UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola Superior d'Enginyeries Industrial, Aeroespacial i Audiovisual de Terrassa

A study of flight cancellation and delays in the UK

Document:
Memòria

Autor:
Alejandro R. Vázquez Ibáñez

Director:
Daniel García-Almiñana

Titulació:
Màster en Ingeniería Industrial

Convocatòria:
Primavera, 2022.

TREBALL DE FI D'ESTUDIS

Resumen

Los retrasos y cancelaciones de vuelos siempre han sido un problema para la industria de la aviación. Sin embargo, la diferente naturaleza de ambos fenómenos ha hecho que la investigación se centre casi exclusivamente en estudiar y predecir retrasos. Esto se debe al hecho de que, en última instancia, es la aerolínea quien decide si un vuelo se cancela, mientras que los retrasos son el resultado involuntario de una amplia gama de causas diferentes, muchas veces debido a las malas prácticas de gestión por parte de los aeropuertos y las aerolíneas.

La literatura ha estudiado los retrasos desde una amplia gama de perspectivas, teniendo en cuenta varios factores que influyen en ellos. Algunos estudios han predicho retrasos desde una perspectiva de aprendizaje automático, mientras que otros han tenido en cuenta la importancia del componente de series temporales de los datos. Sin embargo, la investigación muestra que en realidad son las cancelaciones de vuelos el determinante más importante para la insatisfacción y las quejas de los consumidores, siendo perjudiciales para la reputación de las aerolíneas y dando como resultado que los pasajeros cambien de aerolínea. Por lo tanto, se necesita un estudio y una comprensión más cuidadosos de lo que impulsa y afecta las cancelaciones de vuelos.

Analizando la investigación que se ha centrado en comprender los patrones subyacentes de cancelaciones, lo que más se puede encontrar son enfoques teóricos y de aprendizaje automático. Se han hecho algunos hallazgos para determinar qué aumenta o ayuda a reducir el número de cancelaciones, como la importancia de una capacidad aeroportuaria bien administrada para mejorar la calidad del servicio en términos de cancelaciones \citep{mead2000flight}. Como se mencionó, también hay investigaciones de comportamiento sobre las consecuencias que las cancelaciones tienen en las aerolíneas (Yanying et al., 2019), apuntando hacia un aumento de la insatisfacción y la desconfianza de los clientes, lo que resulta en graves daños para la reputación corporativa de la aerolínea y la lealtad de los pasajeros.

Sin embargo, hay componentes de la comprensión de las cancelaciones que no estaban claros. Por un lado, es necesario realizar un análisis exhaustivo de las series temporales de las cancelaciones. De hecho, como señalan Lemke et al. (p. 85, 2009), las diversas características y los procesos subyacentes de generación de datos de las series temporales han dado lugar al hecho de que "parece que ningún método ha demostrado ser exitoso en varios estudios y series temporales". Por otro lado, los retrasos y las cancelaciones son dos fenómenos que no se pueden entender completamente de forma independiente y, aunque existe un gran número de estudios que analizan la propagación de retrasos, no hay resultados concluyentes sobre el impacto de los retrasos en las cancelaciones. Por lo tanto, la investigación debe determinar si tener en cuenta los retrasos al analizar las cancelaciones mejora la precisión de las previsiones de cancelaciones y la relación entre estos parámetros. Por último, como no solo se pueden estudiar por sí solos, también es necesario realizar un estudio más exhaustivo de los factores de capacidad que influyen en el número de cancelaciones.

Además, el brote del COVID-19 en medio del proceso de investigación hizo que la precisión de los pronósticos se desviara. Los retrasos y cancelaciones han evolucionado de manera dramáticamente diferente durante los primeros meses de 2020. Por lo tanto, es necesario tener en cuenta un nuevo parámetro que ayude a dar sentido a las cancelaciones anormales en 2020 y mejorar las precisiones de las previsiones. Para ello, se han ido tomando en consideración los cambios de comportamiento de la población, lo que se ha hecho con Google Trends. Además, abrió una puerta para comprender la reacción conductual de los pasajeros hacia los viajes aéreos en estas circunstancias, teniendo en cuenta factores locales y globales.

Por lo tanto, este estudio se divide en tres secciones. El primero estudia la relación entre retrasos y cancelaciones desde una perspectiva de series temporales, y se encuentra que tener en cuenta los retrasos como parámetro en el estudio de las cancelaciones mejora la precisión de las previsiones de series temporales en diferentes niveles de agregación. El segundo se centra en estudiar la relevancia de la competencia y los factores de red en la distribución de las cancelaciones. Se encuentra que los vuelos que llegan o salen de un aeropuerto central tienen menos probabilidades de ser cancelados, lo que apunta hacia la relevancia de mantener las redes para las aerolíneas, fortaleciendo así la confiabilidad y la confianza de los pasajeros. Sin embargo, se encontró que la competencia de la ruta y el aeropuerto, aunque confirma la naturaleza del impacto, no era estadísticamente significativa para predecir la cancelación de vuelos. Por último, se encontró que la preocupación pública en el contexto de una pandemia global varía según las circunstancias locales, y que poco después de la primera y más impactante noticia, tanto la preocupación como una actitud de consumo positiva disminuyen a un nivel estabilizado, lo que indica un comportamiento pasivo de doble filo, en el que tanto la preocupación como la voluntad de comprar boletos de vuelo o evento (es decir, que requieren viajes o reuniones sociales) se reducen a niveles similares y bajos durante al menos un mes después del *mayhem* inicial.

Abstract

Flight delays and flight cancellations have always been a problem for the aviation industry. However, the different nature of both phenomena has made research focus almost solely on studying and predicting delays. This is due to the fact that, ultimately, it is the airline who decides whether a flight gets cancelled, whereas delays are an involuntary result of a vast array of different causes, many times due to bad management practices by airports and airlines.

The literature has studied delays from a wide range of perspectives, taking into consideration several factors that influence them. Some studies have predicted delays from a machine learning perspective, while others have taken into consideration the importance of the time series component of the data. However, research shows that it is actually flight cancellations that is the most important determinant for consumer dissatisfaction and complaints, being detrimental for airlines' reputation and resulting in passengers switching carriers. Therefore, a more careful study and comprehension of what drives and affects flight cancellations is needed.

Analyzing the research that has focused on understanding the underlying patterns of cancellations, what can mostly be found are theoretical and machine learning approaches. Some findings have been made in determining what further increases or helps reduce the number of cancellations, like the importance of a well-managed airport capacity to improve service quality in terms of cancellations \citep{mead2000flight}. As mentioned, there is also behavioral research on the consequences that cancellations have on airlines (Yanying et al., 2019), pointing towards an increased dissatisfaction and distrust from customers, resulting in serious damages for the airline's corporate reputation and passengers' loyalty.

Nevertheless, there are components of the understanding of cancellations that remained unclear. On the one hand, a thorough time series analysis of cancellations needs to be done. In fact, as Lemke et al. (p. 85, 2009) point out, the diverse characteristics and underlying data generation processes of time series has resulted in the fact that "it seems as if no method has ever proven successful across various studies and time series". On the other hand, delays and cancellations are two phenomena that cannot be completely understood independently and, although there is a vast number of studies analyzing delay propagation, there are no conclusive results on the impact of delays on cancellations. Therefore, research must determine whether taking delays into account when analyzing cancellations improves the accuracy of cancellations forecasts and the relation among these parameters. Lastly, as they cannot only be studied alone, a more thorough study of the capacity factors that influence the number of cancellations also needs to be done.

Moreover, the outbreak of the COVID-19 in the midst of the research process made the accuracy of the forecasts deviate. Delays and cancellations have evolved dramatically differently over the first months of 2020. Hence, there is a need for taking a new parameter into account that would help make sense of the abnormal cancellations in 2020 and improve forecasts accuracies. For this, the behavioral changes of the population have been taking into consideration, which has been done with Google Trends. Also, it opened a door for understanding the passengers' behavioral reaction towards air travel under these circumstances, taking into consideration both local and global factors.

Therefore, this study is divided into three sections. The first one studies the relationship between delays and cancellations from a time series perspective, and it is found that taking delays into account as a parameter in the study of cancellations improves the accuracy of time series forecasts at different levels of aggregation. The second one focuses on studying the relevance of competition and network factors in the distribution of cancellations. Flights arriving or departing from a hub airport are found to be less likely to be cancelled, pointing towards the relevance of maintaining networks for airlines, thus strengthening passenger reliability and trust. However, it was found that route and airport competition, while confirming the nature of the impact, was not statistically significant in predicting flight cancellations. Finally, it was found that public concern in the context of a global pandemic varies according to local circumstances, and that shortly after the first and most shocking news, both concern and a positive consumer attitude decrease to a stabilized level, which indicating double-edged passive behavior, in which both concern and willingness to purchase flight or event tickets (i.e., requiring travel or social gatherings) are reduced to similarly low levels for at least one month after the initial mayhem.

Index

List of Tables

List of Figures

1. Introduction

1.1 Objective

So far, literature has found that cancellations are the biggest motive for passenger dissatisfaction and lack of trust, and that there may exist a relationship between cancellations and delays. Yanying et al. (2019) deepened into the consequences of cancellations, stating that passengers' dissatisfaction and distrust of airlines as a consequence of cancelled flights seriously damage the airlines' corporate reputation and affect passengers' loyalty, making them switch airlines. Rupp and Holmes (2006) found that cancellations are more inconvenient for passengers than flight delays and that, depending on the circumstances, delays and cancellations can behave either like substitute or complementary goods. Xiong and Hansen (2013) explains how the cancellation decision is a trade between a fixed cancellation cost and a duration-dependent delay cost. However, none of them took the time series component of the data into account. Rupp and Holmes (2006) also found that networks and competition in US domestic flights affected cancellations, but did not take into account the effect of congestion (which is a concept that is explored alone by Mead and General (2000)) or international flights.

This case-study consists of two parts. The first part aims to answer the following: Do delays improve cancellations forecasts accuracy over time, and how important are airport factors in determining cancellation percentages? This will help regions and airlines to improve service quality, i.e., decreased cancellations, therefore making them able to trace short- and long-term horizon preventive plans to increase the reliability of flights. Usually, the cancellation decision is made days or even hours prior to the flight departure, so forecasting the distribution of cancellations over time and identifying time patterns in the data will help airlines reallocate resources for fidelity programs or marketing campaigns to make up for the loss of trust of passengers that comes hand by hand with the forecasted levels of cancellations.

Thus, this first part is confirmatory, exploring how delays affect cancellations, and whether delays improve cancellation forecasts. Research shows that there is no clear consensus regarding this issue, and has failed to study this relationship while taking into consideration how time affects it. Lemke et al. (2009) state that no time series method has ever actually proven successful across several studies, as it is hardly possible to make a model that fits every case, year and region. The diverse characteristics and underlying data generation processes of time series can make it impossible to design a method that works well in all cases. Similarly, the effects of networks and competition was studied for US domestic flights, without taking into effect the impact of congestion, which is closely related to these two factors.

The second part takes into accoun the effects of the COVID-19. In December of 2020, a novel virus appeared in the Hubei province in China, rapidly spread- ing internationally and with immediate economic and social consequences. The Economist (2020a, p.1) remarks how "the COVID-19 is a grave threat to the market's poise. News from Italy of the biggest virus outbreak outside Asia led to a 3.4% decline in the S&P 500 index of American stocks on February 24th, the biggest one-day fall for two years". At the same time, whole cities and provinces are being put into quarantine, achieving citizen protection at high costs (The Economist, 2020b). These instances exemplify how the global economy is experiencing changes and that air transport is expected to be one of the most affected industries by the outbreak because of factors such as fear, strict quarantines, national lockdowns, etc., which will increase the uncertainties and irregularities in air travel demand and cancellations (Bogoch et al., 2020).

Airlines are being forced to cancel flights when bookings are low. Low bookings are a result of public fear or concern with regards to the virus and the public consumer optimism or purchasing intention, which can be measured with Google Trends by analyzing the right queries. This increase in cancellations as a result of low bookings translates in bad reputation to the eyes of the customers that do buy tickets, but see their flights cancelled. These dissatisfied customers will later tell the ones that did not book and were the actual source of the cancellations, which will also have in mind a bad image of the airline even if it was ta consequence of their acts. Mainly, the two kinds of activities that were prohibited as a result of the COVID-19 virus were social acts and travel. Therefore, a greater concern can translate in a decreased willingness to engage in these activities. If it is found how the concern develops and the willingness to engage in these activities (as the optimism in thinking that social events and travel will be allowed soon and therefore and increased search for tickets) evolves, airlines can have a better idea about the distribution of cancellations as a result of low bookings. Lastly, regional circumstances can have a bigger impact than global ones (Ribes, 1992), and it can be useful for airlines to know how this concern develops in function of the timeline and intensity of events in different countries. Therefore, the second part of the study will approach the consequences of the COVID-19 spread on public behavior and the public opinion regarding air transport and willingness to engage in events (that would mark the end of restrictions) by means of analyzing changes in Google search queries throughout those months.

Hence, the second part is exploratory. It answers the following questions: "Do the changes in people's behavior help shed light into the abrupt changes in cancellations as a result of the pandemic? What are the underlying reasons for these changes in behavior?".

1.2 Scope

The working packages were the programming language R using the interface R Studio, and the different deliverables are the code (in Annex) and this memoir.

1.3 Requirements

So far, the literature (Abdel-Aty et al., 2007; Manna et al., 2017; Kuhn and Jamadagni, 2017; Mueller and Chatterji, 2002; Pyrgiotis et al., 2013; Xu et al., 2005; Rebollo and Balakrishnan, 2014; Sternberg et al., 2017; Tu et al., 2008; Wu, 2014) has mainly focused on flight delays. Although there are studies analyzing the relationship between delays and cancellations (Rupp and Holmes, 2006; Xiong and Hansen, 2013), they fail to include the time series component of the data, and actually (Rupp and Holmes, 2006) only makes an analysis of US domestic flights. When taking into account the effect of networks and competition, there are two things that seem to be unavoidable for the correct understanding of cancellations and have been ignored – international flights and the effect of airport congestion, which is closely tied to networks and competition and is studied by itself by Mead and General (2000).

This research proposes a time series forecasting method that takes into account the non-stationarity of the delay series. As Lemke et al. (2009, p.85) point out in their research, "it seems as if no method has ever proven successful across various studies and time series. This is mainly due to the fact that time series can have very diverse characteristics and underlying data generation processes, which makes it impossible to design a method working well for all of them". Therefore, by analyzing cancellations and delays data from a time series perspective, the lack of a standardized method for all dataset can be compensated by having specific studies that focus on different data formats. In particular, this study aims to provide a better understanding of cancellations at a monthly level in the UK. This study aims to fill that gap in the existing literature by clearly determining the impact of delays in forecasting cancellations in a time series and by nuancing the research on how networks, competition and congestion affect them. In addition, this will be among the first studies to analyze the impact of the COVID-19 on public behaviors regarding flight cancellations in particular and air travel in general.

1.4 Justification

Flight delays are a great concern for airlines and passengers, and they have been widely studied from various approaches (Sternberg et al., 2017), including probabilistic models, network representation, operational research and machine learning. However, little research has been made on cancellations, probably because it is the airlines' decision to cancel a flight (even if it may be for economic or external reasons), while delays are usually due to poor airport or air carrier management, e.g. maintenance or crew problems, etc.. Yet, some studies (Rupp, 2005) emphasize the importance of analyzing cancellations, as research demonstrates that flight cancellations are the most relevant metrics for passenger dissatisfaction and complaints (Barnhart and Bratu, 2004), and may cause passengers to switch carriers, as well as big detriments for airlines' reputation.

Some research has narrowed its focus to flight cancellations (Rupp and Holmes, 2006; Xiong and Hansen, 2013; Sridhar et al., 2009; Lambelho et al., 2020; Yanying et al., 2019) using classification techniques that fail to include the time-series component of the data series. There have also been studies emphasizing on the consequences of flight cancellations (Yanying et al., 2019), or just theoretical approaches to what affect flight cancellations (Mead and General, 2000; Sridhar et al., 2009; Lambelho et al., 2020).

Despite the research on flight delays and cancellations, there are no conclusive results on the relationship between them both, and even though there are studies that state that there is a relationship among them, there is no consensus on what impact they have on each other. Rupp and Holmes (2006) suggest that cancellations and delays may behave like substitute goods in certain occasions whereas, in other cases, they act as complementary goods that occur in unison. Xiong and Hansen (2013) state that cancellation decisions are a trade between a fixed cancellation cost and a duration-dependent delay cost, thus resulting in a direct non-linear impact of delays in cancellations. However, inconsistently with the findings of Lemke et al. (2009), which says that a time series study is needed and that there is currently no one-solution-fits-all approach, both of these studies fail to take into consideration the time series of the air transport data.

Moreover, delays and cancellations cannot be only studied alone. Even though this focalized study is needed as a first step to isolate their mutual influence, there is a need to contextualize the findings. For that matter, and nuancing the research proposed by Rupp and Holmes (2006), which performs a similar analysis for US domestic flights, a regression analysis is performed of cancellations and delays in which the role of competition, hubs and congestion is measured.

In addition, the global outbreak of the COVID-19 in the first months of 2020 is expected to incur changes in social and economic dynamics (Bogoch et al., 2020). This pandemic has resulted in an exponential increase of flight cancellations, reaching levels that had never been seen before. This makes the previous forecasts obsolete, as no forecast can foresee this rapid increase. Therefore, a need for including an external parameter that helps improve the reliability of the forecasts appears. Because a large reason why so many flights have been cancelled is been the low number of bookings, which make the fleet of planes economically inviable, it is suggested that a good proxy may be changes in people's behavior, which is analyzed with Google Trends data. This will also open a window for better understanding the consuming behaviors under different local and global circumstances.

These different consequences come from the fact that, amidst the crisis created by this pandemic (which has killed more people alone than the MERS and the SARS combined despite its low fatality rate (Mahase, 2020)), countries around the globe have reacted differently. Some countries chose to follow highly strict contention measures to avoid the spread of the virus – specially affected regions, and sometimes even whole countries, were put in quarantine for periods ranging from 3 weeks to 3 months. At the same time, in these and other countries, small businesses, restaurants, gyms, bars, pubs, etc. were closed to prevent public gatherings, and lectures were cancelled as universities and schools also closed.

Many borders were closed, mainly to prevent visitors from regions with higher infection rate – thus, potential carriers of the virus – from entering countries with a lower number of infections. Governments also put restrictions on the occupation rate of planes to avoid short distances among passengers and potential infections. On top of this, the international health organisms like the World Health Organization or the CDC recommended individuals to limit their trips and avoid unnecessary travel, specially to highly affected areas (World Health Organization, 2020; Centers for Disease Control and Prevention, 2020a), making passengers wary. Additionally, travelers were informed that highly crowded spaces, such as airports, train stations, etc. increase the possibilities of being infected with COVID-19, as there is a chance of there being other travelers infected with the virus (Centers for Disease Control and Prevention, 2020b). This means that, on top of the governmental restrictions imposed on travelling, the effect of public fear also had an impact on a further decrease of flight bookings. This public concern about the COVID-19 virus spread and behavioral changes can be measured with Google Trends, and this study evaluates it by analyzing searches related to consuming intention and cancellations in the context of a pandemic. As explained before, these issues and regulations resulted in extremely low occupancy rates, as passengers were not purchasing flight tickets, which in turn led to air carriers cancelling most of their flights from the end of March until the end of June.

Summing up, this research proposes a time series analysis comparing flight cancellations with flight delays in a series of UK airports. Additionally, an analysis of cancellations and how competition, congestion and hubs influence them will help to nuance the findings about delays. It will also aid regional governments in evaluating addressable measures to improve the air transport service offered to travelers. Lastly, the effects of the COVID-19 will be considered, thus providing stakeholders a better understanding of what are the behavioral changes that occur in the population throughout the different phases of a pandemic.

1.5 Managerial relevance

First, regional governments can improve the regions' overall service quality at the expense of particular airports'. By nuancing what is found about hubs, competition and congestion, and comparing it to delays, governments could negotiate with airlines to establish hubs in their region by moving airlines from smaller airports in that same region to the desired hub, at the same time reducing the congestion in the smaller ones. Additionally, airlines and third parties (such as insurance companies) can benefit from an improved under- standing of cancellations. By having a distribution time of the forecasted levels of cancellations, airlines can allocate resources to fidelity programs or marketing campaigns to palliate the loss in trust that they know is likely to occur at a given time, without wasting resources at times where maybe the public opinion is more positive. Insurance companies will have an increased transparency as they are going to be able to better determine the actual causes of cancellations (when faced with insurance claims). Thanks to knowing whether delays are relevant and how they affect cancellations, they will be able to take into account the delay variable into their operations, so that they can better trace the reasons why the flight was cancelled.

Lastly, the behavioral study of the impact of the COVID-19 can benefit airlines by providing them with a detailed analysis of how the concern and the purchasing intention varies over time and in different regions depending on local circumstances after such a drastic event, thus being able to implement new business models (e.g. subscription-based pricing for low-cost airlines) that are able to secure revenue incomes that balance the losses due to cancellations. By knowing how these behaviors evolve over time, they will be able to trace preventive plans beforehand to optimize fleet utilization in times where there is no room for error in such an impacted industry.

2 Theoretical background

In this section, an overview of the current situation of the airline industry is presented, exhibiting the impact of this industry on the global economy and society, and the important role that it has played towards globalisation. Next, the consequences of cancellations in terms of economic losses and customer dissatisfaction and churn will be presented, as well as a summary of the conditions under which a flight can be cancelled and how these differ from those that might delay a flight. Lastly, there will be an analysis of how research has approached these subjects so far, pinpointing the excessive attention that delays have received so far in contrast to cancellations, and how cancellation modelling has been studied thus far.

2.1 The airline industry

Even though the magnitude and impact of globalisation have considerably increased in the last twenty years, the definition of the concept by Tomlinson (1999) is still precise. Tomlinson (1999, p.1) explained that "globalization refers to the rapidly developing and ever-densening network of interconnections and interdependencies that characterize modern social life. At its most basic, global- ization is quite simply a description of these networks and of their implications – for instance in the various 'flows' – of capital, commodities, people, knowledge, information and ideas, crime, pollution, diseases, fashions, beliefs, images and so on – across international boundaries". In summary, it is a process of accelerat- ing global connectivity. It involves rapid and simultaneous social change across many dimensions (such as economy, politics, communications, healthcare, the physical environment and culture), interacting with one another (Tomlinson, 1999).

The rise of globalisation has resulted in an increased need for travelling long distances (Lee et al., 2015). As Frankel (2000) stated, over the period from 1920 to 1990, the average ocean freight and port charges per ton in the US fell from $95 to $29 (in 1995 US dollars), while an increasing share of cargo was being transported by air. Air shipping and refrigeration of goods changed the status of goods that had previously classified as non-tradeable internationally. "Now fresh-cut flowers, perishable broccoli and strawberries, live lobsters and even ice-cream are sent between continents" (Frankel, 2000, p.3).

But globalisation has not only affected the transport of consumable goods. One of the biggest changes that globalisation brought was the possibility of individuals travelling almost anywhere in the world in less than a day in a relatively affordable price. Reasonably, this led to an ever-increasing volume of global passenger transport. The wide set of reasons why people travel has been widely researched, and they include motives such as business, visiting friends and family, to relax or get away, searching of new experiences, or self-developing (Lee et al., 2015). Companies are trading worldwide, and business had to adapt to it, having meetings with companies in different continents, making regular trips abroad, and sending cargo overseas. At the same time, individuals have seen flight tickets' prices fall dramatically over the recent years, due in part to the fact that, in many affluent countries, air travel is subsidised, which creates the basis for the social norm of cheap flight (Gössling et al., 2019), thus increasing demand and makes it possible to travel by air more affordable and accessible. One of the main triggers for this increase in demand is the rise of low-cost carriers which, as displayed by Gössling et al. (2019), keep offering flight tickets at a price that is equal or below to that of fuel and handling fees, and usually far lower than the equivalent price of the much slower train travel, making the amount of air travel taken on by the highest income groups increase considerably.

For these reasons, travelling by air has become the standard in today's society. As stated by Gössling et al. (2019, p.2), "Air travel is regularly presented as a social norm, specifically by aviation organizations and airlines. This creates and fosters various discourses and mechanisms designed to strengthen a social norm of flying". Some of the examples promoting this view are "support of cheap flights, traveller self-promotion through frequent flier programmes, as well as wider issues of economic development, employment or intercultural understanding" (Gössling et al., 2019, p.2). Logically, this boost in demand has made the airline industry increase not only in size, but also in logistic complexity.

However, there are also studies who warn about the overuse of air travel and question the actual need for it (Gössling et al., 2019), as well as research that considers whether today's expectations for growing air travel demand are realistic (Becken and Carmignani, 2020). This concern with the actual need for air travel is due to the fact that this increasing 'primal' need for aviation is in conflict with societal goals to reduce and limit environmental pollution and climate change, and challenges involving air and noise pollution, as well as infrastructure expansion (Heuwieser, 2017; Gössling and Upham, 2009). It contradicts, among others, the United Nations' Sustainable Development Goals' (UN, 2015) SDG 12 (Responsible Production and Consumption) and SDG 13 (Climate Action), in addition to the Paris Agreement (Scott et al., 2016). Emissions are projected to grow by a factor of 2.8–3.9 between 2010 and 2040 (ICAO, 2016) and neither innovative technologies, the aviation sector's own strategies, nor market-based changes in dynamics (such as carbon pricing) are expected to solve this (Lyle, 2018; Markham et al., 2018). Therefore, the only way to avoid unnecessary emissions and reduce the impact of aviation is to make air transport more efficient. And a large part of this efficiency depends on an improved service quality in terms of cancellations and delays.

In the semi-annual report released by IATA (2019), it is stated how consumers are expected to spend 1% of the world's GDP (totalling $908 billion) in airline transport in 2020, as a result of lower real travel costs and more routes, among others. In addition, the number of new destinations is forecasted to boost further this year, with trip frequencies up too; both increasing consumer benefits. This forecasts, however, have been made completely useless after the COVID-19. Iacus et al. (2020), via a series of hypothetical scenarios, forecast the global air passenger volume to fall to a 50% in the best-case scenario, and 20% in the worst case, and expect air travel to stabilize between August and October 2020, reaching the 100% pre COVID-19 air passenger volume in the best case, and a shier 60% in the worst during that period. The stakeholder that has come out worst are airlines, as they have suffered incommensurate losses. Therefore, a meticulous care for efficiency is going to be crucial for their survival, as cancellations suppose economic and reputatinal losses that they are not going to be able to afford anymore.

Table 1 (IATA, 2019) shows a comparison between air transport financial figures in 2018 and 2019 between Europe and North America, plus a prediction of 2020 financial figures. In Europe, the net post-tax benefit decreased from $9.1 billion to $6.2 billion, although it is expected to increase in 2020 to $7.9 billion. One of the reasons of this rebound in profits is the increase in average profit per passenger, which is expected to boost to an amount of $6.40 per passenger after suffering a decrease of a 34.38% ($2.73 per passenger) in 2019 with respect to 2018. These numbers, however, show a performance far from that one in North America, where the airline industry has higher profits and better profits per passenger. As a matter of fact, while in Europe this year's net post-tax profit added up to $6.2 billion, in North America this quantity was almost tripled, with after-tax profits of $16.9 billion in 2019. The profits per passenger show an even steeper difference, being of $16.81 in North America while only $5.21 in Europe.

Table 1. Comparison of airline industry profitability: Europe and North America (IATA, 2019)

	Europe			North America		
	2018	2019	2020P	2018	2019	2020P
Net post-tax profit, $billion	9.1	6.2	7.9	14.5	16.9	16.5
Net profit per passenger, $	7.94	5.21	6.4	14.66	16.81	16.00

As shown in the report by IATA (2019), by the end of 2020 there will be around 4.7 million available seats, as average size of aircraft in the fleet is continuing to rise. Critically for profitability in such a a capital-intensive industry, these seats are going to continue be used intensively although less than before, sine passenger load factors are expected to ease from all-time high levels to 82.0% on average in 2020. The intensity of flown aircrafts is nevertheless expected to increase to surpass 40 million next year, making an average of 77 aircraft departing each minute of 2020. This means a higher environmental impact, because whereas the load factor is forecasted to decrease, the total number of air transport movements is expected to rise. Additionally, and more relevantly for this research, this relation implies an increase in logistic complexity, as more flights are going to need to be coordinated. Table 2 displays these figures, along with other important ones, for the years 2018 and 2019, in addition to the quantities forecasted for the upcoming year 2020.

Table 2. Worldwide airline industry aircraft figures and change over year (%) (IATA, 2019)

Worldwide airline Industry	2018	2019	2020P
Aircraft fleet	29,507 (4.4%)	29,805 (1.0%)	31,375 (5.3%)
Available seats, million	4.4 (6.1%)	4.5 (1.8%)	4.7 (6.0%)
Average aircraft size, seats	149 (1.6%)	150 (0.8%)	151 (0.7%)
Scheduled flights, million	38.1 (4.5%)	39.0 (2.3%)	40.3 (3.4%)
Passenger load factor, %	81.9%	82.4%	82.0%
Freight load factor, %	49.3%	46.7%	46.3%

Summing up, aviation is one of the industries with greater economic impact, contributing with a 3.6% of the world GDP (US$ 2.7 trillion) and generating a total of 29 million jobs globally (Air Transport Action Group, 2019; ICAO, 2011). Therefore, a careful study of the phenomena driving its dynamics and efficiency optimization is of high relevance.

2.2 Flight delays and cancellations

Cancellations may happen for a number of reasons. However, ultimately it is the flight operators' (in this case, airlines') decision to cancel their flights (Xiong and Hansen, 2013). But conditions differ.

