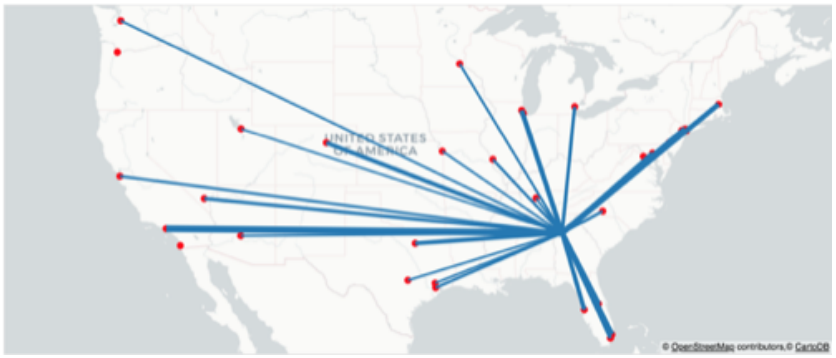


Figure B.7: Map of delay links between Atlanta airport (ATL) and the other airports. The larger the link, the more flights departed with a delay during 2017 from ATL towards the connecting airport. Only links with more than 1000 delayed flights in 2017 were considered.

Figure B.6 shows how ATL is connected to the other airports from this perspective. The larger the link between ATL and another airport, the more features were kept among the features gathering 99% of the total importance for estimating the number of departing delayed flights at ATL. Interestingly,

this airport graph is different from the graph built from the actual BTS values. Figure B.7 shows how ATL is connected to the other airports using the number of delayed departing flights from ATL. For example, although there are many delayed flights departing to Florida, few features from Floridan airports are kept. The opposite observation can be made regarding Portland (PDX): there were less than a thousand delayed flights from ATL to PDX, yet features from PDX were kept.

This example illustrates the possibility of creating a yearly review of airport relationship from a passenger point of view. Future studies should investigate more thoroughly the possible correlation and relation between the passenger connection map and the delay connection map.

B.4 Conclusion

This appendix aimed at investigating further the use of the social media Twitter as an estimator of the US Air Transportation system. Exploiting both raw volume information as well as different levels of content information within the Twitter stream enables to accurately estimate for each airport the number of flights departing with a delay and the number of flights arriving with a delay greater than fifteen minutes. This passenger-based estimation yields a better estimation performance for a majority of airports compared to using a state-of-the-art and off-the-shelf forecasting tool on the flight-centric data alone. Moreover, the methods used to extract relevant features from this passenger-centric data-source can be used to gain additional real-time insight on how passengers relate to the Air Transportation system.

This study confirmed that information contained in passenger-centric datasets are useful for a better understanding of the different stakeholders within the air transportation system, and have the added benefit of being more readily and publicly available than flight centric datasets. Future studies should focus on analyzing cases when the estimation is less accurate, implying differences between the handling of passengers and that of planes. Another direction of study considered is to validate this method to other countries or regions (e.g. the European Union) where sufficient flight-centric data is available.

Appendix C

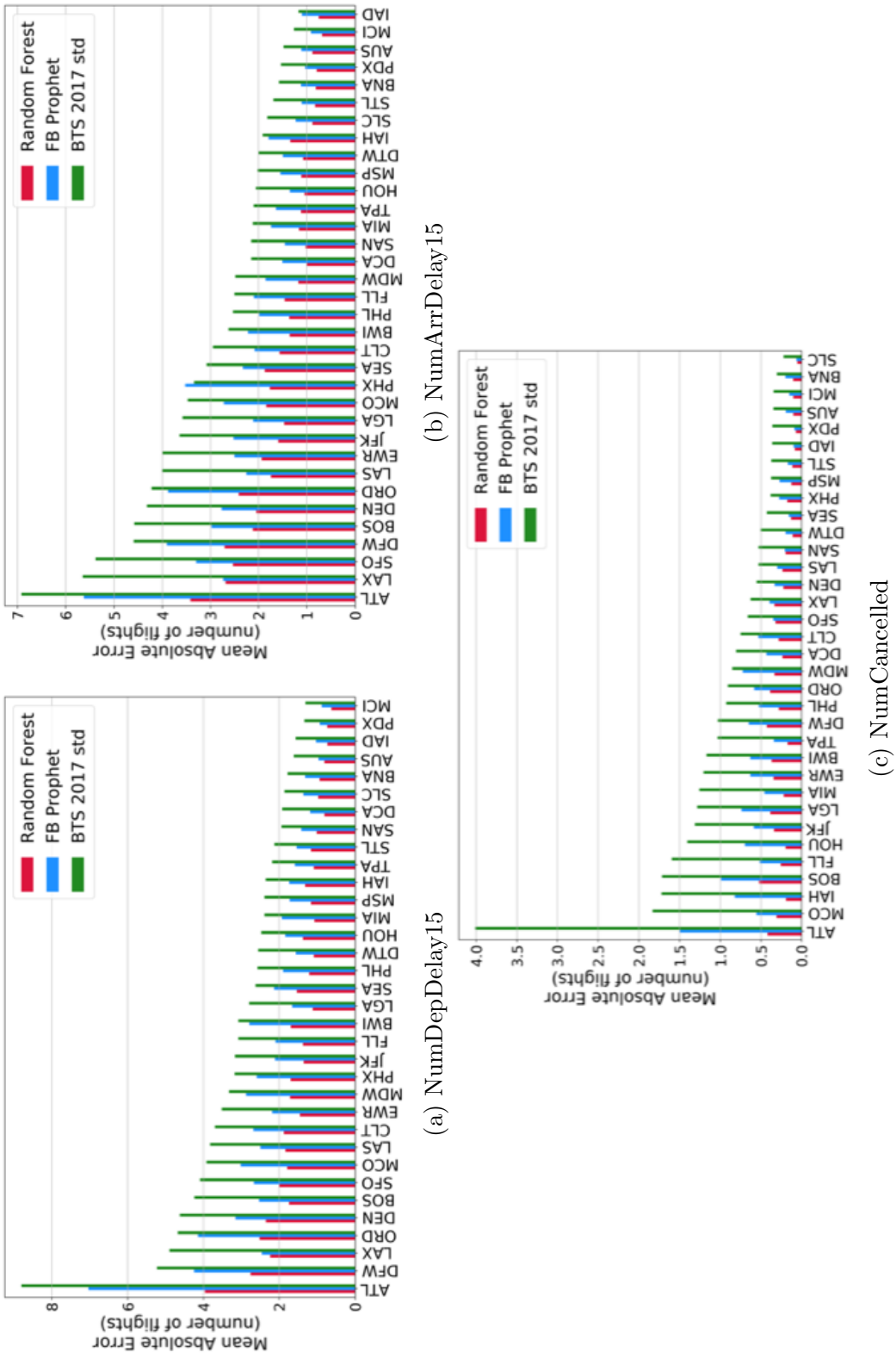
Passengers on social media: A real-time estimator of delays and cancellations in the US air transportation system

C.1 Comparison of the performance of the estimation model with the prediction model

This appendix presents a comparison of the performance of the estimation models versus the prediction models, both presented in Chapter 3, using the mean-absolute error (MAE) as the comparison metric. The mean-absolute error represents the average of the absolute values of the estimation (resp. prediction) errors over the test set. The smaller its value is, the more accurate the estimation (resp. prediction) is. It is calculated using the following formula:

$$\text{MAE} = \frac{1}{n} \sum_{i} |f_i - y_i| \tag{C.1}$$

where y is the value to be estimated (resp. predicted), f is the estimated (resp. predicted) value and n is the number of values being estimated (resp. predicted). The MAE are calculated over the full year of 2018. As explained in Section 3.1, the estimation models based on the features extracted from the Twitter stream are trained once on data from 2017 and then tested on the full year of 2018. The prediction models based on the historic BTS data available predict each month of 2018 separately based on the BTS data from January 2017 to two months before the month to predict, and the predictions are then regrouped for a single application of the MAE formula.



(a) NumDepDelay15 (b) NumArrDelay15 (c) NumCancelled

Figure C.1: Comparison of the mean absolute errors of the estimations of the number of abnormal flights using the features extracted from Twitter with the mean absolute errors of the predictions based on historical BTS values over the period January 1st, 2018 to December 31st, 2018. The standard variation of the BTS values is indicated in green.

Figure C.1 presents the bar plots of the mean-absolute errors of the estimations using on the features extracted from the Twitter stream (in red) along with the mean-absolute errors of the predictions based on the historic BTS values (in blue) for the three types of abnormal flights under consideration: flights departing with a delay greater than 15 minutes (Figure C.1(a)), flights arriving with a delay greater than 15 minutes (Figure C.1(b)) and cancelled flights (Figure C.1(c)). For each type of abnormal flight, the standard deviation of the actual BTS values over the year 2017 is indicated in green for comparison. For each type of abnormal flight, the airports on the x-axis are sorted based on the 2017 standard deviation of the actual BTS value.

From Figure C.1, it is clear that the estimation model based on features extracted from passenger-generated data has a better MAE performance for all airports and for estimating any type of abnormal flight than the prediction model based on historic BTS data. Figure C.1(a) and Figure C.1(b) indicate that for all-but-one airport (ATL), the proposed estimation models have an MAE of less than 3 flights per hour when estimating the number of flights departing or arriving with a delay greater than 15 minutes. Figure C.1(c) indicates that the MAE for estimating the hourly number of cancelled flights is lower than 0.5 flights for all-but-one airport (BOS).

Appendix D

Improving passenger experience at airports, some thoughts

The model presented in Chapter 5 highlights the disproportionate amount of time passengers can spend in airports when traveling by plane. This appendix presents the different works conducted throughout the thesis, in parallel to the works presented in the main body, that focus on analyzing passenger wait time and experience at airports in order to help airports and airlines improve the overall passenger experience.

D.1 Towards a more complete view of air transportation performance combining on-time performance and passenger sentiment

This appendix aims at presenting a novel approach to airline sentiment analysis processing using Twitter data. By transforming trained sentiment classifiers into regressors, the daily sentiment distribution obtained can be represented as a trimodal Gaussian Mixture leading to a simple but efficient classification algorithm. These classes can be considered as daily sentiment scores. This classification applied to passenger generated tweets and airline generated tweets for five major US airlines highlights major difference in experience between passengers and airlines. This methodology also confirms the existing gap between flight performance and passenger experience and the necessity of considering and implementing passenger-centric metrics.

Very few works actually propose an application of the classifiers output. Wang et al. [138] presented a framework to visualize real-time sentiment during political events in the United States using a crowd-sourced labeling method. Samonte et al. [141] proposed a sentiment analysis pipeline with some simple post analysis of the classification results and applied it to local airlines in the Philippines.

The contribution of this appendix is to propose a method to extract the daily sentiment distributions of passengers in such a form that it can then be analyzed to evaluate the airlines performance with respect to passengers, paving the way to a sentiment-based passenger-centric metric for the Air Transportation System.

The rest of the appendix is structured as follows: Section D.1.1 describes the methodology used to extract and process the daily sentiment distributions from the Twitter data. The analysis of the classification results is presented in Section D.1.2. Section D.1.3 concludes this study and discusses possible future steps.

D.1.1 Methodology

Data extraction

The Twitter dataset available for this study consists of all the tweets found using a basic search for each handle of 5 major US airlines, namely @united, @Delta, @AmericanAir, @SouthwestAir, @SpiritAirlines. Each entry consists of a timestamp, a user id, the content of the tweet and the handle used to retrieve the tweet. This dataset spans the entire period from January 1st

2018 to September 30th 2019.

It was then filtered to keep only tweets written in English using a two step process. The language of each tweet is initially taken as the own indicated by Twitter's API. The tweets labelled as "unknown" are then processed through the following language recognition algorithm and their language label are updated accordingly. Using the Natural Language Toolkit NLTK [149] and based on the work of Truica et al. [150], the number of common stop-words contained in a tweet is extracted for each available language in NLTK and the language with the highest count is selected. Due to the limited length of each tweet, a bias towards English has been introduced as well in the count ordering, i.e. if English and another language have the same count of common stop-words, English will have precedence.

Sentiment analysis

A first step in sentiment analysis is to clean the documents analyzed, here the tweets. This cleaning process was already performed in [10] and [11] and consists of the following steps: any reference to websites or pictures was replaced by a corresponding keyword. Every mention to another Twitter user within a tweet (@someone) as well as most emojis were similarly replaced. Since this database contains many replies from airlines to their customers, individual signatures of each agent were also replaced by a keyword. Dates and times were also generically replaced by keywords (e.g. "3rd Jan 2017" becomes "DATE" and "4pm" becomes "TIME"). The resulting text was then filtered from common stop-words and from the generic keywords used during the cleaning process.

Two different datasets were used to train three different classifiers each. The first dataset used was the labelled dataset used in a Kaggle competition [156]. The associated dictionary was created after removing words appearing in less than 20 tweets or in more than 75% of the full dataset. A second dataset and final cleaning process was generated based on the work of Read [115], also known as a distant supervised set used in many sentiment analysis models, with Go et al. [151] creating an impressive training set of 1,600,000 tweets. These tweets are from 2009 and are not specific to airline communication therefore this dataset was not considered here. Emoji filters were used to extract tweets from the initial dataset and automatically label them with a positive or negative sentiment according to Table 3.2 (page 28). The text cleaning process is also improved by merging negation words ("no", "not" and "never") with the word that follows it. The tokens used for the creation of the dictionary are the resulting bigrams, i.e. combinations of two words that follow each other in a tweet, with the same frequency filter as for the Kaggle

dataset.

For both methods, the scikit-learn library [148] was used to train the three classifiers considered, i.e. a random forest classifier, a naive Bayesian classifier and a logistic regressor. Once trained, the sentiment score used is the probability score of a tweet to be classified as positive, transforming in a way the classifiers into regressors. The final sentiment score is then the average of all six regressors and goes from 0 to 1, 0 indicating a negative tweet and 1 indicating a positive tweet.

Classifying using a Gaussian Mixture representation

Once the sentiment score is calculated for each English tweet, it is possible to extract the underlying distribution per day and per airline, assuming a Gaussian Mixture model. Sentiment analysis usually classifies texts as *positive*, *negative* or *neutral*, therefore a trimodal Gaussian Mixture model was assumed for each day of tweets and for each considered airline. Using a Bayesian Gaussian Mixture model [180] enabled to consider uni- and bimodal cases if relevant. A day of tweets can therefore be represented in a 9 dimension vector $(\mu_i, \sigma_i, \omega_i)_{i=1..3}$ such that its sentiment distribution can be approximated as following the following probability function:

$$P = \sum_{i=1}^{3} \omega_i \cdot \mathcal{N}(\mu_i, \sigma_i) \tag{D.1}$$

where $\mathcal{N}(\mu, \sigma)$ is normal gaussian probability function of mean μ and standard deviation σ.

A straight-forward classification method can then be derived based on these gaussian mixtures using the following algorithm. First, the distributions are cleaned from their modes with a weight ω_i smaller than 10% in order to make sure to capture all the uni- and bimodal distributions. Then, the unimodal distributions are split into two classes whether their mean is greater or lower than 0.5. The bimodal distributions are split into three classes depending on the location of their means: both lower than 0.5, both higher than 0.5 or one on each side of 0.5. Trimodal distributions are simply split into two classes depending on the location of its most weighted peak with respect to 0.5. The classes are summarized in Table D.1.

By construction, classes 3 and 6 can be clearly described as representing days with an overall *positive* mood, while classes 4 and 5 clearly represent days when a *negative* mood dominated. Class 2 can be seen as days where sentiments were polarized between *positive* and *negative*. Classes 0 and 1 would represent the normal situation where there are *positive*,*negative* and

Table D.1: Class description

Class	Distribution type	Categorization
0	Trimodal	$\mu_0 \leq 0.5$
1	Trimodal	$\mu_0 > 0.5$
2	Bimodal	$\mu_i \leq 0.5$ and $\mu_j \geq 0.5$
3	Bimodal	$\mu_i > 0.5$ and $\mu_j > 0.5$
4	Bimodal	$\mu_i < 0.5$ and $\mu_j < 0.5$
5	Unimodal	$\mu \leq 0.5$
6	Unimodal	$\mu > 0.5$

neutral tweets in various proportions without necessarily any one or two sentiments taking over.

Visualizing the sentiment space

Each vector $(\mu_i, \sigma_i, \omega_i)_{i=1..3}$ represents a point in the space of trimodal Gaussian Mixture probability functions, space in which the Euclidian distance is not relevant. A useful distance in this space is the Wasserstein distance [181], which can be understood as a transportation problem: The distance between two points P_1 $(\mu_{1i}, \sigma_{1i}, \alpha_i)_{i=1..3}$ and P_2 $(\mu_{2j}, \sigma_{2j}, \beta_j)_{j=1..3}$ in this space is equivalent to the minimal cost of moving the 'pile of earth' P_1 (represented by its probability density function) into the pile P_2. It amounts to solving the following Linear Programming problem:

$$\begin{aligned} & \min \quad \textstyle\sum_{i,j} x_{ij} \cdot d_{ij} \\ \text{s.t.} \quad & \forall j, \textstyle\sum_i x_{ij} = \beta_j \\ & \forall i, \textstyle\sum_j x_{ij} = \alpha_i \\ & \forall (i,j), x_{ij} \geq 0 \end{aligned} \tag{D.2}$$

where d_{ij} represents the Fisher information distance between the two normal distributions $\mathcal{N}(\mu_{1i}, \sigma_{1i})$ and $\mathcal{N}(\mu_{2j}, \sigma_{2j})$. The Fisher information distance d_F between two normal distributions $\nu_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\nu_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ is calculated as follows:

$$\mathcal{F} = \sqrt{((\mu_1-\mu_2)^2+2(\sigma_1-\sigma_2)^2)((\mu_1-\mu_2)^2+2(\sigma_1+\sigma_2)^2)} \tag{D.3}$$

$$d_F(\nu_1, \nu_2) = \sqrt{2} \ln \left(\frac{\mathcal{F} + (\mu_1-\mu_2)^2 + 2(\sigma_1^2+\sigma_2^2)}{4\sigma_1\sigma_2} \right) \tag{D.4}$$

Once this Wasserstein distance is defined, it can be used along with the t-Distributed Stochastic Neighbor Embedding (t-SNE) technique [182] in order to obtain a 2D representation of the space of trimodal Gaussian Mixture probability functions that preserves its implicit structure.

D.1.2 Results

The methodology presented in Section D.1.1 was applied to two different sets of tweets extracted from the initial database. These sets were created based on the writer of the tweets, separating tweets coming from passengers versus tweets coming from the airline account.

Classification results

Counting the number of days related to each airline for every class yields some interesting insights regarding the composition of each class and the difference between passenger tweets and airline tweets. These airline distributions are plotted in Figure D.1 & D.2.

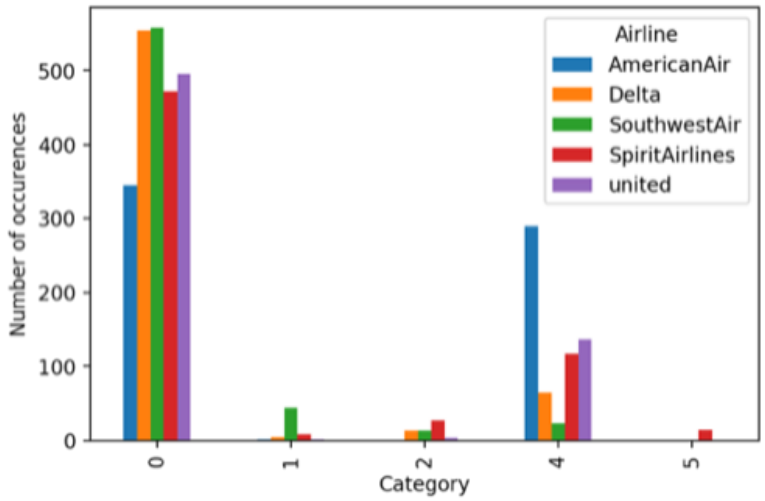


Figure D.1: Airline distribution per class for passenger tweets.

A first takeaway from the passenger perspective in Figure D.1 is that none of the *positive* classes (i.e. classes 3 and 6) are represented during the considered period. One class gathers a total of 76.0% of airline-days: class 0. This indicates that passenger sentiment is usually split between the three modes, although with a bias towards a *negative* mood. The second largest class is class 4, the class with two *negative* modes, with 19.7%. The split between these two classes is similar for four of the five considered airlines with around 500 days in class 0 and 100 days or less in class 4, whereas American Airlines has an rather even split of 300 days for each class. This indicates that American Airlines passengers have the highest ratio of displeasing days, close to 1/2. Spirit Airlines is the only airline with days in class 5, representing days where passengers are overall in a similar *negative* mood.

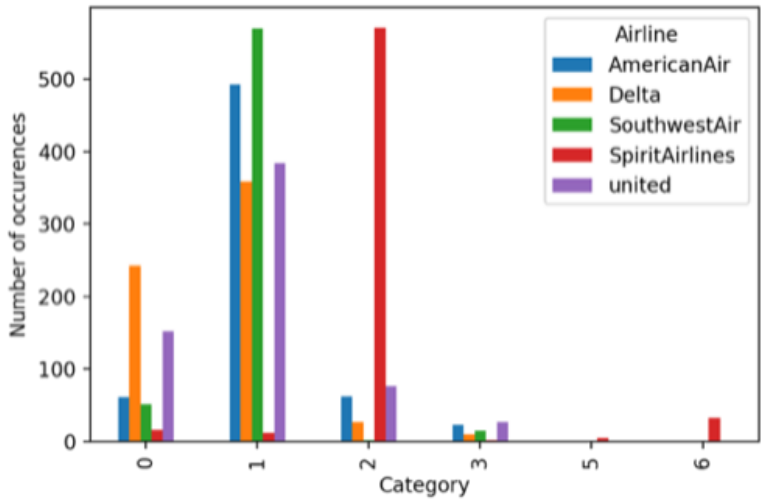


Figure D.2: Airline distribution per cluster for company tweets.

From an airline perspective, Figure D.2 tells a different story: in clear contrast with the passenger class distribution, in the case of airline tweets, the *negative* classes (i.e. classes 4 and 5) are not or barely represented, with only five days in class 5 for Spirit Airlines. This indicates the opposition between how situations are experienced and expressed by passengers and how they are mitigated by the airline communications.

Regarding the main classes for airlines, class 1 concentrates 57.0% of airline-days, followed by class 2 with 23.1% and class 0 with 16.4%. Classes 1 and 2 have however opposite compositions: Spirit Airlines holds for around 75% of class 2 while being almost absent from class 1. This indicates that Spirit's communication contains more tweets conveying a *negative* mood than the other airlines. As for the passenger perspective, Spirit Airlines is also the only airline with days in class 5, which would indicate days when the airline twitter feed were essentially conveying a negative *mood*. Spirit Airlines is however the only airline with days in class 6, indicating that it is able to convey a *positive* mood on certain days.

The total number of airline-days per class is resumed in Table D.2. This representation highlights the quasi-orthogonality of the two perspectives: classes with high representation for airlines are comparatively empty from a passenger perspective and *vice versa*.

It is also possible to compare the daily class of these two perspectives day by day, in order to better visualize the opposition between passenger expressed experience and airline customer communication. Table D.3 shows the correspondence between airline classes and passenger classes. It is worth

Table D.2: Total number of airline-days per class.

