uc3m | Universidad Carlos III de Madrid

Bachelor's Degree in Aerospace Engineering

2017 - 2018

Bachelor Thesis

"Air Traffic Flow Management Regulations: Big Data Analytics"

Ignacio Sánchez Vázquez

Supervisor: Javier García-Heras Carretero

Leganés, September 2018

Abstract

Air traffic in Europe is constantly increasing. Due to this, Air Traffic Management is getting more complex and all stakeholders get affected by that. Among these, air traffic controllers are the ones that suffer the biggest impact in terms of overload of work. Every day, a set of regulations occurs in the regions controlled by these operators, which provokes delays on ground and rerouting in mid-air. All of these variations directly affect the entire ATM network and translates into big expenses for passengers and airlines.

With this project, the aim is to predict these daily contingencies by using big data analysis models, so that costs associated are reduced. Most of the information needed to run the analysis has been very complicated to extract, process and correlate because the data sources are not open to researchers. Therefore, the number of instances available for the prediction is very low (only 18 months of data). Nevertheless, while working with this limitation, a Naive Bayes classifier has been chosen as the analytical algorithm.

In terms of results, the work done does not reveal a high predictive capability due to the amount of data acquired and the simplicity of the temporal variables. This suggests that, in future researches, it could be convenient to intake broader historical data (more years). Moreover, more complex predictive models could be implemented if variables coming from the weather or the number of flights are used.

Keywords: big data, regulations, prediction, temporal variables, ATC sector, airport.

Acknowledgements

First of all, I would like to express my gratitude to my supervisor Javier. He answered that first email showing his passion for embarking this project and has always been by my side to support me.

On the technical part, my friends at Madiva Soluciones have been crucial for me. Special thanks to Guti and Don Carlos that have bear with me and my multiple doubts.

Last but not least, eternally grateful to my family. Gelo with his thinking, Mª Salud with her meticulousness, Marina with her proofreading and Angelito…with his copiloting.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AIRAC	Aeronautical Information Regulation and Control
ANM	ATFCM Notification Message
ANSP	Air Navigation Services Provider
ARP	Airport Reference Point
ATC	Air Traffic Control
ATCC	Air Traffic Control Center
ATFCM	Air Traffic Flow and Capacity Management
ATFM	Air Traffic Flow Management
ATM	Air Traffic Management
CFMU	Central Flow Management Unit
CTOT	Calculated Take Off Time
DDR2	Demand Data Repository 2
DSNA	Direction des Services de la Navigation Aérienne
EASA	European Aviation Safety Agency
ECAC	European Civil Aviation Conference
EDA	Exploratory Data Analysis
ERSA	Elementary Restricted Airspace
ES	Elementary Sector
ESRI	Environmental Systems Research Institute
ETFMS	Enhanced Tactical Flow Management System
FIR	Flight Information Region
FL	Flight Level
FMP	Flow Management Position
GIS	Geographical Information System
HTML	Hyper Text Markup Language
IATA	International Air Transport Association
ICAO	International Civil Aviation Organization
INTUIT	Interactive Toolset for Understanding Trade-offs in ATM
MIT	Massachusetts Institute of Technology
NM	Network Manager
ROC	Receiver Operating Characteristic
SESAR	Single European Sky ATM Research
TV	Traffic Volume
URL	Uniform Resource Locator
WEKA	Waikato Environment for Knowledge Analysis
WKT	Well-Known Text

1. Introduction

1.1. ATFM Regulations

Air traffic is monitored and controlled by air traffic controllers who inform pilots about their mid-air situation and give instructions to avoid accidents. Therefore, airspace is segmented into various regions: Air Traffic Control Centers. These areas are also partitioned into more elementary modules known as ATC sectors. Smaller modules are usually operated by a team of three controllers.

Sectors partitioning may vary throughout the day depending on the air traffic and the controller's workload. ATC sectors may be then split or merged depending on the needs at every specific instant. A general statement is to assume that controller's workload depends largely on the number of flights. However, sector redistribution is not always possible since it depends on the amount of air traffic controllers on duty or the sector cannot be split into more elementary units.

Therefore, in some situations where traffic is too difficult to be conducted by the controllers available, changes have to be made in the network. One way of avoiding dangerous situations is by re-routing those flights that intend to enter the overloaded sector. Another approach is to assign departure delays to the planes causing the excess. For the latter, in the ECAC region, the Central Flow Management Unit is the one in charge of the departure slots allocation. Air Traffic Flow Management slots are periods of time in which take-off must take place.

In Europe, a slot is defined as the period between 5 minutes before and 10 minutes after the Calculated Take-Off Time. Thus, considering the prediction of overloads anticipated by the Flow Management Position operators, CFMU may impose a traffic flow regulation that modifies directly the slot allocation algorithms by giving delayed CTOTs to airplanes. Due to the network behavior, the reason for a regulation to be forced may not only refer to sectors capacity but also various external factors as it may be weather phenomena or malfunctioning aerodrome services (e.g. de-icing equipment).

1.2. State-Of-The-Art

Currently in Europe, the method to decide the pre-tactical sector opening schedules is purely human and does not rely on automatization. There is a database that stores the most common airspace configurations and FMP operators choose the one they think is more adequate for each hour period. Then, assuming that controller's workload is only based on the amount of traffic, a pre-analysis is done by the Enhanced Tactical Flow Management System.

For the pre-tactical explorations, the estimated traffic demand is compared to each sector hourly capacity to see if it is exceeded. Sector hourly capacity is "the maximum number of flight entries in an hour that can safely be assigned to sector controllers" by EUROCONTROL [1]. So as to determine the amount of work considered as overload, Table 1.1 is defined by the authorities.

Threshold	*Interpretation*	*Recorded Working Time during 1 hour*
70 % or above	Overload	42 minutes +
54 % - 69 %	Heavy Load	32 - 41 minutes
30 % - 53 %	Medium Load	18 - 31 minutes
18 % - 29%	Light Load	11 - 17 minutes
0 % - 17 %	Very Light Load	0 - 10 minutes

Table 1.1. Overload Threshold for ATC Controller [1]

After all, this manner of scheduling results in an idealistic approach as it does not consider external factors for the controller's workload and the available set of configurations is limited. Due to this, FMP and the CFMU operators base their forecasts for future overloads on past experiences. That means ATFM regulations are imposed only from historical knowledge of the operators.

With all this in mind, EUROCONTROL (European Organization for the Safety of Air Navigation) and the European Union are promoting a research program that aims to define the new basis for the Air Traffic Management: Single European Sky ATM Research. It constitutes the technological pillar for the Single European Sky project and intends to modernize and harmonize the European ATM systems through the implementation of new technologies. By 2020, SESAR aims to (from [2]):

- Enable a three-fold increase in airspace capacity
- Improve safety by a factor of 10
- Provide a 10% reduction in the environmental impact per flight
- Reduce ATM costs by 50%
- Reduce delays in the air and on the ground

Among SESAR research, the INTUIT project is directly related to ATFM regulations using Machine Learning and Visual Analytics [3]. They focus on identifying cause-effect relationships in the air traffic network and the development of new monitoring and management tools. In the paper, they conclude that visual analytics can be very valuable "to advance in the state-of-the-art in ATM performance modelling" (from [3]). INTUIT also addresses the Data Inventory needed to perform their studies and, in most of the cases, sources are private and only stakeholders may access them. As part of Horizon2020, the proposal is not finished, so no further conclusions may be extracted from the document.

Apart from EUROCONTROL, there are also national studies promoted by other ATM stakeholders. This is the case of the DSNA, the French Air Navigation Services Provider. Based on Neural Networks and Tree Search methods, their R&D team has performed a research to forecast workload and airspace configuration based on the amount of flights and their trajectories [4]. When talking about the study limitations, they refer to the ATM as a "world of uncertainties" (from [4]) and not all of them could be tackled through the work, as it is the case for weather disturbances.

ATM efficiency is not a topic that only concerns European authorities but also the American ones. For the 12th USA/Europe ATM R&D Seminar in 2017, the Department of Aeronautics and Astronautics from the MIT, prepared *A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks* [5]. They look for predicting the delays at airports by considering the Origin-Destination (OD) pair delay variable as the main feature, based on the idea that past delays propagate in the future. Results obtained show that feature selection is key depending on the desired outcome. In this case, time of the day was the most relevant factor when accounting for delays on ground. Again, most of the data comes from private sources and the study is made with American data.

1.3. Motivation & Objectives

Delays on ground may be originated not only from local issues but also from mid-air or destination problems. This study is aimed to understand and predict future regulations that may appear on the ATM Network in Europe, complementing the work done by American researchers as in [5] who based their work on US historical air traffic information. For that purpose, a full Big Data Analytics work is developed.

Big Data Analytics is essentially the process of examining large amounts of data in order to extract knowledge (patterns, trends, preferences…). For this project, the work will be structured in three main steps that follow this order: data extraction, pre-processing and analysis. First phase, data intake, will serve as an exploration of the different available sources and the way of extracting information from them. After that, since the databases explored are varied, a quality assessment and transformation of the data must be made in the stage of data preprocessing. With all the information organized and clean, last step is devoted to extracting knowledge from it by creating prediction models and visualizing. A timeline of the project is shown in Figure 1.1.

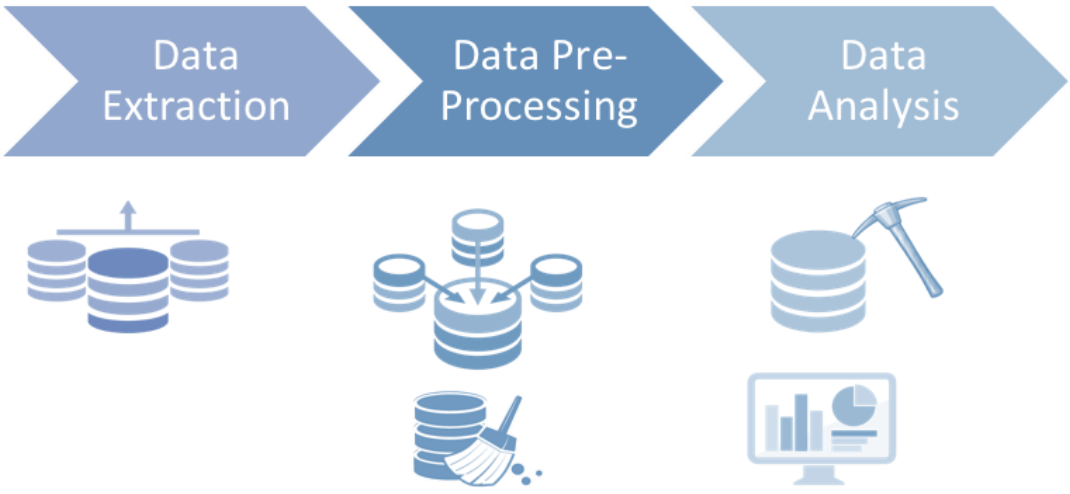


Fig.1.1. Project Timeline

In order to supplement works based on number of flights as [3] and [4], the scope of this study only considers temporal variables (season, day of the week, time of the day…) as the selected features. The expected outcome intends to ameliorate the pre-tactical forecast performed by the FMP and the CFMU operators. As a next step on their approach of

anticipating traffic overload with historical experiences, the results of the study may provide them with a valuable model that considers many more processed past scenarios, so the decision-making is improved. Following the same goals as the SESAR, this project pretends to optimize ATM resources and thus, reduce costs.