Rupp and Holmes (2006) envisions two kinds of cancellations: 'stochastic cancels' and 'strategic cancellations'. The first ones take place when external factors make the available short-run aircraft or in-arrival-airport slot number excessively decreases, may be caused by extreme or unusual weather conditions or an equipment failure or maintenance is required. The second ones are situations in which cancellations occur for economic reasons, e.g. low passen- ger booking. This does not mean that economic factors do not play a role in stochastic cancellations, since aircrafts prioritize high-margin flights over their low-margin ones. For example, the delay of a 300-seat flight and that of a 30-seat flight are likely to have different costs to an airline. Just as well, the impact of cancelling a flight in a high-frequency route or segment may be different than that of cancelling a flight on a low-frequency one (Xiong and Hansen, 2013).

Dwelling deeper into arrival slot management and ‘schotastic cancels’, an- other factor that plays a role in cancellations is slot allocation by the Ground Delay Program. The Ground Delay Program (GDP) (Xiong and Hansen, 2013) is a kind of Air Traffic Management Initiative (TMI), which are measures em- ployed by different air traffic administrations to balance demand with capacity – at airports and airspace. Their objective is to change demand to alternative times or different routes, so that delay and cancellations are reduced, while reliability and passenger safety are increased. They are used when conditions are not ideal. Thus, the Ground Delay Programs are airport-specific TMIs that are implemented when the capacity of arrival at that airport is reduced for a sustained period of time. Flights are assigned waiting times before departures (also known as expect departure clearance times, or EDCTs) that ensure that the aircraft only takes off once it is sure that there is a slot for the arrival at the desti- nation airport, which is the one that the GDP refers to. This way, international aviation agencies and the airlines collaborate together to best use the scarce capacity, letting airlines rearrange and cancel flights to best use the arrival slots they were assigned at the beginning of the GDP. Actually, airlines are given the opportunity to manage delays along their own flights, even being able to cancel flights to avoid excess delays or to vacate arrival slots for other flights. In other words, if an airline with a cancelled flight can use this slot for another of their flights, it may do so. Nevertheless, if not, this slot can be allocated to another airline not to be wasted. This open slot, however, is only created with the aforementioned airline has moved up so many flights that no new flights can fit into the open slot. Then, it can be used by another airline.

Additionally, as air traffic at major airports is usually scheduled near their maximum capacity (and sometimes above it), a drop in capacity can easily result in a demand-capacity imbalance. Given that airplanes' turnaround times (the difference between the scheduled time a plane lands from the coming flight and the scheduled time it departs again for its next flight) have been increasingly shortened, they usually fall short compared to the magnitude of the delays (Xiong and Hansen, 2013). Xiong and Hansen (2013, p.1) explains, "scheduled turnaround times (from the scheduled arrival time to the schedule departure time of a given aircraft) are often quite short relative to the magnitudes of the delays incurred". When flights arrive this late and turnaround times are relatively short, this can lead to higher amounts of cancellations. In addition, Pyrgiotis et al. (2013) explain how a flight arriving 1 hour late in the morning can result in a propagation of delays in the following airports which would ultimately lead to delays of 7h in the evening. Therefore, an arrival delay can result in a departure delay. As these effects usually propagate through networks, the effects accumulate and can lead to cancellations at other airports later that day.

Other determinants for flight ’stochastic’ cancellations are the strict rules applied by the aviation agencies for maximum employee working hours. If a flight is delayed or takes more time than expected (prolonged on-duty time), this may cause the pilot to run short of working hours left for that day. This would cause the airline to cancel that pilot's next flight. In addition, other external factors such as weather uncertainty or aircraft maintenance and performance limitations play a role in increasing the probabilities of flight cancellations.

The various international aviation agencies (such as the FAA or the EASA) do not have decision-making authority when it comes to cancelling flights – they can delay a flight (in some cases for extremely long periods of time), but they cannot cancel one. This is the airlines' decision. There are also not rules that forbid a flight cancellation. This may be one of the reasons why most research focuses on delays, since airlines cancel flights voluntarily, whereas delays are mainly involuntarily caused by external factors and bad management. Barnhart and Bratu (2004) suggest that the most relevant metrics to customers are flight cancellations and the percentage of flights that are delayed by more than 45 minutes. Rupp (2005) also discussed that delays and cancellations move in opposite directions, which suggests a trade-off by the carrier between fewer (more) flight cancellations for more (fewer) delays. However, his model does not provide a concise trade-off between delays and cancellations. At the same time, Rupp (2005) proposes that cancellations and delays can either behave like substitute goods or as complementary goods that occur in unison, under varying circumstances and in different situations. Xiong and Hansen (2013, p.75) state that "all else factors being equal, the cancellation decision is a choice between incurring costs from a cancellation and costs from delay if the flight is not cancelled". Therefore, cancellation decisions would be a trade between a fixed cancellation cost and a duration-dependent delay cost, and that delays to a certain flight and potential delay savings to other flights affect flight cancellation decisions.

Delays and cancellations cause: (1) passengers' dissatisfaction and distrust; and (2) detriments to airline profit and passenger welfare. Nonetheless, data Office of Aviation Enforcement and Proceedings (2020) shows that even when, from November 2018 to November 2019, the number of complaints due to flight problems (cancellations, delays and misconnections, which are deviations from schedule, whether planned or unplanned) reduced from 301 to 226, the percentage of those complaints due to cancellations compared to those due to delays increased from a 47.9% to a 56.2%. Additionally, the percentage of complaints due to flight problems increased in the same period, as shown in Table 3, even when the total amount of computed complaints decreased. This shows an overall improvement of customer satisfaction, although a relative minor improvement in flight problems management. Given that the complaints caused by flight cancellations now make up for the biggest percentage, this illustrates a considerable low performance in managing flight cancellations in particular, which translates to poor customer satisfaction and potential switch of carriers. Indeed, research shows that cancellations are more inconvenient to passengers than flight delays (Rupp and Holmes, 2006). Illustratively, delayed flights take an average of 52 minutes to departure from their origin while, if a flight gets cancelled, the fortunate passenger who is able to get a ticket for the next on-route scheduled flight waits an average of 5 hours.

Table 3. Percentages of complaints due to flight problems in the US. Comparison 2019-2018 (Office of Aviation Enforcement and Proceedings, 2020)

	Flight problems	Total
Aircraft fleet	**165**	**533**
% of total complaints	*31.0%*	
Available seats, million	**206**	**697**
% of total complaints	*29.6%*	

Research on cancellations is of major relevance for a number of reasons. First, research shows that cancellations are more detrimental and inconvenient for passengers than delays. Second, airline performance is a high priority for travellers, airlines and lawmakers. Third, recent social changes increase the need of solid models that can predict flight cancellation. The recent Coronavirus outbreak in Wuhan (Hubei), China has changed the dynamics of air travel during the first months of 2020, making governments impose a curfew on all civilians within the affected areas (Oliphant and Carpani, 2020). This influenza- like pandemic may affect cancellations since novel pathogens have a high chance of rapid appearance and global spread, "with potentially serious global consequences" (Bogoch et al., 2020, p.2).

As pointed out before, the main objective for airlines to cancel flights is profit maximization – in the short term, carriers will tend to avoid canceling a high-profit flight in order to avoid the expensive reimbursements to those customers who decide to abort their trip due to excessive service disruptions. As Rupp and Holmes (2006, p.753) remark, "if the theoretical switching model, proposed by Suzuki (2000) and calibrated with aggregate US DOT data, is accurate and passengers who experience poor service quality are more likely to switch carriers, then the effect of a flight cancellation may also be felt by the carrier long-term". As usually travelers tend to blame airlines for all flight cancellations (even when it is out of their control due to unusual conditions such as extreme weather), this becomes especially relevant. The research by Rupp and Holmes (2006) finds support on maximizing revenue being the main objective of airlines when taking flight cancellation decisions, as they find that there is a significant reduction in cancellations on routes with higher average margin. Additionally, this research also shows that airlines try to minimize passenger inconvenience by not cancelling fuller planes, more infrequently served routes and the final flight of the day.

2.3 Machine learning and theoretical approaches to flight cancellation

Research has increasingly sought to understand the underlying phenomena that drive flight delays and, to a lesser extent, flight cancellations. The amount of papers exploring this subject has grown in the late 2000s since 87.5% of the works had been published between 2007 and 2017 (Sternberg et al., 2017). As previously noted, most research focuses on predicting flight delays, probably due to the fact that, while the airlines are the ones who decide whether to cancel a flight (even though sometimes it may be due to external conditions, such as weather), this is not the case for delays, which usually occur because of aviation agencies' management.

Flight delays have been approached by many researchers, which have pro- posed a number of different methods and models to predict and classify them. Wu (2014) proposes a probabilistic distribution of Airport arrival and departure delays over a selected period of eight months at the Beijing Capital International Airport, whereas Pyrgiotis et al. (2013) study delay propagation in a network of airports using an Approximate Network Delays model (AND). It is described as a stochastic and dynamic queuing model which computes approximate delays of each of the individual airports in that network. This model requires three inputs (complete daily demand schedules at each airport, expected service rates at each airport, and aircraft itineraries) and explains how these delays propagate from one airport to another over the course of a selected time period. Pyrgiotis et al. (2013) also approaches the subject of delay propagation in a network of airports but using Bayesian networks instead.

Sternberg et al. (2017), who studied the prediction of flight delay, delay propagation and cancellation, found that machine learning approaches experienced a major increase in the late 2000s, especially in root delay. They state how machine learning and data management are positively correlated, as the more machine learning is used, the more data management is required, due to the fact that currently, data is gathered in extensive amounts from sensors and IoT devices. Sternberg et al. (2017), studying papers published between 2015 and 2017 related to flight delays and machine learning, find that there is a trend among researchers and, with text-mining methods analyzing these studies, they find that the terms algorithm, learn, big data, data model or train-test are becoming more frequent, which graphically demonstrates the changing trend in delays researching methods.

This illustrates the increasing use of machine learning algorithms to provide flight delay predictions. Methods like k-Nearest Neighbors, neural networks, SVM, fuzzy logic, and random forests are the ones used in relevant research for classification and prediction. The time horizons of these predictors range between 15 minutes before the scheduled time of departure and 24 hours. Rebollo and Balakrishnan (2014) use a random forest model to predict flight delay within a time horizon of 2, 4, 6 and 24h, although their test error increased as did the forecast horizon. Additionally, Lu et al. (2008) employs a k-NN algorithm approach to propagation effects, while Khanmohammadi et al. (2014) created an adapted network based on a fuzzy inference system to predict root delays at New York City's JFK International Airport, and Balakrishna et al. (2010; 2008) used a reinforcement learning algorithm to predict taxi-out delays (15 minutes before the scheduled time of departure) at New York City's JFK International Airport and Tampa Bay International Airport. Manna et al. (2017) also proposes a gradient boosting machine model to predict flight delays.

As Sternberg et al. (2017) notes, there have been temporal approaches to explain phenomena such as seasonality or periodic patterns of data. These works contain characteristics regarding date (season, month, and day of the week) and time (the day or time of the day). Indeed, part of the research aims to provide statistical approaches to flight delays and cancellations. Mueller and Chatterji (2002) make an analysis of arrivals and departures by using Poisson and Normal distributions – models that are improved by adjusting the mean and standard deviation values by implementing a least-squares method, designed to minimize the fit error between the raw distribution and the model. They examine the correlation between the number of departures, number of arrivals and departure delays from a time-series modeling perspective, although no clear concluding evidence is provided. Meyn (2002) employs a probabilistic method to forecast air traffic demand using a probability distribution of an aircraft's location about a nominal location or as a distribution in time about a reference time (i.e. the sector boundary crossing time). Given the likelihood of unanticipated events preventing the plane to take off at the scheduled time, this stochastic approach may be beneficial, even if early and accurate intent information is provided. Tu et al. (2008) identify and study major factors influencing flight delays at the Denver International Airport, while developing a departure delay prediction model that employs nonparametric methods for daily and seasonal trends. With a statistical approach, the model uses a mixture distribution to estimate the residual errors. On the other hand, Abdel-Aty et al. (2007) perform a frequency analysis to detect departure delays parameters at Orlando International Airport as a function of cyclic variations in both air travel demand and weather at that airport. It makes a two-stage approach in which, in the first stage, utilizes a frequency analysis technique to detect periodicities within the dataset. In the second stage, statistical methods are used to identify the factors correlated with the detected frequencies of delay. To study the effect of seasons and months, they use a binary logistic regression model.

Some studies have focused on flight cancellation at airports. Rupp and Holmes (2006); Xiong and Hansen (2013) employ logit models to explain the influence of several variables on a flight being cancelled. Alternatively, Sridhar et al. (2009) aims at predicting the total aggregate number of flight cancellations with a neural network approach, achieving an accuracy of 0.79 in the obtained predictions. Lastly, Lambelho et al. (2020) develops machine learning classifiers to predict delays and cancellations with a 6-month prediction horizon. They however fail to include the time component of the series.

Further analyzing these studies, Rupp and Holmes (2006) uses a regression model that suggests that, at the airport level, route competition improves service quality. They find significant fewer cancellations at hub airports, highlighting the importance that airlines give hub flights in maintaining a flight network. They also present results that confirm that carriers minimize passenger inconvenience, as well as empirical evidence linking revenue with flight cancellations. (Xiong and Hansen, 2013) use the Random Utility Theory (RUT) and mixed logit regression models to address the issue. Three groups of explanatory variables are used, i.e. (1) delay factors (GDP-assigned initial delay (GID), Internal delay, Delay savings from a hypothetical cancellation); (2) Flight characteristics (distance, hub destination, if a flight is operated by a major airline, frequency) and (3) segment characteristics (aircraft size, average size, load factor and market fare). Lambelho et al. (2020) provide a machine learning approach that uses classification algorithms to predict whether flights scheduled in the strategic phase (6 months prior to the day of the execution) are subject to arrival/departure de- lays and cancellations during execution at London Heathrow Airport. The aim is to support strategic flight schedules – the slot allocation process at airports. Rupp (2005) also discussed that "delays and cancellations move in opposite directions suggesting that the carrier is trading-off fewer (more) flight cancellations for more (fewer) flight delays" However, his model does not provide a concise trade-off between delays and cancellations. Yanying et al. (2019) strives to predict delays and cancellations with logit regression, support vector ma- chine, naïve Bayes and decision trees models based on Spark. It only states that decision trees are the best tool from among those ones to predict flight cancellation. Precision approaches a value of 0.9 but AUC = 0.558. The rest of the methods did not provide good results. The predictions were based on the number of nodes in the spark cluster. It is stated that, if we are able to use computer classifications to predict whether flights are cancelled or delayed, we can save resources and reduce passengers' anxiety.

2.4 Hypotheses and exploratory question

This study may be divided into two main parts. The first part takes into consideration the research that has been carried out so far in delays and cancellations and expands it, and is confirmatory. Firstly, although there have been attempts, there has been no concluding insights into how the forecasts of cancellations can be improved by taking into consideration flight delays, while taking into account the time series nature of the data. Thus, the following first hypothesis is tested:

H_1: Cancellations and delays are not independent, and delays improve the accuracy of the cancellations forecasts over time.

Secondly, we do not focus on the time component of the data, but rather contextualize the relationship of delays and cancellations with a regression analysis, which further develops the research by Rupp and Holmes (2006). It only applies the research to domestic US flights and leaves out the impact of airport competition, when exceeding capacity is a common factor for increased delays and cancellations (Mead and General, 2000). It is expected that regions with hub airports and with airports that offer high competitive routes will have a better service quality (i.e. less cancellations), while more intense airport competition (or congestion) has a negative effect in cancellations, as capacity may be exceeded. Competition in an airport can be measured as the number of airlines operating in that specific airport at a given month. Simultaneously, the network effect is measured by the fact of an airport being a hub or not, i.e., whether more than 5 airlines have 26 or more connections from that airport. Therefore, the following hypothesis is tested:

H_2: The presence of hubs, increased route competition and decreased competition in airports are determinants of lower flight cancellations among UK regions.

The second part of the research, as explained, is exploratory. With the out- break of the COVID-19, this subject acquires special relevance, as it shows how UK's aviation market reacted to the virus, as well as the how the concerned public opinion affected air transport. The specific exploratory question would be: "What are the behavioral changes that the population suffers regarding air travel and cancellations as a result of a global pandemic, and how do they help improve the cancellations forecasts during the COVID-19?".

This is one of the first studies to take these factors into consideration, and it will help future researchers better understand the effects a potential future pandemic on air transport – particularly during the first months, which are the most confusing for both institutions and the people.

3 Data and methods

In this section we will discuss the process of data collection, cleansing, the research methods and the final datasets. First, we explain the flight cancellation data, and its obtention and preparation. Then, we present a descriptive time series analysis of cancellation data to understand the underlying patterns behind the data. Next, we dwell into the time series analysis and the comparison between the univariate autoregressive modelling and the multivariate one, to evaluate the effect of delays on cancellations. For addressing the second hypothesis, we perform a Random Forest and a multivariate regression, and compare the results, as the time series is no longer taken into consideration. For the last part, we discuss the process to evaluate change of behaviors as a result of the COVID-19 with Google Trends. Table 4 describes the process followed for this study.

Table 4. Steps followed in the research process

1	Data collection
2	Data compilation
3	Data cleansing
4	Data enrichment
5	Descriptive time series analysis
6	Autoregressive modelling analysis to fix the type of ARIMA model
7	Multivariate time series analysis to evaluate the effects of delays on cancellations and fix the type of ARIMAX model
8	Comparison of both models, fitting them to a training sample and evaluating their performance on a holdout sample, on three levels of aggregation
9	Perform a multivariate regression to evaluate the effects of hubs and competition on service quality and compare results
10	Reevaluation of previous time series models with Google Trends data as exogenous variable

3.1 Data collection and description

This section elaborates on the nature and collection of the flight delays and cancellation data, as well as a description of the original datasets.

3.1.1 Cancelled air transport movements

The data was gathered from the UK Civil Aviation Authority's website (CAA, 2020a). It is provided by various sources and validated by the CAA. They have an extensive open database, with movement, cargo and passenger statistics of over 60 UK airports. Data is retrieved monthly from 1973. However, until 1990 the data simply consists of scanned copies of paper records, making its analysis unfeasible. Thus, the data used in this paper consists on a monthly collection of different parameters from 1990 to 2020.

The statistics are compiled for a maximum of 60 UK airports, although for the data covering flight cancellations this number is reduced to 27 (see the complete list of airports in Table 9).

The variables from the dataset are shown in Table 5. Data on historical flight cancellation in UK airports is classified as 'Total number of air transport movements', 'Of which cancelled', and 'Total air transport movements without cancelled'. These datasets exclude air taxi operations, which are movements by an aircraft of less than 15 tonnes MTWA operating on a non-scheduled service. Therefore, these are predominantly sole-use charter operations. On the contrary, Air transport movements are landings or take-offs of aircraft engaged on the transport of passengers, cargo or mail on commercial terms.

Table 5. Original variables of the Cancellations dataset

Variable name	Variable description
rundate	Date in which the data were uploaded or updated to the website
reporting_period	Period to which that observation refers to (year and month)
reporting_airport_group_name	Variable determining whether the airport belongs to London Area, Non UK Airports, or Other UK Airports
reporting_airport_name	UK Airport
total_atms	Total number of air transport movements
total_cancelled_atms	Total number of cancelled air transport movements
total_atms_excl_cancel	Total number of air transport movements excluding those cancelled

As seen in Table 5, the cancellations dataset only has information on total cancellations relative to total air transport movements. In addition, this dataset only includes information about different airports. However, the main objective of this research is to evaluate the intrinsic relationship between delays and cancellations. Additionally, when analysing flight cancellations from a machine learning perspective, additional parameters need to be evaluated, and this dataset neither includes them nor provides the means to calculate them. Thus, the cancellations dataset fell short to provide the different parameters that this study demanded.

3.1.2 Flight punctuality statistics (Delays)

The data regarding flight punctuality, i.e. delays, was also gathered from the UK Civil Aviation Authority's website (CAA, 2020b). Data is retrieved monthly from 2004 to 2019. The statistics are compiled for a reduced number of airports. Until 2015, only 10 airports were analysed, adding up to a total of 26 airports from 2015 to 2020 – the last update was at March 2020. The variables that conform this dataset (2000-2017) are shown in Table 6. The Flight Punctuality (Delays) dataset includes a wide range of additional information that this study makes use of. From 2018, it already includes the percentage of flights cancelled, in addition to broken down information about delay intervals, average delay minutes, origin/destination airport and country, etc.

The data analyses delays by airport, and by airport of origin/destination. However, it does not specify which airport is the origin or destination airport. CAA (2020b) provides an explanation for the existing of unmatched flights, saying in their website that "the reasons for this would normally be: (a) the flight was a diversion from another airport; (b) the flight was not recorded with Airport Coordination Ltd or airport; (c) the flight was a short-haul flight more than one hour before the planned time; (d) the flight was planned to take place in the previous month". It is also important to state that, from 2018 on, the format of the dataset changed. On the one hand, a new column flights_cancelled_percent (cancellations as a proportion of all flights) was included. Moreover, additional on-time performance bands as a proportion of all flights were included in the dataset (more levels of splits). Firstly, the existing "Early to 15 minutes late" band is split into three sub bands: "Flights more than 15 minutes early percent", "Flights 15 minutes early to 1 minute early percent", "Flights 0 (zero) to 15 minutes late percent". Secondly, the existing "Between 61 and 180 minutes late" band is split into "Between 61-120 minutes late" and "121-180 minutes late". These new parameters and bands are shown in Figure 7.

Cancellations are defined by the CAA as "the non-operation of a previously planned flight, announced less than 24 hours before or after its scheduled departure time". In addition, it is worth clarifying the definition that we give to the variable 'Origin/Destination country'. In the website, it is noted that "the aircraft origin/destination represents the final point on the service. An aircraft serving more than one point on the route is therefore shown once only in the tables". Thus, we interpret it as the destination point of the service, having the British airport as origin. This is obviously not optimal and supposes a limitation for the research, which would be improved in the future if clarified.

Table 6. Original variables of the Delays dataset (2000-2017)

Variable name	**Variable description**
rundate	Date in which the data were uploaded or updated to the website
reporting_airport	UK Airport
reporting_period	Period to which that observation refers to (year and month)
origin_destination_country	Division of origin_destination by countries
origin_destination	Final point on the service. An aircraft serving more than one point on the route is therefore shown once only in the table
airline_name	Name of the airline corresponding to that observation
scheduled_charter	Dummy with value 1 if the flights were scheduled and 0 if it is chartered
number_flights_matched	Air transport movements that took place and for which a corresponding planned flight was found
actual_flights_unmatched	Air transport movements which actually took place at the airport but for which no corresponding planned flight was found
early_to_15_mins_late_percent	% of flights early to 15 minutes late

flts_16_to_30_mins_late_percent	% of flights between 16 to 30 minutes late
flts_31_to_60_mins_late_percent	% of flights between 31 to 60 minutes late
flts_61_to_180_mins_late_percent	% of flights between 61 to 180 minutes late
flts_181_to_360_mins_late_percent	% of flights between 181 to 360 minutes late
more_than_360_mins_late_percent	% of flights more than 360 minutes late
average_delay_mins	Average delay (in minutes)

Table 7. New fields and bands for the Delays dataset (from 2018)

Cancellations as a proportion of all flights
Flights more than 15 minutes early percent
Flights 15 minutes early to 1 minute early percent
Flights 0 (zero) to 15 minutes late percent
Flights between 16 and 30 minutes late percent
Flights between 31 and 60 minutes late percent
Flights between 61 and 120 minutes late percent
Flights between 121 and 180 minutes late percent
Flights between 181 and 360 minutes late percent
Flights more than 360 minutes late percent

3.1.3 Google Trends

To explore people's reaction to COVID-19, their concern with flight cancellations and the evolution of their consuming intentions, data is gathered from Google Trends, using the gtrendsR package from R statistics.

3.2 Data preparation

In this section, we first briefly describe how the cancellations dataset was cleansed and enrich and, next, we deepen on how the delays dataset was prepared for analysis.

3.2.1 Cancellations dataset

Both the data for flight delays and cancellations were collected as a series of independent monthly datasets. Thus, after they were downloaded, they were compiled and merged to obtain a time series of delays and cancellation data.

Focusing on the latter, the data on flight cancellations was reported on absolute terms (see Table 5), i.e. "Total number of air transport movements", "Of which cancelled" (total_atms, total_cancelled_atms). This supposed a limitation, since it would not be accurate to evaluate service quality, regarding frequency of cancellations in different airports, in terms of the total number of flights cancelled. The problem relies in the fact that every airport has a different capacity and a different volume of air transport movements. Hence, we opted for adding a new column, cancelled_percent, calculated as total_cancelled_atms/total_atms∗100 (see Table 8).

Table 8. New variable for the Cancellations dataset

Variable name	Variable description
cancelled_percent	Percentage of flights cancelled relative to the total number of air transport movements, and calculated as total_cancelled_atms/total_atms∗100

This column represents the amount of flights cancelled relative to the total air transport movements, which allows for a fair comparison among airports. The evolution of relative cancellations over time can be seen in Figure 1. After this addition, a column specifying the specific region that a particular airport belonged to was also added. This will be further explained next.

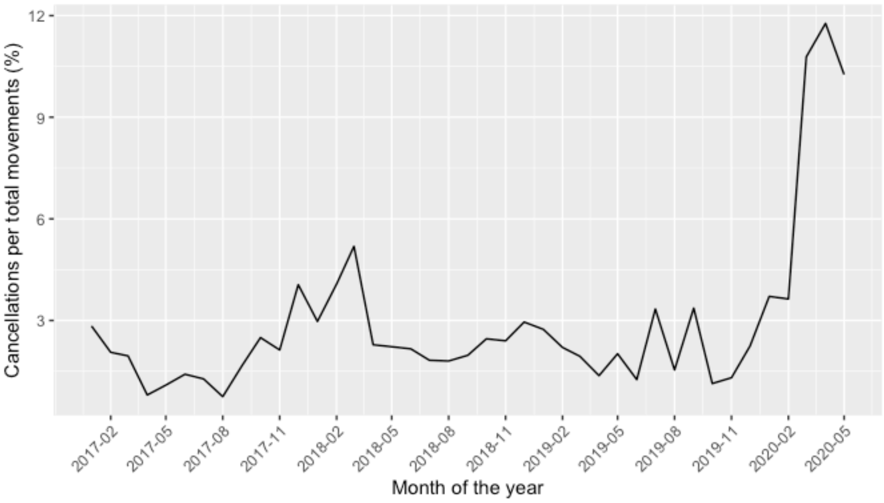

Figure 1. Evolution of cancellations as a percent of total air transport movements over time (2017/01 – 2020/05)

3.2.2 Delays dataset

There were significant differences in the delays datasets before 2017 and the ones after 2018. On the one hand and, most importantly, data from 2018 included information regarding cancellations. On the other, delays intervals were broken down into more on-time performance bands. First, for non-delayed flights, the Figure 1: Evolution of cancellations as a percent of total air transport movements over time (2017/01 – 2020/05) existing "Early to 15 minutes late" band is split into three sub bands: "Flights more than 15 minutes early percent", "Flights 15 minutes early to 1 minute early percent", "Flights 0 (zero) to 15 minutes late percent". Additionally, the existing "Between 61 and 180 minutes late" band is split into "Between 61-120 minutes late" and "121-180 minutes late". For these reasons, the delays data was split into two datasets. One from 2000 until today, by adding up the split columns into the old unsplit version and merging the most recent data with the old dataset, and a new one, with more broken-down data.

At first, there were some columns that were deleted, such as the 'rundate' column. Additionally, the 'reporting_period' column was divided into 'reporting_year', 'reporting_month', and 'reporting_yearmonth', which was converted into a POSIXts date variable.

Next, the regions were added. The airports were divided into a subset of 13 regions, as per List of airports in the United Kingdom and the British Crown Dependencies (2020). The airports are presented by countries and regions in Table 9.

Table 9. UK airports by country and region

Country	Region	Code	Airport
England	Greater London Area	LDN	Heathrow London City Gatwick Southend Stansted
	North East	NE	Newcastle Durham Tees Valley
	North West	NW	Blackpool Liverpool (John Lennon) Manchester
	East	EAS	Norwich
	East Midlands	EM	East Midlands International
	West Midlands	WM	Birmingham
	South East	SE	Lydd Southampton Oxford (Kidlington)
	South West	SW	Bournemouth Bristol Isles Of Scilly (St. Marys) Lands End (St Just) Newquay
	Yorkshire & the Humber	YH	Doncaster Sheffield Humberside Leeds Bradford
	British Crown Dependencies	BCD	Isle of Man Jersey

Northern Ireland	NIR	Belfast City (George Best) Belfast International City Of Derry (Eglinton)
Scotland	SCO	Aberdeen Barra Benbecula Campbeltown Edinburgh Glasgow Prestwick Inverness Islay Kirkwall Lerwick (Tingwall) Scatsta Sumburgh Stornoway Tiree Wick John O Groats
Wales	WAL	Cardiff (Wales)

After this, a new column specifying the route was added. This was done in line with determining the competition of airports and routes. In order to have a deeper understanding of the cancellations and delays, and in order with the literature, the quality of the service (i.e. amount of cancellations or delays) is also evaluated in relation to the competition of the routes offered by a specific airport. In this study, the competition for certain airport's routes can be used as a proxy to measure the competitiveness of that airport. Hence, two new columns "route" and "n_airlines_route" were added. "Route" contains the British airport, a dash, and the origin/destination airport (e.g. "MANCHESTER - VALENCIA"), and "n_airlines_route" depicts the number of airlines offering that same route in that month. One last factor column "Monopoly" was added, with the code "M" in case there is a monopoly, "C" in case that there are more than 20 airlines offering that route, and "N" in the rest of the cases. To determine the competitiveness of airports, a new column "n_airlines_airport" was added, that counts the total number of airlines operating in that specific airport in that month.