Class	0	1	2	3	4	5	6
Airline	521	1816	736	75	0	5	32
Passengers	2422	61	58	0	630	14	0

noting that for five days where the airlines are in class 6 (i.e. a unimodal *positive* mood), the passenger daily sentiment is in class 5 (i.e. a unimodal *negative* mood), another example of the opposite perception between airlines and passengers. Similarly, days when airlines are in class 3 (i.e. a bimodal *positive* mood) are perceived and expressed by passengers as belonging to mood classes with a *negative* bias (classes 4 and 0). On the opposite, days when airlines express a more *negative* mood in class 0 are also perceived as mainly negative by passengers with 80.8% in class 0 and 17.3% in class 4.

Table D.3: Class correspondences between passenger and airline perspectives.

Airlines \ **Passengers**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
0	421	4	6	0	90	0	0
1	1384	46	24	0	361	1	0
2	536	8	26	0	159	7	0
3	56	3	1	0	15	0	0
5	3	0	0	0	1	1	0
6	22	0	1	0	4	5	0

Class visualization

The 2D representation of the daily sentiment distributions using the distance introduced in Section D.1.1 along with a color code for their associated classes are shown in Figure D.3 for passengers and in Figure D.4 for airlines. In these figures each point represents a day of tweets for one of the considered airlines.

In Figure D.3, as expected from the previous results, the dominant class 0 spans the full space and encircles the other classes. Though the classes were not constructed by clustering, all classes are clearly separated from the others, with the exception of class 1 and a few outlying points of the other classes. The fact that class 1 is scattered within the class 0 cluster advocates toward a sensitive frontier between these two classes from a passenger perspective. Class 5 is concentrated in a small area in this representation space, whereas class 4 is more spread out. This indicates that the days with a distribution

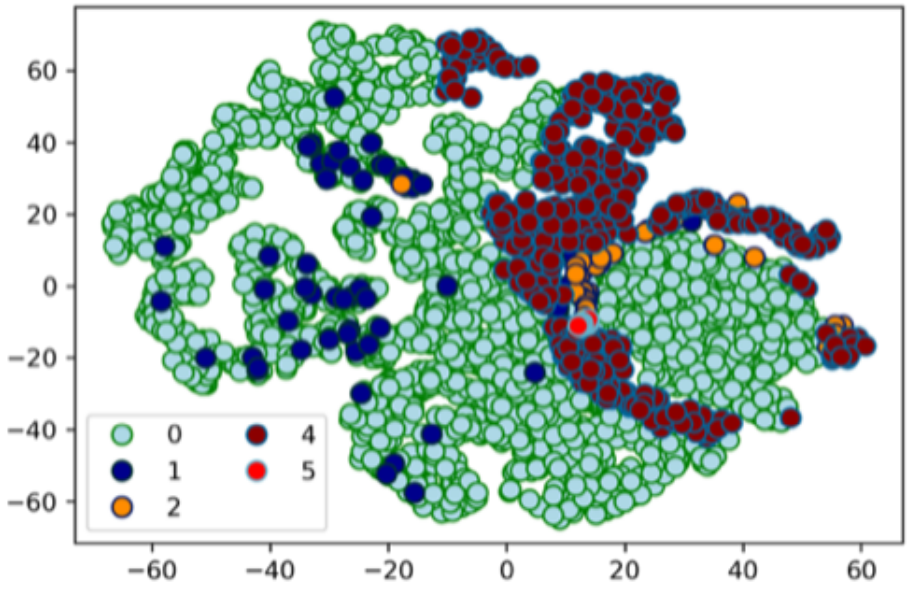


Figure D.3: A 2D clustered representation of daily sentiment distribution of passenger tweets in a reduced dimension based on the Wasserstein distance.

mood unimodal and *negative* (class 5), this distribution did not vary much from one day to another. In other words, it is sufficient to look at the mean of one of these days to have a good estimation of the other class 5 day means. Class 4 being more spread out, the most representative day of the class has to be found by another mean.

From an airline perspective, shown in Figure D.4, the frontier between classes 0 and 1 is clearly defined. Further investigations should look into this frontier to know which tweet formulations should be avoided by airlines in order to stay in the better of the two classes, class 1. Class 2 is also clearly separated from classes 0 and 1, but is overlapped by the most positive classes, classes 3 and 6. Recalling that class 2 was dominated by Spirit Airlines, this overlapping suggests that the airline is aiming for a *positive* messaging but fall shorts of achieving it.

In order to find the day best representing each class, the Wasserstein distance can be used again to compute the central distribution of each class, i.e. the distribution that has the smallest average distance to all the other points. These distributions are plotted in Figure D.5 for the passenger dataset and in Figure D.6 for the airline dataset. The distribution equation, with each parameter rounded at 10^{-3}, is indicated on top of each subfigure for information.

Comparing the central distribution of a same class but from the two

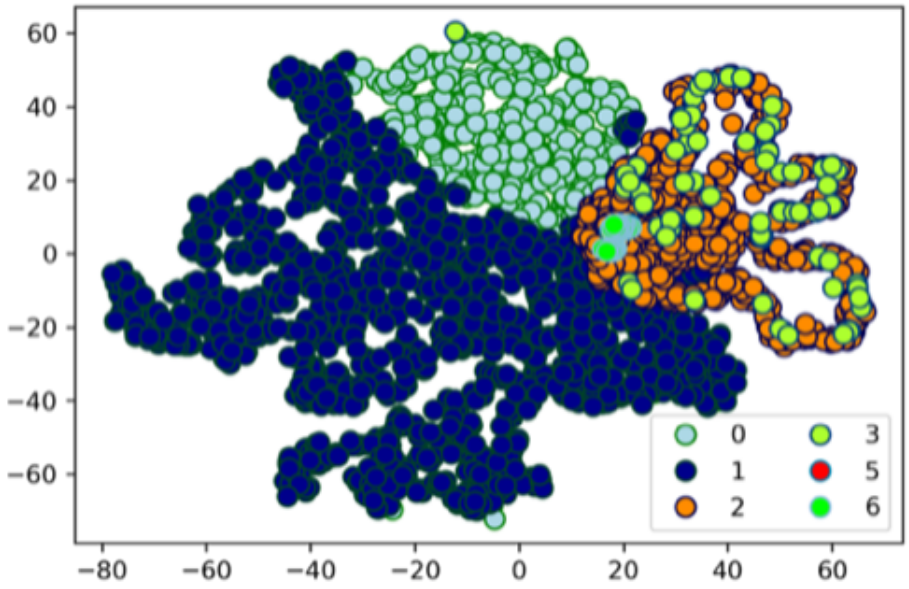
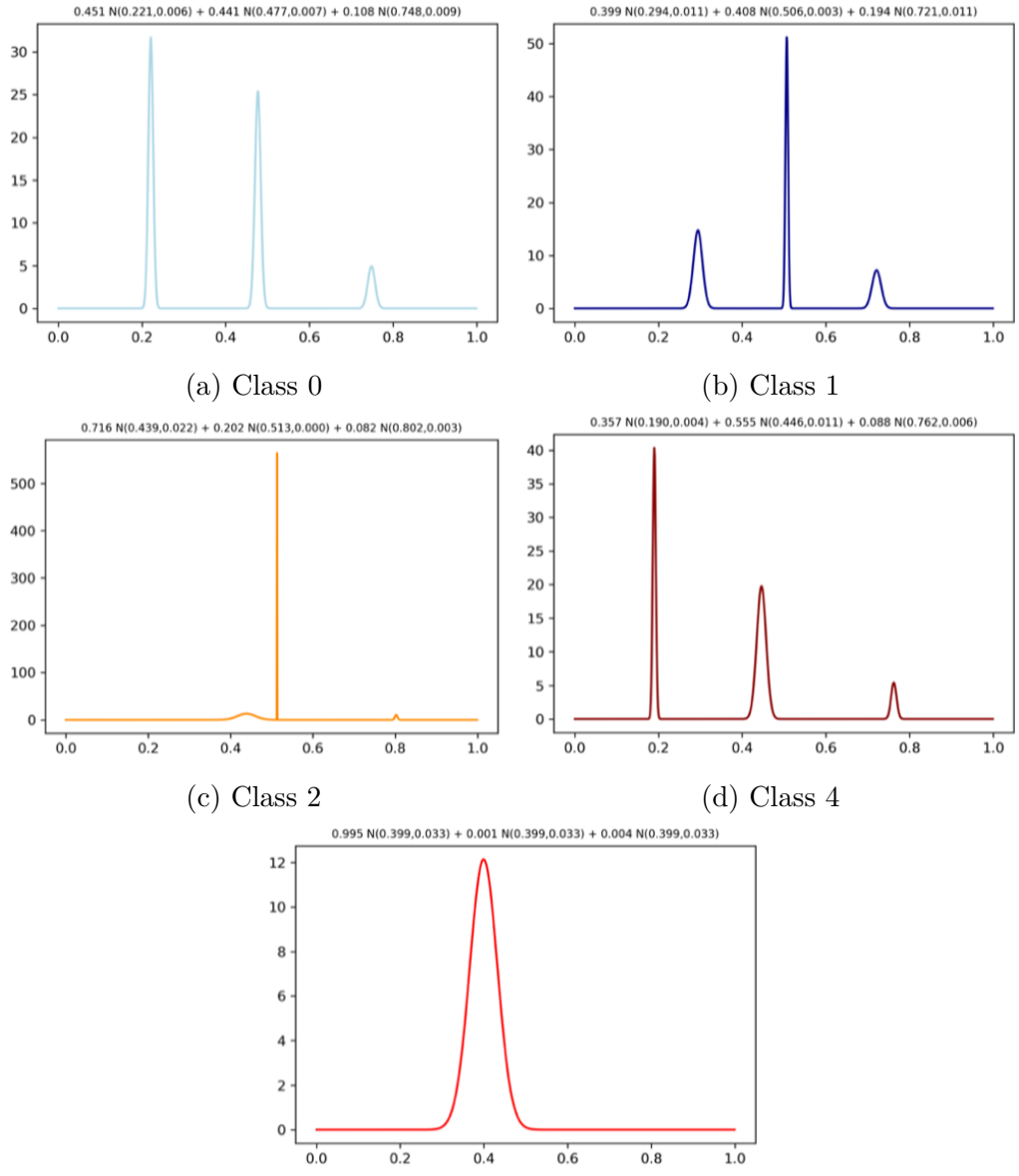


Figure D.4: A 2D clustered representation of daily sentiment distribution of airline tweets in a reduced dimension based on the Wasserstein distance.

available perspectives draws the conclusion that though the class definition does not change, its representation varies drastically from one perspective to another. For example, the centroid of class 0 for passenger tweets has two modes on the *negative* side, whereas the the centroid for the airline tweets has two modes on the *positive* side, though the mode with the highest weight is the negative one by construction. Regarding the unimodal and *negative* class 5, it's mean is closer to the *positive* side for airlines than it is for passengers. Similarly, for class 1 the main mode mean is closer to the *negative* side for the passenger class centroid than for the airline one. The same can be said for class 2 and it's main positive mode.

Passenger experience versus flight performance

Currently the air transportation system is essentially evaluated using flight-centric metrics such as flight delay, and lacks passenger-centric metrics. The class defined in this appendix can help put in perspective the difference between these two approaches. Flight departure information over the considered period were extracted from the Bureau of Transportation Statistics (BTS) website. After analyzing and testing different distributions, the Student's T continuous distribution was kept as best fitting the daily delay distributions. Here a delay can be negative, meaning that the flight left earlier



(a) Class 0

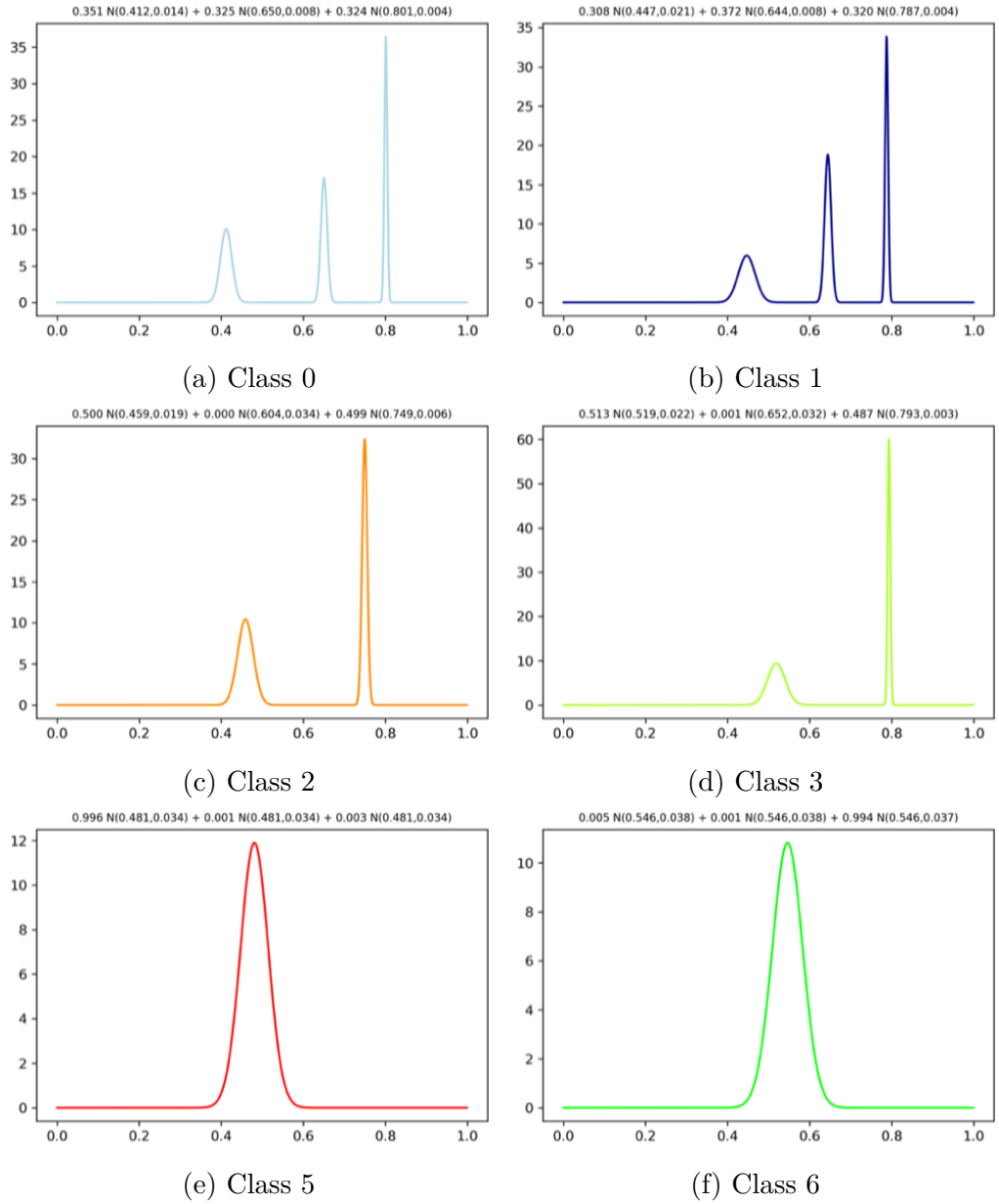
(b) Class 1

(c) Class 2

(d) Class 4

(e) Class 5

Figure D.5: Gaussian mixture representation of the class centroids for passenger tweets.



(a) Class 0 (b) Class 1

(c) Class 2 (d) Class 3

(e) Class 5 (f) Class 6

Figure D.6: Gaussian mixture representation of the class centroids for company tweets

than the scheduled departure time. It is then possible to plot in a 2D plane the different days in the delay space using the location and scale parameters associated. The location parameter represents how much the distribution is shifted from 0 and the scale parameter gives an information on the width of the distribution. Figure D.7 shows the airline daily delay distributions in this 2D plane along with a color code associating each day to its passenger sentiment class.

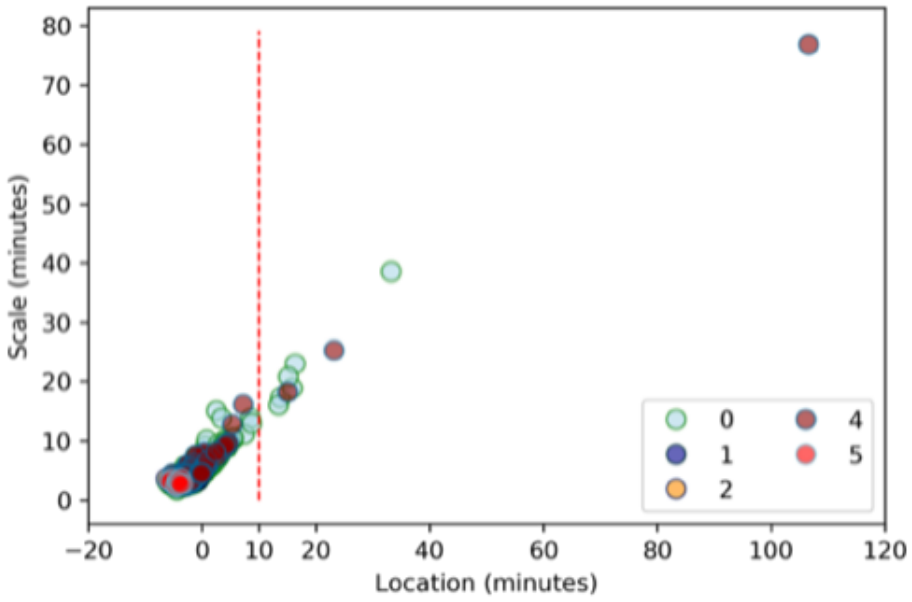


Figure D.7: A 2D representation of the daily distributions of the amount of delay with the associated passenger sentiment class color code as in Figure D.5.

Looking at Figure D.7, there are nine days with a location greater than ten minutes separated in two classes, with three days in the clearly *negative* class 4 and six days in the main class 0. This indicates that airlines managed to mitigate the effect of delays on passenger mood for six of these nine days. On the opposite spectrum, Figure D.8 zooms into days with a delay location of less than ten minutes. What appears clearly here is that days with good flight performance, e.g. days with a negative average delay and a low scale are not necessarily experienced as positive for passengers. More precisely, all the class 5 days are located in this good flight performance zone, indicating that leaving early is not necessarily well perceived by passengers. Most of the class 4 days (89.5%) are days with a negative location and a scale lower than 5 minutes, highlighting the opposition between flight performance and

passenger experience.

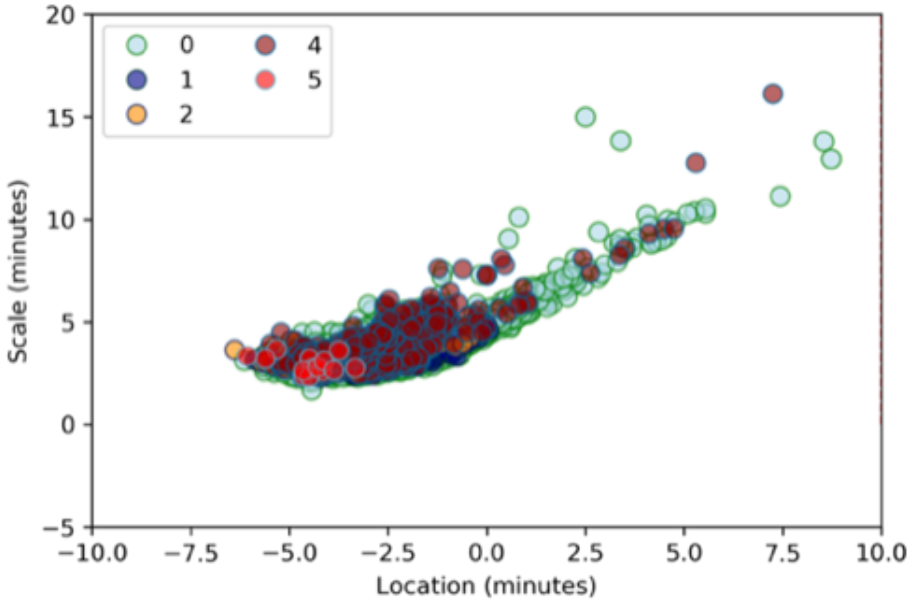


Figure D.8: Zoom into the 2D representation of the daily distributions of the amount of delay with the associated passenger sentiment class color code as in Figure D.5.

A similar representation is shown in Figure D.9 using the airline sentiment class color code. The near totality (97.1%) of the two positive classes 3 and 6 concern days with a negative location and a scale lower than 5 minutes. This concentration suggests that important delays does have an impact on airline communication, in the sense that they cannot afford to express a mood too *positive* with respect to their customers.

D.1.3 Conclusion

This appendix aimed at presenting and leveraging a novel method for processing results from airline sentiment analysis applied to Twitter. Once sentiment classifiers are trained on well defined datasets, transforming them into regressors allows to obtain a Gaussian Mixture representation of the daily sentiment distribution. This representation can then be easily categorized in seven classes clearly defined and with an understandable signification. Separating and comparing the analysis of passenger generated tweets with airline generated tweets highlights the opposition in perception and experience of air travel between passengers and airlines. This opposition is even more visible

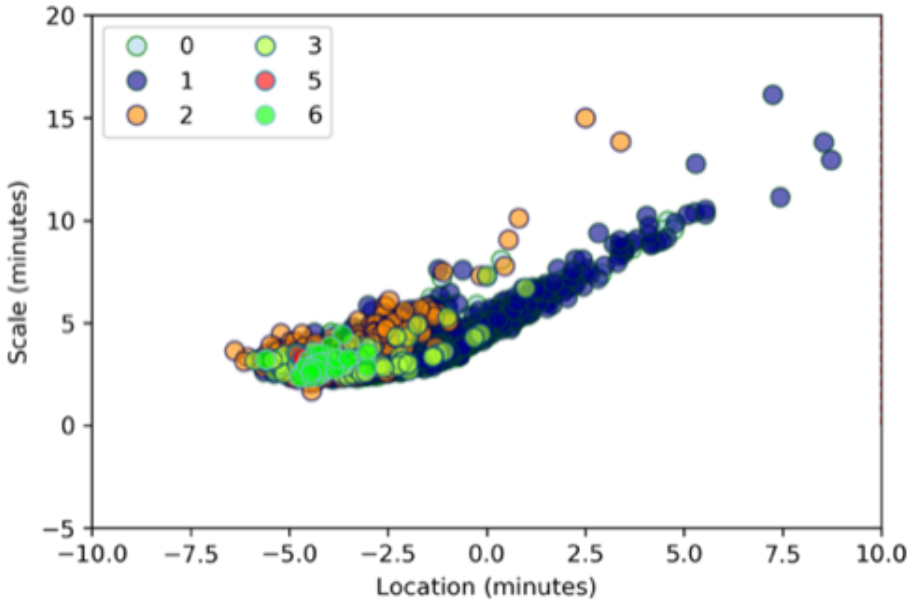


Figure D.9: Zoom into the 2D representation of the daily distributions of the amount of delay with the associated airline sentiment class color code as in Figure D.6.

when comparing these sentiment classes to the usual flight-centric metrics, since it clearly shows that on time and early departures are not a sufficient condition for a *positive* passenger experience.