Additionally, as a secondary goal, the analysis and the resulting dataset expect to serve as a guideline for further researches on this topic. Adding geographical information as well as number of flights to the dataset, the model to be developed could be fed with new features so deeper knowledge is obtained. This could be the case of weather variables that could be combined and provide a better explanation for weather-caused regulations. These analyses fall out of scope for this project since they require high computational capabilities due to the datasets size. And the treatment of those datasets is complex and necessitates a longer period of time for researching.

Lastly, this project looks for promoting the idea of open-data and open-source tools. While many of the studies done are promoted and data-supported by ATM stakeholders as the ones explained in State-Of-The-Art section, it is also important to let external researchers develop their solutions, which may contribute to the common good. In order to function as baseline, all datasets generated through this analysis come from a public source and open-source software is used for the treatment and presentation of the data.

Regarding to the tools, big data environment allows to choose from a wide variety. In terms of simplicity and scalability, Python is the programming language selected for this project as it is becoming the universal language for Data Science. In the same lines, when dealing with large datasets containing geographical objects, Carto seems to be the most popular software and will be used when needed for data manipulation and visualization. As resources are limited and time is a constraint, WEKA is the software used for the prediction algorithms since it is fairly simple to implement.

2. Data Extraction

To begin with, useful and large amounts of data are needed. Following the scope of the project, it is clear that the main focus is on obtaining information about regulations occurring in Europe. Thus, the first aim is on the organization that manages the data for air traffic across Europe: the European Organization for the Safety of Air Navigation (EUROCONTROL).

EUROCONTROL offers a wide variety of datasets on their DDR2 (Demand Data Repository 2) that would deliver enough information to make this data intake fast and easy. In principle, it used to be open data that would serve for NEST and SAAM tools. However, with the data revolution, EUROCONTROL has limited the number of downloads to five per month for standard users. For this very reason, it is necessary to intake implicit information straight from the webpage as well as from other sources. The heterogeneity of these varied datasets implies different methodologies for the data extraction to occur and they are deeply explained by source:

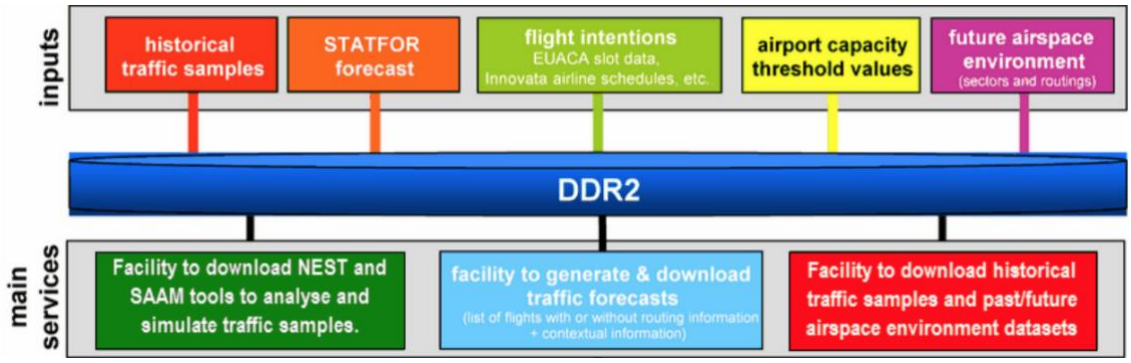


Fig.2.1. DDR2 inputs and services [6]

a) Regulations: (extracted from [7])

Located in the Network Manager section of EUROCONTROL's webpage, it consists in a document that serves as a way of notifying European authorities about any issue related to ATFM delays in the ECAC area. In order for it to work, EUROCONTROL relies on (airport/ANSP) ATC units to validate and report any delay issue through the Post Ops Performance Adjustment Process Change Request Form.

Updated every month, Network Manager publishes a spreadsheet with all the approved data. Again, as most of EUROCONTROL's dataset, the availability is limited. There is public access on the portal to current year's document and it requires being a stakeholder in order to check for the archives. For the analysis of this thesis, it has been possible to retrieve the entire 2017 and the first half of 2018 (June 30th).

Therefore, two different spreadsheets are obtained. In the case of 2017's, it is a table containing 52133 different regulations. For the first semester of 2018, there are 26601 delay issues registered. In regard to the fields that classify these data, the documents provide the following (note: variables with ORIGINAL_ have an additional field with the values before any changes are approved by Network Manager):

- REGU_ID:

  Unique name for identifying each regulation. It is usually composed by the name of the Reference Location followed by a number indicating the starting date of the issue. There would be a letter added if various regulations occur the same day, indicating the period in the day: M = Morning, A = Afternoon, N = Night, E = Early morning, X = Other. E.g.:

REGU_ID	START_DATE	REF_LOC
LFPOA01	1/1/17 16:20	LFPO
LFPOA03	3/1/17 18:00	LFPO
LEMDA02	2/3/17 8:20	LEMD
LEMDA01	1/9/17 8:00	LEMD

Table 2.1. Example of REGU_ID Format

- START_DATE:

  Starting date of the regulation with the format dd/mm/yy and starting time as hh:mm.

- END_DATE:

  Ending date of the regulation with the format dd/mm/yy and ending time as hh:mm.

- (ORIGINAL_)TVS:

  Unique name given to a list of traffic volumes associated to a Flow Management Position: Traffic Volume Set. [8]

- TV:
  Unique name given to the element that defines a specific volume of air traffic, in order to compare the amount of traffic and the given official values. TVs are used by the NM and the FMPs for monitoring and applying ATFM regulations. Each Traffic Volume has an associated Reference Location. [9]
- (ORIGINAL_)REF_LOC:
  Unique name given by the Central Flow Management Unit for the ATFCM systems to describe either an aerodrome or an ATC sector. [10]
- (ORIGINAL_)REF_LOC_TY:
  Categorical field that indicates the Reference Location Type: Airport or En Route.
- (ORIGINAL_)REGU_REASON(_CODE):
  Categorical field that indicates the reason for the regulation occurring (having a code associated to each of them). There are several causes that are explained in Table 2.2.
- ANM_REMARK:
  Brief explanation given by ATFCM Notification Message in the cases where regulation reason needs to be explained deeper. E.g.: “BLOCKED RWY” / “ADDITIONAL TFC DUE TO FRENCH STRIKE” / “BOMB ALERT”.
- ALL_TRAFFIC:
  Number of total traffic affected by the regulation, either delayed or not.
- MP_TRAFFIC:
  Number of total traffic for which the regulation is the most penalizing, either delayed or not.
- MP_DELAYED_TRAFFIC:
  Number of delayed traffic for which the regulation is the most penalizing.
- (ORIGINAL_)REGU_FLIGHT_DELAY:
  Sum of the delay in minutes of all the delayed traffic by the regulation.
- IS_CHANGED:
  Informs whether the regulation has been changed or not. {Y/N}

- POST_OPS_CHANGE:

  Informs whether the changed regulation has been updated in the Post Ops Report. {Y/N}

- REA_TDPIS_CHANGE:

  Informs if the change has been applied due to a Ready-to-depart message and/or a Target Departure Panning Information. {Y/N}

- ORIGINATOR_REF:

  Reference of the originator for the change. {ENAIRE , SKYGUIDE , UKNATS...}

- NM_REF:

  Case number given by the Network Manager to each change made.

- RATIONALE_FOR_CHANGE:

  Brief explanation for the change made. E.g.:

  “Correction procedural error: Regulation reason changed during the regulation.”

  “Correction reference location: agreed DLA-Shift due to Network Optimization.”

REGULATION REASON	CODE	GUIDELINES
ATC Capacity	C	En Route: Demand exceeds, or complexity reduces declared or expected ATC capacity. Airport: Demand exceeds declared or expected ATC capacity.
ATC Industrial Action	I	Reduction in any capacity due to industrial action by ATC staff.
ATC Routeing	R	Network solutions / scenarios used to balance demand and capacity.
ATC Staffing	S	Unplanned staff shortage reducing expected capacity.
ATC Equipment	T	Reduction of expected or declared capacity due to the non- availability or degradation of equipment used to provide an ATC service.
Accident / Incident	A	Reduction of expected ATC capacity due to an aircraft accident / incident.
Aerodrome Capacity	G	Reduction in declared or expected capacity due to the degradation or non-availability of infrastructure at an airport. e.g. Work in Progress, shortage of aircraft stands etc. Or when demand exceeds expected aerodrome capacity.
Aerodrome Services	E	Reduced capacity due to the degradation or non-availability of support equipment at an airport e.g. Fire Service, De-icing / snow removal equipment or other ground handling equipment.
NON-ATC Industrial Action	N	A reduction in expected / planned capacity due to industrial action by non-ATC personnel.

Airspace Management	M	Reduction in declared or expected capacity following changes in airspace / route availability due to small scale military activity.
Special Event	P	Reduction in planned, declared or expected capacity or when demand exceeds the above capacities as a result of a major sporting, governmental or social event. It may also be used for ATM system upgrades and transitions. Large multinational military exercises may also use this reason. This category should only be used with prior approval during the planning process.
Weather	W	Reduction in expected capacity due to any weather phenomena. This includes where weather impacts airport infrastructure capacity, but where aerodrome services are operating as planned / expected.
Environmental Issue	V	Reduction in any capacity or when demand exceeds any capacity due to agreed local noise, runway usage or similar procedures. This category should only be used with prior agreement in the planning process.
Other	O	This should only be used in exceptional circumstances when no other category is sufficient. An explanatory ANM remark MUST be given to allow post ops analysis.

Table 2.2. Regulation Reasons and Codes [11]

b) Number of Daily Flights: (extracted from [12])

Navigating through EUROCONTROL's DDR2, in the section where Historical Traffic may be downloaded (only 5 daily files per month), there is summarized information with the number of flights per day in the ECAC region:

Events	29	30	31	31	30	29	32	31	29	30	31	31	31	36	36	36	36	36	35	36	39	40	41
MONTH	MAY-2018																						
AIRAC	1805																						
	Tue 1	Wed 2	Thu 3	Fri 4	Sat 5	Sun 6	Mon 7	Tue 8	Wed 9	Thu 10	Fri 11	Sat 12	Sun 13	Mon 14	Tue 15	Wed 16	Thu 17	Fri 18	Sat 19	Sun 20	Mon 21	Tue 22	Wed 23
NESTO.	1805																						
EXP2																							
SO6 m1																							
SO6 m3																							
ALL_FT+																							
Ranking	28	15	13	6	30	25	18	24	11	23	19	31	21	7	17	12	10	2	29	27	16	22	8
Nb Flights	30144	32847	33226	34162	28995	31040	32802	31101	33320	31663	32458	28920	32014	34097	32811	33296	33606	35318	29557	30584	32822	31739	33971

Fig.2.2. Historical Traffic section in DDR2

However, there is no direct way of extracting the shown values in each cell for "No. Flights" so that an automation must be made in order to intake that data in an efficient manner. Due to the simple HTML structure of the page, the recommended approach is to perform Web Scraping.

Web Scraping is a technique used for extracting data directly from websites and saving it to a specific file. For the case of this project, a JavaScript scraper has been developed for DDR2 webpage. In order to build it, it is required to find within the HTML code, the desired feature. In this occasion it is straightforward and looking on the source code it is seen that number of flights attribute has the following form:

div[id*=nbflights]

Therefore, all the fields with that HTML element are extracted and stored into an array. With that, we apply the .map() method in order to create a new array with the results of implementing the split function on every element in the calling array, obtaining an output with the form shown in Figure 2.3.