Lastly, to explore the effect of networks and the hub and spoke system, a binary variable called "Hub" was added, that gets the value 1 when that airport is a hub and 0 when it is not. To determine this, a new variable counting the number of connections by each airline from each airport was created. As Rupp (2005) mention in their research, an airline uses an airport as a hub if they have 26 or more connections from there. Thus, at first we had whether, for each observation, that airport was a hub for the airline in that observation. After this, for each airport, we counted for how many airlines it was used as a hub, and obtained the following distribution (Figure 2). Hence, it was decided that, for the sake of uniformity of the analysis, an airport would be considered a hub if more than 5 airlines used it as such, adding then the aforementioned binary variable "Hub". Table 10 briefly describes the new added fields.

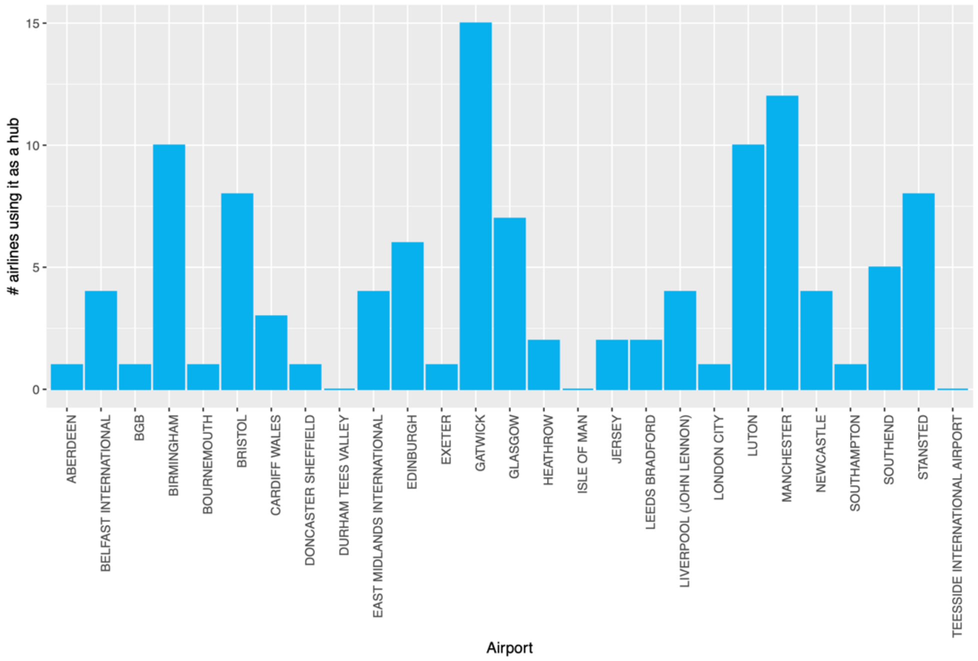

Figure 2. Number of airlines using each UK airport as a hub

Table 10. New added fields and bands for the Delays dataset

Variable name	Variable description
Route	Route between reporting_airport and origin_destination
Monopoly	Categorical variable with three categories: M for monopolist routes, C for competitive routes and N for the rest
n_airlines_airport	Number of airlines operating at an airport that given month
Hub	Dummy variable with value 1 when an airport is considered a Hub and 0 otherwise

3.3 Data analysis methods

Different methods were used to analyze the data. As Marsland (2014) explains, machine learning consists in programming computers to optimize performance criterion using example or past data. The goal of machine learning is to develop models and methods that can automatically detect patterns in data and use them to predict future outcomes and probabilities. It is used, among others, in the fields of statistics and data mining.

At the same time, around the world computers store terabytes of data every hour (Marsland, 2014). Entities such as banks, government agencies, hospitals, laboratories, universities, cities, etc. are incessantly storing data. The size and complexity of this data makes it impossible for humans to analyze it without help of computers. As stated by Jordan and Mitchell (2015), for intelligent beings, many of their skills are acquired or refined through learning from the experience, rather than following explicit instructions. Machine learning methods enable programs and computers to 'learn' and adapt.

This research aims to solve the aforementioned air traffic problems with machine learning techniques. For the first section, in which the relationship between delays and cancellations will be addressed, the main methods are ARIMA, to test the autoregressive model with cancellations data, and ARIMAX, to compare the previous univariate results with a time series model that takes into account the effect of delays.

To evaluate the second hypothesis, the focus is not to evaluate the effect of time on the series, but rather have a clear picture of the predictors of cancellations across different regions, namely networks, competition and congestion of airports. The models are therefore more complex and will be analyzed with a multivariate regression analysis.

Finally, the last section consists of an exploratory case-study, that will mainly focus on the discussion on how behaviors have changed as a result of the COVID-19 virus outbreak and spread. Therefore, we will use descriptive methods for analyzing these changes in behaviors, and then we will apply the fixed time series model types from the first section, taking into account Google Trends data as an exogenous variable.

3.3.1 Time series analysis

This research will be carried out in the following way. Due to the short length of the dataset, the type of ARIMA and ARIMAX models will be fixed with the whole sample of data. However, to estimate the performances, the time series data is divided into a training and holdout samples, and these model types will be fitted to the training data. With the test sample, the forecast accuracy will be measured for three levels of aggregation: UK aggregated data, London region data and Manchester airport data. The training sample comprises the observations from January 2017 until June 2020, and the test sample does from July 2020 to December 2020. The COVID-19 period will be left out of the estimation of performances because the high peaks in the start of 2020 can lead to wrong estimations of accuracies, because they will be part of the test set and no model would forecast such an exponential increase in cancellations. Next, we deepen into the technicalities and explanations of time series modelling.

Time series are sequences of data points organized in time order. They are usually discretely sequenced in equal spaces of time. On the other hand, forecasting is the process of obtaining predictions of the future based on different present and past data. Hence, time series forecasting consists on predicting future values of an outcome based on the observation of its past values.

This forecasting processes are commonly performed in machine learning when analyzing big sets of data. This allows for identifying historical trends and patterns that otherwise would have been inviable to analyze.

ARIMA is the abbreviation for Auto Regressive Integrated Moving Average. A model is defined as Auto Regressive (AR) if the variable under a time period is explained by its own observations in a past time period, plus an error term. This error term is commonly known as white noise when it fulfills the following three conditions: (1) its mean is 0, (2) it has a constant variance, and (3) the covariance corresponding to different observations' errors is also 0 (De Arce and Mahía, 2003). It is also independent of explanatory variables in the model.

The Moving Average (MA) terms refer to the lags of errors – the value of the outcome at a period t is explained by the independent term plus a series of conveniently-pondered errors from previous periods. The Integrated (I) term is the number of differences used to make the time series stationary.

Analyses with the ARIMA model make two assumptions. First, the data must be stationary – the time when the data is captured does not affect the properties of the series. Second, ARIMA works with a single variable. Hence, data should be univariate.

Working with time series models like ARIMA has concrete benefits and drawbacks (De Arce and Mahía, 2003). The main advantage is that there is no need for different series of data (different variables) referred to the same time period, while this is the common characteristic of every univariate model.

A full ARIMA model can be written as

$$y'_t = c + \varphi_1 y'_{t-1} + \ldots + \varphi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

where y'_t is the differenced series. The predictors include both lagged values of y_t and lagged errors. This is call an ARIMA (*p*, *d*, *q*) model, where *p* refers to the order of the autoregressive part; *d* is the degree of first differencing involved; and *q* is the order of the moving average part.

However, this method also comes with an inconvenience. Given that an analysis with a wider set of explanatory variables is rejected, the ARIMA model does not pay attention to the potential relations with further economic, air traffic, or alternative variables. This implies that the time series study of flight cancellation comes with a cost – the loss of a degree of analysis capacity.

There are models that can measure the effect of exogenous variables in the model, like Transfer Function Models or as ARIMAX. Assuming two time series, these models into account the output series (dependent variable), the input series (independent variable), the stochastic and the disturbance (the noise series of the system that is independent of the input series). The transfer function (or impulse response function) allows the first time series to affect the otheroneviaadistributedlag (Durka and Pastoreková, n.d.). When the input series and the stochastic disturbance are assumed to follow an ARMA model, the transfer function model is known as the ARMAX model.

For the time series descriptive analysis, the series is decomoposed, and its components are analyzed. When analyzing time series, it is important to differentiate between different types of patterns. This is called time series decomposition, and consists of decomposing a series into components, which usually are three: trend, seasonality and cycles. However, trend and cycles are usually combined into the so-called single *trend-cycle* component, which is normally called *trend* for simplicity. Therefore, a time series is usually thought as comprising three main components: a trend-cycle component, a seasonal component and a remainder component, which contains anything else in the time series (Hyndman and Athanasopoulos, 2018).

A decomposition may be additive or multiplicative. In this study, we will assume that we have a multiplicative decomposition, since the trend is non-linear. It can be decomposed as

$$y_t = S_t \times T_t \times R_t,$$

where S_t is the seasonal component, T_t is the smoothed trend component, and R_t is a remainder component.

There are a some common ways to decompose a time series onto these three components, such as the classical, X11, SEATS or STL decomposition. In this research, we will compare the classical and STL decomposition.

The classical decomposition is the first method that originated for decompos- ing time series, and it dates back to the 1920s (Hyndman and Athanasopoulos, 2018). Not surprisingly, it is the simplest one, and it is not recommended to use as now there are much better methods.

The seasonal period *m* = 12, as it is monthly data. There are some drawbacks, like the lack of trend-cycle estimates for the first and last 6 observations, respectively. It also assumes the seasonal component to be constant from year to year.

STL ("Seasonal and Trend decomposition using Loess") decomposition, de- veloped by Cleveland et al. (1990), has several advantages when compared with methods such as classical decomposition, X11 or SEATS. Most importantly for this study, it can work with changing seasonalities and can be very robust to outliers, and therefore occasional unusual observations do not affect the estimates of the trend-cycle and seasonal components. However, these will appear on the reminders.

3.3.2 Regression analysis

Once that we have set clear the role of delays on cancellations, a further study that nuances and expands the literature is needed. The aim of this section is to contextualize the relationship between delays and cancellations, by testing the joint effect of three parameters in cancellations in the UK – in particular, hub airports, more competitive routes and less congestion in airports. In this section, we do not evaluate the effect of time. A multivariate regression analysis will be made.

Two models will be used, one with only control variables and another with also the rest of the parameters. The models include as control variables the region of origin/destination, the airline name, the British airport, the average delay in minutes, and whether it is a scheduled or chartered flight.

The unrestricted model serves to infer about the joint effect of hubs, compe- tition and congestion. As independent variables, the variables that are selected are the ones that are the objective of this research, namely a dummy representing whether the airport is a Hub, the competition in the airport (as the number of airlines operating in that airport), and the competition of routes. Therefore, the unrestricted model is:

$$\begin{aligned} flights\_cancelled\_percent =& \alpha + \beta_1 Scheduled + \sum_{i=1}^{n_{RA}} \beta_{2,i} reporting\_airport_i + \\ & \sum_{j=1}^{n_{OD}} \beta_{3,j} ODregion_j + \sum_{k=1}^{n_{AL}} \beta_{4,k} airline\_name_k + \\ & \beta_5 average\_delay\_mins + \beta_6 Hub + \\ & \beta_{7M} Monopoly_M + \beta_{7N} Monopoly_N + \\ & \beta_8 n\_airlines\_airport + \varepsilon, \qquad \varepsilon \sim n(0, \sigma) \end{aligned}$$

Therefore, the null and alternative hypotheses are:

$$H_0 : \beta_6 = \beta_{7M} = \beta_{7N} = \beta_8 = 0$$

$$H_2 : not\ all\ \beta_6, \beta_{7M}, \beta_{7N}, \beta_8\ equal\ 0$$

The restricted (or control) model is obtained by constraining the effects of variables *Monopoly*, *Hub*, and *n_airlines_airport* in model to be equal to zero. These linear models will be run on R statistics, and the regression coefficients will be compared once the models are estimated. Additionally, we will perform the F-tests for these variables, as we want to assess the joint contribution of part of the explanatory variables. The F-test evaluates the null hypothesis that the variable has no effect, against the alternative that at least one of the categories deviates markedly from the other. Thus, we see if the research model contributes to the explanation of the dependent variable *flights_cancelled_percent*.

3.3.3 Methods to analyze the influence of COVID-19

For this section, at first a descriptive analysis of several terms has been made, making use of the gtrendsR package of R statistics. With it, the time series analysis is made. It has been seen how differently delays and cancellations behave after the start of 2020. This makes the forecasts with only delays perform worse than before, and the levels of cancellations cannot be predictive as accurately. As it is later explained, the most searched terms for each subset are tickets and cancelled, respectively. Thus, we estimate whether the terms *tickets* and *cancelled* contribute to the forecasts. We create a training and a test sample, including the Google trends data, which are downloaded with the gtrendsR package in R statistics and then converted to monthly data (they were automatically downloaded as weekly data). These samples range from 2017/01 – 2019/06 and 2019/07 – 2019/12, respectively. The data are included in the delays and cancellations dataset, and the ARIMA(1,0,0) and ARIMAX(2,0,0)(1,0,0) models are fitted to the training data. Then, they are evaluated with the test data. After this, the same procedure is done but changing the training and test samples to observations ranging from 2017/01 – 2019/10 for the training sample and 2019/11 – 2020/03 for the holdout sample. The three models are fitted to the training data, evaluated with the test data, and the performance mesures and forecasts are compared.

4 Analysis

4.1 Time series analysis of delays and cancellations

In this section, the effect of delays on cancellations will be analyzed from a time series perspective. First, a descriptive analysis will be made, to understand the underlying patterns in the data and familiarize with it. Secondly, two different forecasts will be made – one with an autoregressive ARIMA model, and another one with an ARIMAX model. They will be compared to evaluate the influence of delays on the percentage of cancelled flights.

4.1.1 Descriptive time series analysis

Before starting to dwell into the evolution over time of the series and the time series descriptive analysis, it may be good to make sense of the data first, to draw some initial conclusions.

Firstly, an analysis of the cancellations and delays by regions is made. The average delay by region between the year 2000 and 2020 is presented in Figure 3. It can be seen how, even though most of the regions range between 12 and 15 minutes, there are considerable differences among them. This points towards different managerial disciplines that affect the delays distribution. For example, the London area has a substantially larger average delay than the British Crown Dependencies, where the weather is much more adverse and is not such an important economic area.

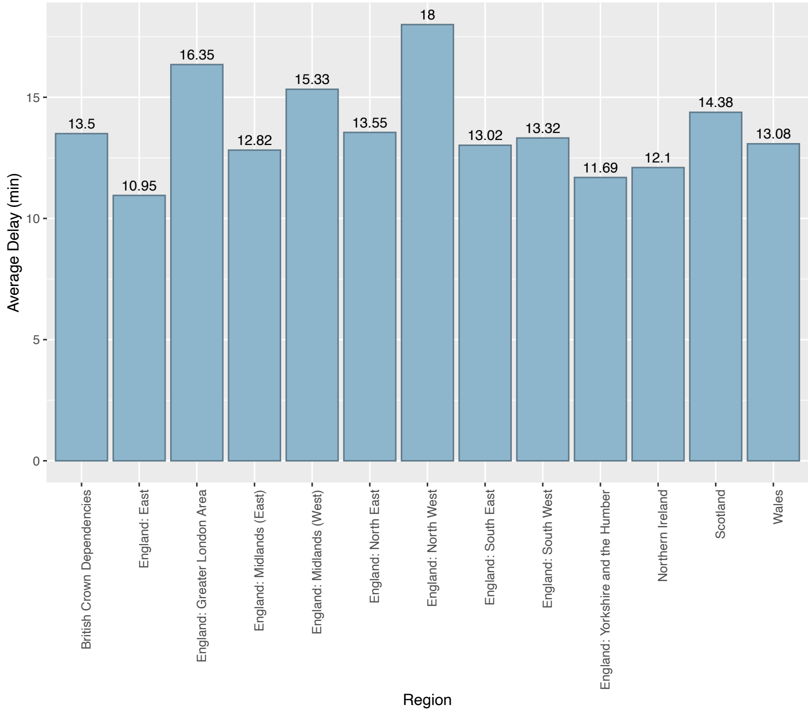

Figure 3. Average delay across different UK regions (2000-2020)

To evaluate how the different regions evolve over time, the values for average delays over time are shown in Figure 4. This graph is consistent with the literature, as it portrays the two components of cancellations that have been discussed over section 3. On the one hand, the common patterns can be explained with the findings of Pyrgiotis et al. (2013); Xu et al. (2005), which estimate how delays propagate through a network of airports, and how a delay of 1h in the morning can lead to delays of 7h in the evening across several airports within that network. On the other hand, it also shows that there are differences over time across regions, as can be seen by individual peaks at certain dates or consistently better performances, as is the case of the East of England region (which already had the lowest value in Figure 3).

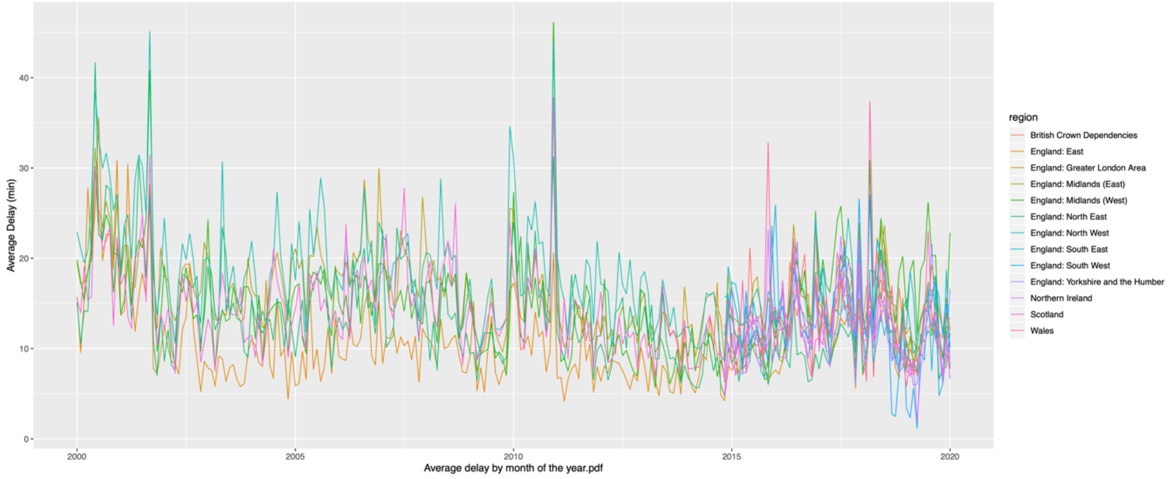

Figure 4. Time evolution of average delay across different UK regions (2000- 2020)

If we analyze these values for the period for which we have values for cancellations (2018-2020), and compare these two parameters, we obtain Figure 5.

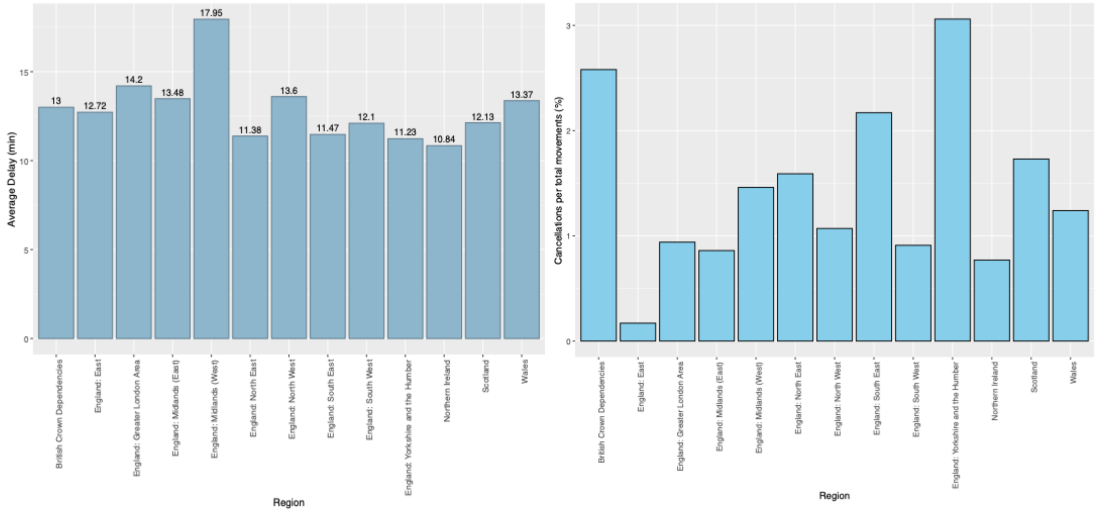

Figure 5. Average delay and average percent cancellations across different UK regions (2018-2020)

First, it is interesting to analyze how delays have evolved over time, and see which regions have improved or worsened. As could be deducted from Figure 4, there is a slight downwards trend. We see an improvement in London Area of over 2 mins in average delay, which is remarkable. The most impressive improvement is that of the North West region, which improves its service quality in terms of delays by 4.4 mins. Nevertheless, it is still among the top three regions with worst performances. On the other side of the spectrum, we see a considerable increase in average delay in West Midlands and the East. As this is only for the last two years, it cannot be said that these changes in performance have been the result of punctual weather effects of other external causes. They point towards changes in managerial practices in different UK regions.

Secondly, if we compare it to the cancellations bar plot, we see some similarities and differences. We can observe how the distribution is more irregular for cancellations. We can also observe that some regions show better performances, as they do for average delay, which are in favour of our hypothesis (see East of England), but we also can observe the opposite happening for other regions. It is thus difficult to infer any conclusions yet from these plots.

Lastly, to make this analysis more visual, a representation of the UK map with the average delay and number of cancellations by region, along with the standard deviations, are shown in Figure 6.

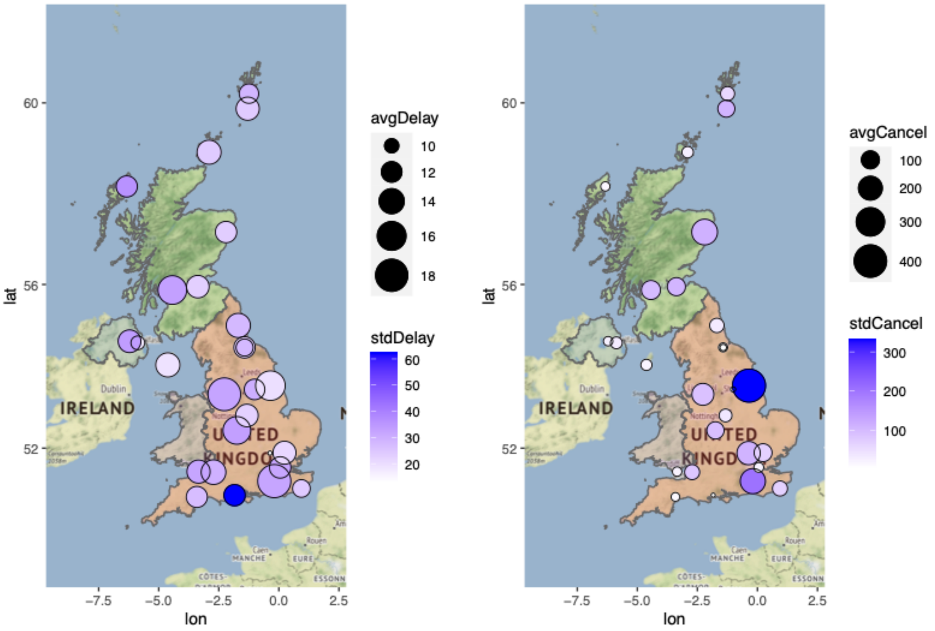

Figure 6. Map of the average delay and average number of cancelled flights in the UK (2018-2020)

From this initial descriptive analysis, the two main takes that can be deduced are: (1) delays and cancellations seem to be affected by both external and internal/managerial factors, that affect their performance; (2) No clear relationship has been able to be made from this simple analysis. Hence, a more thorough time series analysis will be made next.

From here, this section offers the reader a visual interpretation and description of the time series data. This descriptive analysis will be made with three levels of aggregation:

aggregate cancellations and delays data from the UK, aggregate data from the London region, and data from Manchester airport. Given that the Cancellations datasets starts one year prior to the new Delays datasets (cancellations alone start to be measured in 2017, but together with delays they start to be counted in 2018), this section will make use of a merge between a subset of the long Delays dataset ranging from 2017 to 2020 and the Cancellations one, which already finds itself in that range, specifically from January 2017 to April 2020 – however, as the Delays dataset was last updated for March 2020, this will be the last month considered. The other Delays dataset, that automatically includes the percentage of cancelled flights in it, only covers the 2018-2020 period. Therefore, the data consists of 39 observations by airport, i.e., 27 time series. The data is described at a nation-aggregate level. Because of having one less month of observations, the aggregate time series distribution of percentage of flights cancelled does not include the last observation that can be observed in Figure 1. The evolution of cancelled flights (%) vs. the average delay in minutes looks as presented in Figure 7.

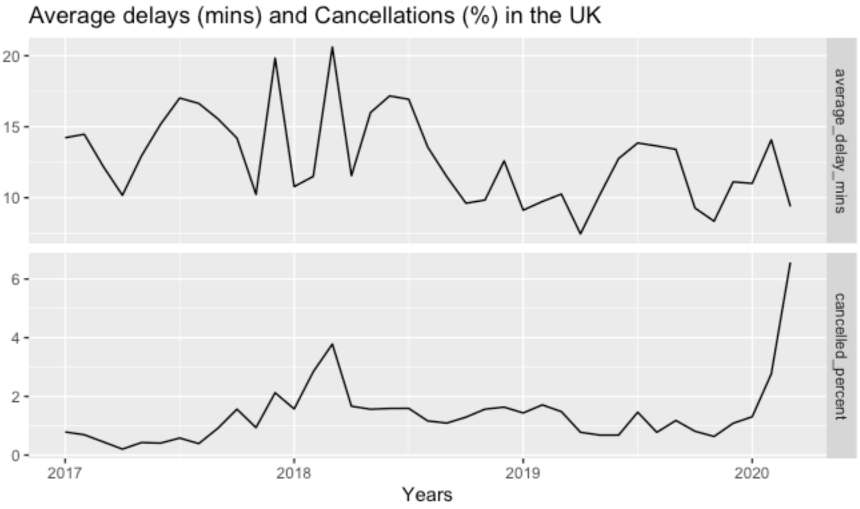

Figure 7. Evolution of the average delay and cancellations percent (2017-2020)

As it can be seen, most of the spikes in the cancellations plot have a counterpart (of similar or different magnitude) in the plot regarding delays. On the counterpart, these two time series seem to follow different trends. Interestingly enough, the natural difference among time series can also be observed, obtaining a more predominant difference from the start of 2020, coinciding with the outbreak of the COVID-19 virus. Whereas the amount of flights cancelled relative to the total number of air transport movements increases exponentially, the average delay in minutes actually decreases. This can be attributed to a better operational and material resource availability for the non-cancelled flights. In other words, given the increasing number of cancelled flights, the ones that are operated enjoy the complete availability of airport and airline resources, therefore decreasing congestion and overcapacity, because they can make use of the resources that are 'freed-up' by the cancelled flights. Not only that, but also flight routes and runways are emptier, leading to a less congested and thus improved air traffic. This combination of factors leads to a better service quality in terms of delays, while cancellations suffer a steep increase.

As a matter of fact, this difference can be better observed in Figure 8. Year by year, the peaks can been observed at the same moments in time. This is consistent with the theory of delay propagation across a network of airports (Pyrgiotis et al., 2013; Xu et al., 2005). The concept of short scheduled turnaround times explained by Xiong and Hansen (2013) may also explain why there is a similar pattern of peaks in the data, even though they may be of different magnitudes. Additionally, the differences in spikes can also be better observed, as can be seen with the start of 2020. In March 2020, while there is an exponential increase in cancellations, there is a substantial fall in average delay (purple lines). This difference can also be observed in the reversed-U shape of the delays pattern during the second and third quarters in delays every year, which do not find a similar counterpart in the cancellations evolution over time (see from April to October).