Future studies should focus on the frontier between the different sentiment class, in order to better understand when and how a day shifts between *positive* and *negative* classes, enabling airlines to prevent unwanted class shifts and thus improving passenger experience.

D.2 Doorway to the United States: An Exploration of Customs and Border Protection Data

This appendix presents a data-driven study of wait time patterns for international arriving passengers across all 61 terminals from the 44 airports of entry of the United States and that was presented in [16]. Each airport is an independent entity which operates with various airlines and handles demand volumes differently. This induces seasonal variation in service quality from one airport to another. Exploring six years worth of data, this appendix investigates the current and long-term performance trends - an increasing number of flights versus a decreasing number of customs booths - of all airports of entry from a passenger perspective. A performance analysis is then conducted that compares average wait times of incoming passengers, considering incoming traffic ratios and allocated resources. Leveraging machine learning algorithms, six regression algorithms are trained and tested to accurately predict passenger wait times through customs at selected airports. An analysis of the performance of these models shows that the best approach - using a Gradient Boosting regressor for each terminal of entry - can capture the daily and seasonal variations of traffic patterns and immigration booth availabilities with a mean absolute error of less or equal to 5 minutes for twenty-eight terminals of entry and less than 10 minutes for all terminals. Observations show significant disparities across airports that may be explained by the foreign/US passenger ratio and the quality of booth management.

D.2.1 Introduction

Foreign international air travelers arriving in the U.S. spend billions of dollars while visiting.

Wait times at security and border patrol play a big role in assessing passenger satisfaction : capacity constraints and inefficiencies at airport entry roads, parking, security, immigration, customs, gates, ramp areas, runways are the primary causes of congestion and of the ensuing delays. Since September 11, 2001, airport screening procedures in the U.S. have been continuously evolving. For example, the passenger screening process is now trying to strike a balance between security and customer service (i.e. minimizing wait times). Using data from 2002 and 2003, Gkritza et al. [54] showed that, while wait times at security screening points are significant determinants of passenger satisfaction, many other factors come into play. Torres et al. showed that

consumption of goods and services grows with the time spent by passengers in the leisure areas. [183]. Thus less time spent queuing at various control points may result in time and financial benefits for everyone.

Moreover, delays on the ground have a disproportionate impact for passengers who often experience lengthy delays before being re-booked if they miss their connecting flights [65]. For international passengers arriving from overseas to the United States, immigration checks are mandatory, whether they are U.S. citizens or foreigners.

This appendix focuses on the passenger experience while going through U.S. customs and border protection. Roberts et al. studied the evolution of wait times at airports over a few years [184]. They showed that average passport inspection wait time at 24 U.S. airports rose by 25% during 2010-2013. They focused on JFK airport. At JFK, nearly 3 million passengers (25 percent of total arrivals) experienced a delay of more than 1 hour, putting them at risk of missing a connecting flight, with the 11 percent who had a total delay of more than 2 hours, possibly missing connections at a higher rate. Extended passport inspection waits were the sole source of missed connection risk for 13 percent of passengers and one of the reasons behind missed connections for many more passengers.

There has been little research in the systematic analysis of passenger wait times at customs across airports. Besides the works presented in Section 2.2, Sankaranarayanan et al. [185] performed an exploratory analysis of airport wait times on customs, border protection data taken from top 3 busiest airports (Atlanta, Chicago, and Los Angeles) from the United States, highlighting the effects of seasonality. Johnstone et al. [186] presented a dynamic queue controller to generate realistic queue formation and behavior within a discrete event environment at airports in Australia.

However, to the best of the authors' knowledge, no work is available on comparing performance across airports or predicting performance at any airport.

The present appendix leverages publicly available data from the United States Customs and Border Protection (CBP) [165]. As stated on their website, "CBP closely monitors the flight processing times, commonly referred to as wait times, for arriving flights at the busiest international airports." The data provided in the online reports show the number of passengers processed on flights arriving in each hour based on how long it took for those passengers to clear Passport Control.

This appendix tackles the following research questions:

- Which are the best airports to enter the U.S. for U.S. citizens and foreigners, in terms of wait times?

- Which airports best manage their customs area?
- Can wait times per hour at airports be reliably predicted from historical data?

The appendix is organized as follows. Section D.2.2 explores the main trends at different scales visible in the CBP data, from average wait times, to passenger volumes and flight volumes. Section D.2.3 compares the performances of the different airports of arrival. Section D.2.4 proposes a machine learning approach to predict passenger wait time per hour at any airport and examines the performance of this approach. Section D.2.5 details the conclusion of the paper and offers future research perspectives.

D.2.2 Exploration

Dataset contents

The data from CBP [165] contains the following fields once a time period of interest is selected:

- Airport Name,
- Terminal number,
- Date,
- Hour,
- Average wait time for U.S. citizens,
- Average wait time for non-U.S. citizens,
- Maximum wait time for U.S. citizens,
- Maximum wait time for non-U.S. citizens,
- Average wait time for all passengers,
- Maximum wait time for all passengers,
- Number of passengers who waited less than 15 minutes,
- Number of passengers who waited 16 to 30 minutes,
- Number of passengers who waited 31 to 45 minutes,
- Number of passengers who waited 46 to 60 minutes,
- Number of passengers who waited 61 to 90 minutes,
- Number of passengers who waited 91 to 120 minutes,
- Number of passengers who waited over 120 minutes,
- Total number of passengers, both U.S. citizens and non-U.S. citizens,
- Number of flights,
- Number of open immigration booths.

Considering the data from 2013 to 2019, the dataset consists of 1,201,181 entries corresponding to 61 terminals within 44 airports. The different terminals are summarized in Table D.4.

Table D.4: Terminal of arrivals abbreviations

Abbreviation	Airport (IATA)	Terminal name	Abbreviation	Airport (IATA)	Terminal name
ATL - CE	ATL	Concourse E	MCO - A1	MCO	Airside 1
ATL - CF	ATL	Concourse F	MCO - A4	MCO	Airside 4
AUS - M	AUS	Main	MDW - MT	MDW	Main Terminal
BOS - TE	BOS	Terminal E	MIA - CT	MIA	Central Terminal
BWI - IA	BWI	International Arrivals	MIA - NT	MIA	North Terminal
CLT - M	CLT	Main	MIA - ST	MIA	South Terminal
DEN - I	DEN	International	MSP - T1L	MSP	Terminal 1 Lindbergh
DFW - TD	DFW	Terminal D	MSP - T2H	MSP	Terminal 2 Humphrey
DTW - MT	DTW	McNamara Terminal	OAK - M	OAK	Main
DTW - NT	DTW	North Terminal	ORD - T5	ORD	Terminal 5
EWR - TB	EWR	Terminal B	PBI - M	PBI	Main
EWR - TC	EWR	Terminal C	PDX - M	PDX	Main
FAT - M	FAT	Main	PHL - TA	PHL	Terminal A
FLL - T1	FLL	Terminal 1	PHX - M	PHX	Main
FLL - T4	FLL	Terminal 4	RDU - T2	RDU	Terminal 2
GUM - MT	GUM	Main Terminal	SAN - M	SAN	Main
HNL - MOT	HNL	Main Overseas Terminal	SAT - M	SAT	Main
IAD - IA	IAD	International A	SEA - SS	SEA	South Satellite
IAH - I	IAH	IAB	SFB - TA	SFB	Terminal A
JFK - T1	JFK	Terminal 1	SFO - TA	SFO	Terminal A
JFK - T4	JFK	Terminal 4 (IAT)	SFO - TG	SFO	Terminal G
JFK - T5B	JFK	Terminal 5 (Jet Blue)	SJC - M	SJC	Main
JFK - T7	JFK	Terminal 7 (British)	SJU - SJA	SJU	San Juan AA
JFK - T8	JFK	Terminal 8 (American)	SLC - M	SLC	Main
LAS - T3	LAS	Terminal 3	SMF - MT	SMF	Main Terminal
LAX - S2	LAX	Satellite 2	SNA - TC	SNA	Terminal C
LAX - S5	LAX	Satellite 5	SPN - IA	SPN	International Arrivals
LAX - S7	LAX	Satellite 7	STL - M	STL	Main
LAX - T4	LAX	Terminal 4	TPA - M	TPA	Main
LAX - TBIT	LAX	Tom Bradley International Terminal			

Long term evolution

Looking at the overall evolution from 2013 to 2019, some clear trends appear as illustrated in Figures D.10 & D.11. Figure D.10 shows the evolution of the total number of arriving international flights per day versus the total number of open booths per day. While the number of flights is steadily increasing

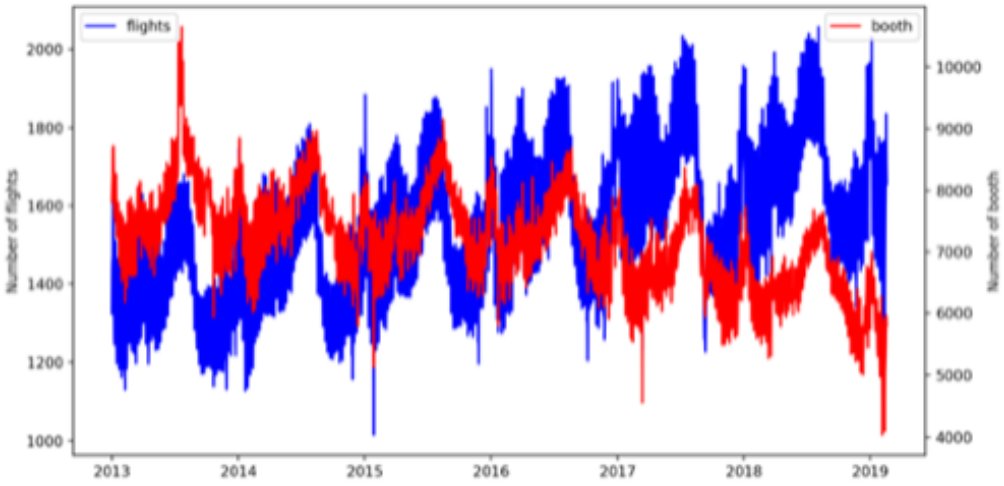


Figure D.10: Comparison of the total number of open booths (red) vs. the total number of arriving flights (blue) per day from January 2013 to January 2019

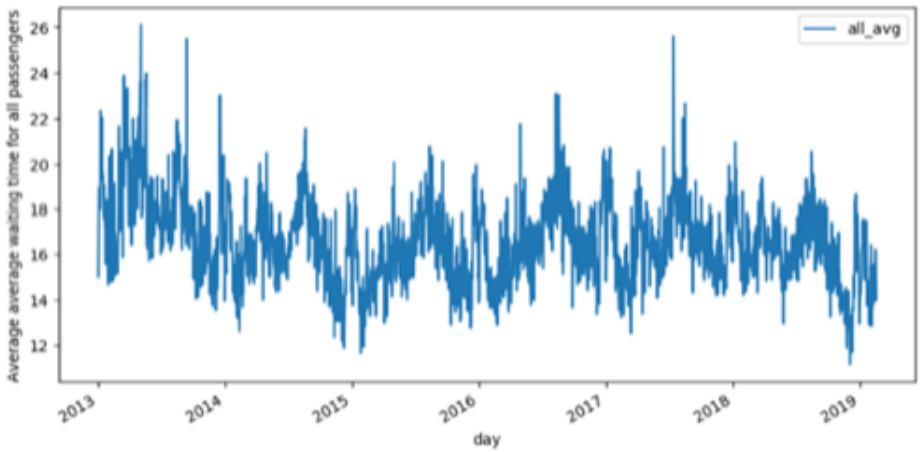


Figure D.11: Average wait time for all passengers per day across all airports from January 2013 to January 2019.

over the years, the number of open booths is slowly decreasing.

Figure D.11 depicts the daily average wait time per passenger across airports. Wide variations can be noted, from a minimum of 11 minutes to a maximum of 26 minutes. Seasonal variations are present during the winter and summer holiday season. However, starting from 2015, the amplitudes of these yearly seasonal variations do not vary much over the years.

This observation is better visualized in Figure D.12, which shows a yearly comparison of the average hourly wait time between 2013 and 2019. With the exception of 2013, this overall average wait time follows the same seasonal variations from one year to another. Longer wait times are observed for the winter (end of December - beginning of January) and summer (August) holidays as well as around April.

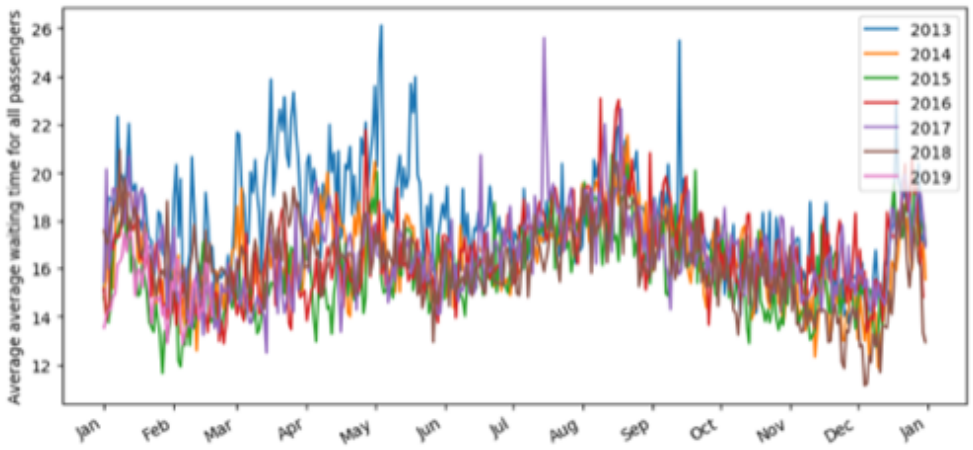


Figure D.12: Yearly comparison of the average wait time for all passengers per day across all airports

Figure D.13 shows the box-and-whisker plot variations of the average wait time across all terminals per month, i.e. for each month it shows the median average wait time and the first and third quartile along with whiskers for a better visualization of the range of the data. This figure highlights the four-month periodic variation discovered in Figure D.12. There are three highs during a year - around the winter holidays from December to January, April and August - interlaced with periods with shorter wait times. This four-month periodic behavior does not seem to be induced by a similar behavior in the number of arriving passengers, as shown in Figure D.14.

Figure D.15 shows the evolution of average wait time for all passengers per hour of the day from 2013 to 2019. As previously observed, the average time for 2013 is higher than all the others, which are tightly packed. The average wait time for 2019 is lower than the previous years since only data

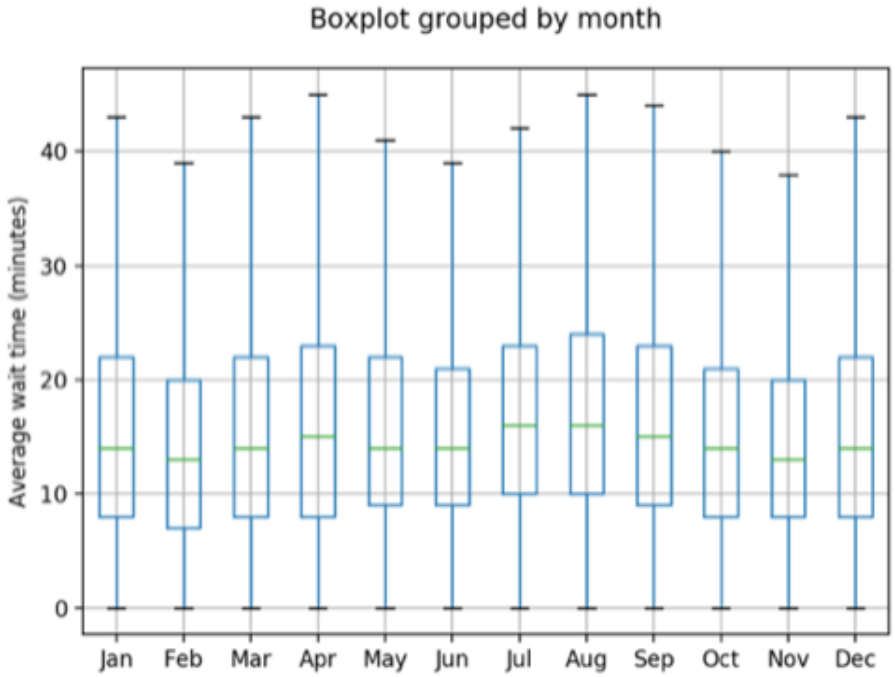


Figure D.13: Boxplots per month of the hourly average wait time for all passengers across all terminals from 2013 to 2019

from January are available at the time of this study.

Figure D.16 shows how long passengers typically wait per day of the week. No clear trend is distinguishable for a particular day of the week.

Wait time differences

The trends presented so far were for any passenger. Yet, U.S. citizens and foreigners enter the U.S. through separate lines, and statistics on wait times are available for each category specifically. For the same number of passengers, processing foreigners at a booth typically takes longer, since it requires checking more paperwork, such as supporting entrance documents and visas, whereas U.S. citizens only need to present their passports. The average wait time for all passengers over the year 2017 is 16.7 minutes and 16.2 minutes over the year 2018. Figure D.17 shows the break down by U.S. passengers and non-U.S. passengers for these two years. Non-U.S. passengers on average spend twice as much time in line at immigration, and experience higher volatility in wait times.

Figure D.18 shows the different wait times distribution per year for US citizens and non-US citizens from 2013 to 2019. The average wait time for

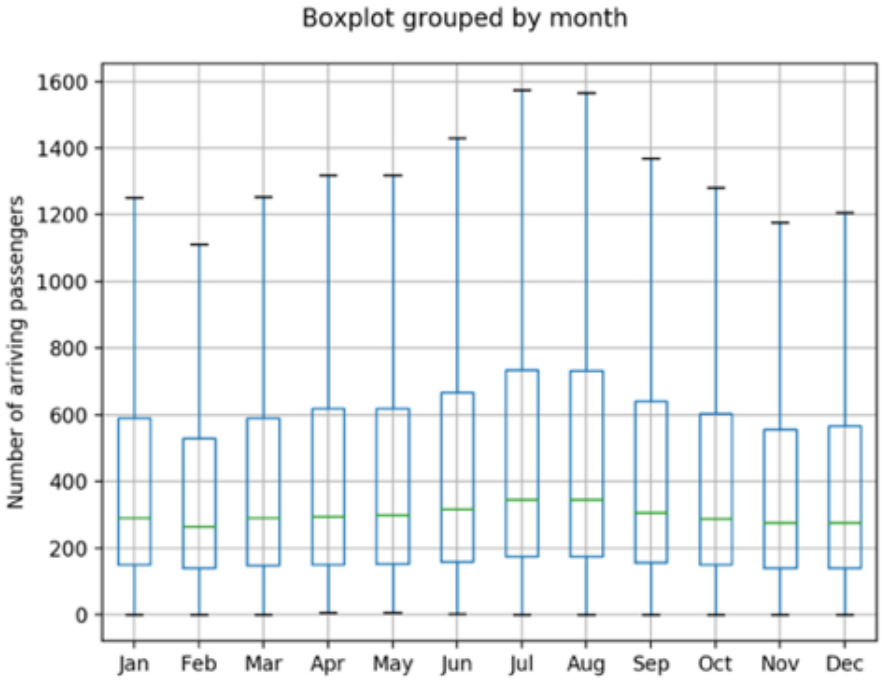


Figure D.14: Boxplots per month of the number of hourly arriving passenger across all terminals from 2013 to 2019

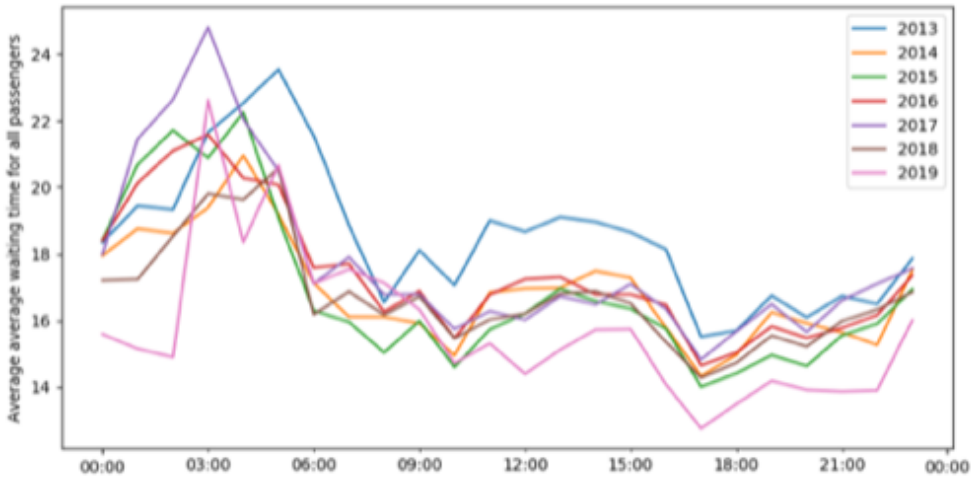


Figure D.15: Yearly comparison of the average wait time for all passengers per hour across all airports

non-US citizens does not vary much over these years, while being twice as more important as for US citizens throughout these years. Though 2019

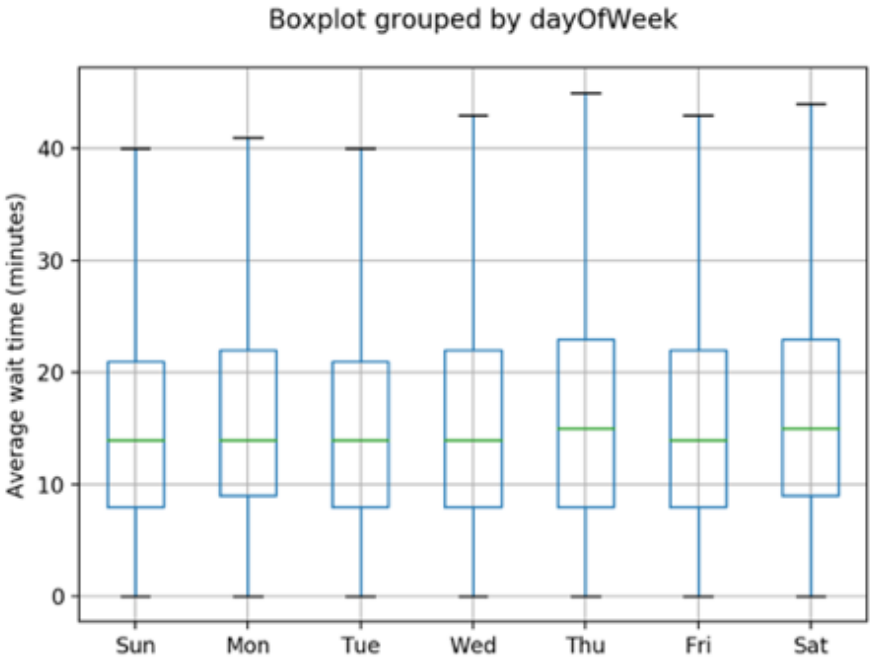
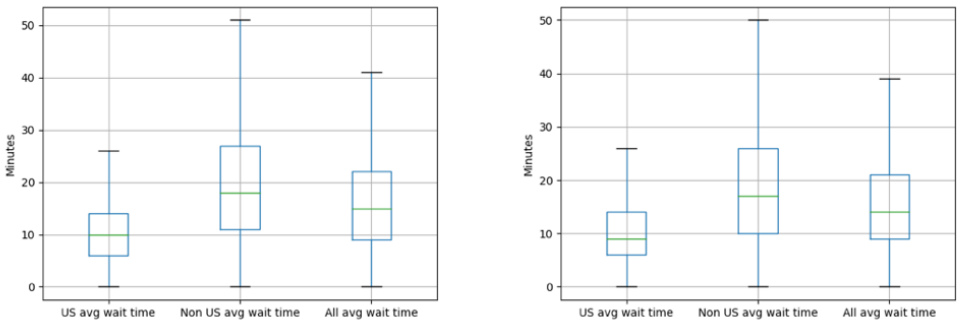


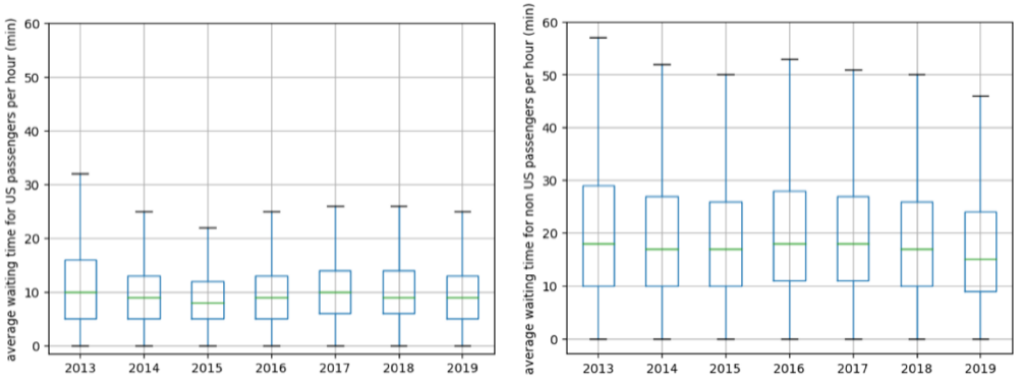
Figure D.16: Average wait time distribution per day of the week from 2013 to 2019



(a) On year 2017 (b) On year 2018

Figure D.17: Average wait time distribution for passengers from January to December for the years 2017 and 2018.

seems to be better, observations from Section D.2.2 indicates that the month of January is not representative of the yearly distribution.