Additionally, EUROCONTROL delivers this information on the screen only per month and the URL does not change from one month to another. Then, so as to extract automatically the entire 2017 and first half of 2018, a simulation of navigation must be

done. For this, a web browser automation tool has been chosen: Selenium. It allows the coder to open a specific web browser and perform tasks as a human being would, such as clicking buttons or filling up forms. Concerning this specific case, it would imitate the clicking button for changing months. With all the set up done, a .csv (comma separated values) file is generated, having one column with the date in dd/mm/yy format and a second column with the amount of flights in that specific day.

```
0: "nbflights;↵
1: "20180101;20794"
2: "20180102;26096"
3: "20180103;25504"
4: "20180104;26056"
```

Fig.2.3. Scraped Data for No. of Flights

c) ATC Sectors: (extracted from [13])

As it has been explained, there are regulations happening in specific regions (Reference Locations) and these areas are not always static. Due to this dynamism, there is no universal dataset with the coordinates of each one. However, EURONTROL, in their DDR2 webpage, shares open datasets relating to this specific topic. Within the PostOPS tag in their dataset files section, they offer the past airspace environment data as completed according to the input from the ETFMS after the official Network Manager AIRAC (Aeronautical Information Regulation And Control) cycle - 28 days - is finished.

To define the 3D sectors, the information is divided into two files. On the one hand, the Gasel Airblock file (.gar) describes the 2D shape of the sector, providing the ID and each of the coordinates (longitude and latitude) for the vertices of the polygon. On the other hand, the Gasel Sector file (.gsl) correlates with the .gar through the unique IDs and complements it by giving the sectors name, the sector type (FIR, ERSA, ES) and the lower and upper Flight Levels. Both files are explained in detail in Tables 2.3 and 2.4.

First line of the file				
#	**Field**	**Type**	**Size**	**Comment**
1	nb	num	1	number of airblocks defined
Header				
#	**Field**	**Type**	**Size**	**Comment**
1	tag	char	1	must be "A"
2	name of airblock	char	~	ID (max 24 characters for SAAM usage). Must be the same as the one found in associated ".gsl" file
3	number of items	num	~	number of items (vertices) composing the airblock (indicate the number of body lines) the first item being repeated at the end of the body
Body (made with Header "number of items" lines)				
#	**Field**	**Type**	**Size**	**Comment**
1	tag	char	1	must be "P"
2	latitude	num	~	in degrees decimals
3	longitude	num	~	in degrees decimals

Table 2.3. Gasel Airblock file format [6]

First line of the file				
#	**Field**	**Type**	**Size**	**Comment**
1	#	char	1	
2	file type	text		= SECTOR
3	version	num		= 2
4	NM AIRAC cycle	num		
5	AIRAC start date	date		format is yyyyMMdd
6	AIRAC end date	date		format is yyyyMMdd
7	number of record	num		
8	datasource	text		
Header				
#	**Field**	**Type**	**Size**	**Comment**
1	S	char	1	S=Sector
2	Sector ID	text		
3	Sector name	text		
4	Nb of airblock	num		= number of BODY lines
5	Airspace category	char	1	
6	Sector type	text		FIR=Flight Information Region ERSA=Elementary Restricted Airspace ES=Elementary Sector
Body				
#	**Field**	**Type**	**Size**	**Comment**
1	A	char	1	A=Airblock
2	Airblock name	text		
3	Operation	char	1	Always "+"
4	Lower FL	num		
5	Upper FL	num		

Table 2.4. Gasel Sector file format [6]

d) Airports: (extracted from [14])

Lastly, not all delays occur because of En Route regulations, there are also Airport ones. Consequently, the coordinates for all airports in ECAC must be extracted. Similar to ATC sectors, EUROCONTROL generates PostOps documents containing airports information but again, splitting it into separate files. However, airports location is not dynamic as it happened with sectors so that a one-time download must be enough. With the purpose of avoiding the tedious task of dealing with various files, an already created spreadsheet is used instead.

OurAirports is a website dedicated to passengers and pilots. It provides with detailed material about worldwide airports and gives additional tools to use their data as a source for diverse use cases. They wanted to free up global aviation data since some countries shut down their public access. In this manner, out of all the datasets offered, the one with airports geographical information.

World-Airports .csv file contains numerous fields, some of which will later be cleansed. In principle, it delivers the following columns for every airport worldwide:

- id
- type {small, medium, large}
- name
- latitude_deg
- longitude_deg
- elevation_ft
- continent
- country
- region
- gps_code
- link
- iata_code

3. Data Pre-Processing

After the large data intake performed, although it mainly comes from reliable sources, it presents some inconsistency and is still incomplete. For that reason, a transformation of this raw information into an understandable format must be done. Once it has been completed, data would be ready to process and analyze.

As regards this project, it is useful to keep and generate fields of data that may help to predict regulations. So, all columns containing geographical and time-related insights will be crucial. Additionally, since the causes for the delays vary, columns containing regulations types and circumstantial records must be kept. With that being said, the data pre-processing is split into three phases:

3.1. Data Cleansing

To begin with, there is a need to get rid of unnecessary information as well as incomplete or inconsistent records that would not contribute to delay prediction. For this matter, data cleansing is separated in different tasks which are:

a) Irrelevant Data:

With reference to the main dataset, the regulations one, there are several fields that do not contribute to the scope of the work. As the source is aimed to deal with PostOPS changes approved by the Network Manager, there are several columns that describe these modifications and do not give useful insights. Then, by using Microsoft Excel to read and manipulate the spreadsheet, the following fields are deleted:

- IS_CHANGED
- POST_OPS_CHANGE
- REA_TDPIS_CHANGE
- ORIGINATOR_REF
- NM_REF
- RATIONALE_FOR CHANGE

In a similar way, due to the purpose of the document, in some cases there is one column devoted to the original information and another one to the actual/corrected one. Consequently, as the adjustment is published once it is finished, the columns with the original information are completely irrelevant and are deleted in Excel as well:

- ORIG_TVS
- ORIGINAL_REF_LOC
- ORIGINAL_REGU_REASON
- ORIGINAL_REF_LOC_TY
- ORIGINAL_REGU_REASON_CODE
- ORIGINAL_REGU_FLIGHT_DELAY

Likewise, in the .csv file downloaded for the Airports dataset, there are plenty of fields that would not add any value. Firstly, by loading the file in Excel, it is possible to filter it (Figure 3.1) and obtain only the airports that are in the European continent since the study is focused on ECAC region. Following this same procedure, EUROCONTROL does not include information about delays in small regional airports or any other kind of aerodrome so that only medium and large size airfields are filtered.

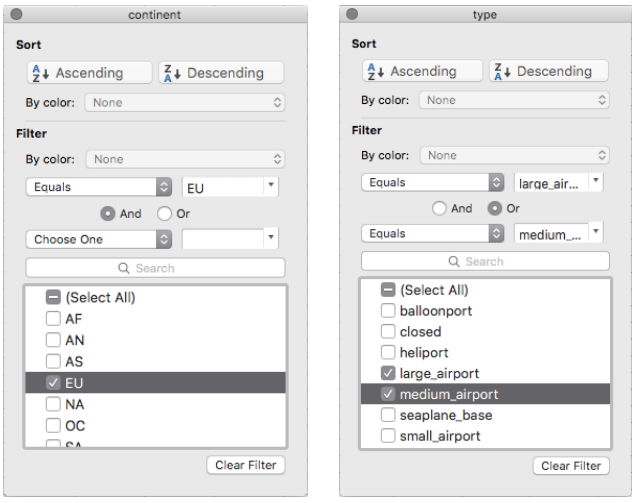


Fig.3.1. Excel filters for Airports.csv

From this file, the most interesting fields to keep are the ones related to the geographical position and characteristics of each airport, consequently, the following columns are discarded:

ƒ id	ƒ continent	ƒ type	ƒ region
ƒ name	ƒ iata_code	ƒ country	ƒ link

b) Inconsistent Data:

Even though Network Manager establishes clear guidelines to submit delay issues, the given format by each ATC unit differs in some cases. This occurs with the field of Regulation IDs. There is a clear structure to follow and it is specified in the ATFCM Users Manual as:

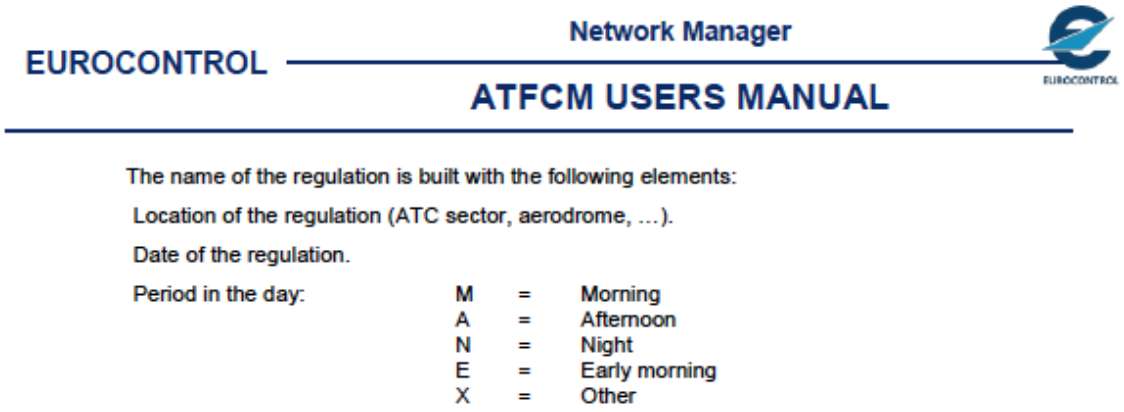


Fig.3.2. Regulation ID in ATFCM Users Manual [11]

Nevertheless, after inspecting the downloaded spreadsheets, it has been found that there are many inconsistencies and the date indicator only deals with the day and not the month (Table 3.1). Due to this, REGU_ID field is not valuable anymore as it is not a unique key that could have been used to filter out the Regulations data frame.

REGU_ID	START_DATE	REF_LOC
LFPOA01	01/01/2017 16:20	LFPO
LFPOA01	01/08/2017 05:00	LFPO
LFPOA01	01/10/2017 19:40	LFPO
LFPOA01	01/11/2017 06:00	LFPO

Table 3.1. REGU_ID Inconsistency Example

Apart from the REGU_ID field, there is another column in the Regulations dataset that presents inconsistencies: ANM_REMARK. Introduced so that ATFCM Notification Messages get recorded, they do not follow a normalized nomenclature and they will not be present in the analysis. Most of the remarks are added for the cases of weather issues and almost 99% of them have a message associated. However, although for the case of weather it may be relevant to analyze the deeper explanation, we encounter inconsistent cases as the following:

LOW VISIBILITY // LOW VIS // LOW VIZ // LOW VSIBILITY // LOW VIBILITY

Among the different Reference Locations, there are either "En Route" or "Airport" ones. Nonetheless, after some inspection, there are locations that have both categories due to some mistakes in the document. This is the case for LSZH, which is an Airport but there are ten issues that refer to it as En Route. As this only occurs for a limited number of cases, the Reference Locations are corrected manually so as to be consistent.

c) Duplicated Data:

While analyzing the data extracted, there are some fields that give repeated information. Following up with the Regulations dataset, there is a column devoted to divide the reasons for the regulations into different categories. Next to it, there is another one that simply recalls the reason by assigning a simple letter (code) as it was shown in Table 2.2. Therefore, there is repetition about this aspect and the duplicated data does not contribute to the work so the REGU_REASON field is removed.

Referring to the ATC sectors information, it has been extracted for every month in the year and a half of the study. Due to European sectorization system, there is dynamism in the distribution of airspace and sectors are not always the same. However, even though new sectors may appear, every month the main ones do not change much. There are then repeated sector names in the files as they are maintained every month. These contain the same coordinates and flight levels information so that that when the ATC sectors dataset is generated, only unique values are kept. It is explained in the Data Integration section.