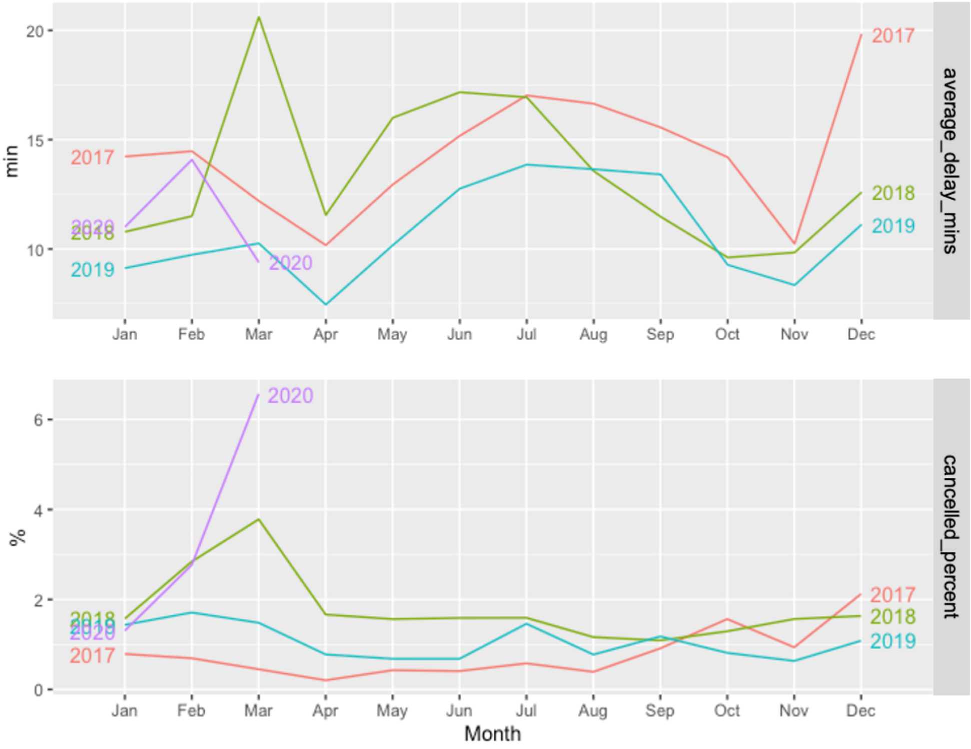

Figure 8. Seasonal comparison between delays and cancellations in the UK

The classical and STL decompositions are presented in Figure 9 and Figure 10, respectively. The drawbacks of the classical decomposition can be easily observed, like the lack of trend-cycle estimates for the first and last 6 observations, respectively. It also assumes the seasonal component to be constant from year to year. Another drawback that can be spotted at a first glance is the misinterpretation of the COVID-19 peak as a change in trend.

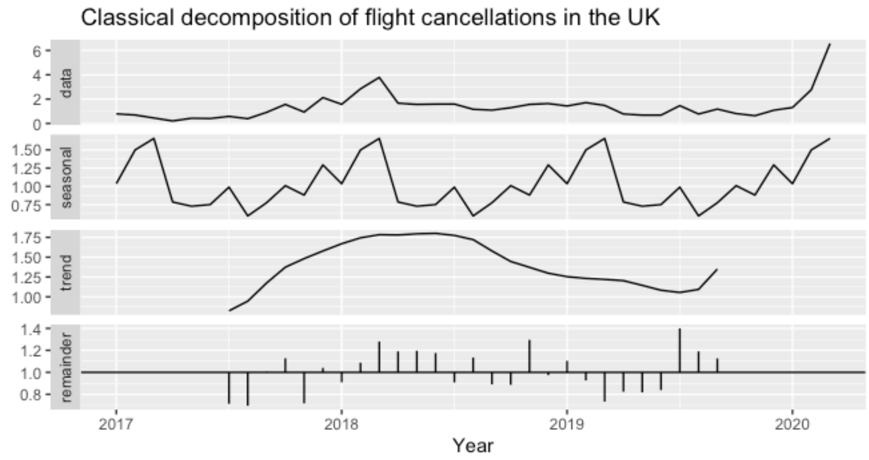

Figure 9. Classical decomposition of flight cancellations time series

The strength of the STL decomposition towards these potential mistakes is actually seen in Figure 10, where it can be observed how the peaks in February and March of 2020 do not affect the trend-cycle component, as it correctly interprets them as outliers, contrary than the classical decomposition method. As expected, they do appear in the remainders plot. The STL trend-cycle component shows a declining trend over time, that goes in line with the airlines' objectives to improve service quality. The STL remainders shows unusual positive cancellations spikes at the start of this year, which are caused by the COVID-19. It also shows negative spikes at 2017 and 2019 seasonal peaks.

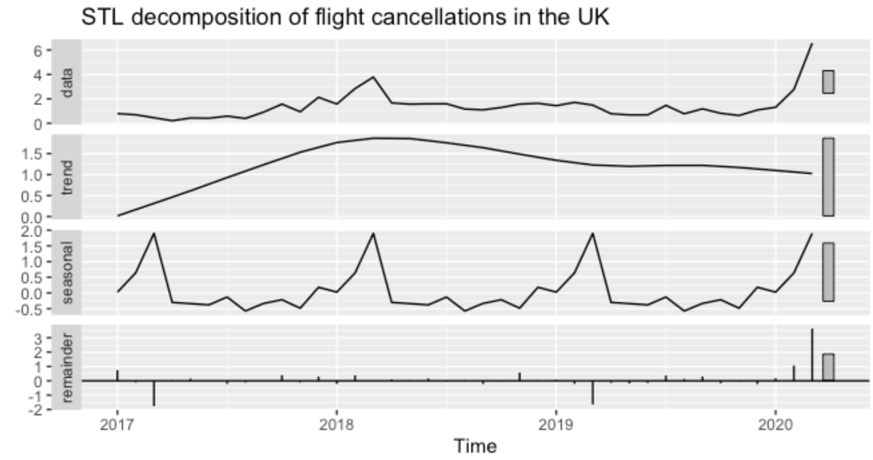

Figure 10. STL decomposition of flight cancellations time series

Both decompositions interpret that there is a seasonality component in the data. However, seasonality does not seem obvious. Rather, the one peak in 2018 seems to be interpreted as the trigger for a seasonal pattern. This seems to suggest that there will be a need for a non-seasonal ARIMA model.

We then perform a stationarity analysis of the cancellations time series with the ndiffs() function in R statistics. It gives the number of differences required to lead to a stationary series based on the KPSS test (Hyndman and Athanasopoulos, 2018). If FS < 0.64, no seasonal differences are suggested; otherwise, one seasonal difference is suggested. In this case, we obtain a value of 0, which confirms the previously-obtained results. It is, however, interesting to remark that, if the same function is used for the complete cancellations dataset (with data until May 2020 and including the airports that were removed because they were not in the delays dataset) we obtain a value of 0. This means that the cancellations time series is stationary, and therefore the time series y_t, for all s, the distribution $(y_t, ..., y_{t+s})$ does not depend on t. The same result is obtained if it is applied to the average delay time series.

There are some key takes from this section: the STL shows a decreasing trend-cycle component, and does take the cancellations due to the COVID-19 as outliers. There is not a pattern of seasonality for the cancellations time series, and there are some similarities and common patterns between the delays and cancellations series. In addition, the delays time series is not stationary, while the cancellations time series is.

4.1.2 Autoregressive modelling (ARIMA)

As explained in Section 4.3, the abnormally small number of observations requires a non traditional approach. First, the whole sample will be used to fix the most adequate type of ARIMA model. Then, the dataset will be divided into training and holdout samples. The models will be fitted to the training sample and, later, we will and evaluate one-month-ahead forecasts for the latter months with the holdout sample, excluding the COVID-19 crisis.

Using the auto.arima() function, we obtain an ARIMA(1, 0, 0) model. When the model is of the order $(p,0,0)$ it is called an autoregressive model. After fitting the model, we obtain an *AIC* = 106.55, and the following coefficients:

$$y_t = c + 0.8458y_{t-1} + \varepsilon_t,$$

where c = 1.8925 × (1 − 0.8458), and ε_t is the white noise with a standard deviation of 0.7871 = $\sqrt{0.8872}$. The forecast for the model can be seen in Figure 11. The forecast seems to follow the same logic as the STL decomposition, understanding that the peaks in cancellations due to the COVID-19 are only outliers, and that they are likely to decrease again. It is important to add that for d = 0, the long-term forecast standard deviation will go to the standard deviation of the historical data, so the prediction intervals are essentially the same for the last few forecast horizons. Additionally, when $c \neq 0$ and d = 0, the long term forecasts will tend to go to the mean of the data. This is what can be seen in the forecast plot.

We can confirm the results obtained woth the auto.arima() function if we plot the ACF and PACF plots for the model. They are plotted in Figure 12. While the ACF plot has a sinusoidal shape, the PACF shows a significant spike at lag 1, bot none afterwrds. This can be interpreted as evidence that the series follows an ARIMA $(p, d, 0)$ model, being d = 0. The PACF plot tends to decrease, and given that there is a significant spike at lag 1 in the ACF plot, we can say that both ACF and PACF lead us to confirm that an ARIMA (1, 0, 0) model is appropiate. Therefore, what was obtained with the auto.arima() function is confirmed.

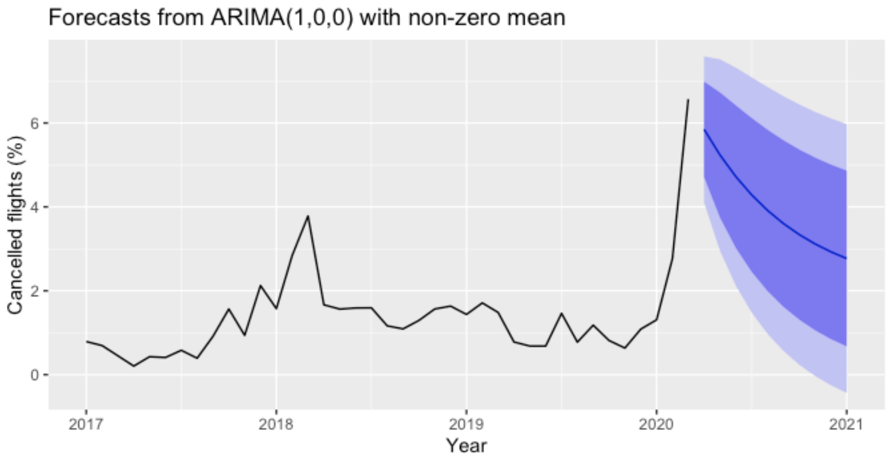

Figure 11. ARIMA forecast of autoregression for percentage of flights cancelled in the UK

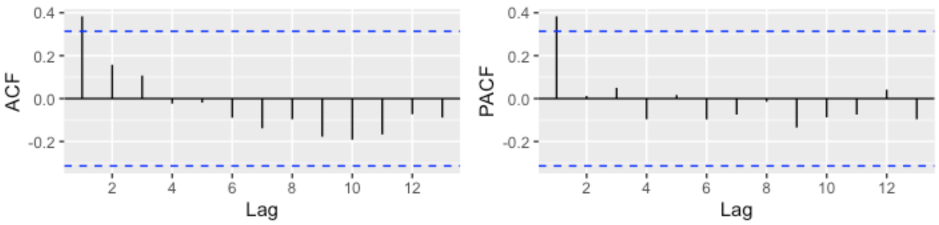

Figure 12. ACF and PACF plots for the cancellations time series

To corroborate the results and leave no room to doubts, we are going to see if nevertheless we can further minimise Akaike's Information Criterion (AIC) by changing p or q, as a verification mechanism. If we fit an ARIMA (2,0,0) model, we observe that the *AIC* slightly increases to a level of 107.49, which is a worse result than before. Moreover, when plotting the forecast, it seems that it interpreted the outlier like a step intervention variable, when it is actually a spike. When fitting an ARIMA (1,0,1) model, we obtain an *AIC* value of 107.64. It also evidences a misinterpretation of the 2020 peak. Consequently, we can infere that the ARIMA (1,0,0) model is the most correct approach to our cancellations data.

4.1.3 Trime series with exogenous variables (ARIMAX)

The model from the past section does not allow for the inclusion of other potentially relevant information apart from past observations. Namely, the effect of delays on cancellations. To take into account how exogenous variables affect the UK cancellations time series, an ARIMAX model needs to be fitted. This is a kind of Transfer Function Model, which can work with multivariate time series.

For that matter, we change the error term ε_t with the error series η_t, which is assumed to follow an ARIMA model. As the mathematical expression in backshift notation includes η_t, the model has two error terms – the error from the regression model (η_t), and the error from the ARIMA model (ε_t), which is the only error assumed to be white noise.

Consistently with the literature, the delays dataset is found to be non-stationary by performing the KPSS test for the exogenous delay variables. After this, we determine whether it would be more beneficial to only take the average delay as exogenous variable or also the time bands. For the first option (only with the average delay in minutes) we get an ARIMA(2,0,0) model, with an *AIC* = 98.59. For the second option, we obtain that the ideal model would be a seasonal ARIMAX(2, 0, 0)(1, 0, 0) model (*AIC* = 79.28), which varies from the previous one, as it now has a seasonal component. It indicates no first or seasonal difference, non-seasonal and seasonal AR(2) and AR(1) components, and no non-seasonal MA(0) component. The *AIC* is considerably lower for the second subset and, consequently, this model is chosen.

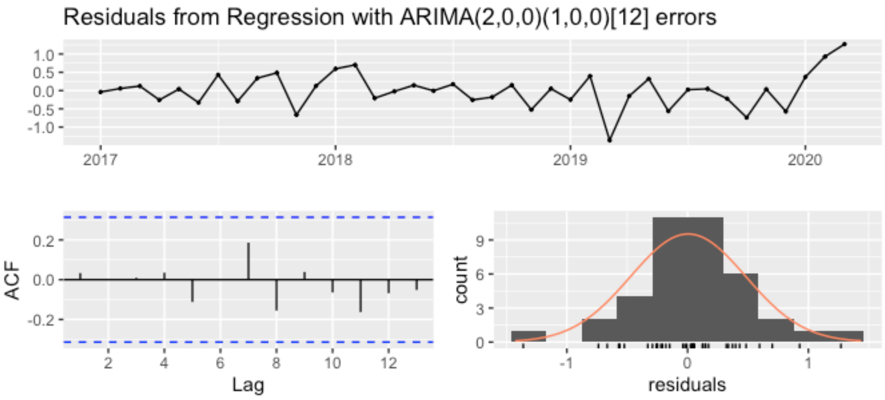

Figure 13. Forecasts of ARIMAX(2,0,0)(1,0,0) model

We can recover the residuals (the ARIMA errors) to estimate whether they resemble a white noise series. We can see in Figure 13 that they are not significantly different from white noise. The forecast is presented in Figure 14.

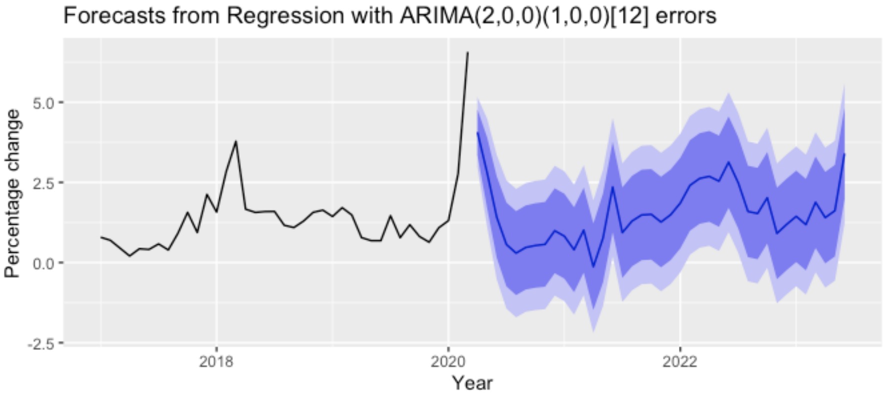

Figure 14. STL decomposition of flight cancellations time series

4.1.4 Comparison of the models

Here, the time series dataset is divided into training and test samples, and the ARIMA models that we fixed in the past subsection are fitted to the training data. Later, the performance is evaluated with the test samples. The training sample comprises observations until June 2019, and the test from July 2019 until December 2020. The models were fitted to the training set and then tested with the test set. Additionally, a training and a test set for the exogenous variables had to be created to perform the comparison.

The coefficients for the ARIMAX(2,0,0)(1,0,0) model are presented in Table 11. They had to be split in three rows because the data did not fit horizontally. The AIC = 37.26 and the σ^2 = 0.1378.

Table 11. Coefficients of the ARIMAX(2,0,0)(1,0,0) model

	coefficient	**s.e.**
ar1	1.0111	0.2650
ar2	-0.2765	0.2581
sar1	0.1098	0.2369
intercept	2.9456	1.4908
average_delay	-0.2118	0.2368
early_to_15	-0.0246	0.0202
16_to_30	-0.0528	0.0927
31_to_60	0.0321	0.1275
61_to_180	0.3356	0.2975
181_to_360	1.4134	0.7141
more_than_360	3.1611	2.4627

After testing both models with the test sets and the exogenous variables' test set, we compare the following performance parameters: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and MASE (Mean Absolute Scaled Error). The first three measures were computed with the function accuracy() in R. However, it does not report the MASE, because it needs historical data to compute the scaling factor. The MASE measures measures the relative reduction in error compared to a naive model by calculating the mean absolute error of the model divided by the mean absolute error of a naïve random-walk-without-drift model (Hyndman, 2006). Thus, a function had to be created to compute the MASE of both models. These parameters are shown in Table 12, and Figure 15 shows the forecasts of both models plotted against the actual values until December of 2020.

Table 12. Accuracy of both time series models for aggregate UK data

	RMSE	MAE	MAPE	MASE
ARIMA(1,0,0)	0.356	0.332	38.44	5.673
ARIMAX(2,0,0)(1,0,0)	0.278	0.234	25.20	5.666

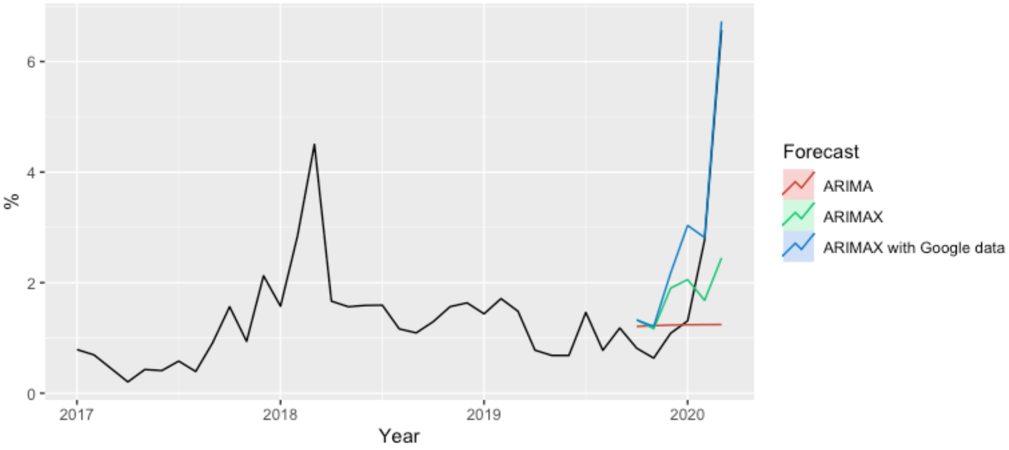

Figure 15. Forecasts of the ARIMA and ARIMAX models plotted against the actual aggregate UK data

The best method is the ARIMAX method (regardless of which accuracy measure is used). The accuracy parameters show that the model including exogenous variables has a better performance than the autoregressive model. The RMSE of both models is relatively good, since the dependent variable is measured in percentage. The RMSE of the ARIMAX model is 0.078% better than the autoregressive model. The MAE improves by 0.098%, and the biggest improvement is found in the MAPE (the model including delays improves the other's performance by 13.24%). The improvement in MASE is almost insignificant.

The fitted ARIMAX(2,0,0)(1,0,0) for UK aggregate data is:

$$\begin{aligned} cancelled\_percent_t = 2.9131 &- 0.0904 average\_delay\_mins_t \\ &- 0.0122 early\_to\_15\_mins\_late\_percent_t \\ &- 0.2831 flts\_16\_to\_30\_mins\_late\_percent_t \\ &- 0.0323 flts\_31\_to\_60\_mins\_late\_percent_t \\ &+ 0.3632 flts\_61\_to\_180\_mins\_late\_percent_t \\ &+ 2.1217 flts\_181\_to\_360\_mins\_late\_percent_t \\ &+ 2.7948 more\_than\_360\_mins\_late\_percent_t + \eta_t \\ \eta_t &= 0.7757\eta_{t-1} - 0.4156\eta_{t-2} - 0.3357\eta_{t-m}, \\ \varepsilon_t &\sim NID(0, 0.1636) \end{aligned}$$

If we perform the same analysis at a different aggregate level, we obtain similar results. The same two models were fitted to a training set consisting on observations only from the London region. Both training and test samples consisted of the same number of observations than in the previous case, with aggregate UK data. The results are shown in Table 13.

Table 13. Accuracy of both time series models for aggregate London region data

	RMSE	MAE	MAPE	MASE
ARIMA(1,0,0)	0.256	0.222	26.38	4.500
ARIMAX(2,0,0)(1,0,0)	0.159	0.234	16.31	4.523

The results are consistent with the previous findings. All the measures appear to improve with the model with exogenous variables, except for the MASE, which roughly stays the same.

Lastly, we will perform the analysis in the smallest aggregate level. We will perform the same steps as before for the Manchester airport (as it is outside of the London region, to widen the spectrum of examination). Once again, the time series was divided in a training and holdout samples, going from January 2017 to June 2020 and from July 2020 to December 2020, respectively. The ARIMA(1,0,0) and ARIMAX(2,0,0)(1,0,0) that we fixed in the previous sections were fitted for the training set, and the following performance measures were obtained (see Table 14).

Table 14. Accuracy of both time series models for aggregate London region data

	RMSE	MAE	MAPE	MASE
ARIMA(1,0,0)	0.593	0.522	75.79	4.934
ARIMAX(2,0,0)(1,0,0)	0.279	0.243	30.29	4.952

Here, the differences are even more salient for the first three measures. Especially, it is intriguing to see what a bad MAPE performance the autoregressive model with no exogenous varibles has. The MASE's are again almost the same for both models. However, the most interesting difference comes when plotting both forecasts against the actual data for Manchester airport until December of 2020 (see Figure 16).

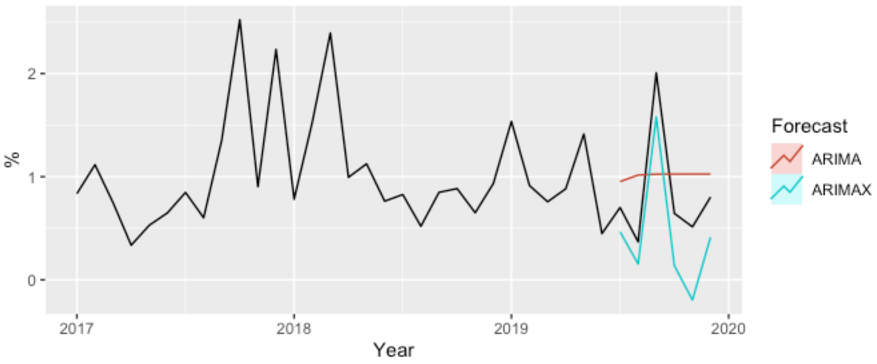

Figure 16. Forecasts of the ARIMA and ARIMAX models plotted against the actual Manchester airport data

Except for an overly optimistic negative peak in November of 2020, it is not difficult to see in the graph that the model taking delays as exogenous variable is best for these data. The shape fits way more accurately than the almost-straight line that is obtained with the forecasts of the ARIMA model. Even in the already mentioned negative peak of December 2020, the shape of the curve is loyal to the time series values from that airport.

Summarizing, it has been demonstrated how the model including different delay parameters has performed better and improved the cancellations forecasts accuracies along different levels of aggregation for UK time series data.

Therefore, there is sufficient evidence to reject H_0, and the results are consistent with H_1.

4.2 Evaluating the effect of networks and competition

Now that we have set clear the role of delays on cancellations, a further study that nuances and expands the literature is needed. The first model includes only control variables, while the second includes the control variables plus the independent target variables. The three models exclude time and delay effects. The standardised coefficients are also reported to show the relative effects of the independent variables on the percentage of flights cancelled.

Regression model including only control variables (restricted model):

$$\begin{aligned} \mathit{flights\_cancelled\_percent} = & \alpha + \beta_1 \mathit{Scheduled} + \sum_{i=1}^{n_{RA}} \beta_{2,i} \mathit{reporting\_airport}_i + \\ & \sum_{j=1}^{n_{OD}} \beta_{3,j} \mathit{ODregion}_j + \sum_{k=1}^{n_{AL}} \beta_{4,k} \mathit{airline\_name}_k + \\ & \beta_5 \mathit{average\_delay\_mins} + \varepsilon, \qquad \varepsilon \sim n(0,\sigma) \end{aligned}$$

Regression model including control and independent variables (unrestricted model):

$$\begin{aligned} \mathit{flights\_cancelled\_percent} = & \alpha + \beta_1 \mathit{Scheduled} + \sum_{i=1}^{n_{RA}} \beta_{2,i} \mathit{reporting\_airport}_i + \\ & \sum_{j=1}^{n_{OD}} \beta_{3,j} \mathit{ODregion}_j + \sum_{k=1}^{n_{AL}} \beta_{4,k} \mathit{airline\_name}_k + \\ & \beta_5 \mathit{average\_delay\_mins} + \beta_6 \mathit{Hub} + \\ & \beta_{7M} \mathit{Monopoly}_M + \beta_{7N} \mathit{Monopoly}_N + \\ & \beta_8 \mathit{n\_airlines\_airport} + \varepsilon, \qquad \varepsilon \sim n(0,\sigma) \end{aligned}$$

A multiple linear regression was calculated to measure the fixed effects of the different features in the model on the percentage of flights cancelled based on the control variables. A significant regression equation was found (F(567,104980) = 14.81, p < 0.01), with an R_2 = 0.074 and adjusted R_2 = 0.069. This adjusted R2 means that the model overall has a very low explanatory power as only 6.9% of the variation in the percentage of cancelled flights is explained by the model. Adjusted R-squared is preferred as it considers the number of variables in the model and only increases if a new term introduced improves the model more than would be expected by chance.

A multiple linear regression was calculated to measure the fixed effects of the different features in the model on the percentage of flights cancelled based on all explanatory variables. A significant regression equation was found (F(571,104976) = 14.87, p < 0.01), with an R^2 = 0.075 and adjusted R^2 = 0.07. This means that 7% of the variance in the percentage of flights cancelled is explained by the model. It does not suppose a remarkable improvement with respect to the model with only control variables.

The F-test are performed and the results are shown in Table 5.2. As can be seen, the new model appears to be statistically significant under any value of significance. There is sufficient evidence to reject the null hypothesis, because the observed F value is larger than the critical f value. The joint contribution of the *Monopoly*, *Hub* and *n_airlines_airport* to the explanation of the variation in the percentage of cancelled flights in the UK is significant at any level of significance. If we perform the F-test of the unrestricted model taking out each of these three variables at a time, we can evaluate the null hypothesis that that particular variable has no effect, against the alternative that at least one of the categories deviates markedly from the rest. We find that *Monopoly* has no significant effect, while *Hub* and *n_airlines_airport do*.

Table 15. ANOVA results for the restricted and unrestricted models

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	104980	9839173.34				
2	104976	9830868.36	4	8304.98	22.17	0.0000

As additional information, we can see which are the five airports that contribute more positively to the percent of cancellations (thus, and paradoxically, most negative β_i), and the top five that have a bigger β_i. The list of the top 5 airports with worst performances are shown in Table 16, while the opposite rank is found in Table 17. It is important to mention that Aberdeen is the offset category, and thus its coefficient is 0. It is interesting to see how two out of these five airports come from the London Region, although in Figure 5 it can be seen how the London Region is not among the regions with highest percent cancellations. From the top 4 worst performing airports, Gatwick, Manchester airport are hubs (Hub = 1), whereas neither of the rest, i.e., Aberdeen, Southampton and London City are (Hub = 0). For the five best performing airports, two of them find themselves in the British Crown Dependencies, which show a relative bad performance in Figure 5. Surprisingly, none of these five airports is a hub (Hub = 0).

Table 16. Coefficients of the 5 airports with worst cancellation service quality

	Airport	Estimate
22	ABERDEEN	0
23	MANCHESTER	0.16
24	LONDON CITY	0.31
25	SOUTHAMPTON	0.80
26	GATWICK	3.26

Table 17. Coefficients of the 5 airports with best cancellation service quality

	Airport	Estimate
1	JERSEY	-4.14
2	BELFAST INTERNATIONAL	-3.42
3	ISLE OF MAN	-3.12
4	EXETER	-2.91
5	CARDIFF WALES	-2.86

4.3 Analysing the impact of COVID-19

Since the beginning of this research, the development of the COVID-19 virus has changed dramatically over time, and so have the papers, publications, articles, etc. that aim to make sense of its economic and social impact. The forecasts have worsened monthly showing less optimistic images as time passed as can be seen in Figure 17, where IATA (2020) demonstrate how economists' forecasts for GDP recovery turned more pessimistic as the virus spread. However, over the last weeks, these forecasts seem to be finally reaching a settling point, although nothing will be certain until the spread is over.

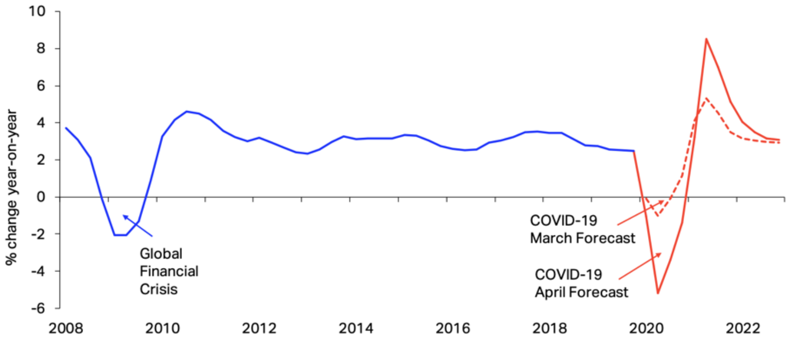

Figure 17. Economists' revised forecasts for global GDP recovery (IATA, 2020)

Particularly, one of the most impacted markets globally has been the aviation industry. Travel bans and restrictions, closing of borders, quarantines, and a number of other factors have brought air travel to a standstill, with many airlines seeing themselves obligated to ground their fleets. The International Air Transport Association (IATA, 2020) released a press conference asking governments to provide airlines with the necessary liquidity and support they need to survive this crisis. Supporting these claims, the International Civil Aviation Organization (ICAO, 2020) estimates that the impact of the COVID-19 on world scheduled passenger traffic for the full year 2020, compared to the status quo for the same year, could range from USD $289 to $387 billion in potential losses for airlines, with an overall reduction of 2,247 to 2,997 million passengers and of 39% to 52% of seats offered. Figure 18 represents the difference in delays and observations since the start of 2020 – the x axis is not divided by month, but by 0.1. Therefore, the last observation corresponds to March 2020.