(a) US passengers (b) Non US passengers

Figure D.18: Average wait time and standard deviation for passengers from 2013 to 2019 across all airports

D.2.3 Airports comparison

Overall comparison

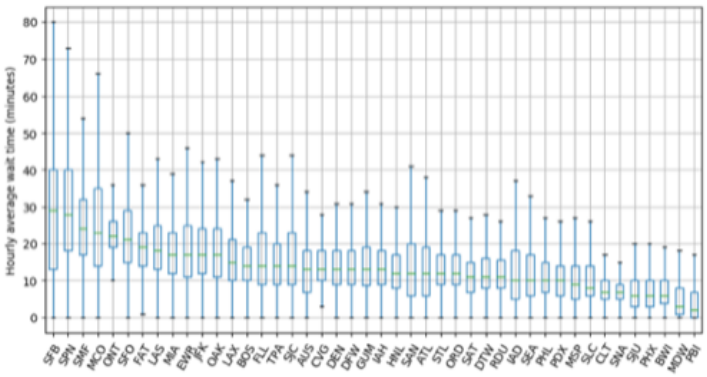
Figure D.19(a) shows the yearly box-plots per airports of the average wait time for all passengers for the year 2018. Interestingly both Orlando airports in Florida (SFB and MCO) show particularly high wait times with high volatility. Second in line for overall longest wait times is Hawaii (SPN), before airports from the San Francisco area (SMF and SFO). Ontario International airport (ONT) is not considered in this top five due to its low volatility compared to SFO.

This high wait time is not necessarily due to lack of means. Figure D.19(b) shows the distribution of the number of open booths per airport for the year 2018. SFO has one of the highest median of open booths, which contrasts with its wait time performance.

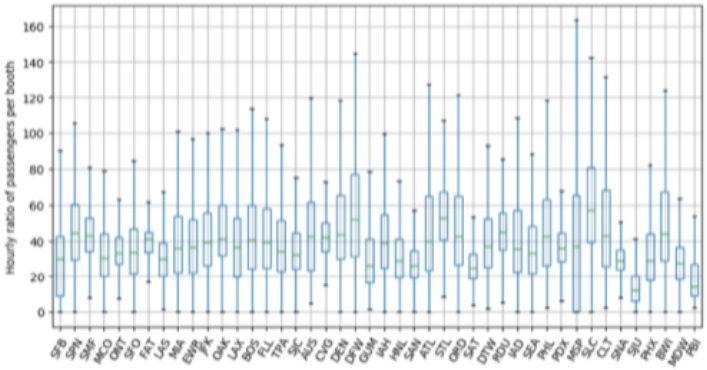
Figure D.19(c) shows the distribution of the number of arriving passengers per airport for the year 2018. A first observation is that airports with high passenger volume volatility have in general a high flexibility regarding the number of open booths.

Worst and Best case scenario

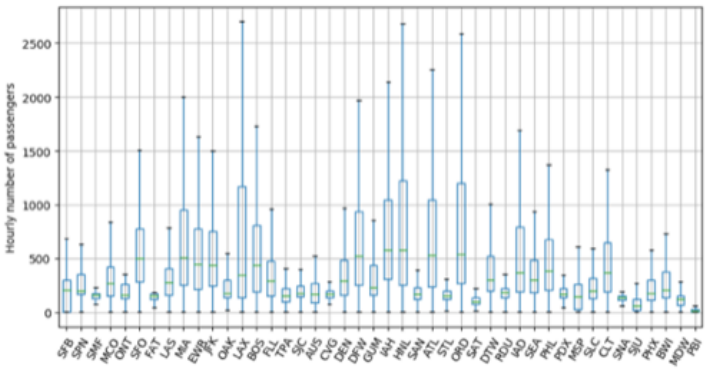
From Figure D.13 & D.14, one can infer that the worst month for entering the United States is August: the highest median average wait time with high volatility combined with one of the largest distribution of arriving passengers. Figure D.20(a) shows the distribution of the average wait times for all passen-



(a) Comparison of the average wait time per hour across airports over the year 2018



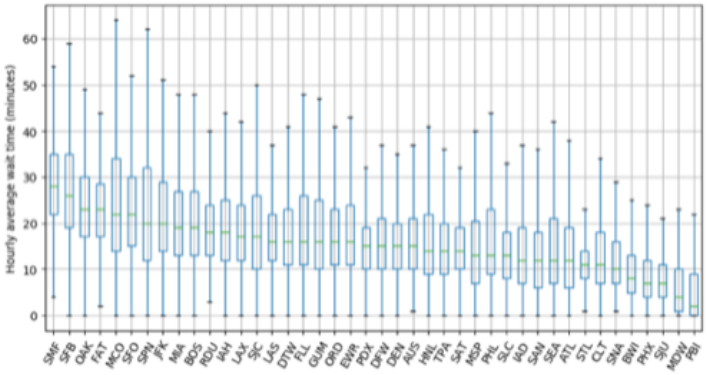
(b) Comparison of the number of open booths per hour across airports over the year 2018



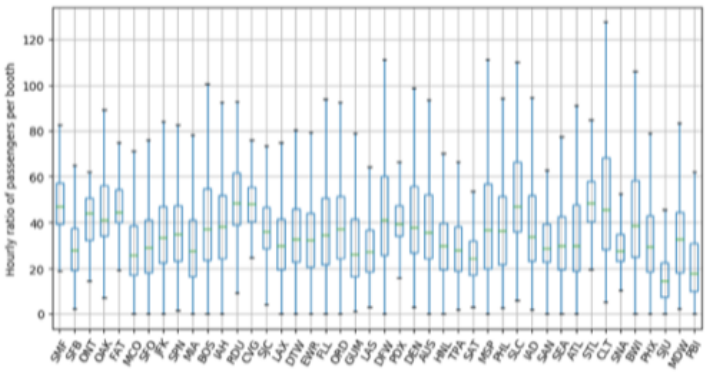
(c) Comparison of the number of arriving passengers per hour across airports over the year 2018

Figure D.19: Airport comparison using boxplots over the year 2018

gers per airports in August over the last six years. Airports ONT and CVG were removed since there is only two years of data for these two airports. The same two airports from Orlando Florida appear in the top five worst performing airports. On the West coast, the previously spotted airports are joined by OAK and FAT join Orlando in the top six worst performing airports.



(a) Comparison of the average wait time per hour across airports



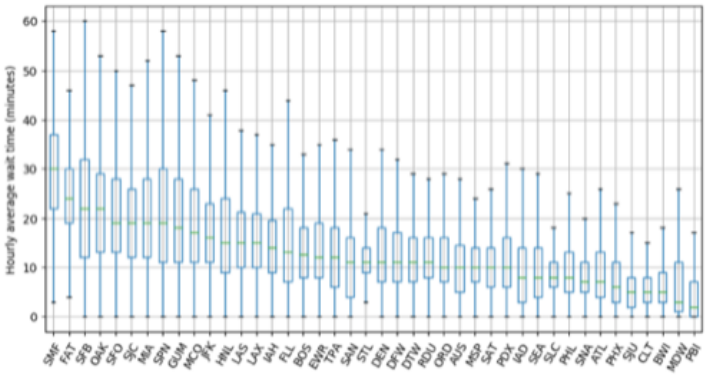
(b) Comparison of the number of arriving passengers per hour across airports

Figure D.20: Airport comparison using boxplots for the month of August from 2013 to 2019

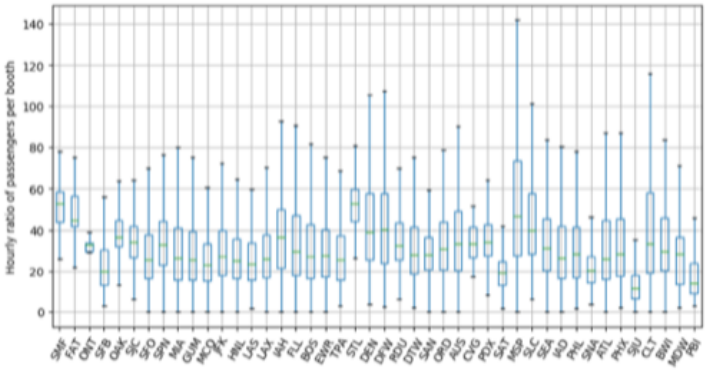
The performance of the first four airports are even worse when considering the number of arriving passengers shown in Figure D.20(b). They do not have the highest median number of arriving passenger per hour nor the highest volatility.

On the other hand, from Figure D.13 & D.14, February appears like the best month to enter the US: lowest average wait time and volatility as well as the lowest volume of arriving passengers. Figure D.21(a) shows the distri-

bution of the average wait times for all passengers per airports in February over the last six years. Palm Beach International airport (PBI) and Chicago Midwest International airport (MDW) have the lowest hourly wait times on average but with a comparatively high volatility. Second best in line are Charlotte Douglas International airport (CLT) and Baltimore/Washington International airport (BWI) with a combination of low average and low volatility.



(a) Comparison of the average wait time per hour across airports



(b) Comparison of the number of arriving passengers per hour across airports

Figure D.21: Airport comparison using boxplots for the month of February from 2013 to 2019

When combining these observations with the number of arriving passengers distribution shown in Figure D.21(b), CLT has more merit seeing how few passengers enter through PBI during that month of the year.

US vs. non-US wait times

Figure D.22(a) & D.22(b) show the distribution of the average wait times for US and non-US passengers per airports during the year 2018. The median average wait time for US citizens is reported in Figure D.22(b) for a better comparison between the two categories. A first observation is that airports with short wait times for US citizens do not necessarily shorter wait times for non-US citizens than airports with high wait times for US citizens. For example, SPN has the second worst wait time for non US passengers while having the third shortest wait time for US passengers. Only two airports (STL and PBI) have a lower median average wait time for non-US passengers than for US passengers. Figure D.22(c), which shows the distribution of the average wait time ratio of non-US citizens over US citizens, indicates that only five airports have a median ratio outside the range [1,2.5]: STL and PBI with a ratio lower than one as noted previously, along with MCO, Guam airport (GUM) and SPN with ratios greater than 2.5.

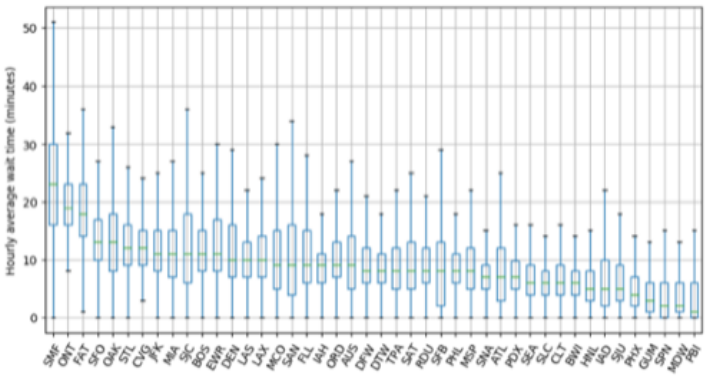
D.2.4 Hourly Wait Time Prediction Across Airports

Typical modeling of queues at airports relies on queuing theory studied in Operations Research to evaluate queue length and service time [187]. In this appendix, we choose to adopt a different approach. Leveraging machine learning techniques, our goal is to predict the average wait time per hour at any airport. ONT and CVG having less data than the other airports, they were not considered in the following study.

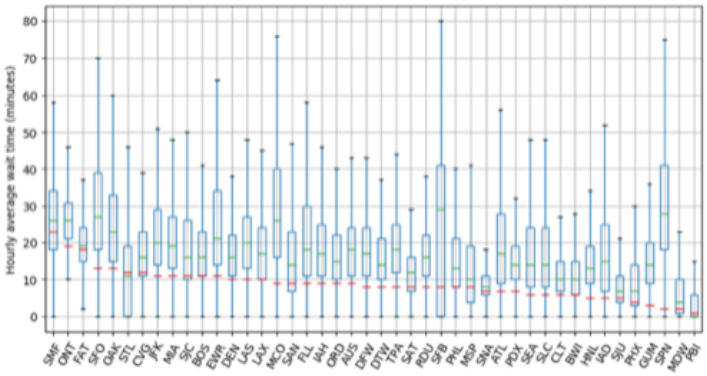
Features and regressors

The problem of interest falls in the category of regression techniques [188], more specifically under time series forecasting. The data set is partitioned into a train set and a test set. Each row in the data set corresponds to a particular hour, and the corresponding label is the average wait time for all passengers for the next hour. The following base set of features is created:

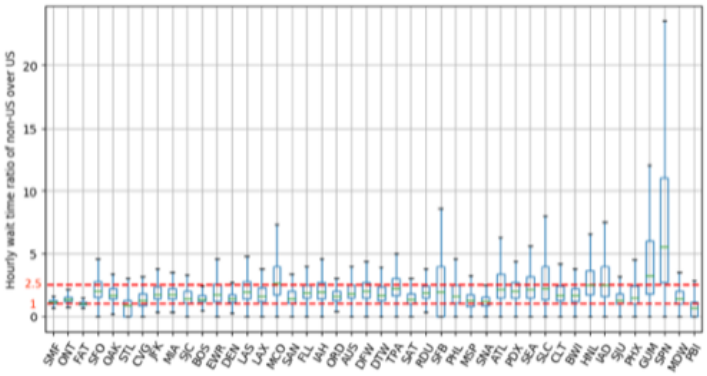
- Month of the year,
- Day of the month,
- Day of the week,
- Hour of the day,
- Number of passengers at this hour,
- Number of open booths at this hour,
- Number of flights at this hour.
- Ratio of passengers per open booth



(a) Comparison of the average wait time for US citizens per hour across airports



(b) Comparison of the average wait time for non-US citizens per hour across airports. In red is indicated the median wait time for US citizens from Figure D.22(a).



(c) Comparison of the ratio of average wait time for non-US citizens over US citizens. In red is indicated the interval [1, 2.5] for a better visualization.

Figure D.22: US vs non-US wait times comparison using boxplots over the year 2018

Intuitively the waiting time at border security seems to depend on the state of the border area in the previous hour as well, e.g. if not all previously arrived passengers were processed, therefore the following features can be added to the base set:

- Number of passengers at the previous hour,
- Number of open booths at the previous hour,
- Number of flights at the previous hour,
- Ratio of passengers per open booth at the previous hour.

To avoid data leakage from the train set to the test set, we do not randomly assign data to either the train or test set, but select a time period for the train set and only assign a later time period to the test set. The models are trained on data from 2013 to 2017 and tested on data from the year 2018.

The different regression models are implemented using Python as the programming language and using the Scikit-learn library [148].

To obtain the best possible performance, we experiment with various algorithms, each having its specific advantages and drawbacks. Below is a brief overview of each algorithm tested:

1. **Linear Regression** assumes that the relationship between the input variables and the measured variable is linear with some noise, and estimates the parameter vector with ordinary least squares minimization. It is the simplest regression method and is easily interpretable.

2. **Ridge Regression**, also known as Tikhonov regularization, extends the ordinary linear regression with a penalty term in the objective function proportional to the error norm. This improves the conditioning of the problem and reduces overfitting.

3. **Lasso Regression** performs a linear regression with regularization as well as a variable selection.

4. **Random Forest Regression** is an ensemble technique that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting

5. **Gradient Boosted Tree Regression** is an ensemble technique relying on decision trees as weak learners and resampling the training samples while assigning weights to the samples. It optimizes a cost function over function space by iteratively choosing a function that points in the negative gradient direction. It is typically more robust than basis learners.

6. **AdaBoost** is also an ensemble of decision trees relying on boosting. However, AdaBoost and Gradient Boosting optimize different loss functions.

Performance measures and prediction benchmarks

In order to measure the quality of the predictions, two different performance measures were computed: the R^2 score and the mean absolute error (MAE) which were already introduced in Section B.2.2.

Four different simple prediction benchmarks were tested in order to have a better understanding of the ease of prediction. For the first three, the prediction consisted in taking the value from the previous year, the previous day or the previous hour. The last one consists in constantly predicting the mean value of the the training set. And similarly to Chapter 3, a more elaborate comparison benchmark was considered: Facebook's time-series forecasting tool Prophet [145] was trained on the actual training set and its performance was measured on the data from the year 2018.

Performance analysis

One-hot encoding analysis

A single model for all terminals was created by adding one-hot encoding features for the different terminals, i.e. 59 binary features were added, each one indicating whether a specific terminal is considered or not. This single model was trained on the data from 2013 to 2017 and tested on the data from 2018. Fifty-nine other models, one for each terminal, were trained and tested on the same data filtered by terminal. Figure D.23 plots the MAE per terminal for each considered regressor. It shows that for each regressor the single terminal method outperforms the one-hot encoding method for a majority of terminals. For example, in the case of the Gradient Boosting regressor, having a different regressor per arriving terminal is better than the one-hot encoding method for forty terminals out of fifty-nine.

Benchmark comparison

The comparison with the chosen benchmarks was done using the single-terminal models, i.e. for each regressor type, one model was trained per arrival terminal. The R^2 score and mean absolute error of these models are aggregated in boxplots presented in Figure D.24 alongside the boxplot performance of the five benchmarks.

Figure D.24(a) shows that he Gradient Boosting regressors have the best R^2 performance, its R^2 scores being greater than 0 for 43 terminals out of 59,

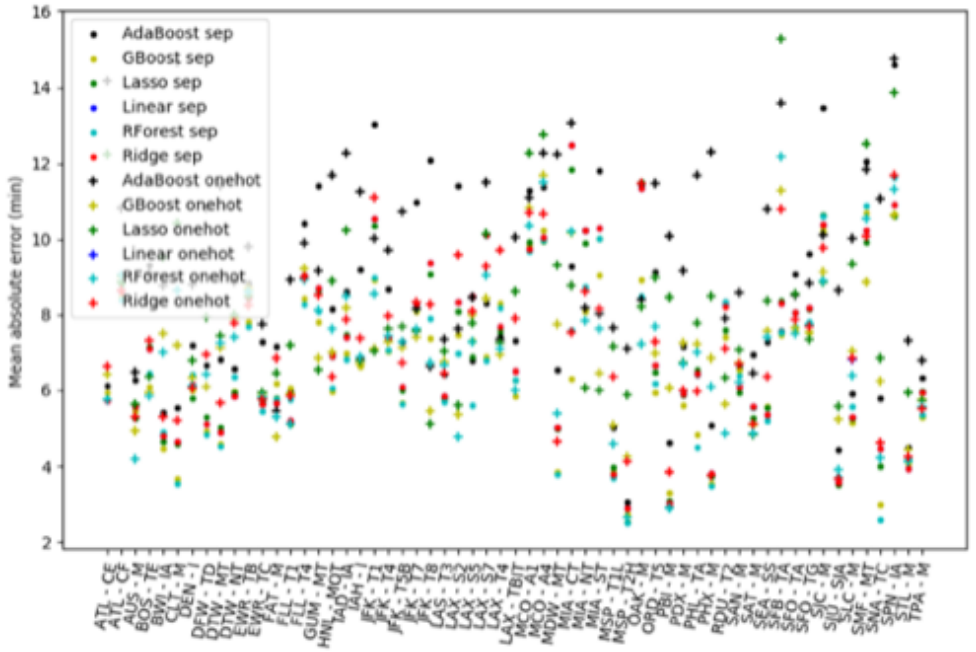
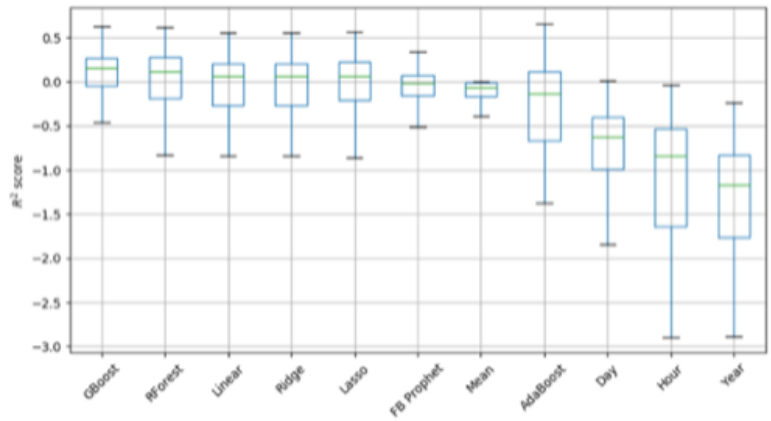


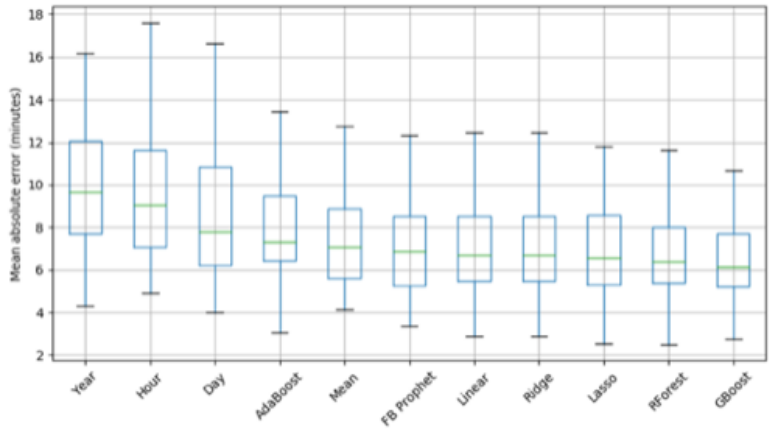
Figure D.23: Mean absolute error comparison between one-hot encoding for airports and different regressors per airports

which is to be compared to 28 out of 59 for the first benchmark, the Prophet tool. Gradient Boosting has also the smallest R^2 performance deviation of all tested models. Though Ada Boost has the highest R^2 score of all models for one terminal (SFB Terminal A), of the six chosen regressors, it is the only model with a median R^2 score less than 0, along with the five benchmarks. The three linear models (Linear, Lasso and Ridge) have similar performances in this study, implying that the overfitting methods added in Lasso and Ridge are not necessarily relevant here. The maximum R^2 difference between Lasso and Linear or Ridge is about 0.64 while this distance is about 0.005 between Ridge and Linear.