3.2. Data Integration

So far it has been explained that several datasets have been extracted. Nevertheless, for the purpose of this analysis, a complete and main dataset is needed. For that, several procedures need to be done so that homogenization is possible.

a) File Generation:

After the extraction, ATC sectors information was split into two files (.gar and .gsl). They are internally related through an ID so that would be the common point to combine both documents. Additionally, there is information that is not needed, so only the key fields are taken for the new file to be generated.

First, from the "sectors.gar" file, it is interesting to extract the polygon coordinates for each of the ATC sectors IDs. For further uses, it is convenient to extract these points in a specific format: Well-Known Text (WKT). It is the ASCII representation of a spatial object and is very useful for visualization software and PostGIS manipulations. An example of WKTs is shown in Figure 3.3.

WKT Example	Description
POINT(1 2)	The point (1,2)
MULTIPOINT(0 0,1 1)	A set made up of the points (0,0) and (1,1)
LINESTRING(1.5 2.45,3.21 4)	The line from the point (1.5,2.45) to the point (3.21,4)
MULTILINESTRING((0 0,–1 –2,–3 –4),(2 3,3 4,6 7))	Two linestrings, one that passes through (0,0), (–1,–2), and (–3,–4), and one that passes through (2,3), (3,4), and (6,7).
POLYGON((1 2,1 4,3 4,3 2,1 2))	The rectangle whose four corners are indicated by (1,2), (1,4), (3,4), and (3,2). A polygon must be closed, so the first and last points in the WKT must match.
POLYGON((0.5 0.5,5 0,5 5,0 5,0.5 0.5), (1.5 1,4 3,4 1,1.5 1))	A polygon (0.5 0.5,5 0,5 5,0 5,0.5 0.5) with a hole in it (1.5 1,4 3,4 1,1.5 1).
MULTIPOLYGON(((0 1,3 0,4 3,0 4,0 1)), ((3 4,6 3,5 5,3 4)), ((0 0,–1 –2,–3 –2,–2 –1,0 0)))	A set of three polygons
GEOMETRYCOLLECTION(POINT(5 8), LINESTRING(–1 3,1 4))	A set containing the point (5,8) and the line from (–1,3) to (1,4)

Fig.3.3. WKTs Examples [15]

In order to manipulate the .gar file, due to the rare format, a text processing tool is used: AWK. With it, by looking at the structure, it is possible to generate a parser that interpret each line and translates it into .csv file. Then, as shown in Table 2.3, it is possible to generate a data frame with two columns: one with the sectors IDs and another with their geometries in WKT.

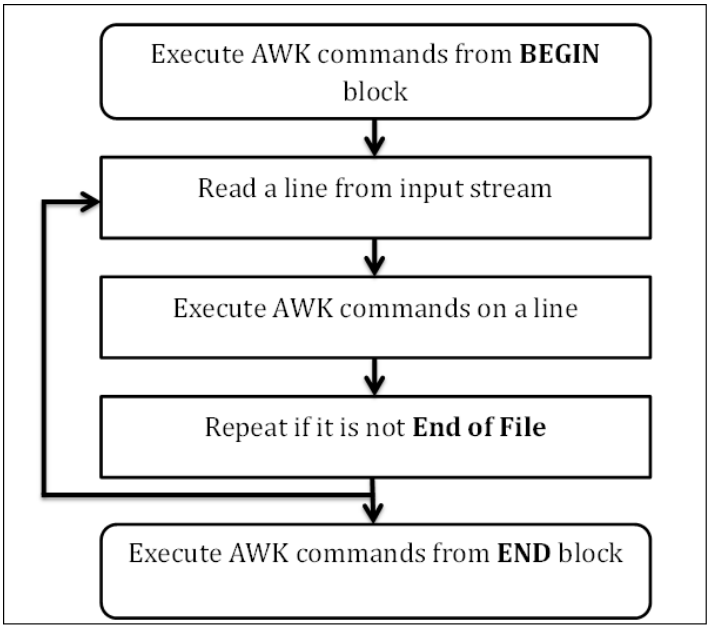


Fig.3.4. AWK Workflow [12]

Additionally, taking advantage of the format WKT, it is possible to boot up QGIS and load the file. It is a GIS (Geographical Information System) software designed to visualize, analyze and manage all types of spatial and geographical data. By doing so, it allows to save the .csv file into shapefile, which is a universal format regulated by ESRI (Environmental Systems Research Institute). That would make future visualizations and manipulations smoother in any kind of geospatial interpreter.

Similarly, for the case of the .gsl file, AWK commands are used as well. By programming a new parser, a final .csv with the following information is obtained: ATC sectors IDs, ATC sectors names, minimum and maximum Flight Levels for the sector.

b) Merge:

While each of the datasets may be valuable on their own, the analysis becomes more robust and relevant if they become a unique data frame. For doing so, it is necessary to merge them all with different approaches; by taking advantage of the shared information among each dataset (key columns) and converge them all into one.

To begin with, ATC sectors information is still in two different .csv files: one with the two-dimensions coordinates and the other with the name and flight levels (which may be considered as the third dimension). Taking advantage that GIS software have the possibility to perform SQL (Structured Query Language) queries, both of them are loaded in Carto. Carto is an open-source software built in PostgreSQL (an Object-Relational DataBase Management System) with an emphasis on PostGIS (an extension for geographic objects in PostgreSQL).

Once the datasets are loaded in Carto, a simple merge is done with an SQL query. This way, a new table is created, containing all columns from both sets. In order to do it, a common column has to be selected and as explained in the files structure, both of them contain an ID column that is used as "key" for this manipulation. As mentioned previously, repetition is to be avoided so unique values for sectors name are selected with SQL as well, which ends up accumulating various IDs in one row.

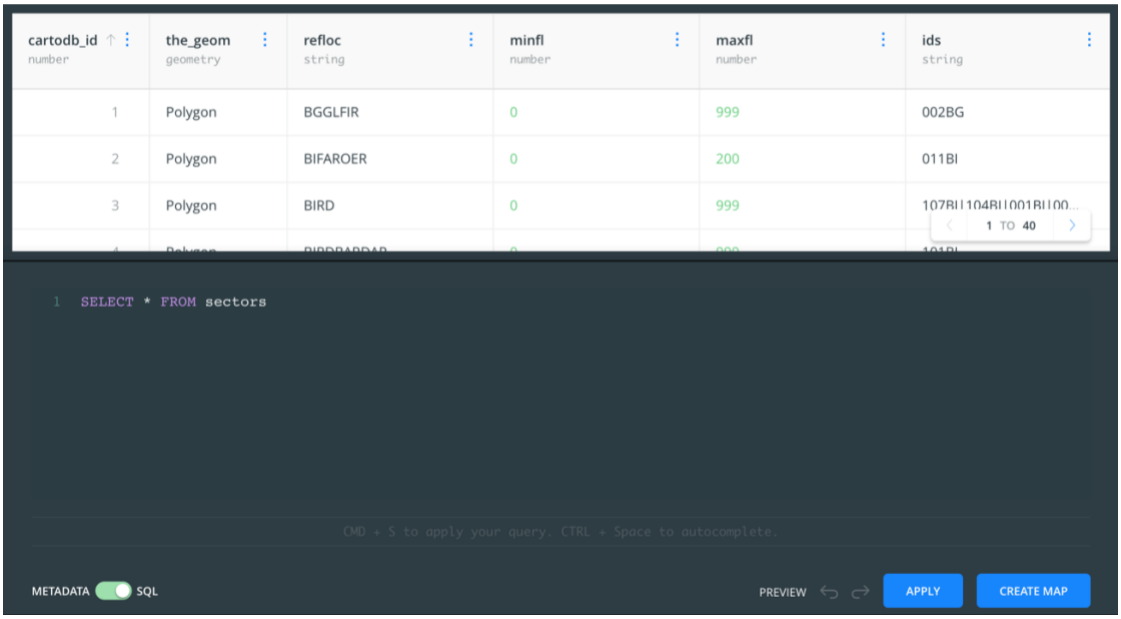


Fig.3.5. Carto SQL Interface

Therefore, there are four datasets to be merged now: Regulations, ATC Sectors, Airports and Number of Flights. The former one would be added in the last place after the other three have been put together. As there are no more merges in which geometries are involved, the tool selected would be the programming language Python, more specifically its Pandas library. Pandas is devoted to data manipulation and analysis with the use of data frame structures.

With reference to the Python interpreter, as the work is devoted to research and educational purposes and not aimed to be scaled up to production, a Jupyter Notebook is used. It is an open-source web application that allows the implementation of both computer code (Python) and rich text elements (paragraph, equations, figures…). Notebooks are human-readable as well as easier to follow as inputs and outputs can be divided into as many pieces as wanted [17].

libraries

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn import preprocessing
from matplotlib.gridspec import GridSpec
import calendar as cl
import Levenshtein as lev
import datetime
import seaborn as sns
```

regulationsTable

In [2]:
```
data = pd.read_excel(io="regulationsTable.xlsx")
```

In [3]:
```
# Show that data was loaded correctly:

data.head()
```

Out[3]:

	startDate	endDate	tvsName	refLoc	locType	reguReason	reguReasonCode	trafficTotal	trafficMost	trafficMostPen	delay
0	2017-01-01 06:00:00	2017-01-01 21:30:00	LFFFAD	LFPN	Airport	ATC Capacity	C	1	1	0	0
1	2017-01-01 06:00:00	2017-01-01 06:45:00	LSAZFMP	LSZH	Airport	Environmental Issues	V	0	0	0	0
2	2017-01-01 06:40:00	2017-01-01 07:40:00	EHAAFMP	EHAM	Airport	Weather	W	9	9	5	50
3	2017-01-01 06:40:00	2017-01-01 08:40:00	LFFFAD	LFPG	Airport	Weather	W	30	30	16	199
4	2017-01-01 07:00:00	2017-01-01 18:00:00	SCENRR1	LCCCW	En route	ATC Routeing	R	0	0	0	0

Fig.3.6. Jupyter Notebook Example

With the selection of tools made, the method for merging is next. After a quick inspection, it is found that notation for ATC sectors names is not consistent between the Regulations table (done by the Network Manager) and the ATC sectors geographical dataset (completed by the ETFMS). However, after a deeper analysis is done, it is shown that inconsistency comes with suffix of the Reference Location name. Then, as the root of the name is the one that declares the geographical region and the suffix gives little additional information (used for dynamic sectorization, to differentiate the divisions), for the analysis is enough to look for quasi-equal sectors names.

While perfect matching is not possible for Regulations and ATC sectors tables, a fuzzy (approximate) match is performed. In order to do so with Python, the extension of Levenshtein is used. It contains the function to determine the Levenshtein distance between two strings, i.e.: it measures the amount of insertion, substitution or deletion of characters from one string to the other. Figure 3.7 shows an example of Levenshtein distance between "Honda" and "Hyundai". There are three permutations done for going from one word to the other so that the distance would be equal to three.

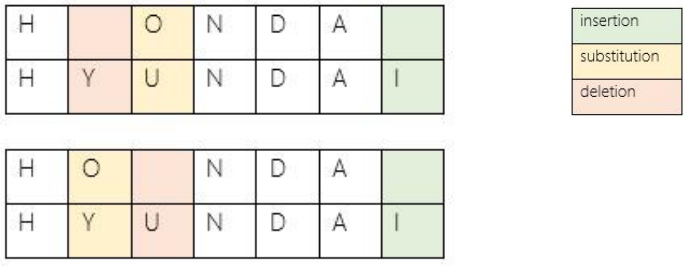


Fig.3.7. Levenshtein Distance [18]

Thus, by applying the "map" function that recalls an anonymous function (lambda) including the Levenshtein distance module, the REF_LOC column (only "En Route" locations) from the Regulations table is looped comparing them with "names" column from ATC sectors dataset. After creating a full matrix with all Levenshtein distances between strings, the minimum for each row is selected, obtaining then the closest match (fuzzy match) for each reference location. Most matches show a distance less or equal to one, which proves that both tables may be merged through these key columns.

ID 1	NAME 1	ID 2	NAME 2	LEV DIST
245	LKAAWLM	3678	LKAAWL	1
246	GCCCNWW	2751	GCCCRWW	1
247	LKAANM	4294	LKAANM	0
248	LFRRQXU	1764	LFRRQU	1
249	LFEEHH	5543	LFEEHH	0
250	LZBBU	4413	LZIBU	1
251	LFEEHN	5545	LFEEHN	0

Table 3.2. Minimum Levenshtein Distance (Regs & Sectors table)

On the other hand, airports names have to be used as well to get the coordinates for each aerodrome. In principle, the names from both tables (Regulations and Airports) must be consistent as it follows ICAO (International Civil Aviation Organization) nomenclature. Nevertheless, after performing the same fuzzy match as with ATC sectors, some errors are found in terms of orthography and nomenclature inconsistency. After ignoring those entries with Levenshtein distance bigger than one, a merge between both datasets may be performed with Pandas commands.

Lastly, after merging Regulations with both ATC Sectors and Airports geographical datasets, it is turn for the No. of Daily Flights information to be added. This last merge is actually done once the "Fill-up" procedure is completed. It simply consists on merging the filled Regulations table with the No. of Flights one by using as key column the "day".

c) Fill-up:

From the extracted data so far, only positive cases (when a regulation occurred) have been stored. Assuming that all delay issues have been properly reported, it can be stated that no regulation was happening for the missing time windows. Therefore, so as to prepare the dataset for data mining, it is convenient to fill-up the table with negative cases. By doing so, supervised learning may be performed since the training set contains all labeled data from both classes (negative and positive).

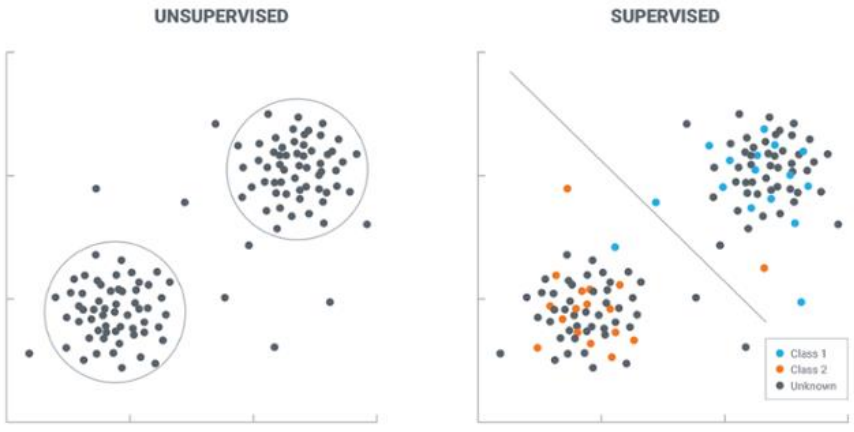


Fig.3.8. Unsupervised vs Supervised Learning [19]

Again, Pandas is used to loop through each ATC sector subset and fill-up the missing time windows with negative cases. Due to computational limitations, an efficient assumption is made so that all non-regulations periods last the mean regulation period. With this, the dataset goes from having 75k (all regulations) to 3.5 million (mixed) entries.

3.3. Data Transformation

Apart from the data extracted directly from the sources, there is more implicit information that may be of interest for the study. For this reason, the information available is translated into new fields. Not only that but also transforming the dataset values into a more suitable scale for the analysis is required.

a) Normalization:

Numerical variables do not usually belong to an absolute scale, so they have to be reduced to a common gradation. This is the case for the amount of regulations happening in a specific period of time. If two periods are compared in absolute amounts, the conclusions are not relevant since the direct relation between regulations and number of flights is distorting the analysis for the temporal variables. In order to achieve relevant results, when analyzing number of regulations, they are normalized with the number of flights happening in that same period of time so temporal variables effect is isolated.

Number of flights maximum level of granularity is in terms of daily quantities. This presents a limitation when variables are fragmented in smaller segments of time, like for the time of the day: night or day. As an estimation for this normalization to be feasible is taken: yearly average for 2016. In that year, a European Aviation Environmental Report was published by EUROCONTROL, EASA (European Aviation Safety Agency) and the European Environment Agency. Taking night time as set by the European Union: from 23:00 to 07:00, Figure 3.9 from report [20] reveals that 17% of the flights occur at night.

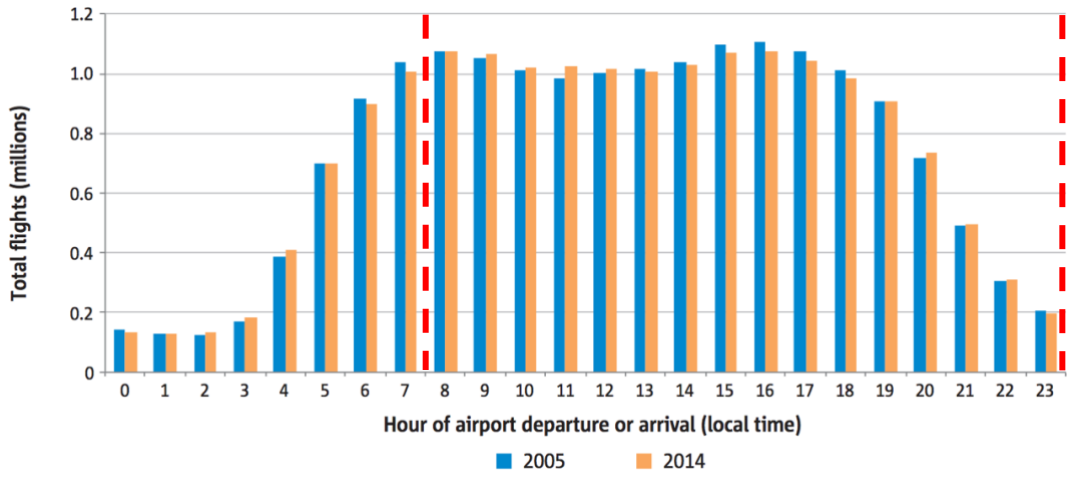


Fig.3.9. No. of Flights vs Hour of the Day [20]

b) Aggregation:

Based on the variables available from the original datasets, it is possible to generate new attributes by grouping or segregating values. This approach is taken for time variables generation from the date information available. By applying anonymous functions to each row with map function, the following fields are generated:

- year:

  Taking the startDate as the regulation date, the year is extracted directly as it is a built-in attribute for the datetime type. {2017 , 2018}

- semester:

  Taking the startDate as the regulation date, the semester is inferred from the month whether it belongs to the first six or not. {First , Second}

- season:

  Taking the startDate as the regulation date, the season is inferred from the month by identifying to which trimester they belong to. {Winter , Spring , Summer , Fall}

- month:

  Taking the startDate as the regulation date, the month is extracted directly as it is a built-in attribute for the datetime type. {January , February , March…}

- weekend:

  Taking the startDate as the regulation date, the day is classified whether it is weekend (6 and 7) or not. {1 , 0}

- weekday:

  Taking the startDate as the regulation date, the weekday is extracted directly as it is a built-in attribute for the datetime type. {Monday , Tuesday , Wednesday…}

- daytime:

  Taking the startDate as the regulation date, the time of the day is inferred from the hour whether it is day (23:00 – 7:00) or night (23:00 – 7:00). {Day , Night}

With reference to the geographical data, there is interest in finding the volume of influence for each Reference Location, so that weather data may be added to the study. For the case of En Route locations, the three dimensions have already been defined by the polygonal area plus the minimum and maximum flight levels. On the other hand, for

Airports, only the altitude of the aerodrome (lower bound) and the ARP (Airport Reference Point: geometric center of the runways) are determined. Then, to get the three dimensions, a proper area must be traced as well as an upper bound.

Taking advantage of PostGIS commands, ST_Buffer is applied so that a geometry covering all points within a given distance from the ARP is created. Following the definition of the Aerodrome Traffic Zone [21], taking the UK measurements as reference, the radius for the geometry is 2.5 Nautical Miles (about 4.6 kilometers). In the same manner, the upper limit is stated to be at 2000 ft (about 600 meters) from ground level.

Reversely, similar to the Airport Reference Points, it is convenient to obtain the centroids for the En Route reference locations for visualization and analysis. Again, from the PostGIS built-in functions, ST_Centroid may be used in order to calculate the geometric center for all ATC Sectors polygons.

Nevertheless, the geometries obtained for the ATC Sectors as well as the generated for the Airports do not perfectly fit in regular point grids as they are not squared. This information may be useful when working with external data sources as the already mentioned weather information. Once more, by using PostGIS and based on its spatial indexing, it is possible to retrieve the geometry of the bounding box of the supplied polygon by using the ST_Envelope function.

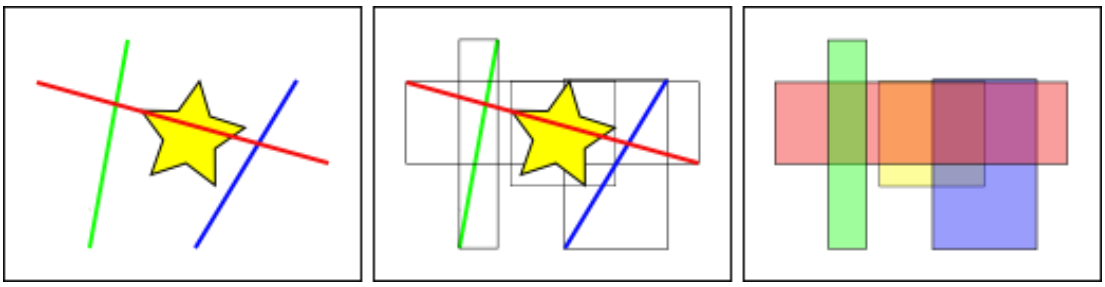


Fig.3.10. PostGIS Spatial Indexing [22]

4. Data Analysis

With all the information organized and clean, the exploration of the dataset is essential so as to gain a good comprehension of the features and potential issues. Not only that, but it will help in generating the hypotheses to be confirmed with a deeper analysis.

By further analyzing the dataset, it is possible to convert what originally was raw data and later was structured information, into knowledge that may serve to various applications and use cases. Lastly, all the retrieved understanding needs to be presented in a user-friendly manner, helped by the proper use of graphs and visualization tools.

4.1. Exploratory Data Analysis

Considered as a fundamental early step after data extraction and pre-processing, Exploratory Data Analysis (EDA) aims to assess the quality and relationship of the data before building models. Although most EDA techniques are graphical, since it is the most powerful and direct way to explore data, there are also quantitative non-graphical methods that provide insights about the variables of interest.

a) Non-Graphical:

To begin, it is relevant to find out the distribution of No. of Flights as it is the variable normalizing the number of regulations and may be useful for the modeling. With the help of Pandas "describe" function, descriptive statistics are generated in a table that informs about central tendency, dispersion and shape of data, this is shown in Table 4.1.

NUMBER OF FLIGHTS	
count	546
mean	29934
std deviation	4334
minimum	12033
quartile 25%	26781
quartile 50%	29857
quartile 75%	33455
maximum	38185

Table 4.1. No. of Flights Descriptive Statistics

Related to the average number of daily regulations, it is possible to normalize them by calculating the quantity per 1000 flights. Pandas has the "groupby" function devoted to data frames, which allows to differentiate between different periods of time as it might be 2017 and 2018. Moreover, with the "aggregate" function it is possible to perform several operations over the specified axis. In order to see the difference between both years of data, these are the numbers calculated:

```
Regulations per day per 1000 flights in 2017: 4.75
Regulations per day per 1000 flights in 2018: 4.95
```

After this quick analysis, it is stated that average daily regulations have increased from one year to another. A deeper exploratory analysis is made with a graphical approach so as to understand better the tendency on the amount of regulations.

b) Graphical:

As it has been previously addressed, to gain a better understanding on the continuous variables, graphical representation allows to detect hidden tendencies and characteristics. For the case of No. of Flights, a daily representation is made in Figure 4.1 to see if there are errors since the descriptive statistics revealed a difference of 16000 flights between the least and most crowded day. All EDA graphics are generated with matplotlib library for Python, which provides any format of 2D plot that may be needed.

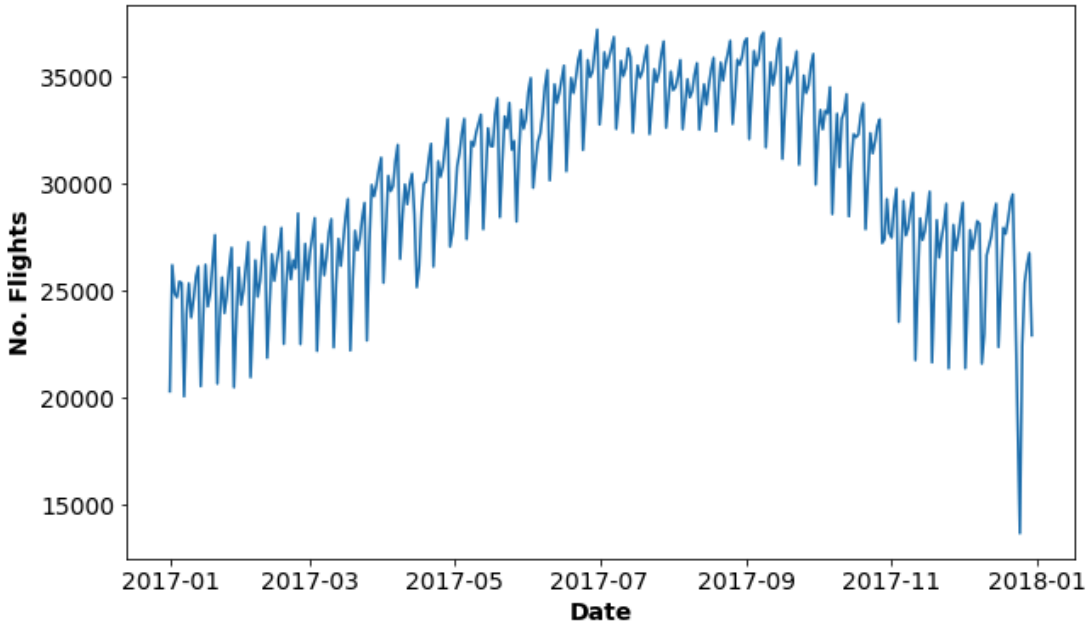
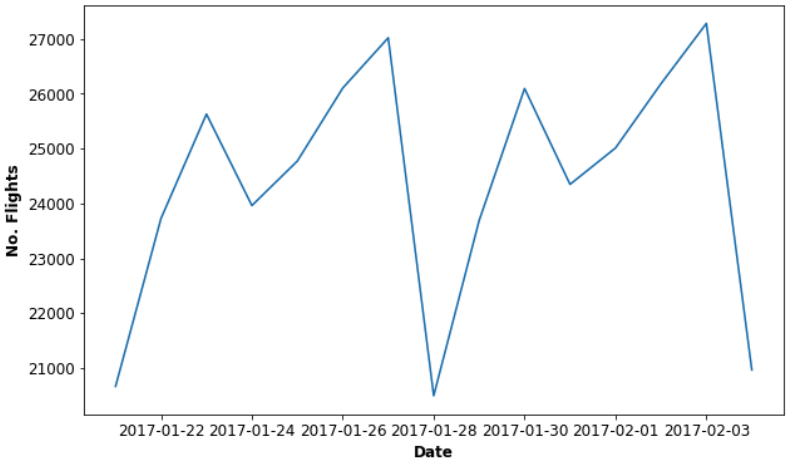


Fig.4.1. No. of Flights vs Date (2017)

From Figure 4.1, it may be seen that there are significantly more flights during the summer period, with the lowest phase in the winter months. With a simple exploration, the anomaly about the minimum number of flights is detected to be around late December. After researching, some news suggests that those days with lower flights may be due to a strike made by various airlines around Christmas days [23]. Further examination exposes a clear pattern that occurs periodically, in principle every week. Therefore, a zoomed representation is made in Figure 4.2 to better understand this cyclic tendency.



startDay	weekdayF
2017-01-21	Saturday
2017-01-22	Sunday
2017-01-23	Monday
2017-01-24	Tuesday
2017-01-25	Wednesday
2017-01-26	Thursday
2017-01-27	Friday
2017-01-28	Saturday
2017-01-29	Sunday
2017-01-30	Monday

Fig.4.2. Weekly Pattern for No. of Flights

With the zoomed picture and checking the days of the week, it can be identified that Saturdays are the least busy days of the week, while the amount of flights constantly increases, except for Tuesdays, where it suddenly decreases, reaching the weeks maximum on Fridays.

In terms of Regulations exploratory analysis, different pie charts as well as bar charts are represented due to the variables categorical type. Firstly, in Figure 4.3 the yearly distribution of the regulations shows that, as happened with the amount of flights, there is a noticeable increase of regulations during the summer, and a minimum is reached during the winter. It is concludable then, just by comparing Figure 4.1 and 4.3, that the quantity of flights and the amount of regulations are strongly related and have direct correlation. This is supportive to the fact that many delay issues are due to ATC Routeing and Capacity, which come from overdemand in a reference location.

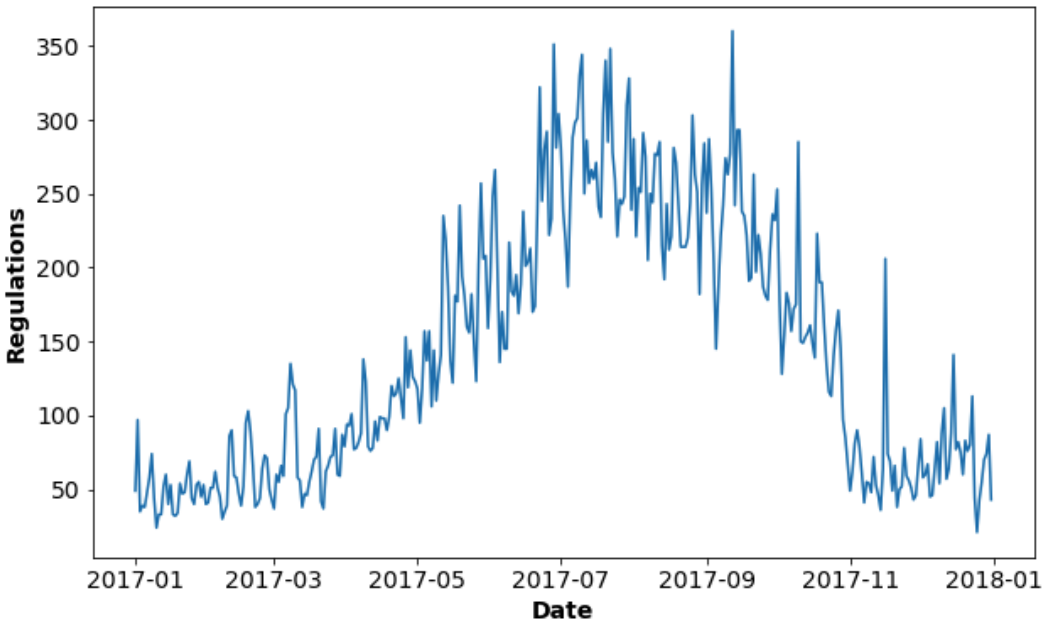


Fig.4.3. Regulations vs Date (2017)

So as to obtain some insight from the season of the year, it is plotted by counting the amount of regulations per flight occurring on each. For a clear representation, a pie chart is chosen, in which each slice is one season. From Figure 4.4, it can be established that there is a relationship between season of the year and the regulation. Summer, while being the busiest season, it is also the one with more regulations per flight, more than 40% of them. On the other hand, winter is by far the season with the lowest amount of delay issues, representing only a 12% of the regulations per flight in 2017.

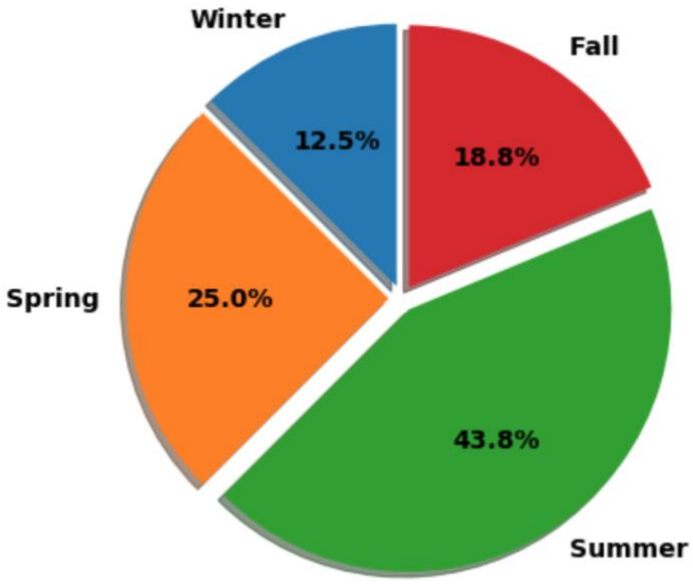


Fig.4.4. Regulations per Flight by Season (2017)

Moreover, distinction between workdays and weekends may be determinant on the study. To get a further perception about this issue, another pie chart is generated in Figure 4.5 with the amount of regulations whether on weekends or not. With this exploration, there is a clear difference on the amount of regulations, 85% of them happen during the workdays (they are also 5 out of the 7 days of a week).

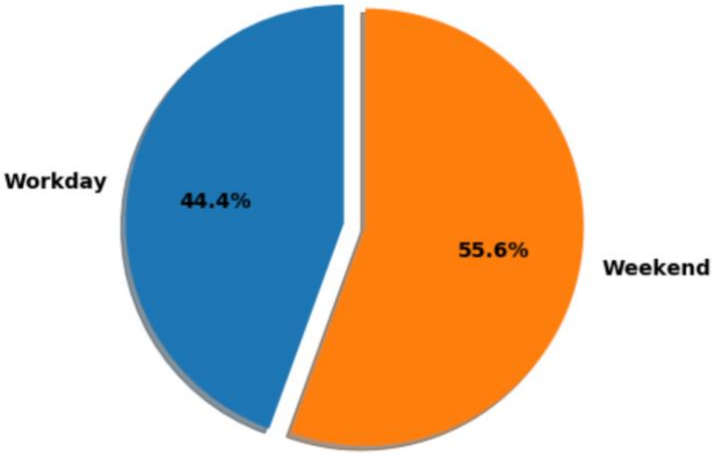


Fig.4.5. Regulations per Flight by Weekday Type (2017)

To be more specific, as it was shown with the first representations, there is an erratic behavior depending on the specific day of the week (Saturdays and Tuesdays decreases). For that, a bar chart is drawn with the amount of regulations per 1000 flights arising each day of the week throughout 2017. The result is addressed on Figure 4.6. It follows a similar structure to the week distribution of number of flights, with valleys on Tuesdays and Saturdays but, in this case, there is a higher density of regulations on Wednesdays.

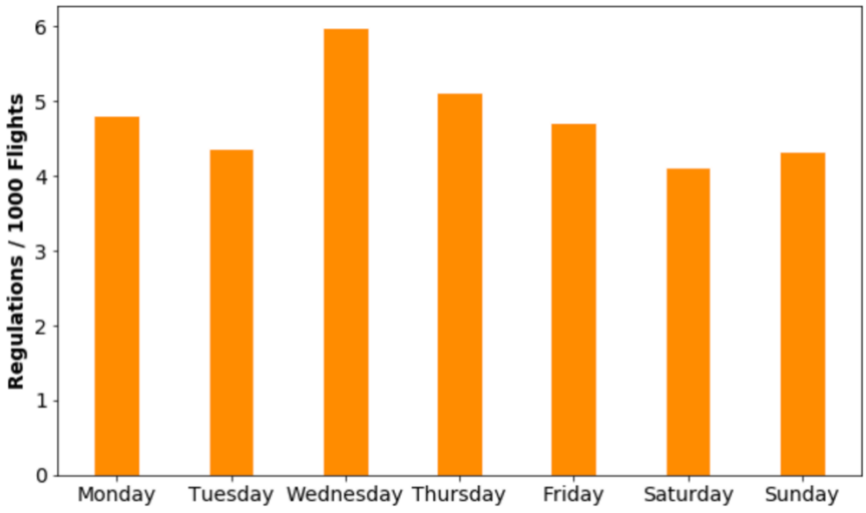


Fig.4.6. Regulations per 1000 Flights by Weekday (2017)

Comparable to the analysis for the weekend days, a quick examination for the daytime is done. Taking the estimation calculated in section 3.3, regulations are normalized with the proportion of flights happening at night (17%) and day (83%). The results shown in Figure 4.7 suggest that even though daytime condenses more flights per hour, it is at night when more regulations per flight occur, about 60%.

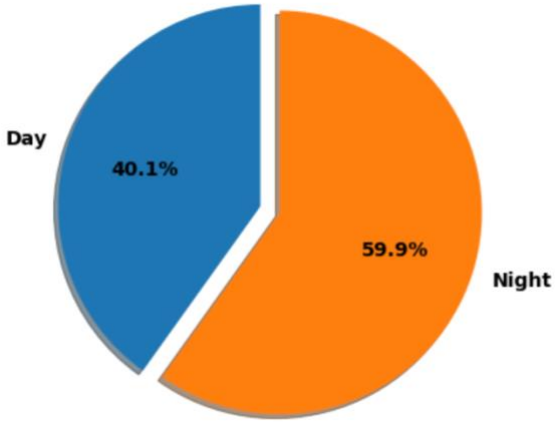


Fig.4.7. Regulations per Flight by Daytime (2017)

Then, if it was assumed that regulations have direct dependence on the amount of flights, there has to be additional reasons for night regulations to occur more often. Figure 4.8 brings some interesting insights when normalized with the average number of regulations by daytime. For the "Accident" reason, the vast majority occur during the day. Very similar situation for the "Airspace Management" reason, most of the aerial military activity happens with daylight. Opposite case for "Aerodrome Services" and "Non-ATC Industrial Action" that are more prevalent at night time, which may explain the higher density of regulations.