The two main reasons why airlines saw themselves forced to cancel this number of flights despite the losses that they incurred are, on the one hand, the travel bans and some closing of borders, which meant that some flights would not be allowed to enter certain airports or even countries – as it happened in Guayaquil (Ecuador), when local authorities made security bodies blockade the landing of two planes coming from Spain and the Netherlands with barricades in the runway (Gant, 2020). On the other hand, an increasing public concern for the disease converged with the travel limitations imposed by governments to lower the demand for air travel to almost 0 (ICAO, 2020). Thus, there is a need to include a new variable that helps researchers to make sense of the new cancellations and come up with accurate forecasts.

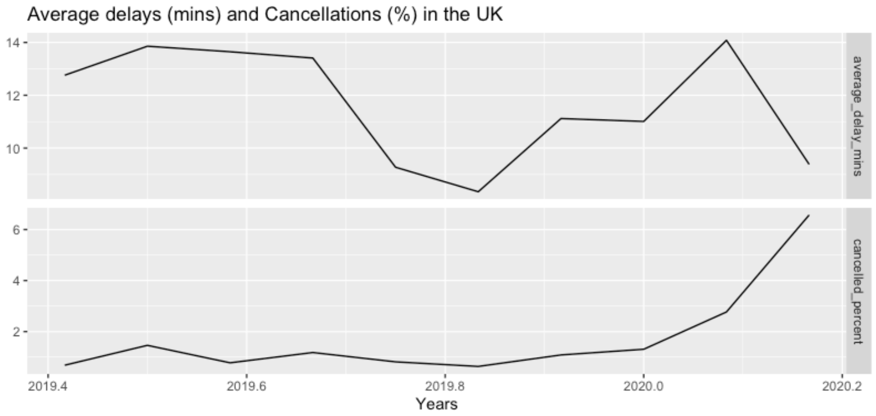

Figure 18. Comparison between average delays (mins) and cancellations(%) since June 2020

To analyze this, we will need to look at Google trends search queries. Three sets of keywords were chosen – two different sets portraying different behaviors, and a third one that serves as a comparison between the most searched terms in both previous subsets. The first subset consists of the keywords "Flights", "Tickets", "Cheap flights" and "Ryanair". This subset represents the consumer behavior, i.e., they are terms that would normally be widely searched but are likely to suffer the consequences of the outbreak, and thus portray a drop in willingness to travel or engage in purchases related to travel or leisure. On the counterpart, the second subset of keywords is formed by the terms "Travel insurance", "cancelled", "Refund", "cancelled flights". This represents the opposite, as usually they are not likely to have irregular or abnormal amount of searches, but this situation has probably increased people's concerns with this kind of topics (compensation and involuntary impediments). These change in long term trends can be seen in Figure 19 and Figure 20.

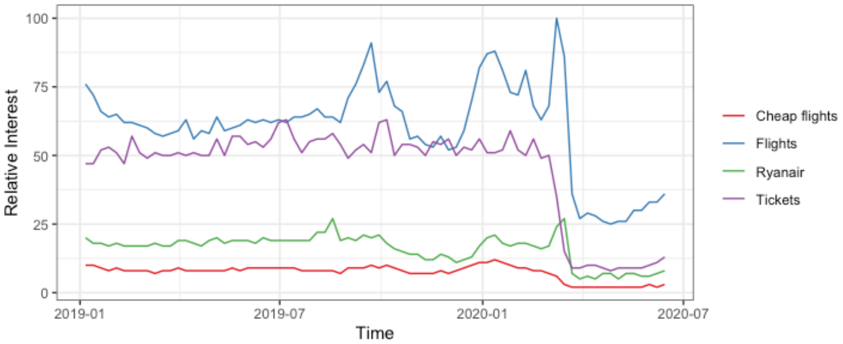

Figure 19. Relative Google search volume for the first subset of keywords

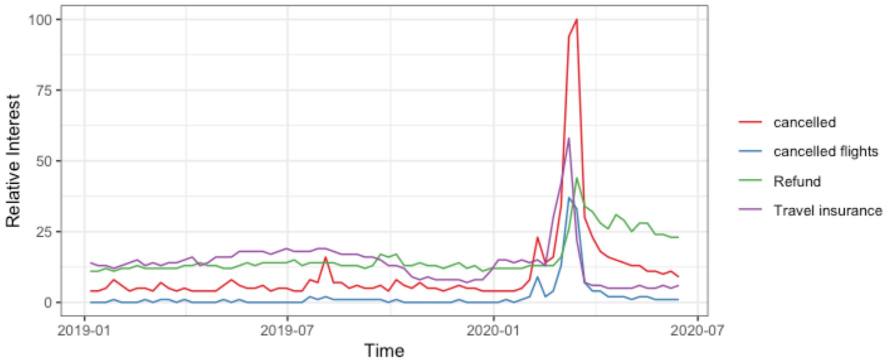

Figure 20. Relative Google search volume for the second subset of keywords

As shown in Figure 21, now we performed a short-term analysis of the two most relevant keywords from the previous subsets – cancelled, because it was included in terms regarding flight cancellations but also event cancellations, and tickets, because it was the term that better portrayed the consumers' intention to engage in purchasing of flight tickets, concert tickets, etc. (i.e., non-physical purchases). Two different profiles for the cancelled curve can be observed: an increasing one, and a decreasing one. Until half of March, both curves presented opposite profiles. Except for a few outliers in February, it can be seen how people started to increasingly search for cancellations at the end of February/ start of March, which was when Italy started to report abnormal numbers of cases – the 26th of February, Italy reported 80 new cases (adding up to a total of 400) of people infected by the COVID-19 (Coronavirus cases surge to 400 in Italy, 2020). However, the same did not happen with the search of tickets, which remained roughly the same until almost the 10th of March. This could mean that people were not aware of the fact that, for such a virus, the concept of 'borders' does not apply, and that even when trips to Italy started to being banned, people were optimistic on the possibility of following a regular lifestyle the upcoming months. However, it seems like the concern grew greatly after some Northern countries started to report the first deaths (van Algemene Zaken, 2020) and the World Health Organisation officially declared it a global pandemic (Ghebreyesus, 2020) on the 11th of March. The peak of cancellation searches, which could be a proxy for the moment of biggest concern, coincides with the day in which the number of deaths in the UK nearly doubled in 24h, Germany announced they were going to close the borders and Ireland and the Netherlands shut down bars (Gelder, 2020), and more voices from the scientific world were raising their concern for the approach that the UK had taken of letting the virus roam free around the country, as 229 specialists in disciplines ranging from mathematics to genetics signed an open letter saying that the UK virus strategy would unnecessarily put many lives at risk (Ghosh, 2020).

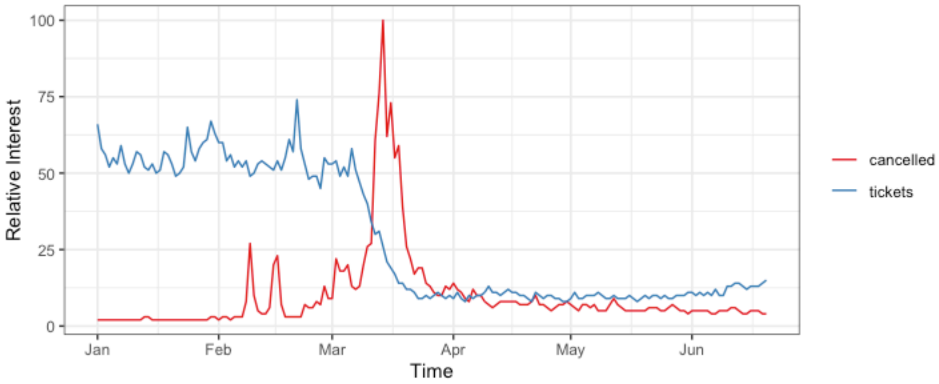

Figure 21. Relative Google search volume short-term comparison between the terms 'cancelled' and 'tickets'

After this, it seems like two interesting things happened. On the one hand, the concern for cancellations dropped again to values slightly above before in a matter of a week, and has remained stable ever since. On the other, the search for tickets continued its steep fall until reaching a stabilization point almost at the same time as cancellations. Since that time, it has seen a slight tendency towards growth, but it is still at around 20% of its previous value. This could mean that, while people's hopes for a normal and social life in which they can travel do not go up, neither does the concern. It somehow portrays a generally passive attitude towards the pandemic. On the one hand, there is not another rise in concern after the main events happen, despite more borders closing or, but neither is there a change in attitude towards the opposite direction, i.e., there is not an increase in optimism. People's attitude seems to remain stagnated in a passive state.

To finish this case study, four plots showing the development in search volume of a series of terms in the UK and Spain can be seen in Figure 22. It can be seen as more crucial searches, i.e., searches regarding health (facemasks) or economy (unemployment) have different profiles for each country and, on top of this, have a different magnitude in countries where the COVID-19 has had different effects at different times. While Spain has almost half of the population than the UK, the volume of searches for these more serious terms is sometimes ten times higher. On the contrary, for more banal searches, such as Ryanair or football, the profiles behave more similarly and, consequently with the population density, show higher volumes of searches for the UK. These anomalies and similarities in searches support the ideas from the start of this section and nuance them at the same time.

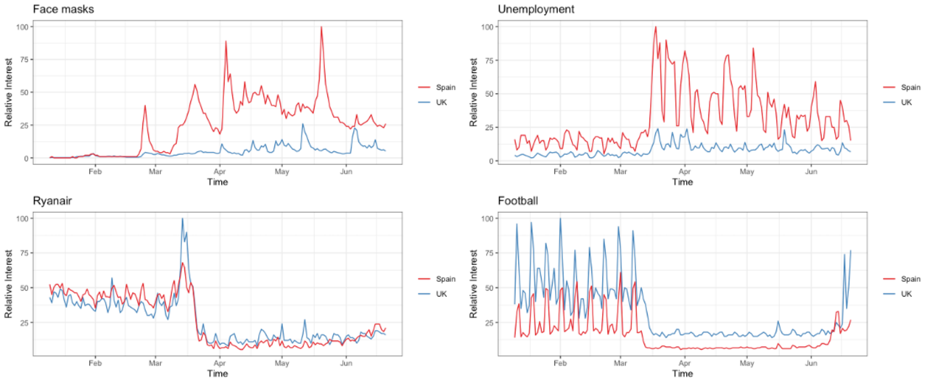

Figure 22. Google search volumes: Comparison between Spain and the UK

Summarizing this behavioral analysis, we first found that, in the context of a global pandemic, consumers tend to have a substantial drop of purchasing intention compared to the previous mean after the local circumstances add up to the global ones, and this drop tends to stay in low numbers for a sustained amount of time. On the contrary, more critical concerns experience an exponential increase that lasts a considerably short amount of time, and then they seem to return to the previous levels. Therefore, global events do not seem to have an impact on citizens until these are amplified by local circumstances, which result in a very intense but short concern and a drop in optimism (as attitude towards buying tickets for events and flights) that goes beyond that of the length of the pandemic, with no sign towards a long-term optimism. Lastly, and nuancing the point by Ribes (1992), as the surroundings determine people's behavior in different ways, depending on the "seriousness" of the topic – concerns with unemployment or health safety depend much more on local circumstances than do terms related to leisure. The countries that are affected first are thus more likely to show higher concern for longer periods of time, even after the situation is similar in the long term.

We now perform the same analysis as in section 4.1, but including the search queries that are found to be relevant in this section as exogenous variables. Here, the level of aggregation is UK aggregated data.

First, we did the necessary data preparation to convert the weekly Google Trends data into monthly data. Then, we add this data to our data frame and split it into a training and a holdout samples. After this, we need to obtain empirical evidence into why using Google Trends data to forecast the values after 2020. For that matter, we part from the assumptions from this section. After this, we fit three models: the ARIMA(1,0,0) model, the simple ARIMAX(2,0,0)(1,0,0) model, and the ARIMAX(2,0,0)(1,0,0) model with also Google data. We fit them with the training data, and test their accuracy with the test sample. All of this is done before 2020.

We obtain the results that appear in Table 18. In Figure 23, the plotted forecasts are presented.

Table 18. Accuracy of the three time series models for UK aggregate data in the COVID-19 period

	RMSE	MAE	MAPE	MASE
Autoregressive model	0.353	0.323	38.18	6.759
Model with delays	0.307	0.250	26.441	6.755
Model with delays and Google Data	0.465	0.405	44.426	6.730

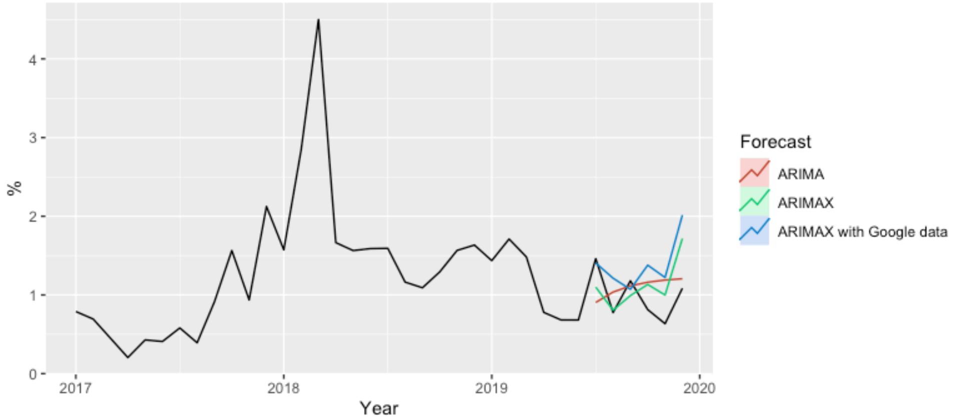

Figure 23. Forecasts of the three models against actual values for flight cancellations before 2020

We can observe how, even though the plotted forecasts do not seem to drastically improve when taking into account the COVID-19 data, the performance measures show different results. The model with delays and Google Data as exogenous variables appears to only perform better in MASE. As there is at least an improvement in one of the performance measures, we can infer that there is evidence to perform a further analysis with Google Trends data, to see if the forecasts improve.

Hence, we proceed to perform the analysis of the COVID-19 period. The new training and test sets comprise the data that ranges in the dates from 2017/01 – 2019/10 and 2019/11 – 2020/03, respectively. The data is aggregated to a UK level. The three models (the ARIMA(1,0,0) model, the simple ARIMAX(2,0,0)(1,0,0) model, and the ARIMAX(2,0,0)(1,0,0) model with also Google data) are fitted with the training data, and the forecasts are obtained. After this, we evaluate their performance and obtain the results in Table 19. Additionally, we have plotted the forecasts of the three models against the actual values of flight cancellations until March 2020 (see Figure 24).

Table 19. Accuracy of the three time series models for UK aggregate data in the COVID-19 period

	RMSE	MAE	MAPE	MASE
Autoregressive model	1.932	1.168	41.937	6.776
Model with delays	1.427	0.955	37.373	6.681
Model with delays and Google Data	1.449	1.262	92.551	6.458

The results obtained in terms of performance parameters are slightly different from before. The model performs better than the only autoregressive one in half of the performance measures (RMSE and MASE). This new model also has a remarkably bad MAPE performance (92.551%). Compared with the model that only comprises the average delay in minutes and the different delay time bands as exogenous variables, it only seems to perform better in MASE. Therefore, it is yet unclear if the model improves the forecasts of the previous one. However, when the forecasts are plotted against one another and against the actual values for the start of 2020, the conclusions are different.

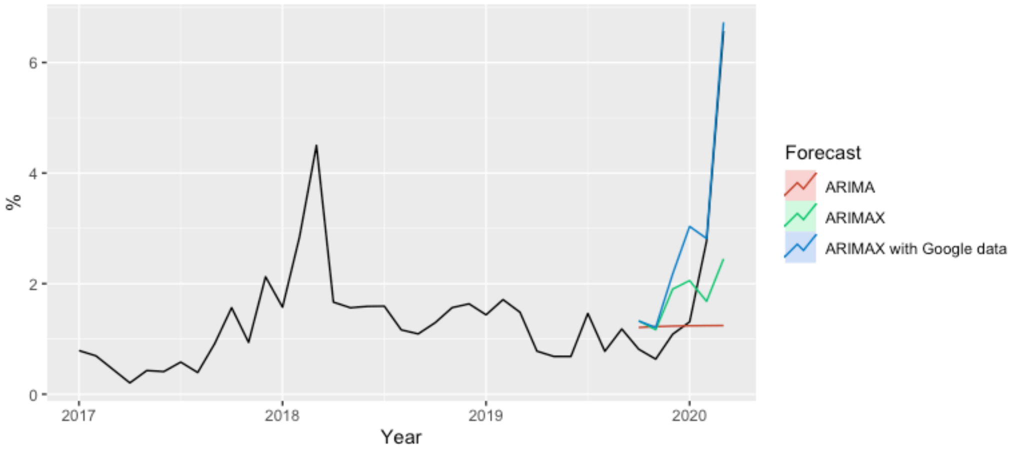

Figure 24. Forecasts of the three models against actual values for flight cancellations after 2020

5 Summary of the budget

The Budget has been calculated approximating the hourly cost of work to 25€/h, given that it took more than 4 months of work on top of the later modifications.

Table 20. Summary of the budget for the project

	Unit	**Amount**
Hourly working capital cost	€/h	25
Days to complete project	d	120
Daily working hours	h	4
Total hours worked	h	480
Total cost	**€**	**12.000**

6 Analysis and assessment of environmental and social implications

On the one hand, the COVID-19 pandemic has affected people all over the globe, but especially those more vulnerable. People on an economic position that does not allow them to have financial flexibility under unexpected circumstances are the ones that have suffered more severely the consequences of this unexpected event.

There are many dimensions in which these individuals and groups have seen their financial plans altered. For instance, when these families saw the flights they had booked several months back canceled due to circumstances outside their reach, those in a more vulnerable position, who had sacrificed part of their income to enjoy their vacations on another country, were suddenly in a situation where part of their annual budget was blocked without use indefinitely. It was only months later when airlines decided to start allowing refunds to those affected, but there were numerous people for whom the moment where that financial aid would have been more helpful had already passed. Additionally, this was only or the fraction of the people who either was aware o this option and knew how to ask or the refund, and on top of this the people who actually had it back.

A careful and thorough planification of flight cancellations is needed to better support these people in the future, so airlines can allocate budget portions and establish strategic plans to face potential refunds. This would not only have a social impact by helping these individuals but also a positive PR effect for airlines implementing it, as customer satisfaction would increase by having their refunds earlier.

On the other hand, an accurate model predicting light cancellations could improve the efficiency of airlines and save both costs and emissions by e.g., better scheduling passenger land transport, improving airports efficiency, more accurately ordering food amounts for the only the flights that are going to depart, and not for all of those planned, etc.

Shortly, having accurate fight cancellation models not only can have a positive social impact for those in financial needs (and its positive marketing implications) but also the potential to reduce CO2 emissions and improve efficiency.

7 Discussion and conclusions

This research aims to provide insight into the dynamics of flight cancellations in the UK in the context of the COVID-19 virus outbreak. In particular, it aims to: (1) deepen into the relationship between delays and cancellations from a time series perspective and improve cancellations forecasts by introducing delays in the model; (2) analyse how hub airports, route competition and airport competition impact the percentage of flights cancelled; and (3) understand behavioral changes regarding air transport that occur as a result of the COVID-19 spread and improve cancellations forecasts in the first months of 2020 by including Google Trends data in the models.

7.1 Main findings

The first section aims to evaluate the impact of delays in cancellations from a time series perspective. Particularly, it aims to answer the question on whether delays improve the forecasts of flight cancellations. It was found that, for 2019 data, the model that included the average flight delay in minutes and different delay time gaps performed better. Every performance measure (RMSE, MAE, MASE and MAPE) was improved by the new ARIMA(2,0,0)(1,0,0) model.

The second section focuses on examining the effects of networks, congestion and competition in service quality performance. In particular, it evaluates whether the presence of hubs, a higher route competition and lower competition in airports result in a lower percentage of flight cancellations, while including flight delays in the model. The multivariate regression analysis nuanced the literature, as it was found that these parameters have a significant effect on the dependent variable flights_cancelled_percent under any level of significance.

In the last section there were two main findings. First, it was found that the pandemic resulted in a situation of passiveness regarding air travel and purchasing intentions. Whereas there was a sudden very high volume of searches for the keyword *cancelled*, that refers to cancelled flights, events, etc., there was an equally sudden drop to a stagnated level, almost at the level that it was prior to the virus outbreak. On the other hand, terms related to purchasing tickets for events or flights also saw a very big drop in terms of search volume. However, instead of returning to the previous level, they have stayed in a very much lower level than before, showing a lack of optimism for rejoining the previous social lifestyle even in the medium term.

Lastly, the Google search volume of the terms *cancelled* and *tickets* were taken into account as exogenous parameters to forecast the relative amount of cancelled flights, and it was found that the MASE improved with respect to the autoregressive model and the model with only the effect of delays. Although the rest of the performance parameters were not improved, this might be due to a strong deviation in the prediction of the value from January 2020, as the plotted forecasts show that the forecast of the model that includes the Google search queries is more faithful to the data.

7.2 Discussion

Most of the studies which analyse air transport service quality focus solely on delays (Abdel-Aty et al., 2007; Manna et al., 2017; Kuhn and Jamadagni, 2017). The studies which do analyse flight cancellations have several flaws. As Lemke et al. (2009) points, no method had ever proven successful across various studies and time series, mainly due to the difficulty to design a *one-solution-fits-all* method that comes from diverse characteristics and underlying data generation processes of the data. On top of this, there is no conclusive evidence on the influence of delays in cancellations Rupp (2005); Xiong and Hansen (2013), and the studies that have taken into account the time series nature of the data point towards the need of analyzing the non-stationarity of the delays time series. In terms of how delays affect cancellations, this study contributes to the existing literature that analyzes flight cancellations and expands it by, on the one hand, studying it from a time series perspective and, on the other hand, establishing the best method to forecast cancellations while taking into account the role of flight delays. It also nuances the findings of Rupp and Holmes (2006) by expanding their contextualization of the role of delays on cancellations. They had found that competition and network effects in US domestic flights are significant, but had not taking into account the effect of airport congestion (in terms of airlines operating in that airport at a given month).

Moreover, this study is among the first ones to examine the role of the COVID-19 on flight cancellations. We have analyzed the behavioral changes of the population in the context of a pandemic, evaluating how their consuming behavior or optimism in reengaging in travel and social events, and their concern for flight cancellations evolved over the first three months of 2020. It was found that both tendencies have different profiles and tend towards a passive behavior. The previous models are improved by taking into account these Google Trends data as an additional exogenous parameter to help predict the relative amount of cancellations relative to total air transport movements.

7.3 Managerial implications

The findings from the research outcome draw several implications for business practices. The first section can benefit airlines and third parties by providing a more thorough understanding of the underlying patterns of cancellations. As it is found by that cancellations are the most relevant metrics for passenger dissatisfaction and switch of carriers, airlines can allocate resources to fidelity programs or marketing campaigns to prevent and reduce the loss in trust that they know is likely to occur at a given time, without wasting resources at times where maybe the public opinion is more positive, by having a distribution over time of the forecasted levels of cancellations. These forecasts can also be used to predict expected airspace congestion levels and lead to more accurate decisions. This acquires particular salience when thinking of insurance companies. By taking into account the delay variable into their operations, they can check the data from that flight, or even airline and airport historical data, and they can better trace the causes of cancellations. When a flight gets cancelled, insurance companies can optimize their claims handling by having an exact understanding of the events under which a cancellation is more likely to occur. This improved transparency eliminates information asymmetries between them and their clients, therefore preventing insurance companies from paying unnecessary amounts to unpleased passengers.

The main managerial relevance of the second section is aimed towards regional governments and institutions. Given the findings, they can consider deviating resources to a specific airport in a region by promoting its use as a hub for more than 5 airlines, thus deriving airlines from other smaller airports towards this one. If capacity is not surpassed, e.g., by building an additional runway, it would come with two main achievements from a regional perspective: increased service quality in new hubs and decongestion in the rest of the airports for the region.

These two improvements, by means of benefitting passengers, would come with advantages for the different regions. On the one hand, by improving the service quality, even more airlines would be interested in using the regional hub as a hub for their own, which would mean that many flights that before would not pass by that region, would now do. This can promote the local economy by having a wider flight network, as it is the case of Frankfurt, in Germany. On the other hand, by deviating flights from that region and concentrating them in a single airport, this spoke decongestion would decrease those airports capacity potential constraints, hence improving service quality as well. Inhabitants of that region would also see their trust on the reliability of air transport in that region increased, and therefore the need for travelling to other parts of the country instead would not be necessary anymore, which would promote more arrival flights from the region.

The last part of the study encounters its main beneficiary in the airlines. After the COVID-19 crisis, airlines have suffered a big reputational detriment, as passengers have logically given airlines responsibility for the cancellations, even when sometimes cancelling the flight was the only option that airlines could take to avoid major losses. On top of this, the forecasts for flight cancellations do not provide any insightful information during these months, as they are unable to forecast the coming values. As managerial tendencies shift towards turning transactions into relationships (Bertini and Gourville, 2012), understanding the factors that shape customers' intention to change from a one-time use to a potential relationship with the company and taking action accordingly is crucial. Thus, as now there is a clear picture on behavioral customer reactions in times of disruptive events such as a pandemic, airlines could implement fidelity programs to secure the relationships in case of unfortunate events such as this one. Companies with heavy volumes of passengers and with generally low fares (such as Ryanair or Easjyet) could benefit from establishing new business models and maybe go to new pricing systems such as subscription plans. That way, airlines would assure themselves the cash inflow in the unlikely case of a new pandemic/event that makes cancellations increase in such a steep pace. Additionally, airlines can know in advance which Google Trends data to take into consideration when intending to forecasts cancellations under similar but unusual circumstances.

7.4 Limitations and further research

As the data series had to be grouped by airport, and each airport supposes a different time series, the real number of data points for the first analysis was reduced to 39. This may bias the results, as it is harder to estimate the components of a time series if the dataset is not lengthy. The problem comes from the lack of cancellations data, as they have only been recorded from 2017. This was certainly one of the biggest limitations in the research, as the algorithms would have interpreted the COVID-19 more easily as an outlier if the data expanded longer in time. Another data limitation is the low frequency of observations. If instead of monthly data, the data were daily or weekly, the study would be more concrete and would not need for very long timespans.

There are also some problems regarding variables in the original dataset. On the one hand, it was a very limited dataset, which also had categorical variables with more than 400 categories at certain cases. Therefore, for predictive methods such as Random Forests, a less arbitrary grouping of these variables (origin/destination region, airlines, etc.) would be beneficial for the research. Some of these variables were also vaguely defined, and more specificity on their meaning would be useful. For example, distinguishing between origin and destination countries, i.e., stating whether it was an origin or destination airport.

What this study predominantly advises in regard to future research is repeating and expanding this study when the COVID-19 is not relevant anymore and some time has passed. Then, it would be very beneficial to make sense of the data that was recorded after the pandemic. For instance, it would be interesting to expand the research and see how cancellations patterns behaved in countries that adopted different measures, and analyze the differences tying it to the diverse institutional pandemic politics that were employed. One thing that research would benefit from is the post-facto cancellations analysis. If *return-to-normal* curves have different profiles in varying countries, e.g., in some countries the return to normality is as abrupt as the increase in cancellations whereas in others it takes more time, that could indicate managerial differences at the airport or local level, airline managerial preference towards certain regions, and even expand the support on the role of hubs and competition on cancellations performance. There were also missing observations over the latest months due to the effects of the virus spread. Sometime after the virus *mayhem* is over, this data would be filled, and a more thorough analysis could be made. If this study is repeated in some time, it would also be able to better evaluate the performance of both time series models, since the presence of a sudden bump in cancellations harms the performance of both models.

In terms of future research, the last recommendation is to examine the dissipation patterns of behavioral changes as a result of the COVID-19. Namely, do these effects dissipate immediately (which would demonstrate the short memory of consumers), or do they take a long time to disappear (which would mean that there is a certain psychological momentum to these kinds of stress situations)? This could have very fruitful consumer psychological and behavioral economics implications, and would allow to build measures to mitigate these effects under other scenarios. It is also recommended to further improve the model from the last section by analyzing the predictive power of these and more terms after the COVID-19 pandemic is over.

The last limitation is a misfortunate consequence of the COVID-19 regulations. Isolation measures and reclusion has made it impossible to have the fluent contact with the tutor that would have been optimal. Even though he was most of the time available for a call, it would have been better to be able to be present to clarify some doubts and better explain some ideas.

8 References

Abdel-Aty, M., Lee, C., Bai, Y., Li, X. and Michalak, M. (2007). Detecting periodic patterns of arrival delay, *Journal of Air Transport Management* **13**(6): 355–361.

Air Transport Action Group (2019). Adding value to the economy.
URL: *https://aviationbenefits.org/economic-growth/adding-value-to-the-economy/*

Balakrishna, P., Ganesan, R. and Sherry, L. (2010). Accuracy of reinforcement learning algorithms for predicting aircraft taxi-out times: A case-study of Tampa Bay departures, *Transportation Research Part C: Emerging Technologies* **18**(6): 950–962.