Figure D.24(b), showing the MAE distribution, has a different ranking with respect to the linear regressors. With this performance measure, Lasso yields better results than Linear and Ridge. Otherwise the conclusions for this performance measure are the same as for the R^2 score. This performance measure has however a more tangible interpretation: of all the models tested, only the Gradient Boosting regressors predict the average waiting time with an average error of 10 minutes or less for every terminal of entry. The Gradient Boosting regressors have a MAE of 5 minutes or less for 28 terminals.



(a) R^2 score



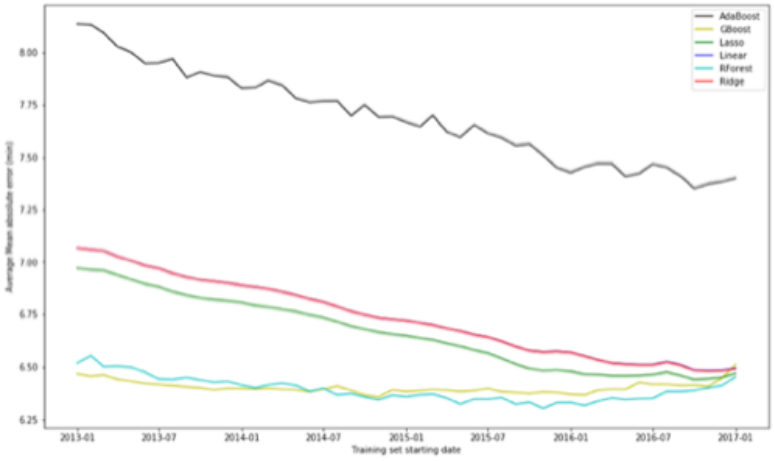
(b) Mean absolute error

Figure D.24: Box plot comparison of the performance measures for the five chosen benchmarks and for the six chosen regressors when predicting the average wait time for all passengers

Training size analysis

In order to visualize the impact of the training set size, the regressors were trained using training sets of different lengths (i.e. different starting dates) and their performance was tested on the data from the year 2018. As shown in Figure D.25, decreasing the size of the training set does not have a major impact on the performance of the Random Forest regressor and the Gradient Boosting regressor. For the four other regressors, their average and median performances improve when the training set size decreases. Their performance is however still not better than nor comparable with the Gradient Boosting regressor trained on the full training set.

(a) Median mean squared error



(b) Average mean squared error

Figure D.25: Evolution of the median and average regressors performance with the beginning of the training set

Comparison with standard deviation

Figure D.26 shows a comparison per terminal of entry between four of the best performing models' mean absolute error and the standard deviation of the average wait time over the test year 2018. This plot shows that except for one exception (SNA Terminal C), the Gradient Boosting models mean absolute errors are significantly better than the standard deviation of the values to predict.

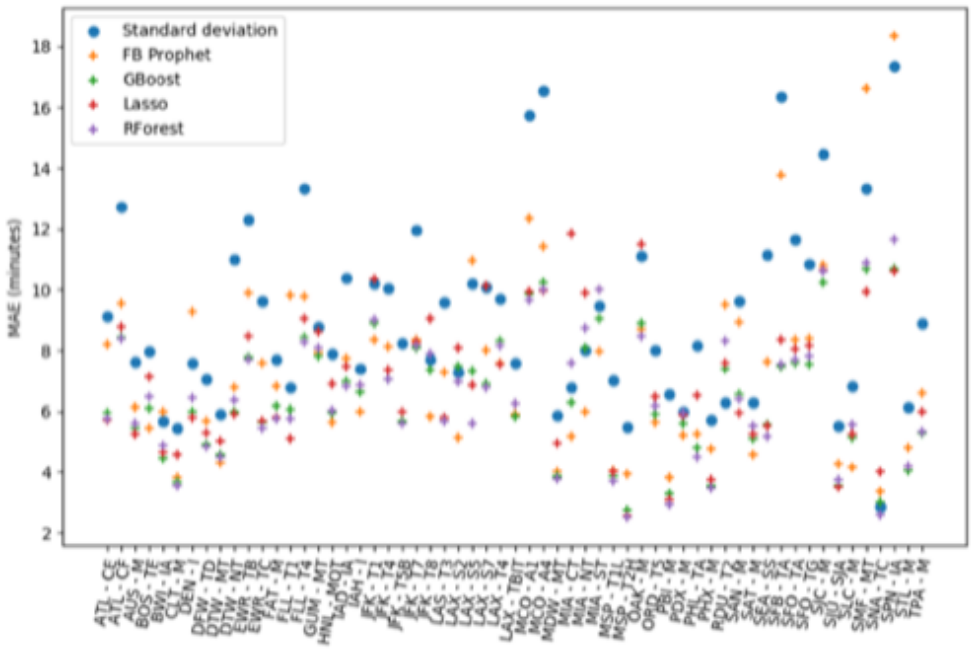


Figure D.26: Comparison per terminal of entry of the regressors' mean absolute error with the average wait time standard deviation over the year 2018

D.2.5 Conclusion

This appendix is a first step in a systematic analysis of passenger wait times at customs across all airports of entry in the United States using publicly available data. This analysis makes it easier to uncover some long-term trends, i.e. an increase in the number of arriving flights coupled with a decrease in the number of open custom booths, while also enabling a per airport comparison. This analysis also laid the ground to implementing machine learning regression models in order to predict the average wait time per airport of entry. These models could be used to better anticipate the number of required booths once the number of arriving passengers is known.

D.3 Predicting Passenger Flow at Charles De Gaulle Airport Security Checkpoints

Airport security checkpoints are critical areas in airport operations. Airports have to manage an important passenger flow at these checkpoints for security reason while maintaining service quality. The cost and quality of such an activity depend on the human resource management for these security operations. An appropriate human resource management can be obtained using an estimation of the passenger flow. This appendix investigates the prediction at a strategic level of the passenger flows at Paris Charles De Gaulle airport security checkpoints using machine learning techniques such as Long Short-Term Memory neural networks. The derived models are compared to the current prediction model using three different mathematical metrics. In addition, operational metrics are also designed to further analyze the performance of the obtained models.

D.3.1 Introduction

Motivation

Airport security checkpoints are key areas in airport operations. All passengers are checked at security checkpoint before entering the airside area. This continuous passenger flow implies an appropriate human resource management, which must satisfy two main objectives. A security checkpoint must be reliable in terms of security, while maintaining a predefined standard regarding passenger wait time. In addition, airports try to minimize their cost providing the best possible services.

At Charles De Gaulle airport, the human resources at security checkpoints are managed at two levels. The first level is a strategic level: passenger flows at security checkpoints are predicted 20 days upstream for the following month in order to determine the appropriate number of agents required. The second level is a tactical level: in real time, the agents are distributed at the security checkpoints to provide the service. This appendix investigates new learning methods such as neural networks in order to improve the prediction phase at the strategic level.

These learning methods are applied to the checkpoints within the zone of Charles De Gaulle airport corresponding to Air France's hub and named CDGE (cf. Figure D.27). It contains eight security checkpoints, separated in three different categories, depending on the type of passengers going through:

- checkpoints handling only passengers with local flights: C2F-Centraux;

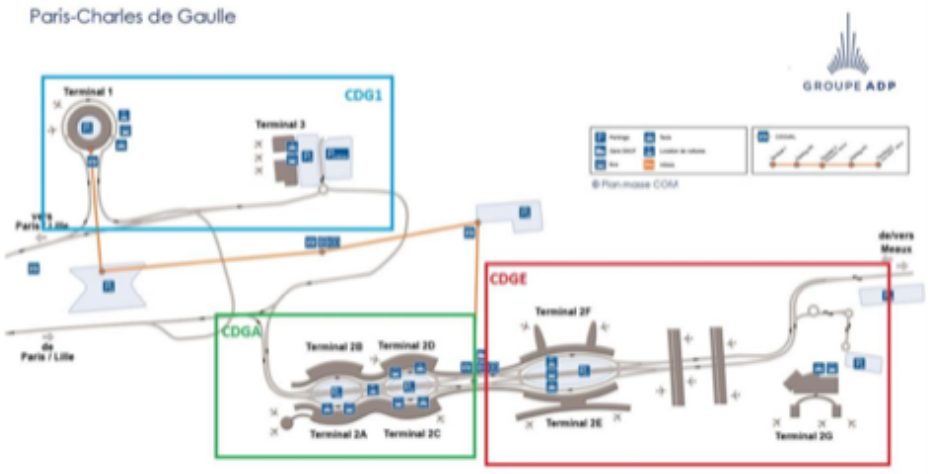


Figure D.27: Overview map of Charles De Gaulle terminals

- checkpoints handling only connecting passengers: C2E-GalerieEF, C2E-Puits2E, C2E-PorteL-CNT;
- checkpoints handling passengers on both local and connecting flights: C2E-PorteK, C2E-PorteL, C2E-PorteM, C2G-Depart.

Checkpoints C2E-GalerieEF and C2E-PorteL-CNT have the added particularity of linking two different terminals (E and F). C2E-Puits2E has the specificity of handling connecting passengers arriving to and leaving from Terminal E.

State of the art

Passenger flow prediction has been investigated for a long time in transportation areas. An exhaustive review was done by Liu et al. [189]. Traffic flow prediction for public transportation was studied in [190, 191], and for air transportation in [192, 193] using various prediction methods. Time series models were developed by Kumar [194] based on Kalman filtering, while Williams and Hoel [195] and then Kumar and Vanajakshi [196] worked on auto-regressive models. In the machine learning field, regression models such as Support Vector Machines [190, 192] or Neural Networks [189, 191] were used to forecast passenger flow. So far, the models derived try to predict the passenger flow using only historical data of the flow. Nevertheless, an airport passenger flow is a complex process. Extra features could be added in order to enhance the model performance. Indeed, a model which includes information relative to the arriving and departing flights should outperform

basic time series models. This motivates the use of machine learning models, that can fit multidimensional inputs.

Optimization of security checkpoints at a tactical level has also been thoroughly investigated. The efficiency of security checkpoint systems and organizations is discussed by Wilson et al. [197] and by Leone and Liu [198]. De Lange et al. [199] suggested creating virtual queuing in order to decrease waiting time at peak periods. However, to the best of the authors' knowledge, no study has been conducted around the airport security checkpoint strategic passenger flow prediction. Usually, each airport has its own process. Yet, the methodology presented in this appendix is generic and could be applied everywhere. The only constraint is the availability of information regarding departing and arriving flight and their expected occupancy.

This appendix is organized as follows: Section D.3.2 describes the data considered, the features extracted from them and the learning models used. In Section D.3.3 the different models are compared using both theoretical and operational performance measures. An in-depth analysis is performed in Section D.3.4 for two chosen checkpoints. Section D.3.5 concludes this study and suggests some possible improvements and future steps.

D.3.2 Model creation

This section presents the machine learning models chosen for the following experiments as well as the data considered.

Machine Learning and Long Short-Term Memory Neural Network

A learning process consists in using data analysis methods and artificial intelligence to predict the behavior of a system. The aim is to define a model that will fit as best as possible the considered system. Machine learning algorithms define learning models h_θ, with parameters θ, that approximate the system function. The learning process is done upon a finite training set $\mathcal{D}$, and aims at minimizing the error over the training set by tuning the parameters θ of the learning model [200, 201, 202].

Various learning models exist in the literature and for various real-world applications, and in this appendix the choice of a particular neural network named Long Short-Term Memory (LSTM) was made and compared to a Random Forest model. LSTM networks were designed as an enhancement of Recurrent Neural Networks (RNN) to perform better supervised learning task on time series data [203, 204, 205]. LSTM are capable of learning long-term dependencies, while simple RNN only learn short term dependencies. LSTM use a cell state that keeps information from the past, and three gates

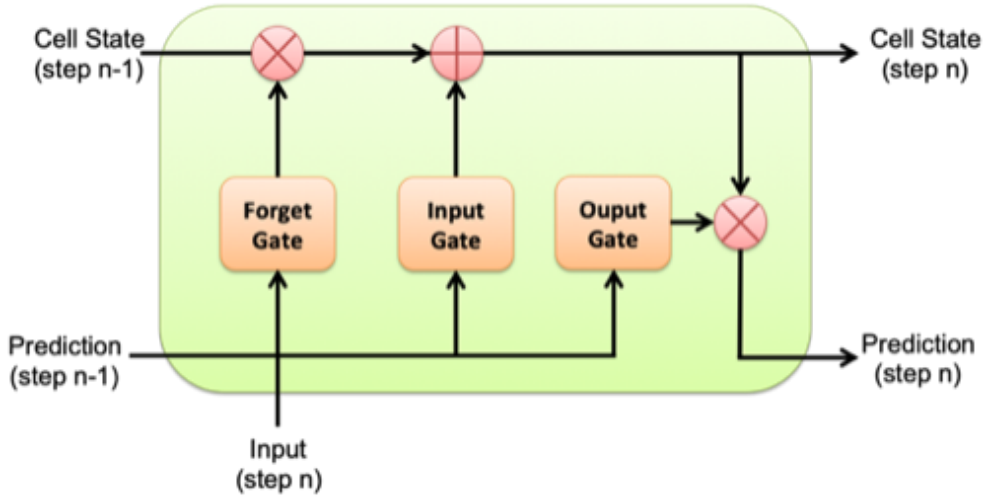


Figure D.28: Simplified illustration of the structure of a LSTM cell

that update the cell state and compute the prediction. First, the forget gate enables updating the cell state in order to forget information that are no longer relevant based on the current input. Second, the input gate enables saving in the cell state relevant information from the current input. Finally, the output gate computes the prediction using the updated cell state and the current input. A simplified illustration of a LSTM structure is depicted in Figure D.28.

From data to features

In this study, the values to predict correspond to the real passenger count at each Safety Checkpoint per ten minute period. As explained in Section D.3.1, the original input dataset used by Charles De Gaulle operational experts is composed with information relative to the schedule and occupancy of arriving and departing flights aggregated per five minute periods.

The dataset starts on February, 1st 2017 and ends on March, 31st 2019. Data from both 2017 and 2018 were used for the training phase, and the data from 2019 for the validation phase. For each flight, there are three passenger count expectations corresponding to:

- the expected number of connecting passengers
- the expected number of local passengers
- the expected total number of passengers

These passenger counts are given by the airlines to the airport. In addition, there are various information such as the date, the status of the flight

(departing or arriving flight), the airport terminal, the airline, the origin airport, the aircraft type, the departure geographic area, the flight range, and the check-in terminal.

Categorical features were represented using one-hot encoding. Additional features were extracted to complete the passenger count expectations. Passenger count expectations were aggregated per terminal, per status, and per terminal and status to create new features. Besides, features relative to the date were created: the month of the year, the day of the month, the day of the week, the hour of the day and the minute of the hour, and categorical variables for weekends, aeronautical weekends (including Fridays), holidays, and public holidays. Additional categories were created to capture whether a day is just before or after a public holiday or is the first or last day of a holiday.

This feature extraction yields a vector of 371 features for every five minutes of data. This vector sums-up the information over all the flights during the corresponding five minute period. The LSTM neural network was then fed with a time series corresponding to the input feature vector ranging from three hours before to five hours after the output 10 minutes time period. This time range was chosen based on two real-world considerations in order to encompass all the relevant flight and passenger information. On the one hand, airlines and airports recommend passengers on international flights to arrive about three hours before their flight departure time. On the other hand, the transfer time between two flights seldom exceeds five hours.

Network Architecture and Learning

This section describes the neural network architecture used in the experiments. The neural network is composed of two layers and a regression output layer. The first layer is a batch normalization. The second layer is a LSTM layer with 200 units and a sigmoid activation function. The layer also contains a dropout to regularize the network. The output layer is a single neuron dense layer with a ReLU activation function. This architecture will be referred to as LSTM200. Figure D.29 illustrates the network architecture.

The learning task was made using Adam optimizer [206] with a decay. The learning rate is 10^{-3} and the decay is 10^{-9}. Networks were trained during 10 epochs over the training set on a multi-GPU cluster. The cluster is composed of a dual ship Intel Xeon E5-2640 v4 - Deca-core (10 Core) 2,40GHz - Socket LGA 2011-v3 with 8 GPU GF GTX 1080 Ti 11 Go GDDR5X PCIe 3.0.

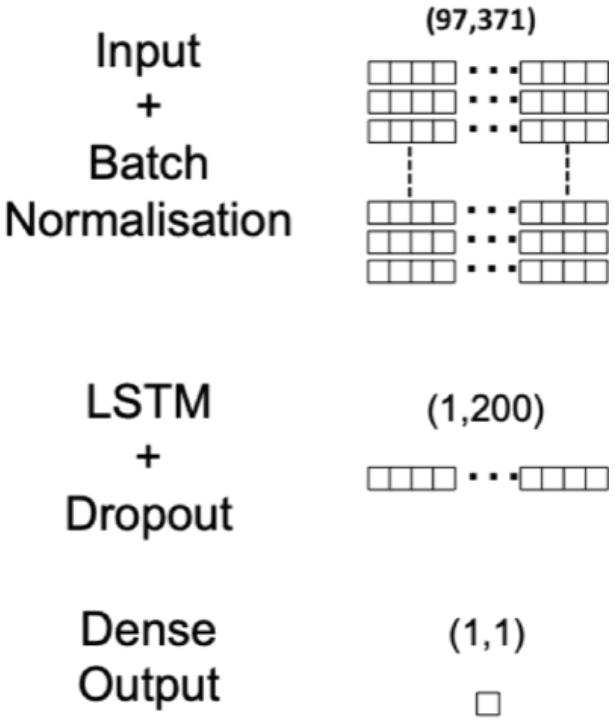


Figure D.29: Description of the neural network LSTM200 architecture used in the experiments

Penalized Loss

In practice, passenger count overestimation is costly. Therefore, a custom loss was designed. The loss aims to minimize overestimation by penalizing the positive part of the mean square error (MSE). As a reminder, the mean square error is the usual loss for regression problems. Let $\mathcal{D}$ be the training set, and h the learning model. The MSE of h over $\mathcal{D}$ is detailed in equation (D.5):

$$\mathbf{MSE}(h, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \cdot \sum_{(x,y)\in\mathcal{D}} (h(x) - y)^2 \tag{D.5}$$

Let $E = h(x) - y$ be the error of a sample $(x, y) \in \mathcal{D}$. E_+ is the positive part of this error, and E_- the negative part. The α-Penalized MSE is defined in equation (D.6) with $\alpha \in \mathbb{R}$:

$$\alpha\text{-}\mathbf{PMSE}(h, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \cdot \sum_{(x,y)\in\mathcal{D}} (E_- + (1 + \alpha) \cdot E_+)^2 \tag{D.6}$$

Model Summary

For this study, three models were used. The first model is a LSTM200 architecture trained with the MSE loss. The second model is a LSTM200 architecture trained with a 0.5-PMSE loss. The last model is a Random Forest model trained with MSE loss using the scikit-learn library [148]. The hyper-parameters of the Random Forest models were set to 40 for the number of estimators, with a max depth of 10, and a minimum sample split of 2. The three models are summarized in Table D.5.

Table D.5: Summary of the three models used in the appendix

Model Name	Model Type	Loss
LSTM (MSE)	LSTM200	MSE
LSTM (0.5-PMSE)	LSTM200	0.5-PMSE
RF	Random Forest	MSE

Additionally, in order to assess the effect of the hour of the day on the robustness of the chosen models, these models were trained twice: a first time with the hour of the day as a feature, and a second time without that feature.

D.3.3 Model comparison

Performance metrics

Theoretical metrics

In order to compare the performance of the different models, three different indicators were used: the R^2 score, the mean-absolute error (MAE) and a daily Pearson correlation score (DPC).

The R^2 score, also known as the coefficient of determination, is defined as the unity minus the ratio of the residual sum of squares over the total sum of squares:

$$\mathbf{R^2}(h, \mathcal{D}) = 1 - \frac{\sum_{(x,y)\in\mathcal{D}}(y - h(x))^2}{\sum_{(x,y)\in\mathcal{D}}(y - \bar{y})^2} \tag{D.7}$$

where y is the value to be predicted, $\bar{y}$ its mean and $h(x)$ is the model prediction and $\mathcal{D}$ the dataset. It ranges from $-\infty$ to 1, 1 being a perfect prediction and 0 meaning that the prediction does as well as constantly predicting the mean value for each occurrence. In the case of a negative R^2, then the model has a worse prediction than if it were predicting the mean value for each occurrence and therefore yields no useful predictions.

Regarding the mean-absolute error, the smaller its value is, the more accurate the prediction is. It is calculated using the following formula:

$$\mathbf{MAE}(h, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(x,y)\in\mathcal{D}} |h(x) - y| \tag{D.8}$$

The daily Pearson correlation score is an average of the usual Pearson correlation score applied to non-overlapping subsets $\mathcal{D}_d$ of $\mathcal{D}$, with each subset $\mathcal{D}_d$ containing the data from an entire day d and $\mathcal{D} = \bigcup_{d\in D} \mathcal{D}_d$. It gives an indication of how well the curve of the predicted number of arriving passenger follows the actual curve of arriving passengers. The closer the score is to 1, the better the prediction is. It is calculated using the following equations:

$$r(h, \mathcal{D}_d) = \frac{\sum_{(x,y)\in\mathcal{D}_d}(h(x) - \bar{h}_d)(y - \bar{y}_d)}{\sqrt{\sum_{(x,y)\in\mathcal{D}_d}(h(x) - \bar{h}_d)^2}\sqrt{\sum_{(x,y)\in\mathcal{D}_d}(y - \bar{y}_d)^2}} \tag{D.9}$$

$$\mathbf{DPC}(h, \mathcal{D}) = \frac{1}{|D|} \sum_{\mathcal{D}_d} r(h, \mathcal{D}_d) \tag{D.10}$$

where $\bar{h}_d$ (resp. $\bar{y}_d$) is the average of $h(x)$ (resp. y) over $\mathcal{D}_d$.

Operational metrics

Airport management being a balance between minimizing costs and maximizing the service given to passengers, two additional metrics were introduced based on these operational considerations. These metrics are simplified versions of reality since the security agent providers do not share their calculation processes and the actual staffing of checkpoints is decided at a tactical level.