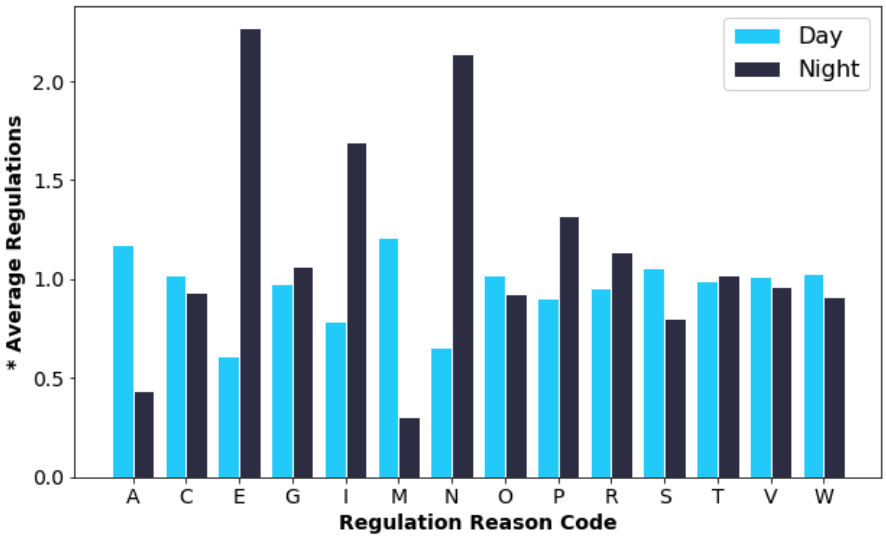


Fig.4.8. Regulation Reasons: Day vs Night (2017)

In terms of absolute values, Figure 4.9 gives a clear idea of the more numerous reasons for regulations. There is a clear predominance of "ATC Capacity" issues which constitute a 30% of the total. With almost a 24%, "ATC Routeing" constitutes a major part as well of the regulations that happened in 2017. "Weather" issues are also an important part for the delays in the ECAC region. From there, the rest of the reasons do not surpass the 10% of the total amount.

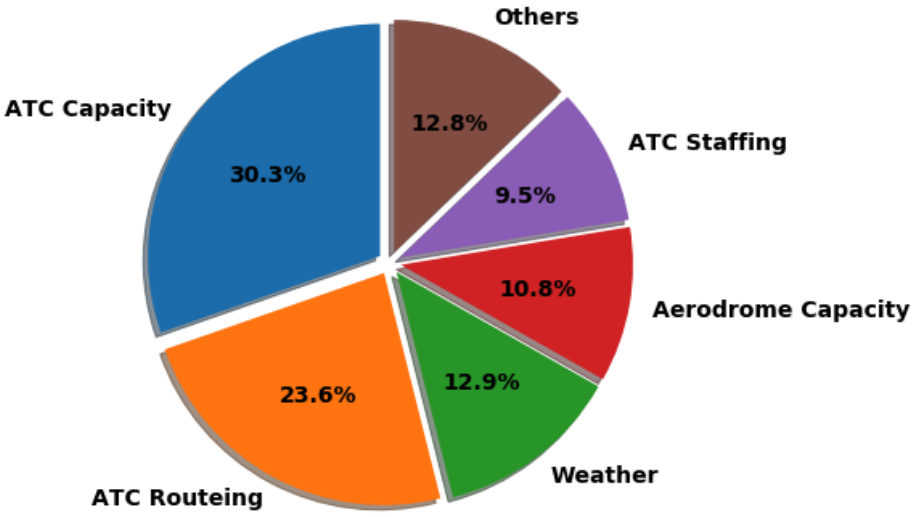


Fig.4.9. Regulations by Regulation Reason

Last relevant insight that is possible to extract from the dataset is the type of reference location where the regulations were imposed. For that, once more, a pie chart is plotted in Figure 4.10, differentiating between "En Route" and "Airport" reference locations.

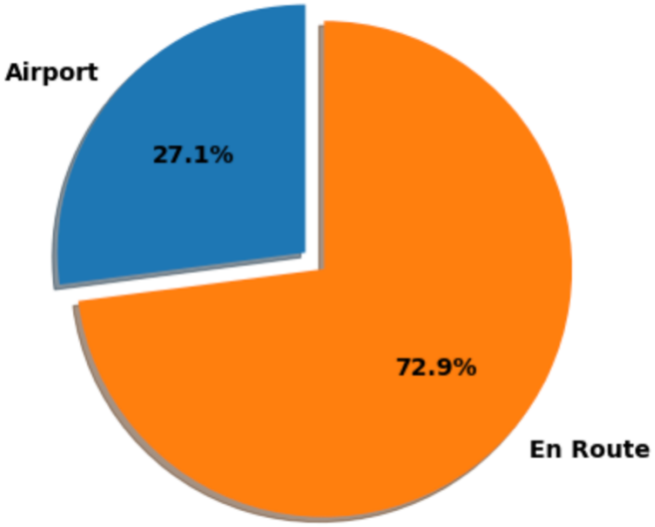


Fig.4.10. Regulation by Reference Location Type

4.2. Data Mining

After a better understanding of the variables with the Exploratory Data Analysis, the next step is to transform the dataset into business value. In this particular study, the aim is to get enough knowledge about regulations and temporal variables dependence, so they can be predicted in a future for each location. Due to this, general conclusions cannot be directly applied to all reference locations. Then, data mining must be done, i.e. a specific prediction model must be constructed, for each particular region.

For model validation, the dataset must be split into a train set and a test one first. Additionally, depending if the features selected are categorical or continuous variables, one model or other would work better. In the scope of this project, all variables used are categorical and the outcome is binary {0 , 1}. In terms of tools, the software chosen is the Waikato Environment for Knowledge Analysis. WEKA is a collection of machine learning algorithms for solving data mining problems. Its powerful user interface permits to train and test datasets by simply clicking.

a) Training:

Firstly, the proper model must be chosen. When looking at the expected outcome, 0 or 1, the problem is easily categorized as a classification one. Among the different classification algorithms, bearing in mind that there are not big amounts of data for each location, decision trees are discarded as they could not break down into many subsets without overfitting. Then, as all temporal variables are intended to be included in the model, the dimensionality may be considered relatively high. With that, Nearest Neighbor is discarded also as it needs big size samples to compare all dimensions.

Therefore, having a classification problem, with high dimensionality (7 features: daytime, weekday, weekend, month, season and semester) and not a large dataset, Naive Bayes Classifier seems the most appropriate and simple algorithm. This technique relies on the Bayesian theorem, which aims to revise existing predictions when adding new features to it. Additionally, Naive Bayes deals with the issue of having many more negative cases (no regulation) than positive ones (regulation) by combining the prior probability with the likelihood of belonging to a class. An example (from [24]) to understand it:

TEXT	TAG
"A great game"	Sports
"The election was over"	Not sports
"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Table 4.2. Naive Bayes Classifier Example

A classifier to determine whether a text is related to sports or not is to be built. For the training set, the 5 sentences used are shown in Table 4.2. With Naive Bayes, the intention is to label the sentence "A very close game". For that, the probability of the outcome is to be calculated from the words' frequencies. In order to do so, Bayes' Theorem states:

$$P(sports|a\,very\,close\,game) = \frac{P(a\,very\,close\,game|sports) \times P(sports)}{P(a\,very\,close\,game)}$$

Since the outcome is binary and the labeling is made based on which tag has a higher probability, the denominator would be common, and it may be discarded:

$$\begin{cases} P(a\,very\,close\,game|Sports) \times P(Sports) \\ P(a\,very\,close\,game|Not\,Sports) \times P(Not\,Sports) \end{cases}$$

However, from the sample given, there are no identical sentence to the one being trained so that no direct probability may be calculated. Here is where the Naive part takes part, which states that every sentence is treated as set of independent words. Thus, the probability for the positive outcome is:

$$P(a\,very\,close\,game|Sports) = P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports)$$

Applying the same principle to the negative tag probability, it is possible to calculate the combined probability for both labels. However, it may occur that some of the probabilities are zero when the word is not found in any text for a specific label. In order to solve that, Laplace smoothing is applied, i.e. every count starts at one instead of zero. So as to balance this, the total number of words is added to the denominator, so probability is never greater than one. Last step is just checking which label probability is bigger and that is the predicted outcome for the tested set.

Once the model is decided, a dataset for each reference location must be generated. However, due to computational limitations and the scarce amount of data, a limited amount of reference locations is taken as a sample. In order to get relevant results, three locations are taken from the total of 1664; those with the higher amount of data collected:

1) LECMDGU – 1408 positive cases (5503 instances)
2) LFFFTE – 1385 positive cases (5645 instances)