Balakrishna, P., Ganesan, R., Sherry, L. and Levy, B. S. (2008). Estimating taxi-out times with a reinforcement learning algorithm, *2008 IEEE/AIAA 27th Digital Avionics Systems Conference*, IEEE, pp. 3–D.

Barnhart, C. and Bratu, S. (2004). Airline passenger delays, *Solan Industry Centers Annual Conference*.

Becken, S. and Carmignani, F. (2020). Are the current expectations for growing air travel demand realistic?, *Annals of Tourism Research* **80**: 102840.

Bertini, M. and Gourville, J. T. (2012). Pricing to create shared value, *Harvard Business Review* **90**(6).

Bogoch, I. I., Watts, A., Thomas-Bachli, A., Huber, C., Kraemer, M. U. and Khan, K. (2020). Pneumonia of Unknown Etiology in Wuhan, China: Potential for International Spread Via Commercial Air Travel, *Journal of Travel Medicine* .

CAA (2020a). Airport data.
URL: https://www.caa.co.uk/Data-and-analysis/UK-aviation- market/Airports/Datasets/UK-Airport-data/Airport-data-2020-01/

CAA (2020b). Punctuality statistics 2020.
URL: https://www.caa.co.uk/Data-and-analysis/UK-aviation-market/Flight-reliability/Datasets/Punctuality-data/Punctuality-statistics-2020/

Centers for Disease Control and Prevention (2020a). Transmission of Coron- avirus Disease 2019 (COVID-19).
URL: *https://www.cdc.gov/coronavirus/2019-ncov/prevent- getting-sick/how-covid-spreads.html?CDCAArefVal = https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F*2019 − *ncov%2Fprepare%2Ftransmission.html*

Centers for Disease Control and Prevention (2020b). Transmission of Coron- avirus Disease 2019 (COVID-19).
URL: *https://www.cdc.gov/coronavirus/2019-ncov/prevent- getting-sick/how-covid-spreads.html?CDCAArefVal = https%3A%2F%2Fwww.cdc.gov%2Fcoronavirus%2F*2019 − *ncov%2Fprepare%2Ftransmission.html*

Cleveland, R. B., Cleveland, W. S., McRae, J. E. and Terpenning, I. (1990). Stl: A seasonal-trend decomposition, *Journal of official statistics* **6**(1): 3–73.

Coronavirus cases surge to 400 in Italy (2020).
URL: *https://www.bbc.com/news/world-europe-51645902*

De Arce, R. and Mahía, R. (2003). Modelos Arima, *Programa CITUS: Técnicas de Variables Financieras* .

Durka, P. and Pastoreková, S.(n.d.). ARIMA vs. ARIMAX – which approach is better to analyze and forecast macroeconomic time series.

Frankel, J. A. (2000). Globalization of the economy, *Technical report*, National Bureau of Economic Research.

Gant, J. (2020). Police in ecuador blockade a runway to stop spanish and dutch jets landing.
URL: *https://www.dailymail.co.uk/news/article-8130183/Police-Ecuador- BLOCKADE-runway-stop-Spanish-Dutch-jets-landing.html*

Gelder, S. (2020). Coronavirus: 15 march at a glance.
URL: *https://www.theguardian.com/world/2020/mar/15/coronavirus-15-march-at- a-glance*

Ghebreyesus, T. A. (2020). Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020.
URL: *https://www.who.int/dg/speeches/detail/who-director-general-s-opening- remarks-at-the-media-briefing-on-covid-19—11-march-2020*

Ghosh, P. (2020). Coronavirus: Some scientists say uk virus strategy is 'risking lives'.
URL: *https://www.bbc.com/news/science-environment-51892402*

Gössling, S., Hanna, P., Higham, J., Cohen, S. and Hopkins, D. (2019). Can we fly less? Evaluating the 'necessity'of air travel, *Journal of Air Transport Management* **81**: 101722.

Gössling, S. and Upham, P. (2009). *Climate change and aviation: Issues, challenges and solutions*, Earthscan.

Heuwieser, M. (2017). The illusion of green flying, *Vienna: Finance & Trade Watch. www. ftwatch.@ flyinggreen, accessed July* **17**: 2017.

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand, *Foresight* **4**(4): 43–46.

Hyndman, R. J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*, OTexts.

Iacus, S. M., Natale, F., Santamaria, C., Spyratos, S. and Vespe, M. (2020). Esti- mating and projecting air passenger traffic during the covid-19 coronavirus outbreak and its socio-economic impact, *Safety Science* p. 104791.

IATA (2019). *Economic Performance of the Airline Industry*.

IATA (2020). Iata thanks governments for support but more need to step up.
URL: *https://www.iata.org/en/pressroom/pr/2020-03-24-02/*

ICAO (2011). The economic social benefits of air transport.

ICAO (2016). *ICAO Environmental Report 2016: Aviation and Climate Change*.

ICAO (2020). *Effects of Novel Coronavirus (COVID-19) on Civil Aviation: Economic Impact Analysis*.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects, *Science* **349**(6245): 255–260.

Khanmohammadi, S., Chou, C.-A., Lewis, H. W. and Elias, D. (2014). A sys- tems approach for scheduling aircraft landings in JFK airport, *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, pp. 1578–1585.

Kuhn, N. and Jamadagni, N. (2017). Application of machine learning algorithms to predict flight arrival delays, *CS229*.

Lambelho, M., Mitici, M., Pickup, S. and Marsden, A. (2020). Assessing strategic flight schedules at an airport using machine learning-based flight delay and cancellation predictions, *Journal of Air Transport Management* **82**: 101737.

Lee, T. M., Markowitz, E. M., Howe, P. D., Ko, C.-Y. and Leiserowitz, A. A. (2015). Predictors of public climate change awareness and risk perception around the world, *Nature Climate Change* **5**(11): 1014–1020.

Lemke, C., Riedel, S. and Gabrys, B. (2009). Dynamic combination of forecasts generated by diversification procedures applied to forecasting of airline can- cellations, *2009 IEEE Symposium on Computational Intelligence for Financial Engineering*, IEEE, pp. 85–91.

List of airports in the United Kingdom and the British Crown Dependencies (2020). **URL:** *https://en.wikipedia.org/wiki/List_of_airports_in_the_United_Kingdom_and _the_British_Crown_Dependencies*

Lu, Z., Wang, J. and Zheng, G. (2008). A new method to alarm large scale of flights delay based on machine learning, *Proceeding of The International Symposium on Knowledge Acquisition and Modeling (KAM*, pp. 589–592.

Lyle, C. (2018). Beyond the ICAO's Corsia: Towards a More Climatically Effec- tive Strategy for Mitigation of Civil-Aviation Emissions, *Climate Law* **8**(1-2).

Mahase, E. (2020). Coronavirus: covid-19 has killed more people than sars and mers combined, despite lower case fatality rate.

Manna, S., Biswas, S., Kundu, R., Rakshit, S., Gupta, P. and Barman, S. (2017). A statistical approach to predict flight delay using gradient boosted decision tree, *2017 International Conference on Computational Intelligence in Data Science (ICCIDS)*, IEEE, pp. 1–5.

Markham, F., Young, M., Reis, A. and Higham, J. (2018). Does carbon pricing reduce air travel? Evidence from the Australian ‘Clean Energy Future’policy, July 2012 to June 2014, *Journal of Transport Geography* **70**: 206–214.

Marsland, S. (2014). *Machine learning: an algorithmic perspective*, Chapman and Hall/CRC.

Mead, K. and General, I. (2000). Flight delays and cancellations, *US Department of Transportation, Report Number CC-2000-356* p. 15.

Meyn, L. (2002). Probabilistic methods for air traffic demand forecasting, *AIAA Guidance, Navigation, and Control Conference and Exhibit*, p. 4766.

Mueller, E. and Chatterji, G. (2002). Analysis of aircraft arrival and departure delay characteristics, *AIAA’s Aircraft Technology, Integration, and Operations (ATIO) 2002 Technical Forum*, p. 5866.

Office of Aviation Enforcement and Proceedings (2020). *Air Travel Consumer Report January 2020*.

Oliphant, R. and Carpani, J.and Vogt, A. (2020). ’Like a wartime curfew’: Inside Italy’s coronavirus quarantine zone. **URL:** *https://www.telegraph.co.uk/news/2020/02/24/nightmare-inside-italys- coronavirus-quarantine-zone/*

Pyrgiotis, N., Malone, K. M. and Odoni, A. (2013). Modelling delay propagation within an airport network, *Transportation Research Part C: Emerging Technologies* **27**: 60–75.

Rebollo, J. J. and Balakrishnan, H. (2014). Characterization and prediction of air traffic delays, *Transportation research part C: Emerging technologies* **44**: 231–241.

Ribes, E. (1992). Factores macro y micro-sociales participantes en la regulación del comportamiento psicológico, *Revista mexicana de Análisis de la Conducta* **18**(3): 39–55.

Rupp, N. G. (2005). Flight delays and cancellations, *Working Paper* .

Rupp, N. G. and Holmes, G. M. (2006). An investigation into the determinants of flight cancellations, *Economica* **73**(292): 749–783.

Scott, D., Hall, C. M. and Gössling, S. (2016). A report on the Paris Climate Change Agreement and its implications for tourism: Why we will always have Paris, *Journal of Sustainable Tourism* **24**(7): 933–948.

Sridhar, B., Wang, Y., Klein, A. and Jehlen, R. (2009). Modeling flight delays and cancellations at the national, regional and airport levels in the United States, *8th USA/Europe ATM R&D Seminar, Napa, California (USA)*.

Sternberg, A., Soares, J., Carvalho, D. and Ogasawara, E. (2017). A review on flight delay prediction.

Suzuki, Y. (2000). The relationship between on-time performance and airline market share: a new approach, *Transportation Research Part E: Logistics and Transportation Review* **36**(2): 139–154.

The Economist (2020a). Markets wake up with a jolt to the implications of COVID-19, *The Economist* .

The Economist (2020b). The virus is coming, *The Economist* .

Tomlinson, J. (1999). *Globalization and culture*, University of Chicago Press.

Tu, Y., Ball, M. O. and Jank, W. S. (2008). 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.

UN (2015). About the Sustainable Development Goals - United Nations Sustain- able Development.
URL: *https://www.un.org/sustainabledevelopment/sustainable-development-goals/*

van Algemene Zaken, M. (2020). Patient with novel coronavirus deceased.
URL: *https://www.government.nl/latest/news/2020/03/06/patient-with-novel- coronavirus-deceased*

World Health Organization (2020). Coronavirus disease (COVID-19) advice for the public.
URL: *https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for- public*

Wu, Q. (2014). A stochastic characterization based data mining implementation for airport arrival and departure delay data, *Applied Mechanics and Materials*, Vol. 668, Trans Tech Publ, pp. 1037–1040.

Xiong, J. and Hansen, M. (2013). Modelling airline flight cancellation decisions, *Transportation Research Part E: Logistics and Transportation Review* **56**: 64–80.

Xu, N., Donohue, G., Laskey, K. B. and Chen, C.-H. (2005). Estimation of delay propagation in the national aviation system using Bayesian networks, *6th USA/Europe Air Traffic Management Research and Development Seminar*, FAA and Eurocontrol Baltimore, MD.

Yanying, Y., Mo, H. and Haifeng, L. (2019). A Classification Prediction Analysis of Flight Cancellation Based on Spark, *Procedia Computer Science* **162**: 480–486.

A study of flight cancellation and delays

Document:
Annexos

Autor:
Alejandro R. Vázquez Ibáñez

Director:
Daniel García-Almiñana

Titulació:
Màster en Ingeniería Industrial

Convocatòria:
Tardor/Primavera/Pròrroga, any.