From a cost perspective, the key figure is the number of security agents necessary for a smooth operation. Agents being paid per hour, the cost metric considered is the total number of agent-hours induced by the predicted passenger arrivals. A smooth operation is here defined as a nominal passenger flow f_N, which has a unit of passengers per line per ten minutes. These flows are specific to each security checkpoint and are determined by the airport management. Airports also define a peak-time passenger flow f_P that security agents should be able to cope with when needed. From these nominal flows and the number of expected passengers p_t at time step t, it is then possible to compute the number of lines n_t required to achieve this flow: $n_t = \frac{p_t}{f_N}$. Assuming that each line is staffed by five security agents yields the number of agents required at each time step t. Each time steps being of ten minutes, it is then necessary to divide the resulting cost by six to obtain the agent-hour cost. The total cost metric C_T can be resumed by the following

equation:

$$C_T = \frac{5}{6} \sum_t \frac{p_t}{f_N} \tag{D.11}$$

From a quality perspective, the key figure is the average waiting time at the security checkpoints. In order to estimate it at each time step, the following simplified queuing model is considered. At time step t, y_t passengers arrive at the checkpoint SC adding to the r_{t-1} passengers not processed during the previous time step. Under nominal conditions, $n_t \cdot f_N$ passengers are processed during a ten minute time step, where n_t is the number of lines estimated for the cost calculation. Peak-time conditions were defined here as time steps where the remaining number of passengers r_{t-1} was greater than the nominal flow f_N. Under peak-time conditions, the number of processed passengers becomes $\max(n_{t-1}, n_t) \cdot f_P$, i.e. the number of lines kept open stays the same if it was initially supposed to become smaller. If the prediction indicated that no lines should be open and that there are in fact passengers, then either the lines open in the previous time step are kept open if any, or one line is opened.

The processed number of passengers π_t at time step t can therefore be calculated as followed:

$$\pi_t = \begin{cases} \max(n_{t-1}, n_t) \cdot f_P & \text{if } r_{t-1} > f_N \text{ and } n_t > 0 \\ n_{t-1} \cdot f_N & \text{if } n_t = 0 \text{ and } n_{t-1} > 0 \\ f_N & \text{if } n_t = 0 \\ n_t \cdot f_N & \text{otherwise} \end{cases} \tag{D.12}$$

The average wait time τ_t during a time step t can be computed using the following equation:

$$\tau_t = \sum_{i=1}^{y_t + r_{t-1}} \frac{i}{\pi_t} = \frac{y_t + r_{t-1} + 1}{2\pi_t} \tag{D.13}$$

The overall quality metric Q_T is then calculated by taking the average of all τ_t.

The passenger flow model at a checkpoint is represented as an automata in Figure D.30.

First results

All three models presented in Section D.3.2 were trained using data from February 2017 to December 2018 and tested on the months of January to March 2019 using the performance metrics presented in Section D.3.3. These metrics were also applied to the current model in use at Charles De Gaulle

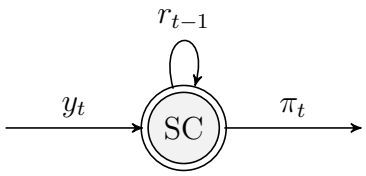


Figure D.30: Model of the passenger flow at a security checkpoint

airport for comparison. Based on operational observations, the output of the neural nets was forced to 0 when the hour of the day was between 00:00 and 04:00.

Hour of the day

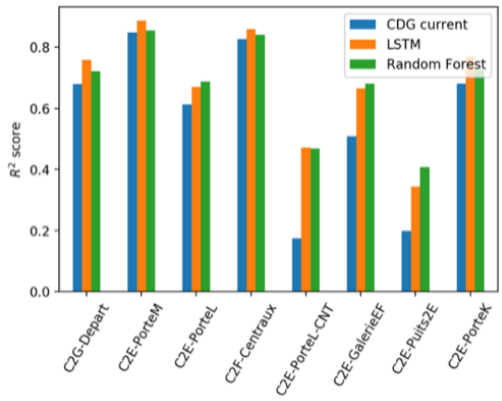
Table D.6 summarizes the performances of the three developed learning models based on two of the three mathematical metrics. This table enables a quick comparison of the use of the hour of the day as a feature. For the upcoming analysis, only one LSTM model and one Random Forest regressor were kept per checkpoint based on their MAE. The kept models have their performance cells highlighted in green, while the best of all models are also highlighted in bold. A first observation is that the influence of the hour of the day is not the same for Random Forests and for neural networks. For seven checkpoints over eight, using the hour of the day highly improves the Random Forest's performance. On the other side, six checkpoints over eight with LSTM (MSE and 0.5-PMSE) have better scores without the hour of the day.

Mathematical Performance Metrics

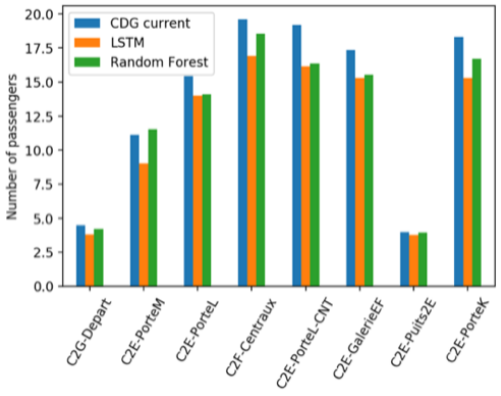
Figure D.31 presents the performance of the current model and the kept models from Section D.3.3 using the mathematical metrics introduced in Section D.3.3. From a R^2 score perspective, both the LSTM and Random Forest models outperform the current prediction model with improvements ranging from 0.01 for C2E-PorteM to 0.3 for C2E-PorteL-CNT. Regarding the mean-absolute error performance, the LSTM nets outperform once more the current model while the Random Forest regressors have higher errors for two of the checkpoints (C2E-PorteM and C2E-Puits2E). The LSTM reduces the mean-absolute errors from 5.6% (C2E-Puits2E) to 18.9% (C2E-PorteM) compared to the current model: LSTM net have a mean-absolute error of less than seventeen passengers per ten minutes for all checkpoints while the current model has an error greater than seventeen passengers per ten minutes for half of the checkpoints. Finally, regarding the daily Pearson correlation score, LSTM model outperforms the current prediction model at every checkpoint, while the Random Forest regressor outperforms it for seven checkpoints out of eight.

Table D.6: Comparison of the models using or not the hour in the training set. Green color cells correspond to the model kept in the following study. Bold cells correspond to the best models

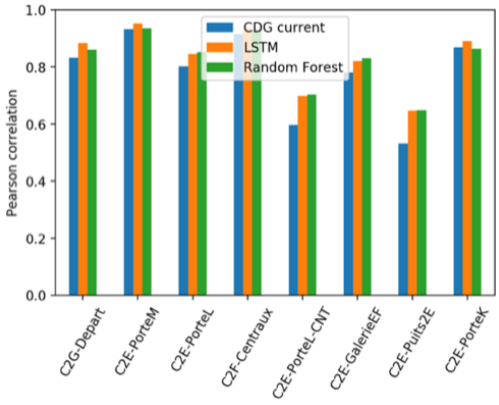
	LSTM (MSE)				LSTM (0.5-PMSE)				Random Forest			
	With Hour		Without Hour		With Hour		Without Hour		With Hour		Without Hour	
	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE	R^2	MAE
C2G-Depart	0.736	4.02	0.732	4	**0.768**	**3.75**	0.754	3.85	0.721	4.27	0.712	4.68
C2E-PorteM	0.822	11.08	**0.887**	**9.02**	0.796	11.92	0.866	9.82	0.853	11.68	0.808	17.34
C2E-PorteL	**0.674**	**13.96**	0.641	14.58	0.578	15.8	0.636	14.67	0.684	14.21	0.558	19.58
C2F-Centraux	0.834	18.34	0.861	16.79	0.851	17.34	**0.86**	**16.62**	0.788	21.48	0.826	21.08
C2E-PorteL-CNT	0.436	16.54	**0.474**	**16.14**	0.426	16.6	0.425	16.49	0.471	16.36	0.475	17.62
C2E-GalerieEF	**0.667**	**15.32**	0.667	15.46	0.656	15.46	0.626	15.95	0.68	15.61	0.608	20.03
C2E-Puits2E	0.411	3.77	**0.37**	**3.76**	0.395	3.78	0.375	3.77	0.405	4.09	0.381	4.56
C2E-PorteK	0.688	18.37	0.758	15.63	0.662	15.67	**0.769**	**15.26**	0.726	16.73	0.647	23.53



(a) Comparison of the R^2 score



(b) Comparison of mean absolute error



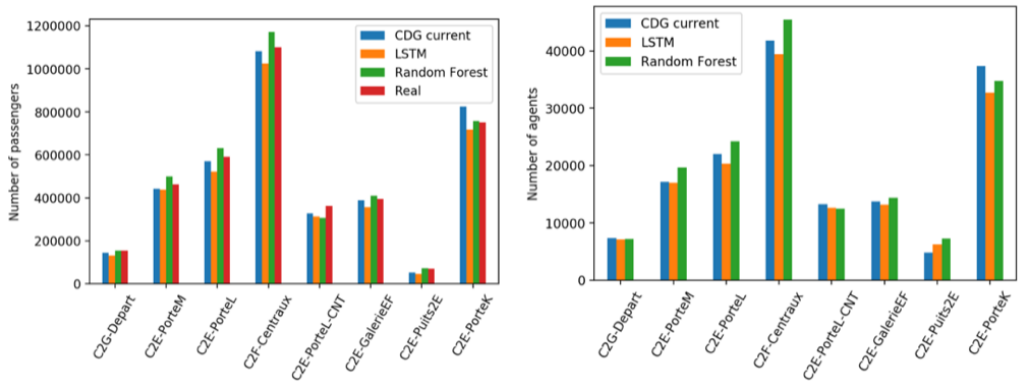
(c) Comparison of daily Pearson correlation score

Figure D.31: Comparison per checkpoints of different mathematical metrics for the three considered models

Operational Performance Metrics

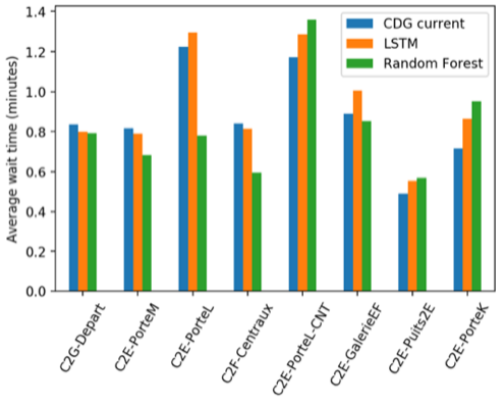
Using the simplified operational metrics introduced in Section D.3.3, the difference in performance is less straightforward. Figure D.32 shows the comparison of the cost metric (i.e. the number of agent-hour over the three months) per checkpoint as well as the comparison of the quality metric. Figure D.32(a) presents the comparison of the total number of predicted passengers per checkpoints along with the actual number of passengers for comparison. A first observation is that the LSTM nets tend to underestimate the number of passengers regardless of the loss function considered, while the Random Forest regressors overestimate the number of passengers.

Since LSTM nets tend to underestimate the number of passengers more



(a) Comparison of the number of predicted passengers

(b) Comparison of the number of estimated hour agents



(c) Comparison of the number of the estimated average wait times

Figure D.32: Comparison per checkpoints of different operational metrics for the three considered models

than the current model, it is also reflected from a cost perspective in Figure D.32(b): For seven of the checkpoints, the number of agent-hours required based on the neural nets is less than the number required based on the current model. For C2E-Puits2E, the number of required agent-hours is greater than the current model, a paradox illustrating the specificity of that terminal and further analyzed in Section D.3.4.

Synthesis

Figure D.33 shows the performance difference between the neural networks and the current prediction model, for all the metrics, and all the security checkpoints. All the metrics are normalized by the current prediction model value, except for the R^2 score, and the correlation score since they

already have consistent magnitude and a norm lower than 1. The normalization enables comparison between security checkpoints. The difference is explained in percentage of improvement relative to the current model, except for the R^2 score and the correlation score where it is the improvement difference in percentage (norm lower than 1). In addition, the performance sign is selected such that a positive sign corresponds to a metric improvement. Finally, the performance difference is displayed with color from green when the improvement is greater than 20% to red when the best model deteriorates the performance more than 20%

Figure D.33: Heatmap visualization of the performance difference between the LSTM models and the current predictive model

For three security checkpoints over eight (C2G-Depart, C2E-PorteM, C2F-Centraux), LSTM models outperforms the current prediction model for all the performance metrics. For four over eight (C2E-PorteL, C2E-PorteL-CNT, C2E-GalerieEF, C2E-PorteK) LSTM models outperform current model for all the metric excepted the waiting time metric, which is deteriorated more than 10% in half of the cases (C2E-GalerieEF, C2E-PorteK). Finally, at security checkpoint C2E-Puit2E, the performance metric is highly

deteriorated for the agent number (-29%) and waiting time (-12.9%). This particular behavior will be explained in Section D.3.4.

D.3.4 Case Study

In this section, two security checkpoints were selected with respect to their performances for a further analysis. C2G-Depart was chosen to illustrate the good results of the LSTM model while C2E-Puits2E was chosen to better understand why the LSTM model does not outperform the current model from an operational perspective.

Daily analysis

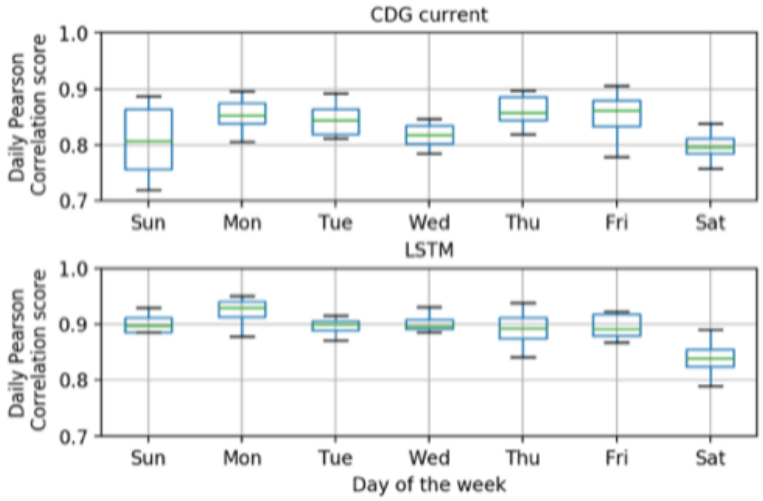
A first step in understanding the differences in performance is to analyze the performances of the two models (LSTM and current) on a less aggregated level such as the different days of the week. Figure D.34 shows the distribution of the daily Pearson correlation score per day of the week for the two chosen security checkpoints. It confirms the previous observation that the LSTM model is overall better than the current model with this metric, while adding some information on how this improvement is structured.

Regarding C2G-Depart, Figure D.34(a) shows that both models are less precise on Saturdays compared to other days, though the LSTM model reduces the score variability on that day. An important improvement can be seen for Sundays: the current model has a lower score with a large variability, whereas the LSTM model reduces drastically that variability and improves the median score of 0.1.

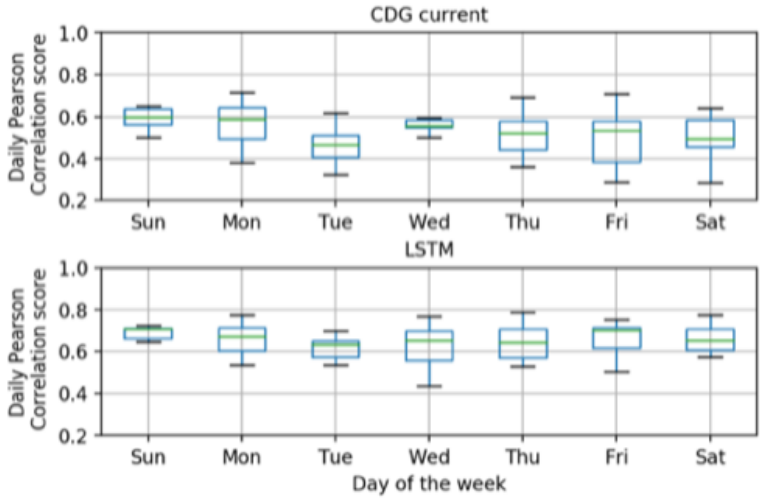
Regarding C2E-Puits2E, Figure D.34(b) shows that the LSTM model manages to reduce variability on most days, with an important reduction on Fridays. Wednesdays show an opposite behavior: though the LSTM model does increase the median correlation score, it also triples the score variability.

Hourly analysis

A similar analysis can be conducted by aggregating the performance metrics per hour of the day. Figure D.35 shows the hourly distribution of the error in predicting the number of arriving passengers for the current model and the LSTM model at the two chosen security checkpoints. It confirms the LSTM tendency to underestimate the number of passengers: All medians are at or below zero for the LSTM while the current model tend to overestimate for five hours out of the sixteen considered hours for C2G-Depart. For C2E-Puits2E, both models have median errors at or below zero, however the



(a) C2G-Depart

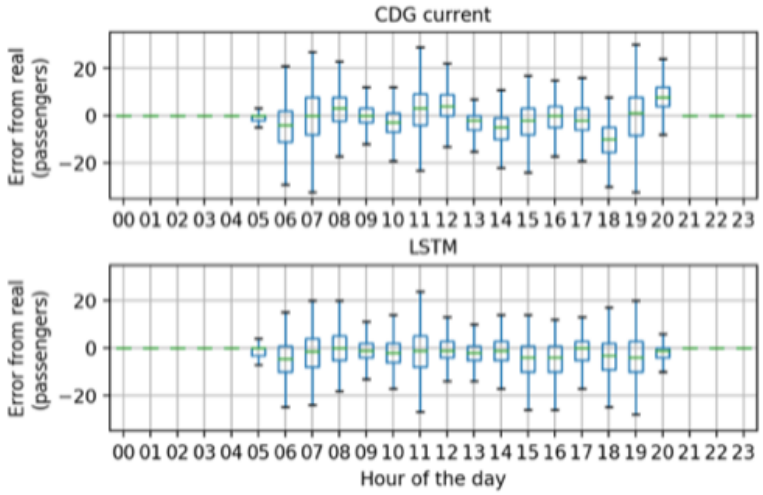


(b) C2E-Puits2E

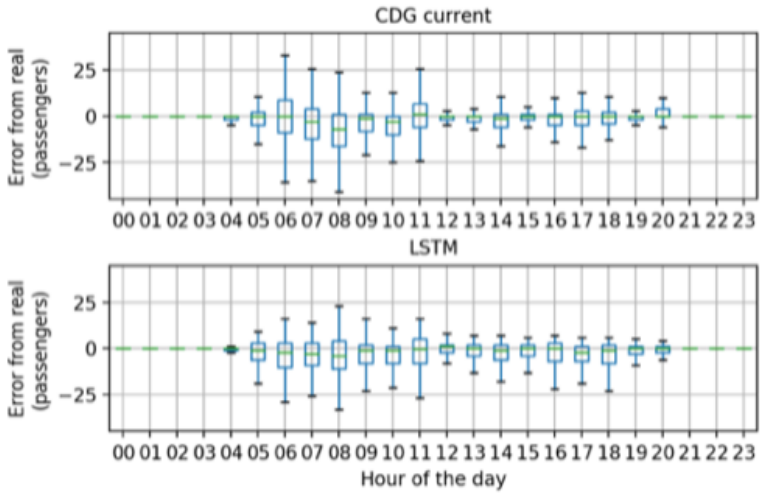
Figure D.34: Daily correlation distribution per day of the week for C2G-Depart and C2E-Puits2E

LSTM model variations are shifted towards the negative with a smaller tendency to overestimation, which is indicated by smaller upper whiskers. This underestimation can be seen as a lower cost, since the predicted number of passengers determines the number of required agents.

Figure D.36 shows the hourly distribution of the average wait time using the predictions from the current model and the LSTM model. Combining Figures D.36 & D.35 makes the impact of underestimation clearer on the qual-



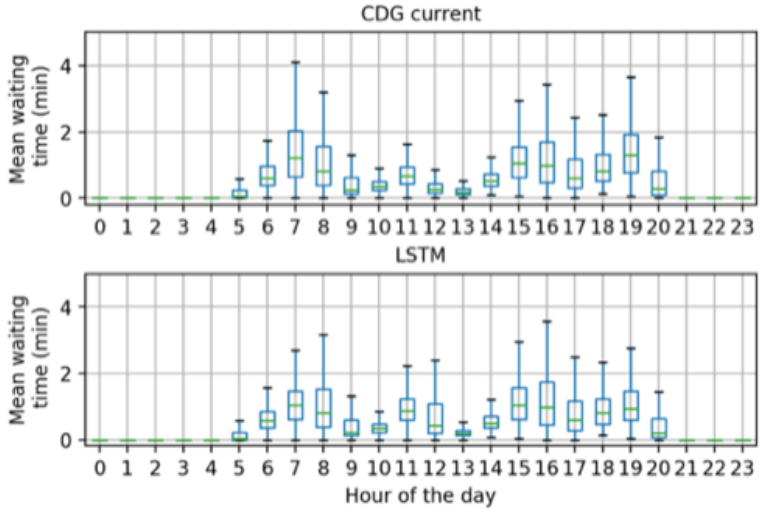
(a) C2G-Depart



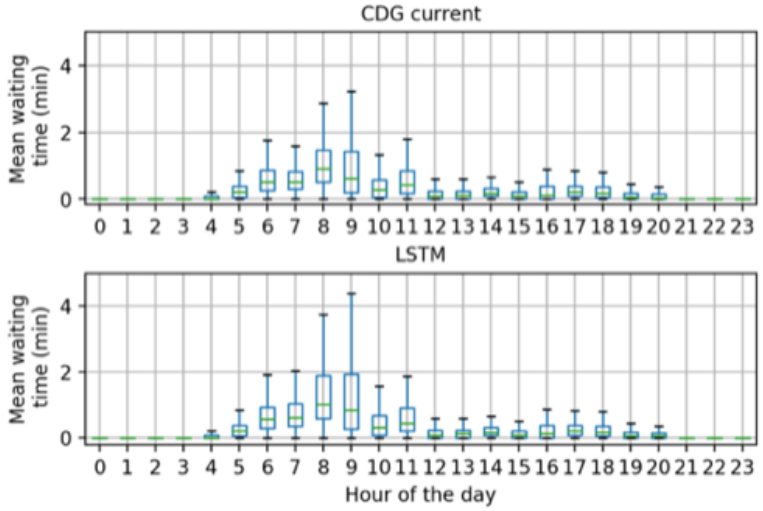
(b) C2E-Puits 2E

Figure D.35: Hourly passenger error boxplots comparison between the current model and the neural net trained with a mean squared error loss function at two different checkpoints

ity of service. Underestimations in the number of passengers is associated with a higher median average wait time, which is then propagated in the following hours. For C2G-Depart, the underestimations at 3pm and 7pm on Figure D.36(a) are clearly associated with a rise and propagation of the average wait time on Figure D.35(a). For C2E-Puits2E, it is most visible for the underestimation at 7am for both models.



(a) C2G-Depart



(b) C2E-Puits 2E

Figure D.36: Hourly average wait time boxplots comparison between the current model and the neural net trained with a mean squared error loss function at two different checkpoints

This analysis could be used to further improve the derived models and the determination of the number of required agents. By highlighting hours of the days where the models are known to underestimate (resp. overestimate) the number of passengers, it should be possible to mitigate this underestimation (resp. overestimation) by adjusting the predicted value or by adapting

accordingly the number of required agents for these specific periods.

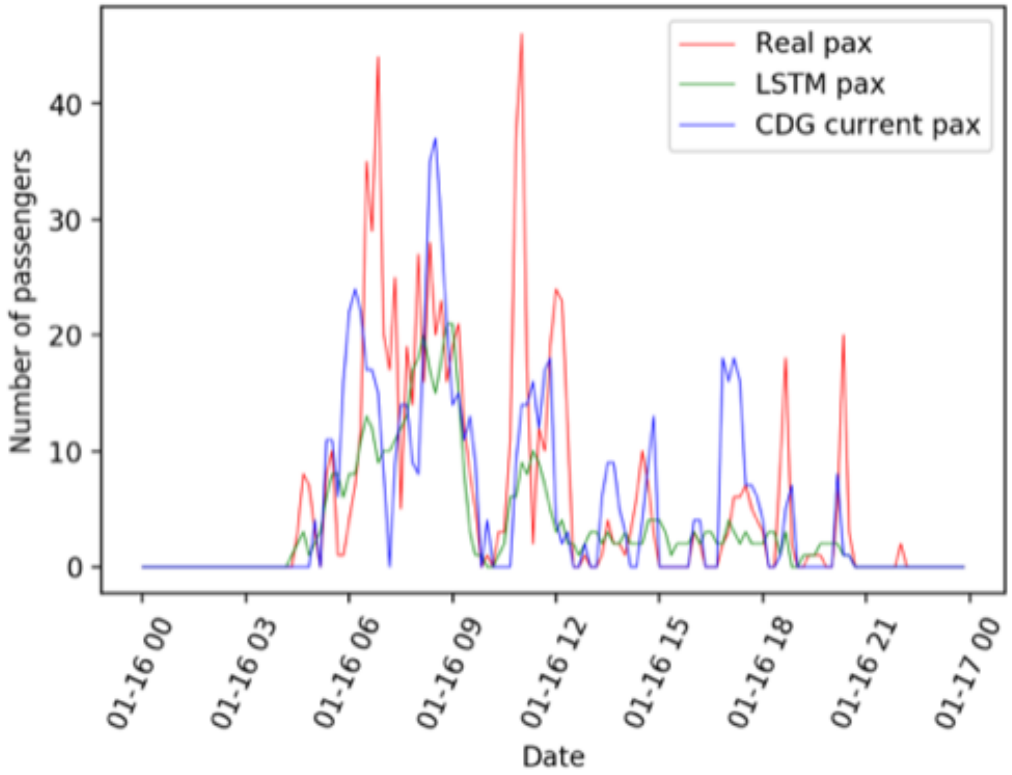
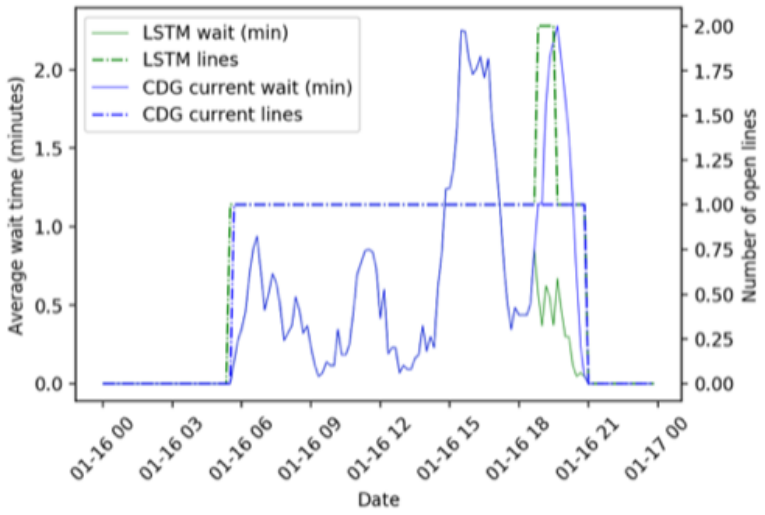
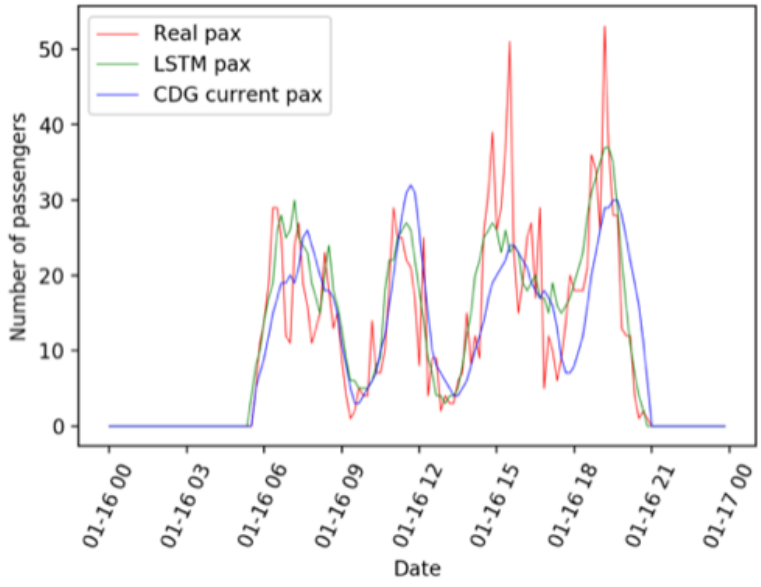


Figure D.37: Hourly comparison of the predicted number of passengers between the current model and the neural net at C2E-Puits2E on January 16th 2019

In order to better understand the differences in performance for these two checkpoints, the estimated number of passengers is plotted over a day (January 16th, 2019) in Figure D.37 for C2E-Puits2E and in Figure D.38(b) for C2G-Depart. Figure D.37 highlights the difficulty of predicting the number of passengers for C2E-Puits2E: There are irregular yet continuous arrival spikes in the early morning (5am-9am) and then the rest of the day is composed of arrival spikes of varying amplitudes with periods with no passengers at all. From a prediction performance perspective, Figure D.37 clearly illustrates the paradox of predicting less passengers while requiring more agents. The LSTM model underestimates more the passenger arrival spikes in the early morning than the current model, and estimates a low number of passengers for the rest of the day though never predicting zero arrivals. This means that agents are required all day long from the LSTM perspective, while the current model captures better the periods with no arrivals, enabling an economy



(a) Comparison of the average wait time and number of open lines



(b) Comparison of the predicted number of passengers

Figure D.38: Hourly comparison between the current model and the neural net trained with a mean squared error loss function at C2G-Depart on January 16th 2019

of agents. A potential improvement of the LSTM model would be to hard-code the periods where operational expertise indicates that transfers within

Terminal E are highly unlikely.

Regarding C2G-Depart, Figure D.38(b) is a good day example to understand the better performance of the LSTM model compared to the current model. There are four daily spikes in passenger arrival with varying amplitude, and though both models capture the number of spikes, the LSTM model yields a better estimation of the amplitude of each spike as well as their initial slope increase. This higher accuracy has a direct impact on the estimated wait time, as shown in Figure D.38(a). The average wait time is identical for both model until the fourth spike, where the better estimation of the increase in passengers triggers the opening of a second line, which reduces the wait time by half compared to the current model.

D.3.5 Discussion & Conclusion

This appendix investigated predicting passenger flow at Paris Charles De Gaulle airport security checkpoints using LSTM neural networks. The models performance was evaluated over several theoretical and operational metrics. The overall results are promising since LSTM models outperform the current model for every checkpoints using the theoretical metrics and for three checkpoints out of eight, LSTM models outperform the current prediction model using all the considered metrics. Though the considered operational metrics were simplified, these results illustrate that implementing a better and accurate strategic passenger flow prediction would surely reduce operational cost while maintaining predefined standard regarding passengers waiting time.

The methodology presented in this study can still be enhanced and tuned to be efficient and dedicated on specific cases. Future works should investigate a more elaborated queuing model or simulation. In addition, the models could be validated with real experimentation in the operations. Further works could be done on the neural network architecture and learning, or with expert to tune the models bringing relevant information to improve the prediction (hybrid models).

Bibliography

[1] Bureau of Transportation Statistics. 2018 Traffic Data for U.S Airlines and Foreign Airlines U.S. Flights, 2020.

[2] Bureau of Transportation Statistics. Bureau of Transportation Statistics, About BTS, 2018.

[3] European Statistical System. eurostat, Your keys to European statistics, 2020.

[4] eurostat. Record number of air passengers carried at more than 1.1 billion in 2018. Technical report, December 2019.

[5] EUROCONTROL. CODA digest - all-causes delay and cancellations to air transport in europe - 2017. `https://www.eurocontrol.int/sites/default/files/publication/files/coda-digest-annual-2017.pdf`, 2017.

[6] Roger Beatty, Rose Hsu, Lee Berry, and James Rome. Preliminary Evaluation of Flight Delay Propagation through an Airline Schedule. *Air Traffic Control Quarterly*, 7(4):259–270, October 1999.

[7] NextGen Integration and Implementation Office. NextGen Implementation Plan. In *Federal Aviation Administration*, 2009.

[8] Maarek Darecki, Charles Edelstenne, Emma Fernandez, Peter Hartman, Jean-Paul Herteman, Michael Kerkloch, Ian King, Patrick Ky, Michel Mathieu, Giuseppe Orsi, Gerald Schotman, Colin Smith, and Johann-Dietrich Wörner. *Flightpath 2050: Europe's Vision for Aviation ; Maintaining Global Leadership and Serving Society's Needs ; Report of the High-Level Group on Aviation Research.* European Commission, Luxembourg, 2011.

[9] EUROCONTROL and Federal Aviation Administration Air Traffic Organization System Operations Services. 2015 Comparison of Air Traffic

Management-Related Operational Performance: U.S./Europe. Technical report, August 2016.

[10] Aude Marzuoli, Philippe Monmousseau, and Eric Feron. Passenger-centric metrics for Air Transportation leveraging mobile phone and Twitter data. In *Data-Driven Intelligent Transportation Workshop - IEEE International Conference on Data Mining 2018*, Singapore, November 2018.

[11] Philippe Monmousseau, Aude Marzuoli, Eric Feron, and Daniel Delahaye. Predicting and Analyzing US Air Traffic Delays using Passenger-centric Data-sources. In *Thirteenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2019)*, Vienna, Austria, 2019.

[12] P Monmousseau, A Marzuoli, E Feron, and D Delahaye. Passengers on social media: A real-time estimator of the state of the US air transportation system. In *ENRI Int. Workshop on ATM/CNS (EIWAC 2019)*, Tokyo, Japan, 2019.

[13] Philippe Monmousseau, Aude Marzuoli, Eric Feron, and Daniel Delahaye. Putting the Air Transportation System to sleep: A passenger perspective measured by passenger-generated data. *arXiv:2004.14372 [physics]*, April 2020.

[14] Philippe Monmousseau, Daniel Delahaye, Aude Marzuoli, and Eric Feron. Door-to-door travel time analysis from Paris to London and Amsterdam using Uber data. In *Ninth SESAR Innovation Days*, Athens, Greece, 2019.

[15] Philippe Monmousseau, Daniel Delahaye, Aude Marzuoli, and Eric Feron. Door-to-door Air Travel Time Analysis in the United States using Uber Data. In *1st International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT 2020)*, Singapore, February 2020.

[16] Philippe Monmousseau, Aude Marzuoli, Christabelle Bosson, Eric Feron, and Daniel Delahaye. Doorway to the United States: An Exploration of Customs and Border Protection Data. In *38th Digital Avionics Systems Conference*, San Diego, California, USA, 2019.

[17] Philippe Monmousseau, Gabriel Jarry, Florian Bertosio, Daniel Delahaye, and Marc Houalla. Predicting Passenger Flow at Charles De

Gaulle Airport Security Checkpoints. In *2020 International Conference on Artificial Intelligence and Data Analytics for Air Transportation (AIDA-AT)*, pages 1–9, Singapore, Singapore, February 2020. IEEE.

[18] Philippe Monmousseau, Stephane Puechmorel, Daniel Delahaye, Aude Marzuoli, and Eric Feron. Towards a more complete view of air transportation performance combining on-time performance and passenger sentiment. In *9th International Conference on Research in Air Transportation (ICRAT '20)*, Tampa, Florida, US, 2020.

[19] Alice Sternberg, Jorge Soares, Diego Carvalho, and Eduardo Ogasawara. A Review on Flight Delay Prediction. *arXiv:1703.06118 [cs]*, March 2017.

[20] L. Schaefer and D. Millner. Flight delay propagation analysis with the Detailed Policy Assessment Tool. In *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, volume 2, pages 1299–1303, Tucson, AZ, USA, 2001. IEEE.

[21] Eric Mueller and Gano Chatterji. Analysis of Aircraft Arrival and Departure Delay Characteristics. In *AIAA's Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, Los Angeles, California, October 2002. American Institute of Aeronautics and Astronautics.

[22] Paul T. R Wang, Lisa A Schaefer, and Leonard A Wojcik. Flight connections and their impacts on delay propagation. In *22nd Digital Avionics Systems Conference Proceedings (Cat No 03CH37449) DASC-03*, pages 5.B.4–5.1, Indianapolis, IN, USA, 2003. IEEE.

[23] Ning Xu, George Donohue, Kathryn Blackmond Laskey, and Chun-Hung Chen. Estimation of delay propagation in the national aviation system using bayesian networks. In *Sixth USA/Europe Air Traffic Management Research and Development Seminar*, page 11, Baltimore, Maryland, USA, 2005. FAA and Eurocontrol.

[24] Yujie Liu and Fan Yang. Initial Flight Delay Modeling and Estimating Based on an Improved Bayesian Network Structure Learning Algorithm. In *2009 Fifth International Conference on Natural Computation*, pages 72–76, Tianjian, China, 2009. IEEE.

[25] Yu-Jie Liu and Song Ma. Flight Delay and Delay Propagation Analysis Based on Bayesian Network. In *2008 International Symposium on*

Knowledge Acquisition and Modeling, pages 318–322, Wuhan, China, December 2008. IEEE.

[26] Shervin AhmadBeygi, Amy Cohn, Yihan Guan, and Peter Belobaba. Analysis of the potential for delay propagation in passenger airline networks. *Journal of Air Transport Management*, 14(5):221–236, September 2008.

[27] Yufeng Tu, Michael O Ball, and Wolfgang S Jank. Estimating Flight Departure Delay Distributions—A Statistical Approach With Long-Term Trend and Short-Term Pattern. *Journal of the American Statistical Association*, 103(481):112–125, March 2008.

[28] Banavar Sridhar and Neil Y Chen. Short-term national airspace system delay prediction using weather impacted traffic index. *Journal of guidance, control, and dynamics*, 32(2):657–662, 2009.

[29] Michael B Callaham, James S DeArmon, Arlene M Cooper, Jason H Goodfriend, Debra Moch-Mooney, and George H Solomos. Assessing NAS Performance: Normalizing for the Effects of Weather. In *Fourth USA/Europe Air Traffic Management Research & Development Symposium*, Santa Fe, December 2001.

[30] Alexander Klein, Chad Craun, and Robert S Lee. Airport delay prediction using weather-impacted traffic index (WITI) model. In *29th Digital Avionics Systems Conference*, pages 2.B.1–1–2.B.1–13, Salt Lake City, UT, USA, October 2010. IEEE.

[31] Banavar Sridhar, Yao Wang, Alexander Klein, and Richard Jehlen. Modeling Flight Delays and Cancellations at the National, Regional and Airport Levels in the United States. In *Eighth USA/Europe Air Traffic Management Research and Development Seminar*, 2009.

[32] Andrew M. Churchill, David J. Lovell, and Michael O. Ball. Flight Delay Propagation Impact on Strategic Air Traffic Flow Management. *Transportation Research Record: Journal of the Transportation Research Board*, 2177(1):105–113, January 2010.

[33] Juan Jose Rebollo and Hamsa Balakrishnan. A Network-Based Model for Predicting Air Traffic Delays. In *Fifth International Conference on Research in Air Transportation*, 2012.

[34] Juan Jose Rebollo and Hamsa Balakrishnan. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies*, 44:231–241, July 2014.

[35] Nikolas Pyrgiotis, Kerry M. Malone, and Amedeo Odoni. Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, 27:60–75, February 2013.

[36] Pablo Fleurquin, José J. Ramasco, and Victor M. Eguiluz. Systemic delay propagation in the US airport network. *Scientific Reports*, 3(1):1159, December 2013.

[37] Abdulwahab Aljubairy, Ali Shemshadi, and Quan Z. Sheng. Real-time investigation of flight delays based on the Internet of Things data. In Jinyan Li, Xue Li, Shuliang Wang, Jianxin Li, and Quan Z. Sheng, editors, *Advanced Data Mining and Applications*, volume 10086, pages 788–800. Springer International Publishing, Cham, 2016.

[38] Karthik Gopalakrishnan and Hamsa Balakrishnan. A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks. In *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*, Seattle, Washington, USA, June 2017.

[39] Sandip Roy, Mengran Xue, and Banavar Sridhar. Vulnerability Metrics for the Airspace System. In *Twelfth USA/Europe Air Traffic Management Research and Development Seminar*, 2017.

[40] Max Z Li, Karthik Gopalakrishnan, Hamsa Balakrishnan, and Kristyn Pantoja. A Spectral Approach Towards Analyzing Air Traffic Network Disruptions. In *Thirteenth USA/Europe Air Traffic Management Research and Development Seminar*, Vienna, Austria, 2019.

[41] Max Z Li, Karthik Gopalakrishnan, Yanjun Wang, and Hamsa Balakrishnan. Outlier Analysis of Airport Delay Distributions in US and China. In *First International Conference on Artificial Intelligence and Data Analytics for Air Transportation*, Singapore, 2020.

[42] D W Conner. Passenger comfort technology for system decision making. Technical report, August 1980.

[43] Andrew C. Lemer. Measuring performance of airport passenger terminals. *Transportation Research Part A: Policy and Practice*, 26(1):37–45, January 1992.

[44] Laurence Matthews. Forecasting peak passenger flows at airports. *Transportation*, 22(1):55–72, February 1995.

[45] Ad Pruyn and Ale Smidts. Effects of waiting on the satisfaction with the service: Beyond objective time measures. *International Journal of Research in Marketing*, 15(4):321–334, October 1998.

[46] C.V. Robertson, S. Shrader, D.R. Pendergraft, L.M. Johnson, and K.S. Silbert. The role of modeling demand in process re-engineering. In *Proceedings of the Winter Simulation Conference*, volume 2, pages 1454–1458, San Diego, CA, USA, 2002. IEEE.

[47] Jeremy R Brown and Poornima Madhavan. Using discrete event simulation to identify choke points in passenger flow through airport checkpoints. *Proc. of the Student Capstone Conference of the Virginia Modeling, Analysis, \& Simulation Center, Norfolk, Virginia*, page 6, 2010.

[48] Jeremy R Brown and Poornima Madhavan. Examining Passenger Flow Choke Points at Airports Using Discrete Event Simulation. Technical report, 2011.

[49] Sheng-Hshiung Tsaur, Te-Yi Chang, and Chang-Hua Yen. The evaluation of airline service quality by fuzzy MCDM. *Tourism Management*, 23(2):107–115, April 2002.

[50] Yu-Hern Chang and Chung-Hsing Yeh. A survey analysis of service quality for domestic airlines. *European Journal of Operational Research*, 139(1):166–177, May 2002.

[51] Safak Aksoy, Eda Atilgan, and Serkan Akinci. Airline services marketing by domestic and foreign firms: Differences from the customers' viewpoint. *Journal of Air Transport Management*, 9(6):343–351, November 2003.

[52] Adival Aparecido Magri Junior and Claudio Jorge Pinto Alves. Convenient airports: Point of view of the passengers. Technical report, 2005.

[53] Airport Council International. *Quality of Service at Airports: Standards and Measurements*. ACI World Headquarters, 2000.

[54] Konstantina Gkritza, Debbie Niemeier, and Fred Mannering. Airport security screening and changing passenger satisfaction: An exploratory assessment. *Journal of Air Transport Management*, 12(5):213–219, September 2006.

[55] Joyce A Hunter. A correlational study of how airline customer service and consumer perception of airline customer service affect the air rage phenomenon. *Journal of Air Transportation*, 11(3):78–109, 2006.

[56] Fatma Pakdil and Özlem Aydın. Expectations and perceptions in airline services: An analysis using weighted SERVQUAL scores. *Journal of Air Transport Management*, 13(4):229–237, July 2007.

[57] Chien-Chang Chou. A model for the evaluation of airport service quality. *Proceedings of the Institution of Civil Engineers - Transport*, 162(4):207–213, November 2009.

[58] Chien-Chang Chou, Li-Jen Liu, Sue-Fen Huang, Jeng-Ming Yih, and Tzeu-Chen Han. An evaluation of airline service quality using the fuzzy weighted SERVQUAL method. *Applied Soft Computing*, 11(2):2117–2128, March 2011.

[59] Vesna Popovic, Ben Kraal, and Philip Kirk. Towards Airport Passenger Experience Models. In *7th International Conference on Design & Emotion*, Chicago, Illinois, USA, October 2010.

[60] Yu-Chiun Chiou and Yen-Heng Chen. Service quality effects on air passenger intentions: A service chain perspective. *Transportmetrica*, 8(6):406–426, November 2012.

[61] Juan de Oña and Rocio de Oña. Quality of Service in Public Transport Based on Customer Satisfaction Surveys: A Review and Assessment of Methodological Approaches. *Transportation Science*, 49(3):605–622, August 2015.

[62] Stephane Bratu and Cynthia Barnhart. An Analysis of Passenger Delays Using Flight Operations and Passenger Booking Data. *Air Traffic Control Quarterly*, 13(1):1–27, 2005.

[63] Stephane Bratu and Cynthia Barnhart. Flight operations recovery: New approaches considering passenger recovery. *Journal of Scheduling*, 9(3):279–298, June 2006.

[64] Danyi Wang, Dr Lance Sherry, and Dr George Donohue. Passenger Trip Time Metric for Air Transportation. In *The 2nd International Conference on Research in Air Transportation*, 2006.

[65] Danyi Wang. *Methods for Analysis of Passenger Trip Performance in a Complex Networked Transportation System.* Doctor of Philosophy, George Mason University, Fairfax, Virginia, USA, 2007.

[66] World Economic Forum. Connected World : Transforming Travel, Transportation and Supply Chains. http://www3.weforum.org/docs, 2013.

[67] World Economic Forum. Smart travel: Unlocking economic growth and development through travel facilitation. http://www3.weforum.org/docs/GAC/2014, 2014.

[68] Yuri O Gawdiak and Tony Diana. NextGen Metrics for the Joint Planning and Development Office. volume 11th AIAA Aviation Technology, Integration, and Operations (ATIO) Conference, including the AIAA Balloon Systems Conference and 19th AIAA Lighter-Than, 2011.

[69] A Cook, G Tanner, S Cristóbal, and M Zanin. Passenger-Oriented Enhanced Metrics. 2012.

[70] SITA. The passenger IT trends survey. www.sita.aero/system/files/Passenger-IT-Trends-Survey-2014.pdf, 2014.