TREBALL DE FI D'ESTUDIS

Annex

```
#--------------------------------------------------------------------------
# NAME  : ReadAndPrepare01 - Delays.R
# TASK  : Read data about delays from multiple csv files
#--------------------------------------------------------------------------

# Empty the memory
remove(list=ls())
cat("\f")

#--------------------------------------------------------------------------
# Install and load packages
#--------------------------------------------------------------------------

# T packages are used in this section: plyr to calculate sub-group
# summaries, and ggplot2 to make plots

# Package ggplot2 is used for making graphics
# install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)

# Package plyr is used to make data summaries, but als
# to make available function rbind.fill, which allows
# one to stack data frames with different columns
# install.packages("plyr", dependencies = TRUE)
library(plyr)

# Package lubridate is isued to handle columns with
# mixed data frames
# install.packages("lubridate", dependencies = TRUE)
library(lubridate)

# install.packages("dplyr", dependencies = TRUE)
library(dplyr)

#--------------------------------------------------------------------------
# Set path and library
#--------------------------------------------------------------------------

# Paths are set according the logic taught in class. Names have been
# taken from the files in the shared Zip file. The name of the covid
# directory has been changed to Covid19 to get rid of the dash in the
# original name. The Rda subfolder has been added to store the
# processed data.
# dir  <- "~/Werk/Education/Other/QuestionAlejandro/" # Thuis

dir  <- "~/Offline Documents/RSM's MScBA/Thesis/QuestionAlejandro/" # Thuis

dirProg     <- paste0(dir, "Programs/")
dirData     <- paste0(dir, "Data/")
dirRslt     <- paste0(dir, "Results/")

dirData.Cancellations <- paste0(dirData, "Cancellations/")
dirData.Covid19       <- paste0(dirData, "Covid19/")
dirData.Delays        <- paste0(dirData, "Delays/")
dirData.Delays2       <- paste0(dirData, "Delays2/")
dirData.Rda           <- paste0(dirData, "Rda/")

#--------------------------------------------------------------------------
# Read delays data
#--------------------------------------------------------------------------

# Find all files in the delays directory
theFiles <- dir(dirData.Delays)

# Select all files that contain delay data (this is just a check,
# but it also allows one to store other named files in the same
# directory without destroying the code)
theFiles <- theFiles[grep("d\\d{4}_\\d{2}\\.csv", theFiles)]
```

```
# Start a loop over the files in the delays directory to import
# all the data. Function rbind.fill from package plyr is used to bind
# the different files, because the column names differ across files.
# This is soething to look into.
dfDelays <- data.frame()
for (i in 1:length(theFiles)) {
  if (i %% 12 == 0) cat("Reading file :", theFiles[i], "\n")

  dfDelays <-
    rbind.fill(
      dfDelays,
      read.csv(paste0(dirData.Delays, theFiles[i]), stringsAsFactors = FALSE)
    )
  } # end of the for-loop

# ---
# AV: Ideally, we would change the column name as we bind them, but as it is not possible,
# I am going to read every cell from the unwanted columns and say that, if it's different
# from N/A, to add it to the 'correct' one. Then, I'll delete the columns with the wrong
names.
# With an ifelse so that if it's N/A, FALSE, otherwise, it adds the value to the right
column

# The correct columns are the ones from 2000 to 2017, just due to the fact that the data
# is divided in a simpler number of timeslots. Namely, there are 2 main differences:

# 1. Until 2017, flihts from 61 to 180 min late are in the same group, while in the newer
#    datasets there's a differentiation (61-120 and 121-180)
# 2. Until 2017, only 'flights early to 15 min late'. From 2018, 3 groups( more than 15 min
#    early; 15 early-0, 0-15 late)

# The best option is to create a big DF (2000-2020), grouping the values of newer years to
# the simpler columns, and then another DF with only data from 2018-2020, but more complete
# ---


#-------------------------------------------------------------------------
# After care
#-------------------------------------------------------------------------

# Delete run_date column
dfDelays$run_date <- NULL

# Delete 'new' columns
dfDelays[20:31] <- list(NULL)


# Handle the 'reporting_period' column. Split into years and
# months. Then, a year month column is made with type Date
# (as in the Challenges data frame)
dfDelays$reporting_year  <- as.numeric(substr(dfDelays$reporting_period, 1, 4))
dfDelays$reporting_month <- as.numeric(substr(dfDelays$reporting_period, 5, 6))
dfDelays$reporting_yearmonth <-
  as.Date(paste(dfDelays$reporting_year, dfDelays$reporting_month, "01"),
          format="%Y %m %d", tz="GMT")

#AV: check airport names
dfAirportNames <- unique(dfDelays$reporting_airport)


#-------------------------------------------------------------------------
# Regions
#-------------------------------------------------------------------------
# AV: create new column 'Region' with the region where each airport is located
dfDelays$region <- NA

# AV: For the rest, I am going to create a vector for each region
England_LDN       <- c("HEATHROW", "LONDON CITY", "GATWICK", "SOUTHEND", "STANSTED")
England_NE        <- c("NEWCASTLE", "DURHAM TEES VALLEY",
                       "TEESSIDE INTERNATIONAL AIRPORT")
England_NW        <- c("BLACKPOOL", "LIVERPOOL", "MANCHESTER")
England_MidlandsE <- c("EAST MIDLANDS INTERNATIONAL")
England_MidlandsW <- c("BIRMINGHAM")
England_East      <- c("NORWICH", "LUTON")
England_SE        <- c("LYDD", "SHOREHAM", "SOUTHAMPTON", "OXFORD")
```

```
England_SW        <- c("BOURNEMOUTH", "BRISTOL", "EXETER",
                       "ISLES OF SCILLY", "LANDS END", "NEWQUAY",
                       "GLOUCESTERSHIRE")
England_Yorkshire <- c("DONCASTER SHEFFIELD", "HUMBERSIDE", "LEEDS BRADFORD")
NIreland          <- c("BELFAST CITY",
                       "BELFAST INTERNATIONAL", "CITY OF DERRY")
Scotland          <- c("ABERDEEN", "BARRA", "BENBECULA", "CAMPBELTOWN", "DUNDEE",
                       "EDINBURGH", "GLASGOW", "PRESTWICK", "INVERNESS", "ISLAY",
                       "KIRKWALL", "LERWICK", "SCATSTA", "SUMBURGH",
                       "STORNOWAY", "TIREE", "WICK JOHN O GROATS")
Wales             <- c("CARDIFF WALES")
BCD               <- c("JERSEY", "ISLE OF MAN")

# AV: Now, we check if they contain any of the character values in the vectors and add
# the name of the region of the vector
dfDelays$region[unique(
  grep(paste(England_LDN,collapse="|"),
       dfDelays$reporting_airport))] <- "LDN"

dfDelays$region[unique(
  grep(paste(England_NE,collapse="|"),
       dfDelays$reporting_airport))] <- "NE"

dfDelays$region[unique(
  grep(paste(England_NW,collapse="|"),
       dfDelays$reporting_airport))] <- "NW"

dfDelays$region[unique(
  grep(paste(England_MidlandsE,collapse="|"),
       dfDelays$reporting_airport))] <- "ME"

dfDelays$region[unique(
  grep(paste(England_MidlandsW,collapse="|"),
       dfDelays$reporting_airport))] <- "MW"

dfDelays$region[unique(
  grep(paste(England_East,collapse="|"),
       dfDelays$reporting_airport))] <- "EAS"

dfDelays$region[unique(
  grep(paste(England_SE,collapse="|"),
       dfDelays$reporting_airport))] <- "SE"

dfDelays$region[unique(
  grep(paste(England_SW,collapse="|"),
       dfDelays$reporting_airport))] <- "SW"

dfDelays$region[unique(
  grep(paste(England_Yorkshire,collapse="|"),
       dfDelays$reporting_airport))] <- "YH"

dfDelays$region[unique(
  grep(paste(NIreland,collapse="|"),
       dfDelays$reporting_airport))] <- "NIR"

dfDelays$region[unique(
  grep(paste(Scotland,collapse="|"),
       dfDelays$reporting_airport))] <- "SCO"

dfDelays$region[unique(
  grep(paste(Wales,collapse="|"),
       dfDelays$reporting_airport))] <- "WAL"

dfDelays$region[unique(
  grep(paste(BCD,collapse="|"),
       dfDelays$reporting_airport))] <- "BCD"


# AV: Alternatively, I could have created a matrix, and say that if it recognised a value
# it should assign the column name to 'Region'

# AV: Check for missing values in the Region creating a new DF
# with the rows with missing values
dfMissingRegion      <- dfDelays[is.na(dfDelays$region),]
dfMissingRegionNames <- unique(dfMissingRegion$reporting_airport)
```

```
dfMissingRegionNames

# Replace Belfast City (George Best) by BGB
dfDelays$reporting_airport[
  dfDelays$reporting_airport=="BELFAST CITY (GEORGE BEST)"] <- "BGB"

dfDelays$origin_destination[
  dfDelays$origin_destination=="BELFAST CITY (GEORGE BEST)"] <- "BGB"

# Convert airline names to type factor
# sort(unique(dfDelays$airline_name)) # check if names are correct before conversion to
factor
dfDelays$airline_name[
  dfDelays$airline_name ==
    "2 EXCEL AVIATION LTD T/A THE BLADES BROADSWORD SCIMITAR SABRE AND T2"] <-
  "2 EXCEL AVIATION"
dfDelays$airline_name <- as.factor(dfDelays$airline_name)

# Add column with bnames (REGAIR = REGion, AIRport)
# First, we create a data frame, selecting the first 3 characters of the
# airport column and the first 3 of the region column
tmp <- data.frame(matrix(ncol = 3, nrow = 649043)) # MUST READ!! : cambiar nrow cada
                                                    # vez que actualizo base de datos
colnames(tmp) <- c("airport", "region", "REGAIR")
tmp$airport <- substr(dfDelays$reporting_airport, 0, 3)
tmp$region <- substr(dfDelays$region, 0, 3)

# Combine both of them into a 3rd column
tmp$REGAIR <- with(tmp, paste0(region, airport))

# Add it to the original dataframe
dfDelays$REGAIRP <- tmp$REGAIR

rm(dfMissingRegion, dfMissingRegionNames)
# rm(BCD, England_East, England_LDN, England_MidlandsE, England_MidlandsW, England_NE,
#    England_NW, England_SE, England_SW, England_Yorkshire, NIreland,
#    Scotland, Wales)
#-------------------------------------------------------------------------
# Define Competition of routes
#-------------------------------------------------------------------------

# 1. Order routes by competition
# 2. Establish a treshold for 'competitive routes'
# 3. Add labels: 'C' for competitive, 'M' for monopolist, 'N' otherwise

# for the whole dataframe, define route 'ENG. CITY - OTHER CITY'
# OR group by origin, destination and airline

# Create new empty data frame

# dfCompetition <-  data.frame("reporting_airport", "origin_destination_country",
#                              "origin_destination", "airline_name",
#                              "scheduled_charter", "average_delay_mins",
#                              "flights_more_than_31_mins_late_percent",
#                              "region")

# Add column on dfDelays with route name
dfDelays$route <- as.factor(paste(
  dfDelays$reporting_airport,"-",dfDelays$origin_destination))

# Create new temporal dataset with airlines per route
tmp <- dfDelays %>% count(route)
colnames(tmp)[colnames(tmp) == 'n'] <- 'n_airlines'

# Create new temporal dataset grouping avg delay by routes
tmp2 <-
  dfDelays %>%
  group_by(route) %>%
  summarize(avg_delay = mean(average_delay_mins))

# Add in tmp dataset and delete tmp2
tmp$avg_delay <- tmp2$avg_delay
rm(tmp2)
```

```
# Add the number of airlines competing in that route in the original dataset
require(data.table)
setDT(dfDelays)[, count := uniqueN(airline_name), by = c("route", "reporting_yearmonth")]

colnames(dfDelays)[colnames(dfDelays) == 'count'] <- 'n_airlines_route'

# I can either leave it quantitative for the regression analysis or add a label
# to create a dummy variable with 'M', 'C', 'N'
# To determine benchmark for competitiveness, I could do the 80th percentile,
# instead of just having it above 20. For the moment' I'll just cut at 20
dfDelays$Monopoly <- ifelse(dfDelays$n_airlines == 1, "M",
                            ifelse(dfDelays$n_airlines >=20, "C", "N"))

colnames(dfDelays)


#--------------------------------------------------------------------------
# Define Competition of airports (dfDelays)
#--------------------------------------------------------------------------

# Add the number of airlines competing in each airport
setDT(dfDelays)[, count := uniqueN(airline_name), by = c("reporting_airport",
"reporting_yearmonth")]
colnames(dfDelays)[colnames(dfDelays) == 'count'] <- 'n_airlines_airport'


#--------------------------------------------------------------------------
# Define hubs (networks)
#--------------------------------------------------------------------------

# Add the number of connections offered by each airline from a specific airport
#require(data.table)
setDT(dfDelays)[, count := uniqueN(origin_destination),
                 by = c("reporting_airport", "airline_name", "reporting_yearmonth")]

# Change name of column to 'number of connections by that airline from that airport'
colnames(dfDelays)[colnames(dfDelays) == 'count'] <- 'n_connections'


# I want to check if 'n_connections' is >26 for more than 'x' airlines
# I would like to see how many airlines have a certain airport as hub

# 1. add a new column 'Hub', where if 'n_connections' > 26 --> 'H'
# 2. Count how many unique H's for each airline by airport
# 3. Order airprts by number of H's


# 1. Add a 'H' in new column, if 'n_connections' >= 26
dfDelays$Hub <- ifelse(dfDelays$n_connections >= 26, "H", "N")

# Add airline ID if there's a hub, otherwise NA
dfDelays$Hub2 <- ifelse(dfDelays$Hub == "H",  dfDelays$airline_name, "NA")

# 2. Count unique values of Hub2
setDT(dfDelays)[, count := uniqueN(Hub2),
                 by = c("reporting_airport")]

# Change name of column to 'number of airlines that use it as a hub, n_hubs'
colnames(dfDelays)[colnames(dfDelays) == 'count'] <- 'n_hubs'
dfDelays$n_hubs <- as.numeric(dfDelays$n_hubs)

# Substract 1 (to counter the effect of NA's)
dfDelays$n_hubs <- dfDelays$n_hubs - 1

# Create tmp dataset with number of airlines that use each airport as a hub
tmp <-
  as.data.frame(
    dfDelays %>%
      group_by(reporting_airport) %>%
      summarize(n_hubs = mean(n_hubs))
  )

# Plot n-hubs by airport to see the distribution among airports
ggplot(tmp, aes(x=reporting_airport, y=n_hubs)) +
  geom_bar(stat = "identity",fill = "deepskyblue2", col = "deepskyblue2") + #col = outline
color
  xlab("Airport") + ylab("# airlines using it as a hub") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
# ggsave(paste0(dirRslt, "Airports as hubs.pdf"))


# Delete temporal columns from dfDelays
dfDelays$Hub           <- NULL
dfDelays$Hub2          <- NULL
dfDelays$n_connections <- NULL

# Create new Column 'Hub' that reads '1' if more than 5 airlines use it as a hub
dfDelays$Hub <- ifelse(dfDelays$n_hubs >= 5, "1", "0")

# Delete temporal n_hubs column from dfDelays
dfDelays$n_hubs <- NULL


colnames(dfDelays)


#--------------------------------------------------------------------------
# Save the data to an R system file (.Rda)
#--------------------------------------------------------------------------

save(dfDelays, file= paste0(dirData.Rda, "dfDelays.Rda"))


#--------------------------------------------------------------------------

# READ AND CLEANSE DELAYS DATA 2

#--------------------------------------------------------------------------



# Find all files in the delays directory
theFiles2 <- dir(dirData.Delays2)

# Select all files that contain delay data (this is just a check,
# but it also allows one to store other named files in the same
# directory without destroying the code)
theFiles2 <- theFiles2[grep("d\\d{4}_\\d{2}\\.csv", theFiles2)]

# Start a loop over the files in the delays directory to import
# all the data. Function rbind.fill from package plyr is used to bind
# the different files, because the column names differ across files.
# This is something to look into.
dfDelays2 <- data.frame()
for (i in 1:length(theFiles2)) {
  if (i %% 12 == 0) cat("Reading file :", theFiles2[i], "\n")

  dfDelays2 <-
    rbind.fill(
      dfDelays2,
      read.csv(paste0(dirData.Delays2, theFiles2[i]), stringsAsFactors = FALSE)
    )
} # end of the for-loop

# Delete run_date column
dfDelays2$run_date <- NULL

# Handle the 'reporting_period' column. Split into years and
# months. Then, a year month column is made with type Date
# (as in the Challenges data frame)
dfDelays2$reporting_year  <- as.numeric(substr(dfDelays2$reporting_period, 1, 4))
dfDelays2$reporting_month <- as.numeric(substr(dfDelays2$reporting_period, 5, 6))
dfDelays2$reporting_yearmonth <-
  as.Date(paste(dfDelays2$reporting_year, dfDelays2$reporting_month, "01"),
          format="%Y %m %d", tz="GMT")

# AV: check airport names
dfAirportNames <- unique(dfDelays2$reporting_airport)
```

```
# Now, we will add regions
# AV: create new column 'Region' with the region where each airport is located
dfDelays2$region <- NA

dfDelays2$region[unique(
  grep(paste(England_LDN,collapse="|"),
       dfDelays2$reporting_airport))] <- "LDN"

dfDelays2$region[unique(
  grep(paste(England_NE,collapse="|"),
       dfDelays2$reporting_airport))] <- "NE"

dfDelays2$region[unique(
  grep(paste(England_NW,collapse="|"),
       dfDelays2$reporting_airport))] <- "NW"

dfDelays2$region[unique(
  grep(paste(England_MidlandsE,collapse="|"),
       dfDelays2$reporting_airport))] <- "ME"

dfDelays2$region[unique(
  grep(paste(England_MidlandsW,collapse="|"),
       dfDelays2$reporting_airport))] <- "MW"

dfDelays2$region[unique(
  grep(paste(England_East,collapse="|"),
       dfDelays2$reporting_airport))] <- "EAS"

dfDelays2$region[unique(
  grep(paste(England_SE,collapse="|"),
       dfDelays2$reporting_airport))] <- "SE"

dfDelays2$region[unique(
  grep(paste(England_SW,collapse="|"),
       dfDelays2$reporting_airport))] <- "SW"

dfDelays2$region[unique(
  grep(paste(England_Yorkshire,collapse="|"),
       dfDelays2$reporting_airport))] <- "YH"

dfDelays2$region[unique(
  grep(paste(NIreland,collapse="|"),
       dfDelays2$reporting_airport))] <- "NIR"

dfDelays2$region[unique(
  grep(paste(Scotland,collapse="|"),
       dfDelays2$reporting_airport))] <- "SCO"

dfDelays2$region[unique(
  grep(paste(Wales,collapse="|"),
       dfDelays2$reporting_airport))] <- "WAL"

dfDelays2$region[unique(
  grep(paste(BCD,collapse="|"),
       dfDelays2$reporting_airport))] <- "BCD"

# AV: Alternatively, I could have created a matrix, and say that if it recognised a value
# it should assign the column name to 'Region'

# AV: Check for missing values in the Region creating a new DF
# with the rows with missing values
dfMissingRegion      <- dfDelays2[is.na(dfDelays2$region),]
dfMissingRegionNames <- unique(dfMissingRegion$reporting_airport)
dfMissingRegionNames

# Replace Belfast City (George Best) by BGB
dfDelays2$reporting_airport[
  dfDelays2$reporting_airport=="BELFAST CITY (GEORGE BEST)"] <- "BGB"

dfDelays2$origin_destination[
  dfDelays2$origin_destination=="BELFAST CITY (GEORGE BEST)"] <- "BGB"

# Convert airline names to type factor
# sort(unique(dfDelays$airline_name)) # check if names are correct before conversion to
factor
```

```
dfDelays2$airline_name[
  dfDelays2$airline_name ==
    "2 EXCEL AVIATION LTD T/A THE BLADES BROADSWORD SCIMITAR SABRE AND T2"] <-
  "2 EXCEL AVIATION"
dfDelays2$airline_name <- as.factor(dfDelays2$airline_name)

# Add column with bnames (REGAIR = REGion, AIRport)
# First, we create a data frame, selecting the first 3 characters of the
# airport column and the first 3 of the region column
tmp <- data.frame(matrix(ncol = 3, nrow = 148355)) # CAMBIAR numero con el que d√© en el
error
colnames(tmp) <- c("airport", "region", "REGAIR")
tmp$airport <- substr(dfDelays2$reporting_airport, 0, 3)
tmp$region <- substr(dfDelays2$region, 0, 3)

# Combine both of them into a 3rd column
tmp$REGAIR <- with(tmp, paste0(region, airport))

# Add it to the original dataframe
dfDelays2$REGAIRP <- tmp$REGAIR

#--------------------------------------------------------------------------
# Define Competition of routes (dfDelays2)
#--------------------------------------------------------------------------

# 1. Order routes by competition
# 2. Establish a treshold for 'competitive routes'
# 3. Add labels: 'C' for competitive, 'M' for monopolist, 'N' otherwise

# for the whole dataframe, define route 'ENG. CITY - OTHER CITY'
# OR group by origin, destination and airline

# Create new empty data frame

# dfCompetition <-  data.frame("reporting_airport", "origin_destination_country",
#                              "origin_destination", "airline_name",
#                              "scheduled_charter", "average_delay_mins",
#                              "flights_more_than_31_mins_late_percent",
#                              "region")

# Add column on dfDelays with route name
dfDelays2$route <- as.factor(paste(
  dfDelays2$reporting_airport,"-",dfDelays2$origin_destination))

# Create new temporal dataset with airlines per route
tmp <- dfDelays2 %>% count(route)
colnames(tmp)[colnames(tmp) == 'n'] <- 'n_airlines_routes'

# Create new temporal dataset grouping avg delay by routes
tmp2 <-
  dfDelays2 %>%
  group_by(route) %>%
  summarize(avg_delay = mean(average_delay_mins))

# Add in tmp dataset and delete tmp2
tmp$avg_delay <- tmp2$avg_delay
rm(tmp2)

# Create new tmp2 dataset with cancellations, add and delete
tmp2 <-
  dfDelays2 %>%
  group_by(route) %>%
  summarize(cancelled_percent = mean(flights_cancelled_percent))
tmp$cancelled_percent <- tmp2$cancelled_percent
rm(tmp2)

# Add the number of airlines competing in that route in the original dataset
require(data.table)
setDT(dfDelays2)[, count := uniqueN(airline_name), by = c("route", "reporting_yearmonth")]

colnames(dfDelays2)[colnames(dfDelays2) == 'count'] <- 'n_airlines_route'

# I can either leave it quantitative for the regression analysis or add a label
# to create a dummy variable with 'M', 'C', 'N'
# To determine benchmark for competitiveness, I could do the 80th percentile,
# instead of just having it above 20. For the moment' I'll just cut at 20
dfDelays2$Monopoly <- ifelse(dfDelays2$n_airlines == 1, "M",
                            ifelse(dfDelays2$n_airlines >=20, "C", "N"))

colnames(dfDelays2)


#-------------------------------------------------------------------------
# Define Competition of airports (dfDelays2)
#-------------------------------------------------------------------------

# Add the number of airlines competing in each airport
setDT(dfDelays2)[, count := uniqueN(airline_name), by = c("reporting_airport",
"reporting_yearmonth")]
colnames(dfDelays2)[colnames(dfDelays2) == 'count'] <- 'n_airlines_airport'

colnames(dfDelays2)

#-------------------------------------------------------------------------
# Define hubs (networks)
#-------------------------------------------------------------------------

# Add the number of connections offered by each airline from a specific airport
require(data.table)
setDT(dfDelays2)[, count := uniqueN(origin_destination),
                 by = c("reporting_airport", "airline_name", "reporting_yearmonth")]

# Change name of column to 'number of connections by that airline from that airport'
colnames(dfDelays2)[colnames(dfDelays2) == 'count'] <- 'n_connections'


# I want to check if 'n_connections' is >26 for more than 'x' airlines
# I would like to see how many airlines have a certain airport as hub
# 1. add a new column 'Hub', where if 'n_connections' > 26 --> 'H'

# 2. Count how many unique H's for each airline by airport
# 3. Order airprts by number of H's


# 1. Add a 'H' in new column, if 'n_connections' >= 26
dfDelays2$Hub <- ifelse(dfDelays2$n_connections >= 26, "H", "N")

# Add airline ID if there's a hub, otherwise NA
dfDelays2$Hub2 <- ifelse(dfDelays2$Hub == "H",  dfDelays2$airline_name, "NA")

# 2. Count unique values of Hub2
setDT(dfDelays2)[, count := uniqueN(Hub2),
                                    by = c("reporting_airport")]

# Change name of column to 'number of airlines that use it as a hub, n_hubs'
colnames(dfDelays2)[colnames(dfDelays2) == 'count'] <- 'n_hubs'
dfDelays2$n_hubs <- as.numeric(dfDelays2$n_hubs)

# Substract 1 (to counter the effect of NA's)
dfDelays2$n_hubs <- dfDelays2$n_hubs - 1

# Create tmp dataset with number of airlines that use each airport as a hub
tmp <-
  as.data.frame(
    dfDelays2 %>%
      group_by(reporting_airport) %>%
      summarize(n_hubs = mean(n_hubs))
  )

# Plot n-hubs by airport to see the distribution among airports
ggplot(tmp, aes(x=reporting_airport, y=n_hubs)) +
  geom_bar(stat = "identity",fill = "deepskyblue2", col = "deepskyblue2") + #col = outline
color
  xlab("Airport") + ylab("# airlines using it as a hub") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
# ggsave(paste0(dirRslt, "Airports as hubs.pdf"))


# Delete temporal columns from dfDelays2
dfDelays2$Hub           <- NULL
```

```
dfDelays2$Hub2          <- NULL
dfDelays2$n_connections <- NULL

# Create new Column 'Hub' that reads 'Hub' if more than 5 airlines use it as a hub
dfDelays2$Hub <- ifelse(dfDelays2$n_hubs >= 5, "1", "0")

# Delete temporal n_hubs column from dfDelays2
dfDelays2$n_hubs <- NULL


colnames(dfDelays2)


#------------------------------------------------------------------------
# Save the data to an R system file (.Rda)
#------------------------------------------------------------------------

# save(dfDelays2, file= paste0(dirData.Rda, "dfDelays2.Rda"))

#------------------------------------------------------------------------
# AV: Add the values from the new columns to the old ones
# (2018-2020 -> 2000-2017)
#------------------------------------------------------------------------
# Sum the values from 15' early to 15' late to flights_0_to_15_min_late_percent
dfDelays2$flights_0_to_15_minutes_late_percent <-
  dfDelays2$flights_0_to_15_minutes_late_percent +
  dfDelays2$flights_more_than_15_minutes_early_percent +
  dfDelays2$flights_15_minutes_early_to_1_minute_early_percent

# Delete columns
dfDelays2$flights_more_than_15_minutes_early_percent <- NULL
dfDelays2$flights_15_minutes_early_to_1_minute_early_percent <- NULL

# Sum the values from 61-120 to 121-180
dfDelays2$flights_between_121_and_180_minutes_late_percent <-
  dfDelays2$flights_between_121_and_180_minutes_late_percent +
  dfDelays2$flights_between_61_and_120_minutes_late_percent

# Delete columns
dfDelays2$flights_between_61_and_120_minutes_late_percent <- NULL
dfDelays2$flights_unmatched_percent <- NULL
dfDelays2$flights_cancelled_percent <- NULL
dfDelays2$number_flights_cancelled <- NULL

dfDelays$planned_flights_unmatched <- NULL


# Rename columns
names <- c(colnames(dfDelays))
colnames(dfDelays2) <- names

# Delete rows from 2018-2020 from dfDelays
dfDelays <- dfDelays[!(dfDelays$reporting_year==2018),]
dfDelays <- dfDelays[!(dfDelays$reporting_year==2019),]
dfDelays <- dfDelays[!(dfDelays$reporting_year==2020),]

# Use rbind to merge both datasets
dfDelays <-
  rbind.fill(dfDelays, dfDelays2)


#------------------------------------------------------------------------
# Save the data to an R system file (.Rda)
#------------------------------------------------------------------------

# save(dfDelays, file= paste0(dirData.Rda, "dfDelays.Rda"))

#------------------------------------------------------------------------
# Delete unnecessary data from the Global Environment
#------------------------------------------------------------------------

rm(dfAirportNames, dfMissingRegion, dfMissingRegionNames)
rm(BCD, England_East, England_LDN, England_MidlandsE, England_MidlandsW, England_NE,
   England_NW, England_SE, England_SW, England_Yorkshire, NIreland,
   Scotland, Wales)
```

```
rm(tmp, i)
rm(names)




#--------------------------------------------------------------------------
# NAME  : ReadAndPrepare01 - Cancellations.R
# TASK  : Read data about cancellations from multiple csv files
#--------------------------------------------------------------------------

# Empty the memory
remove(list=ls())
cat("\f")

#--------------------------------------------------------------------------
# Install and load packages
#--------------------------------------------------------------------------

# Two packages are used in this section: plyr to calculate sub-group
# summaries, and ggplot2 to make plots

# install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)

# install.packages("plyr", dependencies = TRUE)
library(plyr)

#--------------------------------------------------------------------------
# Set path and library
#--------------------------------------------------------------------------

# Paths are set according the logic taught in class. Names have been
# taken from the files in the shared Zip file. The name of the covid
# directory has been changed to Covid19 to get rid of the dash in the
# original name. The Rda subfolder has been added to store the
# processed data.
# dir  <- "~/Werk/Education/Other/QuestionAlejandro/" # Thuis

dir  <- "~/Offline Documents/RSM's MScBA/Thesis/QuestionAlejandro/" # Thuis

dirProg     <- paste0(dir, "Programs/")
dirData     <- paste0(dir, "Data/")
dirRslt     <- paste0(dir, "Results/")

dirData.Cancellations <- paste0(dirData, "Cancellations/")
dirData.Covid19       <- paste0(dirData, "Covid19/")
dirData.Delays        <- paste0(dirData, "Delays/")
dirData.Delays2       <- paste0(dirData, "Delays2/")
dirData.Rda           <- paste0(dirData, "Rda/")

#--------------------------------------------------------------------------
# Read cancellation data
#--------------------------------------------------------------------------

# Find all files in the cancellations directory
theFiles <- dir(dirData.Cancellations)

# Select all files that contain cancellation data (this is just a check,
# but it also allows one to store other named files in the same
# directory without destroying the code)
theFiles <- theFiles[grep("c\\d{4}_\\d{2}\\.csv", theFiles)]

# Start a loop over the files in the Cancellations directory to import
# all the data. Function rbind.fill from package plyr is used to bind
# the different files, because the column names differ across files.
# This is something to look into.
dfCancellations <- data.frame()
for (i in 1:length(theFiles)) {
  if (i %% 12 == 0) cat("Reading file :", theFiles[i], "\n")

  dfCancellations <-
    rbind.fill(
      dfCancellations,
      read.csv(paste0(dirData.Cancellations, theFiles[i]), stringsAsFactors = FALSE)
    )
  } # end of the for-loop


#-------------------------------------------------------------------------
# After care
#-------------------------------------------------------------------------

# Change the character rundate information to POSIXct, and a
# regular date variable. The yearmon variable indicates the
# year-month anchored at the first day of the month
dfCancellations$rundate <- strptime(dfCancellations$rundate,
                                    format = "%d/%m/%Y %H:%M", tz="GMT")
dfCancellations$date    <- as.Date(dfCancellations$rundate)
dfCancellations$yearmon <- as.Date(paste0(format(dfCancellations$rundate, "%Y-%m"),"-01"))

# Delete rundate column
dfCancellations[1] <- list(NULL)

# Handle the 'reporting_period' column. Split into years and
# months. Then, a yearmonth column is made with type Date
# (as in the Challenges data frame)
dfCancellations$reporting_year  <- as.numeric(substr(dfCancellations$reporting_period, 1,
4))
dfCancellations$reporting_month <- as.numeric(substr(dfCancellations$reporting_period, 5,
6))
dfCancellations$reporting_yearmonth <-
  as.Date(paste(dfCancellations$reporting_year, dfCancellations$reporting_month, "01"),
          format="%Y %m %d", tz="GMT")


# Most counts are stored as character, which is fine but inconvenient.
# It may be caused by the thousands separator, but also by incidental
# dashes. The three 'total' variables will be converted to numeric after
# removing the comma. Anuthing non-numeric will be assigned NA.
dfCancellations$total_atms <-
  as.numeric(sub(",", "", dfCancellations$total_atms))

dfCancellations$total_cancelled_atms <-
  as.numeric(sub(",", "", dfCancellations$total_cancelled_atms))

dfCancellations$total_atms_excl_cancelled <-
  as.numeric(sub(",", "", dfCancellations$total_atms_excl_cancelled))

# JvD: you may want to set the missing values equal to zero.

# AV: Changing missing values to 0.
dfCancellations$total_atms[
  is.na(dfCancellations$total_atms)] <- "0"

dfCancellations$total_cancelled_atms[
  is.na(dfCancellations$total_cancelled_atms)] <- "0"

dfCancellations$total_atms_excl_cancelled[
  is.na(dfCancellations$total_atms_excl_cancelled)] <- "0"

# AV: create new column 'Region' with the region where each airport is located
dfCancellations$region <- NA

# AV: Airports in London and BCD already indicated in Airport Group (Except for Biggin)
dfCancellations$region[grep("London Area Airports",
                            dfCancellations$reporting_airport_group_name
                            )] <- "England: Greater London Area"
dfCancellations$region[grep("Non UK Reporting Airports",
                            dfCancellations$reporting_airport_group_name
                            )] <- "British Crown Dependencies"

#-------------------------------------------------------------------------
# Regions
#-------------------------------------------------------------------------

# AV: Biggin Hill
dfCancellations$region[grep("BIGGIN HILL", dfCancellations$reporting_airport_name
                            )] <- "England: Greater London Area"
```

```
# AV: For the rest, I am going to create a vector for each region
England_NE        <- c("NEWCASTLE", "DURHAM TEES VALLEY",
                       "TEESSIDE INTERNATIONAL AIRPORT")
England_NW        <- c("BLACKPOOL", "LIVERPOOL", "MANCHESTER")
England_MidlandsE <- c("EAST MIDLANDS INTERNATIONAL")
England_MidlandsW <- c("BIRMINGHAM")
England_East      <- c("NORWICH")
England_SE        <- c("LYDD", "SHOREHAM", "SOUTHAMPTON", "OXFORD")
England_SW        <- c("BOURNEMOUTH", "BRISTOL", "EXETER",
                       "ISLES OF SCILLY", "LANDS END", "NEWQUAY",
                       "GLOUCESTERSHIRE")
England_Yorkshire <- c("DONCASTER SHEFFIELD", "HUMBERSIDE", "LEEDS BRADFORD")
NIreland          <- c("BELFAST CITY",
                       "BELFAST INTERNATIONAL", "CITY OF DERRY")
Scotland          <- c("ABERDEEN", "BARRA", "BENBECULA", "CAMPBELTOWN", "DUNDEE",
                       "EDINBURGH", "GLASGOW", "PRESTWICK", "INVERNESS", "ISLAY",
                       "KIRKWALL", "LERWICK", "SCATSTA", "SUMBURGH",
                       "STORNOWAY", "TIREE", "WICK JOHN O GROATS")
Wales             <- c("CARDIFF WALES")

# AV: Now, we do the same as with Biggin, but with the vectors
dfCancellations$region[unique(
  grep(paste(England_NE,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: North East"

dfCancellations$region[unique(
  grep(paste(England_NW,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: North West"

dfCancellations$region[unique(
  grep(paste(England_MidlandsE,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: Midlands (East)"

dfCancellations$region[unique(
  grep(paste(England_MidlandsW,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: Midlands (West)"

dfCancellations$region[unique(
  grep(paste(England_East,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: East"

dfCancellations$region[unique(
  grep(paste(England_SE,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: South East"

dfCancellations$region[unique(
  grep(paste(England_SW,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: South West"

dfCancellations$region[unique(
  grep(paste(England_Yorkshire,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "England: Yorkshire and the Humber"

dfCancellations$region[unique(
  grep(paste(NIreland,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "Northern Ireland"

dfCancellations$region[unique(
  grep(paste(Scotland,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "Scotland"

dfCancellations$region[unique(
  grep(paste(Wales,collapse="|"),
       dfCancellations$reporting_airport_name))] <- "Wales"

# AV: Alternatively, I could have created a matrix, and say that if it recognised a value
# it should assign the column name to 'Region'

# AV: Check for missing values in the Region creating a new DF
# with the rows with missing values
dfMissingRegion <- dfCancellations[is.na(dfCancellations$region),]
dfMissingRegion$reporting_airport_name
unique(dfMissingRegion$reporting_airport_name)
dfMissingRegionNames <- unique(dfMissingRegion$reporting_airport_name)
```

```
# At first, I got errors for the names that included a parenthesis - like
# 'Liverpool (John Lennon)'. After trying many solutions, I found that I just had
# to remove all the parentheses

#-------------------------------------------------------------------------