[71] Harold Nikoue, Aude Marzuoli, John-Paul Clarke, Eric Feron, and Jim Peters. Passenger Flow Predictions at Sydney International Airport: A Data-Driven Queuing Approach. *arXiv:1508.04839 [cs]*, August 2015.

[72] Jereon Van den Heuvel, Danique Ton, and Kim Hermansen. Advances in measuring pedestrians at dutch train stations using bluetooth, wifi and infrared technology. In *Traffic and Granular Flow'15*, pages 11–18. Springer, 2016.

[73] Weixin Huang, Yuming Lin, Borong Lin, and Liang Zhao. Modeling and predicting the occupancy in a China hub airport terminal using Wi-Fi data. *Energy & Buildings*, 203:19, 2019.

[74] Xiaoqian Sun, Volker Gollnick, and Sebastian Wandelt. Robustness analysis metrics for worldwide airport network: A comprehensive study. *Chinese Journal of Aeronautics*, 30(2):500–512, April 2017.

[75] Xiaoqian Sun and Sebastian Wandelt. Complementary strengths of airlines under network disruptions. *Safety Science*, 103:76–87, March 2018.

[76] Eric Pels, Peter Nijkamp, and Piet Rietveld. Access to and competition between airports: A case study for the San Francisco Bay area. *Transportation Research Part A: Policy and Practice*, 37(1):71–83, January 2003.

[77] Jan-Willem Grotenhuis, Bart W. Wiegmans, and Piet Rietveld. The desired quality of integrated multimodal travel information in public transport: Customer needs for time and effort savings. *Transport Policy*, 14(1):27–38, January 2007.

[78] *NextGen Priorities - Joint Implementation Plan Update Including the Northeast Corridor*. Federal Aviation Administration, October 2017.

[79] Jere Meserole and John Moore. What is System Wide Information Management (SWIM)? In *2006 IEEE/AIAA 25TH Digital Avionics Systems Conference*, pages 1–8, Portland, OR, USA, October 2006. IEEE.

[80] Alvin Sipe and John Moore. Air traffic functions in the NextGen and SESAR airspace. In *2009 IEEE/AIAA 28th Digital Avionics Systems Conference*, pages 2.A.6–1–2.A.6–7, Orlando, FL, USA, October 2009. IEEE.

[81] Ryan Klock, David Owens, Henry Schwartz, and Robert Plencner. Integrated Intermodal Passenger Transportation System. page 20, 2012.

[82] Isabelle Laplace, Aude Marzuoli, and Eric Feron. META-CDM: Multimodal, Efficient Transportation in Airports and Collaborative Decision Making. In *Airports in Urban Networks 2014 (AUN2014)*, Paris, 2014.

[83] Sang Hyun Kim, Aude Marzuoli, John-Paul Clarke, Daniel Delahaye, and Eric Feron. Airport Gate Scheduling for Passengers, Aircraft, and Operation. In *Tenth USA/Europe Air Traffic Management Research and Development Seminar (ATM2013)*, Chicago, Illinois, June 2013.

[84] Lynnette Dray, Aude Marzuoli, and Antony Evans. Air Transportation and Multimodal, Collaborative Decision Making during Adverse Events. In *Eleventh USA/Europe Air Traffic Management Research and Development Seminar (ATM2015)*, Lisboa, Portugal, June 2015.

[85] Aude Marzuoli, Emmanuel Boidot, Pablo Colomar, Mathieu Guerpillon, Eric Feron, Alexandre Bayen, and Mark Hansen. Improving Disruption Management With Multimodal Collaborative Decision-Making: A Case Study of the Asiana Crash and Lessons Learned. *IEEE Transactions on Intelligent Transportation Systems*, 17(10):2699–2717, October 2016.

[86] Aude Marzuoli, Emmanuel Boidot, Eric Feron, Paul B. C. van Erp, Alexis Ucko, Alexandre Bayen, and Mark Hansen. Multimodal Impact Analysis of an Airside Catastrophic Event: A Case Study of the Asiana Crash. *IEEE Transactions on Intelligent Transportation Systems*, 17(2):587–604, February 2016.

[87] Lynnette Dray, Isabelle Laplace, Aude Marzuoli, Eric Feron, and Antony Evans. Using Ground Transportation for Aviation System Disruption Alleviation. *Journal of Air Transportation*, 25(3):95–107, July 2017.

[88] Stefanie Peer, Jasper Knockaert, Paul Koster, Yin-Yen Tseng, and Eric T Verhoef. Door-to-door travel times in Revealed Preference departure time choice models: An approximation method using GPS data. *Transportation Research Part B: Methodological*, 58:134–150, 2013.

[89] Maria Salonen and Tuuli Toivonen. Modelling travel time in urban networks: Comparable measures for private car and public transport. *Journal of Transport Geography*, 31:143–153, July 2013.

[90] Elsa Durán-Hormazábal and Alejandro Tirachini. Estimation of travel time variability for cars, buses, metro and door-to-door public transport trips in Santiago, Chile. *Research in Transportation Economics*, 59:26–39, November 2016.

[91] Sebastian Wandelt, Zezhou Wang, and Xiaoqian Sun. Worldwide Railway Skeleton Network: Extraction Methodology and Preliminary Analysis. *IEEE Transactions on Intelligent Transportation Systems*, 18(8):2206–2216, August 2017.

[92] Wolfgang Grimme and Sven Maertens. Flightpath 2050 revisited – An analysis of the 4-hour-goal using flight schedules and origin-destination passenger demand data. *Transportation Research Procedia*, 43:147–155, 2019.

[93] Xiaoqian Sun, Sebastian Wandelt, and Eike Stumpf. Competitiveness of on-demand air taxis regarding door-to-door travel time: A race through Europe. *Transportation Research Part E: Logistics and Transportation Review*, 119:1–18, November 2018.

[94] Andrew Cook, Graham Tanner, S Cristóbal, H Ureta, D Perez, and A Paul. DATASET2050 D2.2 - Data-driven Model. Technical report, The Innaxis Foundation and Research Institute, 2016.

[95] Daniel B Work and Alexandre M Bayen. Impacts of the Mobile Internet on Transportation Cyberphysical Systems: Traffic Monitoring using Smartphones. In *Physical Systems*, pages 18–20, Washington DC, USA, November 2008.

[96] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[97] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1):10, December 2015.

[98] Vincent D. Blondel, Markus Esch, Connie Chan, Fabrice Clerot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for Development: The D4D Challenge on Mobile Phone Data. *arXiv:1210.0137 [physics, stat]*, September 2012.

[99] Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *arXiv:1407.4885 [physics]*, July 2014.

[100] Yves-Alexandre de Montjoye, Luc Rocher, and Alex Sandy Pentland. Bandicoot: A Python Toolbox for Mobile Phone Metadata. *Journal of Machine Learning Research*, 17(1):6100–6104, 2016.

[101] Rex W Douglass, David A Meyer, Megha Ram, David Rideout, and Dongjin Song. High resolution population estimates from telecommunications data. *EPJ Data Science*, 4(1):4, December 2015.

[102] Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240–250, September 2015.

[103] Miguel Picornell, Tomás Ruiz, Maxime Lenormand, José J. Ramasco, Thibaut Dubernet, and Enrique Frías-Martínez. Exploring the potential of phone call data to characterize the relationship between social network and travel behavior. *Transportation*, 42(4):647–668, July 2015.

[104] Jameson L. Toole, Serdar Colak, Bradley Sturt, Lauren P. Alexander, Alexandre Evsukoff, and Marta C. González. The path most traveled:

Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162–177, September 2015.

[105] Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, and Jakob Puchinger. Inferring dynamic origin-destination flows by transport mode using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 101:254–275, April 2019.

[106] Aude Marzuoli, Emmanuel Boidot, Eric Feron, and Ashok Srivastava. Implementing and Validating Air Passenger–Centric Metrics Using Mobile Phone Data. *Journal of Aerospace Information Systems*, 16(4):132–147, April 2019.

[107] Pedro García-Albertos, Oliva G Cantú Ros, Ricardo Herranz, and Carla Ciruelos. Understanding Door-to-Door Travel Times from Opportunistically Collected Mobile Phone Records. In *SESAR Innovation Days 2017*, 2017.

[108] Javier Burrieza, Rita Rodríguez, Pablo Ruiz, María José Sala, Javier Torres, Pedro García, Oliva García-Cantú, and Ricardo Herranz. Enhanced Passenger Characterisation through the Fusion of Mobile Phone Records and Airport Surveys. In *Ninth SESAR Innovation Days*, Athens, Greece, 2019.

[109] Pedro García-Albertos, Olivia G Cantú Ros, and Ricardo Herranz. Analyzing door-to-door travel times through mobile phone data. *CEAS Aeronaut Journal*, 2019.

[110] Statista. Number of monthly active mobile social media users in europe as of january 2018, by country (in millions). `https://www.statista.com/statistics/299496/active-mobile-social-media-users-in-european-countries/`.

[111] Statista. Leading countries based on the number of active Twitter users as of April 2020, 2020.

[112] Akshay Java, Xiaodan Song, Tim Finin, and Belle Tseng. Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis - WebKDD/SNA-KDD '07*, pages 56–65, San Jose, California, 2007. ACM Press.

[113] Balachander Krishnamurthy, Phillipa Gill, and Martin Arlitt. A few chirps about Twitter. In *Proceedings of the First Workshop on Online*

Social Networks - WOSP '08, page 19, Seattle, WA, USA, 2008. ACM Press.

[114] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *arXiv:0812.1045 [physics]*, December 2008.

[115] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL Student Research Workshop on - ACL '05*, page 43, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

[116] Luiz Fernando Sommaggio Coletta, Nadia Felix Felipe da Silva, Eduardo Raul Hruschka, and Estevam Rafael Hruschka. Combining Classification and Clustering for Tweet Sentiment Analysis. In *2014 Brazilian Conference on Intelligent Systems*, pages 210–215, Sao Paulo, Brazil, October 2014. IEEE.

[117] Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Topical Clustering of Tweets. page 8, July 2011.

[118] Oren Tsur, Adi Littman, and Ari Rappoport. Efficient Clustering of Short Messages into General Domains. page 10, 2013.

[119] Janette Lehmann, Bruno Gonçalves, José J. Ramasco, and Ciro Cattuto. Dynamical Classes of Collective Attention in Twitter. *arXiv:1111.1896 [physics]*, November 2011.

[120] Kirill Kireyev, Leysia Palen, and Kenneth M Anderson. Applications of Topics Models to Analysis of Disaster-Related Twitter Data. page 4, 2009.

[121] Sarah Vieweg, Amanda L Hughes, Kate Starbird, and Leysia Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. page 10, 2010.

[122] Leysia Palen, Kate Starbird, Sarah Vieweg, and Amanda Hughes. Twitter-based information distribution during the 2009 Red River Valley flood threat. *Bulletin of the American Society for Information Science and Technology*, 36(5):13–17, June 2010.

[123] Teun Terpstra and R Stronkman. Towards a realtime Twitter analysis during crises for operational crisis management. page 10, 2012.

[124] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: Real-time event detection by social sensors. page 10, 2010.

[125] Cynthia Chew and Gunther Eysenbach. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS ONE*, 5(11):e14118, November 2010.

[126] David A. Broniatowski, Michael J. Paul, and Mark Dredze. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. *PLoS ONE*, 8(12):e83672, December 2013.

[127] J. Brian Houston, Joshua Hawthorne, Mildred F. Perreault, Eun Hae Park, Marlo Goldstein Hode, Michael R. Halliwell, Sarah E. Turner McGowen, Rachel Davis, Shivani Vaid, Jonathan A. McElderry, and Stanford A. Griffith. Social media and disasters: A functional framework for social media use in disaster planning, response, and research. *Disasters*, 39(1):1–22, January 2015.

[128] Bruno Takahashi, Edson C. Tandoc, and Christine Carmichael. Communicating on Twitter during a disaster: An analysis of tweets during Typhoon Haiyan in the Philippines. *Computers in Human Behavior*, 50:392–398, September 2015.

[129] Shalini Priya, Manish Bhanu, Sourav Kumar Dandapat, Kripabandhu Ghosh, and Joydeep Chandra. TAQE: Tweet Retrieval-Based Infrastructure Damage Assessment During Disasters. *IEEE Transactions on Computational Social Systems*, 7(2):389–403, April 2020.

[130] Harshit Srivastava and Ravi Sankar. Information Dissemination From Social Network for Extreme Weather Scenario. *IEEE Transactions on Computational Social Systems*, 7(2):319–328, April 2020.

[131] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. *Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, 10:79–86, May 2002.

[132] Alexander Pak and Patrick Paroubek. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. *LREc*, 10(2010):1320–1326, May 2010.

[133] Nádia F.F. da Silva, Eduardo R. Hruschka, and Estevam R. Hruschka. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66:170–179, October 2014.

[134] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2, 2008.

[135] Jeffrey Oliver Breen. Mining twitter for airline consumer sentiment. *Practical text mining and statistical analysis for non-structured text data applications*, 133, 2012.

[136] Yun Wan and Qigang Gao. An ensemble sentiment classification system of Twitter data for airline services analysis. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 1318–1325, Atlantic City, NJ, USA, November 2015. IEEE.

[137] Fotis Misopoulos, Miljana Mitic, Alexandros Kapoulas, and Christos Karapiperis. Uncovering customer service experiences with Twitter: The case of airline industry. *Management Decision*, 52(4):705–723, May 2014.

[138] Hao Wang, Dogan Can, Abe Kazemzadeh, Fran{\c{c}}ois Bar, and Shrikanth Narayanan. A system for real-time Twitter sentiment analysis of 2012 us presidential election cycle. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, July 2012.

[139] Keng Siau. An Approach to Sentiment Analysis – The Case of Airline Quality Rating. *Pacific Asia Conference on Information Systems (PACIS 2014)*, 2014.

[140] Brent D Bowen, Dean E Headley, and Jacqueline R Luedtke. *Airline Quality Rating*, volume 91. Wichita State University, National Institute for Aviation Research, 1991.

[141] Mary Jane C. Samonte, John Michael R. Garcia, Valerie Jade L. Lucero, and Shayann Celine B. Santos. Sentiment and opinion analysis on Twitter about local airlines. In *Proceedings of the 3rd International Conference on Communication and Information Processing - ICCIP '17*, pages 415–422, Tokyo, Japan, 2017. ACM Press.

[142] Simone Gitto and Paolo Mancuso. Brand perceptions of airports using social networks. *Journal of Air Transport Management*, 75:153–163, March 2019.

[143] Rupinder Paul Khandpur, Taoran Ji, Yue Ning, Liang Zhao, Chang-Tien Lu, Erik R Smith, Christopher Adams, and Naren Ramakrishnan. Determining Relative Airport Threats from News and Social Media. In *Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications*, page 7. Association for the Advancement of Artificial Intelligence, 2017.

[144] Priyanga Gunarathne, Huaxia Rui, and Avi Seidmann. Customer Service on Social Media: The Effect of Customer Popularity and Sentiment on Airline Response. In *2015 48th Hawaii International Conference on System Sciences*, pages 3288–3297, HI, USA, January 2015. IEEE.

[145] Sean J. Taylor and Benjamin Letham. Forecasting at Scale. *The American Statistician*, 72(1):37–45, January 2018.

[146] Facebook. Prophet - Forecasting at scale, 2018.

[147] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[148] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, and David Cournapeau. Scikit-learn: Machine Learning in Python. *Machine Learning in Python*, 2011.

[149] Edward Loper and Steven Bird. NLTK: The Natural Language Toolkit. *arXiv:cs/0205028*, May 2002.

[150] Ciprian-Octavian Truica, Julien Velcin, and Alexandru Boicea. Automatic Language Identification for Romance Languages Using Stop Words and Diacritics. In *2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 243–246, Timisoara, Romania, September 2015. IEEE.

[151] Alec Go, Richa Bhayani, and Lei Huang. Twitter Sentiment Classification using Distant Supervision. *CS224N project report, Stanford*, 1, 2009.

[152] T. F. Chan, G. H. Golub, and R. J. LeVeque. Updating Formulae and a Pairwise Algorithm for Computing Sample Variances. In H. Caussinus, P. Ettinger, and R. Tomassone, editors, *COMPSTAT 1982 5th Symposium Held at Toulouse 1982*, pages 30–41. Physica-Verlag HD, Heidelberg, 1982.

[153] Yoav Freund and Robert E Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[154] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[155] Hsiang-Fu Yu, Fang-Lan Huang, and Chih-Jen Lin. Dual coordinate descent methods for logistic regression and maximum entropy models. *Machine Learning*, 85(1-2):41–75, 2011.

[156] Kaggle. Twitter US airline sentiment. `https://www.kaggle.com/crowdflower/twitter-airline-sentiment`, 2018.

[157] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, pages 993–1022, 2003.

[158] Radim Rehurek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010.

[159] Business Insider. JFK, LaGuardia airports closed as 'bomb cyclone' shuts down thousands of flights and strands travelers. `https://www.businessinsider.com/new-york-airports-cancel-flights-bomb-cyclone-winter-storm-2018-1`, 2018.

[160] LaGuardia Airport. `https://twitter.com/LGAairport/status/948747505763840001`, 2018.

[161] Leo Breiman. *Classification and regression trees*. Routledge, 2017.

[162] New York Times. Coronavirus Travel Restrictions, Across the Globe, 2020.

[163] WorldAtlas. Which countries are in mandatory lockdown due to COVID-19?, 2020.

[164] US Department of State. Global Level 4 Health Advisory - Do Not Travel, 2020.

[165] United States Customs and Border Protection. Airport wait times, 2020.

[166] Kari Edison Watkins, Brian Ferris, Alan Borning, G. Scott Rutherford, and David Layton. Where Is My Bus? Impact of mobile real-time information on the perceived and actual wait time of transit riders. *Transportation Research Part A: Policy and Practice*, 45(8):839–848, October 2011.

[167] SafeGraph. SafeGraph COVID-19 data consortium, 2020.

[168] Se-Yeon Jung and Kwang-Eui Yoo. Passenger airline choice behavior for domestic short-haul travel in South Korea. *Journal of Air Transport Management*, 38:43–47, June 2014.

[169] Uber Technologies Inc. Uber, About us, 2020.

[170] Ziru Li, Yili Hong, and Zhongju Zhang. Do On-demand Ride-sharing Services Affect Traffic Congestion? Evidence from Uber Entry. 2016.

[171] Gregory D. Erhardt, Sneha Roy, Drew Cooper, Bhargava Sana, Mei Chen, and Joe Castiglione. Do transportation network companies decrease or increase congestion? *Science Advances*, 5(5), May 2019.

[172] Jonathan D. Hall, Craig Palsson, and Joseph Price. Is Uber a substitute or complement for public transit? *Journal of Urban Economics*, 108:36–50, November 2018.

[173] Mingshu Wang and Lan Mu. Spatial disparities of Uber accessibility: An exploratory analysis in Atlanta, USA. *Computers, Environment and Urban Systems*, 67:169–175, January 2018.

[174] Mackenzie Pearson, Javier Sagastuy, and Sofia Samaniego. Traffic Flow Analysis Using Uber Movement Data. 2018.

[175] INSEE. Institut national de la statistique et des études économiques, 2020.

[176] Gérald Gurtner, Andrew Cook, Anne Graham, and Samuel Cristóbal. The economic value of additional airport departure capacity. *Journal of Air Transport Management*, 69:1–14, June 2018.

[177] Wikipedia. January 2018 North American Blizzard. `http://en.wikipedia.org/wiki/January_2018_North_American_blizzard`, 2018.

[178] Twitter. Twitter developper API. `http://dev.twitter.com`.

[179] Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, pages 399–408, Shanghai, China, 2015. ACM Press.

[180] S.J. Roberts, D. Husmeier, I. Rezek, and W. Penny. Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1133–1142, Nov./1998.

[181] SS Vallender. Calculation of the wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications*, 18(4):784–786, 1974.

[182] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of machine learning research*, 9:2579–2605, 2008.

[183] E. Torres, J.S. Domínguez, L. Valdés, and R. Aza. Passenger waiting time in an airport and expenditure carried out in the commercial area. *Journal of Air Transport Management*, 11(6):363–367, November 2005.

[184] Bryan Roberts, Steve McGonegal, Fynnwin Prager, Dan Wei, Adam Rose, Charles Baschnagel, Timothy Beggs, and Omeed Baghelai. Analysis of primary inspection wait time at U.S. ports of entry. Technical report, 2014.

[185] Hari Bhaskar Sankaranarayanan, Gaurav Agarwal, and Viral Rathod. An exploratory data analysis of airport wait times using big data visualisation techniques. In *2016 International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*, pages 324–329, Bengaluru, India, October 2016. IEEE.

[186] Michael Johnstone, Vu Le, Saeid Nahavandi, and Doug Creighton. A dynamic architecture for increased passenger queue model fidelity. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 3129–3139, Austin, TX, USA, December 2009. IEEE.

[187] Leonard Kleinrock. *Queueing systems, volume 2: Computer applications*, volume 66. Wiley New York, 1976.

[188] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[189] Lijuan Liu and Rung-Ching Chen. A novel passenger flow prediction model using deep learning methods. *Transportation Research Part C: Emerging Technologies*, 84:74–91, 2017.

[190] Yuxing Sun, Biao Leng, and Wei Guan. A novel wavelet-svm short-time passenger flow prediction in beijing subway system. *Neurocomputing*, 166:109–121, 2015.

[191] Yu Wei and Mu-Chen Chen. Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, 21(1):148–162, 2012.

[192] Gang Xie, Shouyang Wang, and Kin Keung Lai. Short-term forecasting of air passenger by using hybrid seasonal decomposition and least squares support vector regression approaches. *Journal of Air Transport Management*, 37:20–26, 2014.

[193] Chaug-Ing Hsu and Yuh-Horng Wen. Improved grey prediction models for the trans-pacific air passenger market. *Transportation planning and Technology*, 22(2):87–107, 1998.

[194] Selvaraj Vasantha Kumar. Traffic flow prediction using kalman filtering technique. *Procedia Engineering*, 187:582–587, 2017.

[195] Billy M Williams and Lester A Hoel. Modeling and forecasting vehicular traffic flow as a seasonal arima process: Theoretical basis and empirical results. *Journal of transportation engineering*, 129(6):664–672, 2003.

[196] S Vasantha Kumar and Lelitha Vanajakshi. Short-term traffic flow prediction using seasonal arima model with limited input data. *European Transport Research Review*, 7(3):21, 2015.

[197] Diane Wilson, Eric K Roe, and S Annie So. Security checkpoint optimizer (sco): an application for simulating the operations of airport security checkpoints. In *Proceedings of the 38th conference on Winter simulation*, pages 529–535. Winter Simulation Conference, 2006.

[198] Kelly Leone and Rongfang Rachel Liu. Improving airport security screening checkpoint operations in the us via paced system design. *Journal of Air Transport Management*, 17(2):62–67, 2011.

[199] Robert De Lange, Ilya Samoilovich, and Bo Van Der Rhee. Virtual queuing at airport security lanes. *European Journal of Operational Research*, 225(1):153–165, 2013.

[200] V.N. Vapnik. *Statistical learning theory*. Adaptive and learning systems for signal processing, communications, and control. Wiley, 1998.

[201] Trevor Hastie, Robert Tibshirani, Jerome Friedman, and James Franklin. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85, 2005.

[202] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[203] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[204] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[205] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[206] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.