# Add a column for percent cancellations
#-------------------------------------------------------------------------

# Make columns numeric
dfCancellations$total_cancelled_atms <- as.numeric(dfCancellations$total_cancelled_atms)
dfCancellations$total_atms <- as.numeric(dfCancellations$total_atms)

# Define a column for relative cancellations (cancelled/total_atms)
dfCancellations$cancelled_percent <-
  dfCancellations$total_cancelled_atms/dfCancellations$total_atms*100

#-------------------------------------------------------------------------
# Save the data to an R system file (.Rda)
#-------------------------------------------------------------------------

save(dfCancellations, file= paste0(dirData.Rda, "dfCancellations.Rda"))





#-------------------------------------------------------------------------
# NAME  : RAnalysis01 - Delays2.R
# TASK  : Visualize and analyze data from Delays2
#-------------------------------------------------------------------------

# Empty the memory
remove(list=ls())
cat("\f")

#-------------------------------------------------------------------------
# Install and load packages
#-------------------------------------------------------------------------

# Two packages are used in this section: plyr to calculate sub-group
# summaries, and ggplot2 to make plots

# install.packages("ggplot2", dependencies = TRUE)
library(ggplot2)

# install.packages("plyr", dependencies = TRUE)
library(plyr)

# install.packages("dplyr", dependencies = TRUE)
library(dplyr)

library(stargazer)

# install.packages("plm", dependencies = TRUE)
library(plm)

# Load packages for Regression Tree
library(rpart)  # Build the trees
library(rpart.plot)  # Plot the trees

# install package for Neural Networks
# install.packages("forecast", dependencies = TRUE)
library(forecast)

# Install (once) and load package ROCR
# install.packages("ROCR", dependencies = TRUE)
library(ROCR)

# Install and load random forest package
# install.packages("randomForest", dependencies = TRUE)
library(randomForest)

# Install and load GBM package
# install.packages("gbm", dependencies = TRUE)
library(gbm)

# Install and load data.table package
# install.packages("data.table", dependencies = TRUE)
library(data.table)

library(xtable)
library(rpart)
library(rpart.plot)
library(psych)
library(forecast)

# Install and load the lm.beta package
# install.packages("lm.beta", dependencies = TRUE)
# library(lm.beta)

# Install and load the GGally package
# install.packages("GGally", dependencies = TRUE)
library(GGally)

#-------------------------------------------------------------------------
# Set path and library
#-------------------------------------------------------------------------

# Paths are set according the logic taught in class. Names have been
# taken from the files in the shared Zip file. The name of the covid
# directory has been changed to Covid19 to get rid of the dash in the
# original name. The Rda subfolder has been added to store the
# processed data.
# dir  <- "~/Werk/Education/Other/QuestionAlejandro/" # Thuis
dir  <- "~/Offline Documents/RSM's MScBA/Thesis/QuestionAlejandro/" # Thuis

dirProg     <- paste0(dir, "Programs/")
dirData     <- paste0(dir, "Data/")
dirRslt     <- paste0(dir, "Results/")

dirData.Cancellations <- paste0(dirData, "Cancellations/")
dirData.Covid19       <- paste0(dirData, "Covid19/")
dirData.Delays        <- paste0(dirData, "Delays/")
dirData.Delays2       <- paste0(dirData, "Delays2/")
dirData.Rda           <- paste0(dirData, "Rda/")

#-------------------------------------------------------------------------
# Read Delays2 data
#-------------------------------------------------------------------------

# Load the data
#load(file= paste0(dirData.Rda, "dfDelays.Rda"))
#dfDelaysPast <- dfDelays

#load(file= paste0(dirData.Rda, "dfDelays2.Rda"))
load(file= paste0(dirData.Rda, "dfCancellations.Rda"))

#dfDelays <- dfDelays2
#rm(dfDelays2)

# Read dfDC
load(file= paste0(dirData.Rda, "dfDC.Rda"))
dfDC$cancelled_percent <- as.numeric(dfDC$cancelled_percent)
dfDC$cancelled_percent <- ifelse(is.na(dfDC$cancelled_percent), 0, dfDC$cancelled_percent)

# Define a date variable signifying the month of the cancellations,
# which is identified as the first of the month
# AV: I changed it to reporting_yearmonth
# dfDelays$reporting_yearmonth <-
#   as.Date(paste0(format(dfDelays$reporting_yearmonth, "%Y-%m"), "-01"))

# dfDelaysPast$reporting_yearmonth <-
#   as.Date(paste0(format(dfDelaysPast$reporting_yearmonth, "%Y-%m"), "-01"))

# Check missing values (none)
# colSums(is.na(dfDelays))

# Identify complete cases
# dfDelays[complete.cases(dfDelays),]

# For dfDelaysPast, delete all observations before 2017
# dfDelaysPast <- dfDelaysPast %>% filter(reporting_yearmonth >= "2017-01-01")

#--------------------------------------------------------------------------
# 2. Time series
#--------------------------------------------------------------------------

#-----
# Create subset and plot
#-----

dfGroup <-
  ddply(dfDC, .(reporting_yearmonth), summarise,
        cancelled_percent = mean(as.numeric(cancelled_percent), na.rm = TRUE))

# Make a plot of the monthly cancellations (2)
ggplot(dfGroup, aes(x=reporting_yearmonth, y=cancelled_percent)) +
  geom_line() +
  xlab("Month of the year") + ylab("Cancellations per total movements (%)") +
  scale_x_date(breaks = "3 months", date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

# Create dataset by grouping by yearmonth
dfGroupYM <-
  as.data.frame(
    dfDC %>%
      group_by(reporting_yearmonth) %>%
      summarize(average_delay_mins                = mean(average_delay_mins),
                early_to_15_mins_late_percent     = mean(early_to_15_mins_late_percent),
                flts_16_to_30_mins_late_percent   = mean(flts_16_to_30_mins_late_percent),
                flts_31_to_60_mins_late_percent   = mean(flts_31_to_60_mins_late_percent),
                flts_61_to_180_mins_late_percent  = mean(flts_61_to_180_mins_late_percent),
                flts_181_to_360_mins_late_percent =
mean(flts_181_to_360_mins_late_percent),
                more_than_360_mins_late_percent   = mean(more_than_360_mins_late_percent),
                cancelled_percent                 = mean(cancelled_percent)      )
  )

# Change outlier in 2018-3
dfColumn <- dfGroupYM$cancelled_percent
dfColumn[15] = 3.78
dfGroupYM$cancelled_percent <- dfColumn

ggplot(dfGroupYM, aes(x=reporting_yearmonth, y=cancelled_percent)) +
  geom_line() +
  xlab("Month of the year") + ylab("Cancellations per total movements (%)") +
  scale_x_date(breaks = "3 months", date_labels = "%Y-%m") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))

# Create univariate dataset with all cancellations until May
dsCan <-
  ddply(dfCancellations, .(reporting_yearmonth), summarise,
        cancelled_percent = mean(as.numeric(cancelled_percent), na.rm = TRUE))



# Create new dataset with only one airport's time series (Gatwick)
# sub <- colnames(dfDC)
# columnNames <- sub[-1:-2]

# dsSub <- dfDC[dfDC$reporting_airport == "MANCHESTER", columnNames]
# dsSub$cancelled_percent <- as.numeric(dsSub$cancelled_percent)

dsSub <- dfGroupYM[-1]
dsCan <- dsCan[-1]
dsCan <- dsCan[-c(55),]

# Convert all columns to ts
dsSubts <- ts(dsSub, frequency=12, start=c(2017,1))
dsCants <- ts(dsCan, frequency=12, start=c(2017,1))

# Same but only Auto-regressive
dsSubA <- dsSub$cancelled_percent
```

```
dsSubtsA <- ts(dsSubA, frequency=12, start=c(2017,1))
plot.ts(dsSubtsA)


#-----
# Time series analysis
#-----

# Plot
autoplot(dsCants) +
  ggtitle("Flights cancelled in the UK") +
  ylab("%") +
  xlab("Year")

autoplot(dsCants, dsSubts[,c("average_delay_mins")]) +
  xlab("Year") + ylab("") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))


# Seasonal plots ---
# Cancellations
ggseasonplot(dsSubtsA, year.labels=TRUE, year.labels.left=TRUE) +
  ylab("%") +
  ggtitle("Seasonal plot: Flights cancelled in the UK")
# Delays
ggseasonplot(dsSubts[,c("average_delay_mins")], year.labels=TRUE, year.labels.left=TRUE) +
  ylab("min") +
  ggtitle("Seasonal plot: Flights Delayed in the UK")

# Comparison Delay and cancellations
autoplot(dsSubts[,c("average_delay_mins","cancelled_percent")], facets=TRUE) +
  xlab("Years") + ylab("") +
  ggtitle("Average delays (mins) and Cancellations (%) in the UK")

window(dsSubts[,c("average_delay_mins","cancelled_percent")], end =c(2021,06))

autoplot(window(dsSubts[,c("average_delay_mins","cancelled_percent")],
                start =c(2019,06), end =c(2021,06)), facets=TRUE) +
  xlab("Years") + ylab("") +
  ggtitle("Average delays (mins) and Cancellations (%) in the UK")

#then make a plot
(plot1 <- ggplot(dsSubts, aes(x=c("average_delay_mins","cancelled_percent"), y=Temp)) +
    geom_point(col="gray",shape=1) +
    theme_bw(24) +
    ggtitle("Stream Name") +
    ylab("Temperature (°C)") +
    xlab("Month"))

# Now set the breaks, labels and x-limits as you please
(plot1 <- plot1 +
    scale_x_datetime(breaks = "1 month",
                     labels=date_format("%b"),
                     limits = as.POSIXct(c("2015-03-30","2015-11-25"),
                                         timezone="CEST")))

# Comparison among different predictors
# First select only some variables (leave out 16-30 and 31-60)
cols <- c(1,8)
autoplot(dsSubts[,cols], facets=TRUE) +
  ylab("Flight delay and cancellations in the UK")

# Comparison among different predictors
GGally::ggpairs(as.data.frame(dsSubts[,cols]))

# Classical decomposition
dsSubtsA %>% decompose(type="multiplicative") %>%
  autoplot() + xlab("Year") +
  ggtitle("Classical decomposition of flight cancellations in the UK")


# STL decomposition
dsSubtsA %>%
  stl(t.window=13, s.window="periodic", robust=TRUE) %>%
  autoplot() +
  ggtitle("STL decomposition of flight cancellations in the UK")

# STL decomposition delays
dsSubts[,c("average_delay_mins")] %>%
  stl(t.window=13, s.window="periodic", robust=TRUE) %>%
  autoplot() +
  ggtitle("STL decomposition of flight cancellations in the UK")


# Unit root tests to see if it needs differentiating
library(urca)
dsSubtsA %>% ur.kpss() %>% summary()

#> KPSS Unit Root Test
#> Value of test-statistic is: 0.2013
# This means that H0 (stationarity) is accepted, and the data are stationary
# There is no need to differentiate

# Actually, runing ndiffs(), which tells you how many differentatings are needed,
# it says that we need 0
ndiffs(dsSubts[,c("average_delay_mins")])

#-----
# 8.5. Non-seasonal ARIMA
#-----

# Use auto.arima() to get the values of (p,d,q)
# Select a model automatically
fit <- auto.arima(dsSubts[,"cancelled_percent"], seasonal=FALSE)
fit
#> Series: dsSubts[, "cancelled_percent"]
#> ARIMA(1,0,0) with non-zero mean

#> Coefficients:
#>          ar1    mean
#>       0.8458  1.8925
#> s.e.  0.1381  0.8951

#> sigma^2 estimated as 0.7871:  log likelihood=-50.27
#> AIC=106.55   AICc=107.23   BIC=111.54

# That means that the model is: yt = c + 0.8458Et-1 + Et,
# where c = 1.8925, and Et is white noise with a standard deviation
# of sqrt(0.7871) = 0.8872

# IMPORTANT: as c ≠ 0 & d = 0, the long-term forecasts will go
# to the mean of the data.
# For d = 0, the long-term forecast standard deviation will go to
# the standard deviation of the historical data, so the prediction
# intervals will all be essentially the same.

# Plot forecast
fit %>% forecast(h=10) %>% autoplot(include=80) +
  ylab("Cancelled flights (%)") +
  xlab("Year")

# ACF plot
ggAcf(dsSubts[,"cancelled_percent"])
ggAcf(dsCants)

# PACF plot
ggPacf(dsSubts[,"cancelled_percent"])
ggPacf(dsCants)

# Both plots
dsSubtsA %>%  ggtsdisplay()

# Try models of different p, q
fit1 <- Arima(dsSubtsA, order=c(2,0,0))
fit2 <- Arima(dsSubtsA, order=c(1,0,1))
fit3 <- auto.arima(dsCants, seasonal=TRUE,
                   stepwise=FALSE, approximation=FALSE)
fit3 <- Arima(dsCants, order=c(2,1,0))
```

```
# Forecast and plot
fit2 %>% forecast(h=10) %>% autoplot(include=80) +
  ylab("Cancelled flights (%)") +
  xlab("Year")

fit3 %>% forecast(h=10) %>% autoplot(include=80) +
  ylab("Cancelled flights (%)") +
  xlab("Year")

#-------
# ARIMAX
#-------

# Create vector with exogenouos variables' names
covariates <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                     "flts_16_to_30_mins_late_percent",
                     "flts_31_to_60_mins_late_percent",
                     "flts_61_to_180_mins_late_percent",
                     "flts_181_to_360_mins_late_percent",
                     "more_than_360_mins_late_percent"
                     )
# Which series are stationary
library(urca)
dsSubts[,covariates] %>% ur.kpss() %>% summary()
ndiffs(dsSubts[,"average_delay_mins"])

dsSubts[,"average_delay_mins"] %>% diff() %>% ggtsdisplay()

# Fit the ARIMAX model
model <- auto.arima(dsSubts[,"cancelled_percent"],
               xreg = dsSubts[,covariates])
model

arimax.fit <- Arima(dsSubts[,"cancelled_percent"],
                  order=c(2,0,0), seasonal =list(order = c(1, 0, 0), period = 12),
                  xreg=dsSubts[,covariates])

# Check errors and residuals
cbind("Regression Errors" = residuals(model, type="regression"),
      "ARIMA errors" = residuals(model, type="innovation")) %>%
  autoplot(facets=TRUE)
checkresiduals(arimax.fit)


# FORECAST with ARIMAX ----
fcast <- forecast(model, xreg=dsSubts[,covariates])
fcastx <- forecast(arimax.fit, xreg=dsSubts[,covariates], h=6)
autoplot(fcastx) + xlab("Year") +
  ylab("Percentage change")

#-------------------
# CROSS-VALIDATION
#-------------------

# Fit an AR(1) model to each rolling origin subset
# far2 <- function(x, h){forecast(Arima(x, order=c(1,0,0)), h=h)}
# e <- tsCV(dsSubts[,"cancelled_percent"], far2, h=1)
# sqrt(sqrt(e^2))

#Fit an ARX(1) model to each rolling origin subset
# fc <- function(y, h, xreg, xreg_ncol)
# {
#   X <- matrix(xreg[1:length(y), ], ncol = xreg_ncol)
#   if(NROW(xreg) < length(y) + h)
#     stop("Not enough xreg data for forecasting")
#   newX <- matrix(xreg[length(y) + (1:h), ], ncol = xreg_ncol)
#   fit <- auto.arima(y, xreg=X)
#   forecast(fit, xreg = newX)
# }

# tsCV(dsSubts[,"cancelled_percent"], fc,
#      xreg=dsSubts[,"average_delay_mins"], xreg_ncol=1)
```

```
#-------------------
# FINAL PART
#-------------------

# BORRADOR: he encontrado esta línea de código para hacer xval pero la dejo
# por si acaso
# modelcv <- CVar(dsSubtsA, k=5, lambda=0.15)
# OTRO BORRADOR: fcast <- forecast(fit, xreg = test[, covariates])

# Split the data
train <- window(dsSubts, end = c(2020,2))
test  <- window(dsSubts, start = c(2020,3), end = c(2021,06))

# Fit models (1st define xreg)
covariates <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                 "flts_16_to_30_mins_late_percent",
                 "flts_31_to_60_mins_late_percent",
                 "flts_61_to_180_mins_late_percent",
                 "flts_181_to_360_mins_late_percent",
                 "more_than_360_mins_late_percent"
                 )

# Train baseline model (arima.fit) and arimax model on train set
arima.fit  <- Arima(train[,"cancelled_percent"], order=c(2,0,0))
arimax.fit <- Arima(train[,"cancelled_percent"],
                    order=c(2,0,0),
                    seasonal =list(order = c(1, 0, 0), period = 12),
                    xreg = train[, covariates])

arima.fit
arimax.fit

# Test baseline model on test data
arima.test <- Arima(test[,"cancelled_percent"], model=arima.fit)
#accuracy(arima.test)
residuals.baseline <- checkresiduals(arima.fit)
fcast <- forecast(arima.fit, h=14)
autoplot(fcast + xlab("Year") +
  ylab("Percentage change"))
#accuracy <- accuracy(fcast, test[,"cancelled_percent"])


# Test arimax model on test data
arimax.test <- Arima(test[,"cancelled_percent"],
                     model=arimax.fit, xreg = test[, covariates])

fcastx <- forecast(arimax.fit, xreg = test[, covariates])
autoplot(fcastx) + xlab("Year") +
  ylab("Percentage change")

# PLOT BOTH FORECAST AGAINST ACTUAL VALUES
autoplot(window(dsSubts[,"cancelled_percent"], end =c(2021,06))) +
  autolayer(fcast, series="ARIMA", PI=FALSE) +
  autolayer(fcastx, series="ARIMAX", PI=FALSE) +
  xlab("Year") + ylab("%") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))


#accuracy.x <- accuracy(fcastx, test[,"cancelled_percent"], )

# Fitted values and residuals for arima
residuals.baseline <- residuals(arima.test)

# Fitted values and residuals for arimax (2 ways)
# Way 1:
fitted.values          <- fitted(arimax.test)
fitted.values.baseline <- fitted(arima.test)
# residuals     <- (test - fitted.values)
# averageAE     <- mean(residuals)

# Way 2 (simpler):
residuals <- residuals(arimax.test)
```

```
# Perform DM test
dm.test(residuals.baseline, residuals, alternative = "greater", h=1, power=2)

# Compare Accuracies
accuracy(arima.test)
accuracy(arimax.test)

# COMPUTE MASE's
# Create function computeMASE
computeMASE <- function(forecast,train,test,period){

  # forecast - forecasted values
  # train - data used for forecasting .. used to find scaling factor
  # test - actual data used for finding MASE.. same length as forecast
  # period - in case of seasonal data.. if not, use 1

  forecast <- as.vector(forecast)
  train <- as.vector(train)
  test <- as.vector(test)

  n <- length(train)
  scalingFactor <- sum(abs(train[(period+1):n] - train[1:(n-period)])) / (n-period)

  et <- abs(test-forecast)
  qt <- et/scalingFactor
  meanMASE <- mean(qt)
  return(meanMASE)
}

# MASE ARIMAX MODEL
MASE.arima <-
  computeMASE(forecast = fitted.values.baseline,
              train = train,
              test = test,
              period = 1)

# MASE ARIMAX MODEL
MASE.arimax <-
  computeMASE(forecast = fitted.values,
              train = train,
              test = test,
              period = 1)
# Compare Accuracies

round(accuracy(arima.test),3)
round(accuracy(arimax.test),3)
MASE.arima
MASE.arimax


# RESULTS:

#> MASE.arima
#[1] 5.673
#> MASE.arimax
#[1] 5.666

#> accuracy(arima.test)
#                  ME  RMSE   MAE   MPE   MAPE MASE  ACF1
# Training set -0.092 0.356 0.332 -18.541 38.44  NaN -0.617
#> accuracy(arimax.test)
#                  ME  RMSE   MAE   MPE   MAPE MASE  ACF1
# Training set -0.094 0.278 0.234 -12.555 25.201  NaN -0.761


# THEREFORE, THE END RESULTS ARE:

#> Accuracy(arima.test)
#                  ME  RMSE   MAE     MPE  MAPE   MASE   ACF1
# Training set -0.092 0.356 0.332 -18.541 38.44  5.673 -0.617

#> Accuracy(arimax.test)
#                  ME  RMSE   MAE     MPE   MAPE   MASE   ACF1
# Training set -0.094 0.278 0.234 -12.555 25.201  5.666 -0.761
```

```
#--------------------
# CROSS VALIDATION FOR REGIONAL LEVEL
#--------------------

# Create dataset by grouping by yearmonth
dfGroupYM <-
  as.data.frame(
    dfDC %>%
      group_by(reporting_yearmonth, region) %>%
      summarize(average_delay_mins                = mean(average_delay_mins),
                early_to_15_mins_late_percent     = mean(early_to_15_mins_late_percent),
                flts_16_to_30_mins_late_percent   = mean(flts_16_to_30_mins_late_percent),
                flts_31_to_60_mins_late_percent   = mean(flts_31_to_60_mins_late_percent),
                flts_61_to_180_mins_late_percent  = mean(flts_61_to_180_mins_late_percent),
                flts_181_to_360_mins_late_percent =
mean(flts_181_to_360_mins_late_percent),
                more_than_360_mins_late_percent   = mean(more_than_360_mins_late_percent),
                cancelled_percent                 = mean(cancelled_percent)      )
  )

# Create dataset with only London
sub <- colnames(dfGroupYM)
columnNames <- sub[-1:-2]

dsLDN <- dfGroupYM[dfGroupYM$region == "LDN", columnNames]
dsLDN$cancelled_percent <- as.numeric(dsLDN$cancelled_percent)

# delete first column because it only says LDN
# dsLDN <- dsLDN[-1]

# Convert all columns to ts
dsLDNts <- ts(dsLDN, frequency=12, start=c(2017,1))

# Split the data
train <- window(dsLDNts, end = c(2019,6))
test  <- window(dsLDNts, start = c(2019,7), end = c(2019,12))


# Train baseline model (arima.fit) and arimax model on train set
arima.fit  <- Arima(train[,"cancelled_percent"], order=c(1,0,0))
arimax.fit <- Arima(train[,"cancelled_percent"],
                    order=c(2,0,0),
                    seasonal =list(order = c(1, 0, 0), period = 12),
                    xreg = train[, covariates])

# Test baseline model on test data
arima.test <- Arima(test[,"cancelled_percent"], model=arima.fit)
fcast <- forecast(arima.fit, h=6)

# Test arimax model on test data
arimax.test <- Arima(test[,"cancelled_percent"],
                     model=arimax.fit, xreg = test[, covariates])
fcastx <- forecast(arimax.fit, xreg = test[, covariates])

autoplot(window(dsLDNts[,"cancelled_percent"], end =c(2019,12))) +
  autolayer(fcast, series="ARIMA", PI=FALSE) +
  autolayer(fcastx, series="ARIMAX", PI=FALSE) +
  xlab("Year") + ylab("%") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))

# Fitted values and residuals for arimax (2 ways)
# Way 1:
fitted.values          <- fitted(arimax.test)
fitted.values.baseline <- fitted(arima.test)
# residuals     <- (test - fitted.values)
# averageAE     <- mean(residuals)

# Way 2 (simpler):
# residuals <- residuals(arimax.test)

# Compare Accuracies
accuracy(arima.test)
accuracy(arimax.test)
```

```
# COMPUTE MASE's
# Create function computeMASE
computeMASE <- function(forecast,train,test,period){

  # forecast - forecasted values
  # train - data used for forecasting .. used to find scaling factor
  # test - actual data used for finding MASE.. same length as forecast
  # period - in case of seasonal data.. if not, use 1

  forecast <- as.vector(forecast)
  train <- as.vector(train)
  test <- as.vector(test)

  n <- length(train)
  scalingFactor <- sum(abs(train[(period+1):n] - train[1:(n-period)])) / (n-period)

  et <- abs(test-forecast)
  qt <- et/scalingFactor
  meanMASE <- mean(qt)
  return(meanMASE)
}

# MASE ARIMAX MODEL
MASE.arima <-
  computeMASE(forecast = fitted.values.baseline,
              train = train,
              test = test,
              period = 1)

# MASE ARIMAX MODEL
MASE.arimax <-
  computeMASE(forecast = fitted.values,
              train = train,
              test = test,
              period = 1)
# Compare Accuracies

round(accuracy(arima.test),3)
round(accuracy(arimax.test),3)
MASE.arima
MASE.arimax




#-------------------
# CROSS VALIDATION FOR AIRPORT LEVEL
#-------------------

# Create dataset by grouping by yearmonth
dfGroupYM <-
  as.data.frame(
    dfDC %>%
      group_by(reporting_yearmonth, reporting_airport) %>%
      summarize(average_delay_mins                = mean(average_delay_mins),
                early_to_15_mins_late_percent     = mean(early_to_15_mins_late_percent),
                flts_16_to_30_mins_late_percent   = mean(flts_16_to_30_mins_late_percent),
                flts_31_to_60_mins_late_percent   = mean(flts_31_to_60_mins_late_percent),
                flts_61_to_180_mins_late_percent  = mean(flts_61_to_180_mins_late_percent),
                flts_181_to_360_mins_late_percent =
mean(flts_181_to_360_mins_late_percent),
                more_than_360_mins_late_percent   = mean(more_than_360_mins_late_percent),
                cancelled_percent                 = mean(cancelled_percent)       )
  )

# Create dataset with only London
sub <- colnames(dfGroupYM)
columnNames <- sub[-1:-2]

dsMAN <- dfGroupYM[dfGroupYM$reporting_airport == "MANCHESTER", columnNames]
dsMAN$cancelled_percent <- as.numeric(dsMAN$cancelled_percent)

# Convert all columns to ts
dsMANts <- ts(dsMAN, frequency=12, start=c(2017,1))
```

```
# Split the data
train <- window(dsMANts, end = c(2019,6))
test  <- window(dsMANts, start = c(2019,7), end = c(2019,12))

# Train baseline model (arima.fit) and arimax model on train set
arima.fit  <- Arima(train[,"cancelled_percent"], order=c(1,0,0))
arimax.fit <- Arima(train[,"cancelled_percent"],
                    order=c(2,0,0),
                    seasonal =list(order = c(1, 0, 0), period = 12),
                    xreg = train[, covariates])

# Test baseline model on test data
arima.test <- Arima(test[,"cancelled_percent"], model=arima.fit)
fcast <- forecast(arima.fit, h=6)

# Test arimax model on test data
arimax.test <- Arima(test[,"cancelled_percent"],
                     model=arimax.fit, xreg = test[, covariates])
fcastx <- forecast(arimax.fit, xreg = test[, covariates])

# PLOT BOTH FORECAST AGAINST ACTUAL VALUES
autoplot(window(dsMANts[,"cancelled_percent"], end =c(2019,12))) +
  autolayer(fcast, series="ARIMA", PI=FALSE) +
  autolayer(fcastx, series="ARIMAX", PI=FALSE) +
  xlab("Year") + ylab("%") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))


# Fitted values and residuals for arimax (2 ways)
# Way 1:
fitted.values          <- fitted(arimax.test)
fitted.values.baseline <- fitted(arima.test)
# residuals     <- (test - fitted.values)
# averageAE     <- mean(residuals)

# Way 2 (simpler):
# residuals <- residuals(arimax.test)

# Compare Accuracies
accuracy(arima.test)
accuracy(arimax.test)

# COMPUTE MASE's
# Create function computeMASE
computeMASE <- function(forecast,train,test,period){

  # forecast - forecasted values
  # train - data used for forecasting .. used to find scaling factor
  # test - actual data used for finding MASE.. same length as forecast
  # period - in case of seasonal data.. if not, use 1

  forecast <- as.vector(forecast)
  train <- as.vector(train)
  test <- as.vector(test)

  n <- length(train)
  scalingFactor <- sum(abs(train[(period+1):n] - train[1:(n-period)])) / (n-period)

  et <- abs(test-forecast)
  qt <- et/scalingFactor
  meanMASE <- mean(qt)
  return(meanMASE)
}

# MASE ARIMAX MODEL
MASE.arima <-
  computeMASE(forecast = fitted.values.baseline,
            train = train,
            test = test,
            period = 1)

# MASE ARIMAX MODEL
MASE.arimax <-
```

```
  computeMASE(forecast = fitted.values,
              train = train,
              test = test,
              period = 1)
# Compare Accuracies

round(accuracy(arima.test),3)
round(accuracy(arimax.test),3)
MASE.arima
MASE.arimax




#------------------------------------
# ARIMAX with google trends data
#------------------------------------

# install.packages("gtrendsR", dependencies = TRUE)
library(gtrendsR)

#--------------
# 1. Setting the keywords, country and time window
#--------------

# Define the keywords
keywords <- c("tickets", "cancelled")

# Set the geographic area: DE = Germany
country   <- c('GB')

# Set the time window
timePast <- ("2017-01-01 2019-12-31")
time <- ("2017-01-01 2021-06-30")

# Set channels
channel <- 'web'


#--------------
# 2. Run query
#--------------

trends   <- gtrends(keywords, gprop = channel, geo = country, time = time )


# Select only interst over time
dsTrend  <- trends$interest_over_time

# Change the values that have <1 to 0 and 'hits' to numeric
dsTrend$hits <- ifelse(dsTrend$hits == "<1", 0, dsTrend$hits)
dsTrend$hits <- as.numeric(dsTrend$hits)

#--------------
# 3. Create ts dataset including Google Trends data
#--------------

# NOW WE HAVE TO SUMMARIZE DATA BY MONTH

# Handle the 'rdate' column. Split into years and
# months. Then, a yearmonth column is made with type Date
# (as in the Challenges data frame)
dsTrend$reporting_yearmonth  <- (substr(dsTrend$date, 1, 7))

#dsTrend$date <- strptime(dsTrend$date,
#                                    format = "%Y-%m-%d", tz="GMT")


# calculate the mean search intensity for each month
dsTrend.m <- dsTrend %>%
  group_by(reporting_yearmonth, keyword) %>%
  summarise(mean_hits = mean(hits))

# Create new dataframe with only 2 columns, one by keyword
```

```
dsGT <- data.frame(matrix(ncol=2,nrow=54, # same nrows as dsSub, ie months observed
                          dimnames=list(NULL, c("tickets", "cancelled"))))

# Create temporal dataset with subset from mean_hits with condition
# (only if it's cancelled)
tmp <- subset(dsTrend.m, keyword == "tickets")

# Add to dsGT
dsGT$tickets <- tmp$mean_hits

# Create temporal dataset with subset from mean_hits with condition
# (only if it's cancelled)
tmp <- subset(dsTrend.m, keyword == "cancelled")

# Add to dsGT
dsGT$cancelled <- tmp$mean_hits

# Add these columns to dfGroupYM
dfGroupYM$GTCancelled <- as.numeric(dsGT$cancelled)
dfGroupYM$GTTickets   <- as.numeric(dsGT$tickets)


# Convert GroupYM into ts
dsSub <- dfGroupYM[-1]

# Convert all columns to ts
dsSubts <- ts(dsSub, frequency=12, start=c(2017,1))

#--------------
# 4. Analysis of pre COVID-19 period
#--------------

# Split the data
train0 <- window(dsSubts, end = c(2020,2))
test0  <- window(dsSubts, start = c(2020,3), end = c(2021,6))

# Create vector with exogenouos variables' names
covariates <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                 "flts_16_to_30_mins_late_percent",
                 "flts_31_to_60_mins_late_percent",
                 "flts_61_to_180_mins_late_percent",
                 "flts_181_to_360_mins_late_percent",
                 "more_than_360_mins_late_percent",
                 "GTCancelled",
                 "GTTickets"
)

# Create vector with exogenouos variables' names
covariatesD <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                  "flts_16_to_30_mins_late_percent",
                  "flts_31_to_60_mins_late_percent",
                  "flts_61_to_180_mins_late_percent",
                  "flts_181_to_360_mins_late_percent",
                  "more_than_360_mins_late_percent"
)

# train0 baseline model (arima.fit) and arimax model on train0 set
arima.fit0  <- Arima(train0[,"cancelled_percent"], order=c(2,0,0))
arimaxG.fit0 <- Arima(train0[,"cancelled_percent"],
                     order=c(2,0,0),
                     seasonal =list(order = c(1, 0, 0), period = 12),
                     xreg = train0[, covariates])
arimaxD.fit0 <- Arima(train0[,"cancelled_percent"],
                     order=c(2,0,0),
                     seasonal =list(order = c(1, 0, 0), period = 12),
                     xreg = train0[, covariatesD])

# test0 baseline model on test0 data
arima.test0 <- Arima(test0[,"cancelled_percent"], model=arima.fit0)
fcast0 <- forecast(arima.fit0, h=16)
autoplot(fcast0) + xlab("Year") +
  ylab("Percentage change")


# test0 arimax model (google) on test0 data
```

```
arimaxG.test0 <- Arima(test0[,"cancelled_percent"],
                      model=arimaxG.fit0, xreg = test0[, covariates])

fcastxG0 <- forecast(arimaxG.fit0, xreg = test0[, covariates])
autoplot(fcastxG0) + xlab("Year") +
  ylab("Percentage change")

# test0 arimax model on test0 data
arimaxD.test0 <- Arima(test0[,"cancelled_percent"],
                      model=arimaxD.fit0, xreg = test0[, covariatesD])

fcastxD0 <- forecast(arimaxD.fit0, xreg = test0[, covariatesD])
autoplot(fcastxD0) + xlab("Year") +
  ylab("Percentage change")

# PLOT BOTH FORECAST AGAINST ACTUAL VALUES
autoplot(window(dsSubts[,"cancelled_percent"], end =c(2021,6))) +
  autolayer(fcast0, series="ARIMA", PI=FALSE) +
  autolayer(fcastxD0, series="ARIMAX", PI=FALSE) +
  autolayer(fcastxG0, series="ARIMAX with Google data", PI=FALSE) +
  xlab("Year") + ylab("%") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))

# Fitted values and residuals for arimax (2 ways)
# Way 1:
fitted.values.G        <- fitted(arimaxG.test0)
fitted.values.baseline <- fitted(arima.test0)
fitted.values.D        <- fitted(arimaxD.test0)

# COMPUTE MASE's
# Create function computeMASE
computeMASE <- function(forecast,train0,test0,period){

  # forecast - forecasted values
  # train0 - data used for forecasting .. used to find scaling factor
  # test0 - actual data used for finding MASE.. same length as forecast
  # period - in case of seasonal data.. if not, use 1

  forecast <- as.vector(forecast)
  train0 <- as.vector(train0)
  test0 <- as.vector(test0)

  n <- length(train0)
  scalingFactor <- sum(abs(train0[(period+1):n] - train0[1:(n-period)])) / (n-period)

  et <- abs(test0-forecast)
  qt <- et/scalingFactor
  meanMASE <- mean(qt)
  return(meanMASE)
}

# MASE ARIMAX MODEL
MASE.arima <-
  computeMASE(forecast = fitted.values.baseline,
              train0 = train0,
              test0 = test0,
              period = 1)

# MASE ARIMAX MODEL
MASE.arimax.G <-
  computeMASE(forecast = fitted.values.G,
              train0 = train0,
              test0 = test0,
              period = 1)

MASE.arimax.D <-
  computeMASE(forecast = fitted.values.D,
              train0 = train0,
              test0 = test0,
              period = 1)

# Compare Accuracies
round(accuracy(arima.test0),3)
round(accuracy(arimaxD.test0),3)
round(accuracy(arimaxG.test0),3)
MASE.arima
MASE.arimax.D
MASE.arimax.G

xtable(round(accuracy(arima.test),3))

#--------------
# 5. Analysis of COVID-19 period
#--------------

# Split the data
train <- window(dsSubts, end = c(2019,9))
test  <- window(dsSubts, start = c(2019,10), end = c(2020,3))

# Create vector with exogenouos variables' names
covariates <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                 "flts_16_to_30_mins_late_percent",
                 "flts_31_to_60_mins_late_percent",
                 "flts_61_to_180_mins_late_percent",
                 "flts_181_to_360_mins_late_percent",
                 "more_than_360_mins_late_percent",
                 "GTCancelled",
                 "GTTickets"
                 )

# Create vector with exogenouos variables' names
covariatesD <- c( "average_delay_mins", "early_to_15_mins_late_percent",
                 "flts_16_to_30_mins_late_percent",
                 "flts_31_to_60_mins_late_percent",
                 "flts_61_to_180_mins_late_percent",
                 "flts_181_to_360_mins_late_percent",
                 "more_than_360_mins_late_percent"
                 )

# Train baseline model (arima.fit) and arimax model on train set
arima.fit  <- Arima(train[,"cancelled_percent"], order=c(1,0,0))
arimaxG.fit <- Arima(train[,"cancelled_percent"],
                    order=c(2,0,0),
                    seasonal =list(order = c(1, 0, 0), period = 12),
                    xreg = train[, covariates])
arimaxD.fit <- Arima(train[,"cancelled_percent"],
                    order=c(2,0,0),
                    seasonal =list(order = c(1, 0, 0), period = 12),
                    xreg = train[, covariatesD])

# Test baseline model on test data
arima.test <- Arima(test[,"cancelled_percent"], model=arima.fit)
fcast <- forecast(arima.fit, h=6)
autoplot(fcast) + xlab("Year") +
           ylab("Percentage change")


# Test arimax model (google) on test data
arimaxG.test <- Arima(test[,"cancelled_percent"],
                     model=arimaxG.fit, xreg = test[, covariates])

fcastxG <- forecast(arimaxG.fit, xreg = test[, covariates])
autoplot(fcastx) + xlab("Year") +
  ylab("Percentage change")

# Test arimax model on test data
arimaxD.test <- Arima(test[,"cancelled_percent"],
                     model=arimaxD.fit, xreg = test[, covariatesD])

fcastxD <- forecast(arimaxD.fit, xreg = test[, covariatesD])
autoplot(fcastxD) + xlab("Year") +
  ylab("Percentage change")

# PLOT BOTH FORECAST AGAINST ACTUAL VALUES
autoplot(window(dsSubts[,"cancelled_percent"], end =c(2020,3))) +
  autolayer(fcast, series="ARIMA", PI=FALSE) +
  autolayer(fcastxD, series="ARIMAX", PI=FALSE) +
  autolayer(fcastxG, series="ARIMAX with Google data", PI=FALSE) +
  xlab("Year") + ylab("%") +
  ggtitle("") +
  guides(colour=guide_legend(title="Forecast"))

# Fitted values and residuals for arimax (2 ways)
# Way 1:
fitted.values.G        <- fitted(arimaxG.test)
fitted.values.baseline <- fitted(arima.test)
fitted.values.D        <- fitted(arimaxD.test)

# COMPUTE MASE's
# Create function computeMASE
computeMASE <- function(forecast,train,test,period){

  # forecast - forecasted values
  # train - data used for forecasting .. used to find scaling factor
  # test - actual data used for finding MASE.. same length as forecast
  # period - in case of seasonal data.. if not, use 1

  forecast <- as.vector(forecast)
  train <- as.vector(train)
  test <- as.vector(test)

  n <- length(train)
  scalingFactor <- sum(abs(train[(period+1):n] - train[1:(n-period)])) / (n-period)

  et <- abs(test-forecast)
  qt <- et/scalingFactor
  meanMASE <- mean(qt)
  return(meanMASE)
}

# MASE ARIMAX MODEL
MASE.arima <-
  computeMASE(forecast = fitted.values.baseline,
              train = train,
              test = test,
              period = 1)

# MASE ARIMAX MODEL
MASE.arimax.G <-
  computeMASE(forecast = fitted.values.G,
              train = train,
              test = test,
              period = 1)

MASE.arimax.D <-
  computeMASE(forecast = fitted.values.D,
              train = train,
              test = test,
              period = 1)

# Compare Accuracies
round(accuracy(arima.test),3)
round(accuracy(arimaxD.test),3)
round(accuracy(arimaxG.test),3)
MASE.arima
MASE.arimax.D
MASE.arimax.G

xtable(round(accuracy(arima.test),3))
```