3) EDGG7 – 1055 positive cases (4041 instances)

After that, the datasets must be divided in two subsets each: one for training the model and another for testing it. Typically, the split of train/test is done as 80/20 or 70/30, for this study, a relation of 80% for the training subset and 20% for the testing one is chosen. Each dataset is then loaded in WEKA and is automatically split and trained by the system.

b) Testing:

While the training is automatically done, WEKA deploys the testing right after it is finished. With the 20% subset, the model is tested versus the actual output since all samples are labeled. Subsequently, WEKA displays on screen the statistical variables that may be relevant for the study. In this case, apart from the accuracy (correct vs incorrect), the recall and the precision are considered. The reason for this is that, in general, reference locations present a no-regulation status, so relative measurements become more relevant than absolute ones.

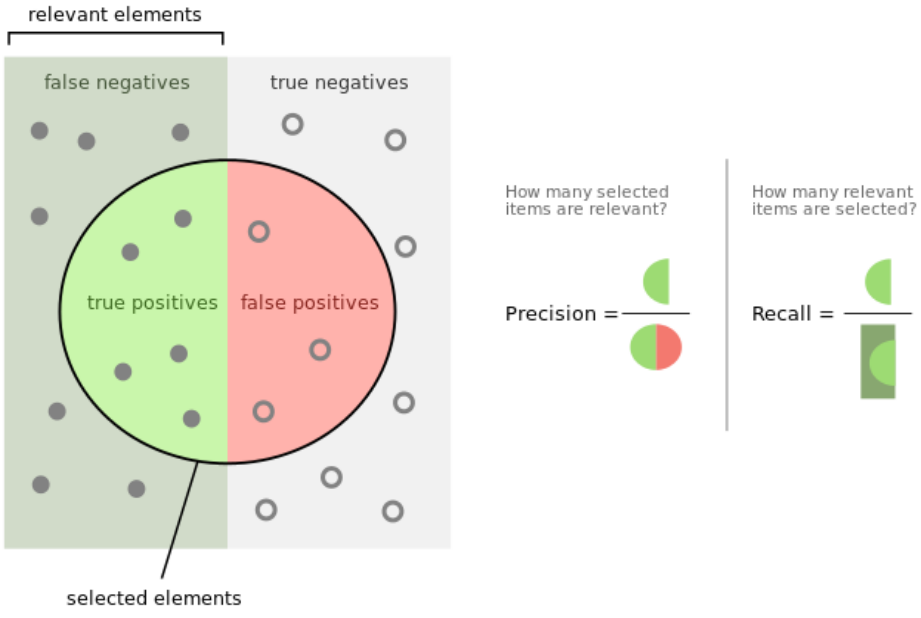


Fig.4.11. Precision and Recall [25]

To gain a better understanding on the results, an explanation of the evaluation metrics is given. Firstly, the accuracy is directly calculated from the overall results: number of instances correctly labeled over total amount of instances. Furthermore, the precision may be computed dividing the quantity of correctly labeled over the retrieved instances for each tag. On the other hand, recall or sensitivity, deals with the ratio of correct instances and the total amount of instances for a label. A graphical explanation of recall and precision may be consulted in Figure 4.11.

REF LOC	**ACCURACY**		**PRECISION**	**RECALL**
LECMDGU	79,1%	Regulation	0,66	0,55
		No Regulation	0,83	0,89
LFFFTE	75,1%	Regulation	0,54	0,39
		No Regulation	0,80	0,88
EDGG7	76,1%	Regulation	0,63	0,41
		No Regulation	0,79	0,90
AVERAGE	**76,8%**	**Regulation**	**0,61**	**0,45**
		No Regulation	**0,81**	**0,89**

Table 4.3. Evaluation Metrics for Naive Bayes Model

Keeping these parameters as the ones for validating the model, the results for the three sectors are assessed in Table 4.3. With reference to the accuracy, an average of 76,8% of the instances are properly classified. As it may be seen, recall results are not that great as the model would be predicting an average of 45% of regulations to occur in a region. Similarly, although higher, the precision is fairly low. An average of 61% of the instances classified with regulation label are actually regulations. With this in mind, it may be stated that the model is better than simply random guess (50% accuracy), thus, it adds value. However, the recall and precision results are not so promising, a fact that may be due to the lack of a bigger open dataset and the simplicity of the selected features.

4.3. Data Visualization

Information is understood better with images, that is why visualization is necessary when dealing with data analysis. Thus, presenting the predicted results in a pictorial format is the last step so that decision makers are capable of identifying issues and solving them. Doing the visualization interactive permits for a deeper comprehension at all times and further insights may be obtained, and more complex situations can be deciphered.

a) Numerical:

Analogously to the EDA, data from the input and output of the model may be useful to visualize through a set of different plots. WEKA, aside from the machine learning algorithms, contains a built-in display for quick graphical checks on the data. For the matter of this study, there are some representations concerning the sensitivity analysis, the most important is the ROC curve. It is a graphical plot that shows the True Positive Rate vs the False Positive Rate.

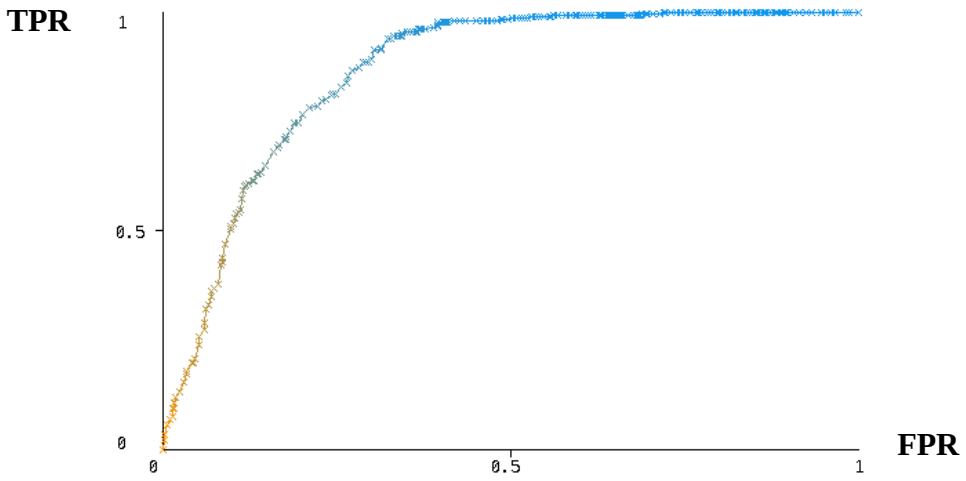


Fig.4.12. ROC Curve for LECMDGU

True Positive Rate is what was defined as recall for the analysis, also known as sensitivity. On the other hand, for the False Positive Rate, it is known as specificity. Therefore, a perfect classification would be on the top left corner of the graph. And the closer to that point, the better the classifier. We may conclude that for Figure 4.12 which is the representation for LECMDGU sector's model, the best method TPR would be around 0.5 and 0.6, that matches the 0.548 obtained mathematically. WEKA has the possibility to plot cost/benefit analysis as well as margin curves for the results of the model.

b) Geographical:

One of the key parameters about regulations is geographical situation; their locations. Then, a map representation seems appropriate for the study. Taking advantage of Carto's various uses, it is possible to generate interactive maps with PostGIS logic behind the dataset. With the results about regulations happening on each sector, it would be possible to perform relevant visualizations with the centroids for each reference location for example. It could also serve as a way of checking the sectorization at all levels in time.

Since Carto is built keeping SQL structure in the background, it allows to query the data and directly plot it in the map. As an example, it is possible to check all the ATC Sectors stored as FIR, which are the largest divisions of the airspace. The result from the query is shown in Figure 4.13.

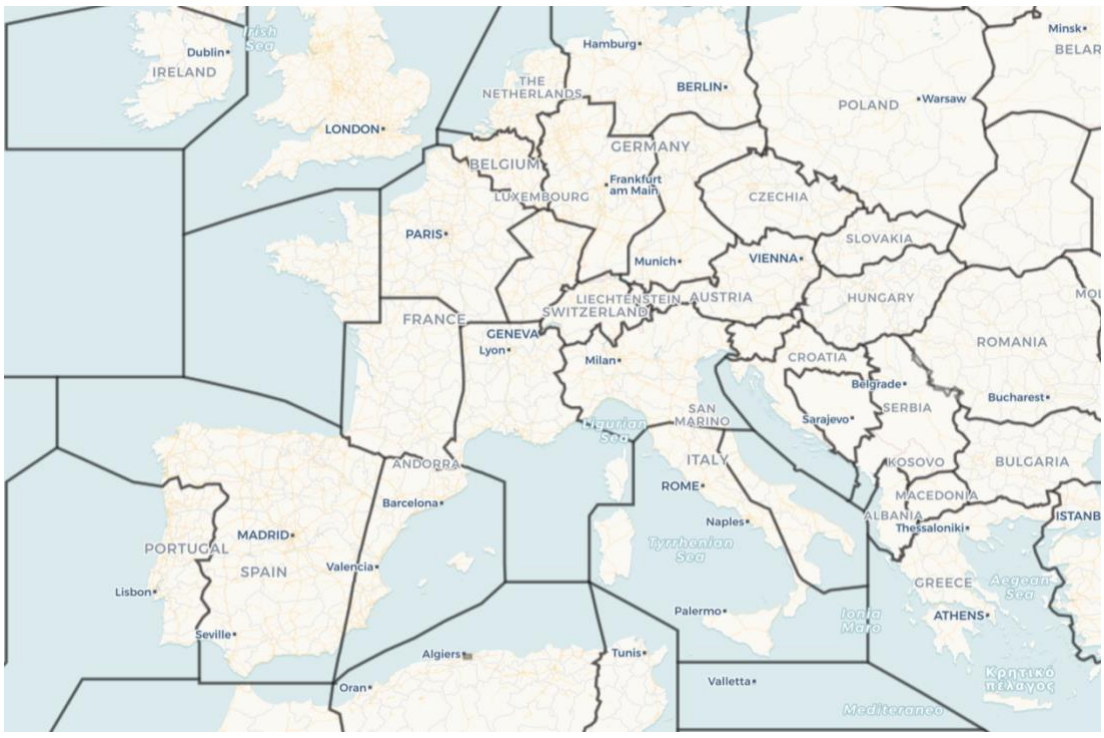


Fig.4.13. FIRs in Carto (from SQL query)

Moreover, Carto has several built-in widgets that are automatically generated from the dataset and serve as an instant filter for representation. In this case, it may be interesting to check the sectors available at each Flight Level or the regulations distribution for a specific period of time. In Figure 4.14, a widget for the minimum Flight Level is used and the result is shown in real time with the sectors that have that characteristic.

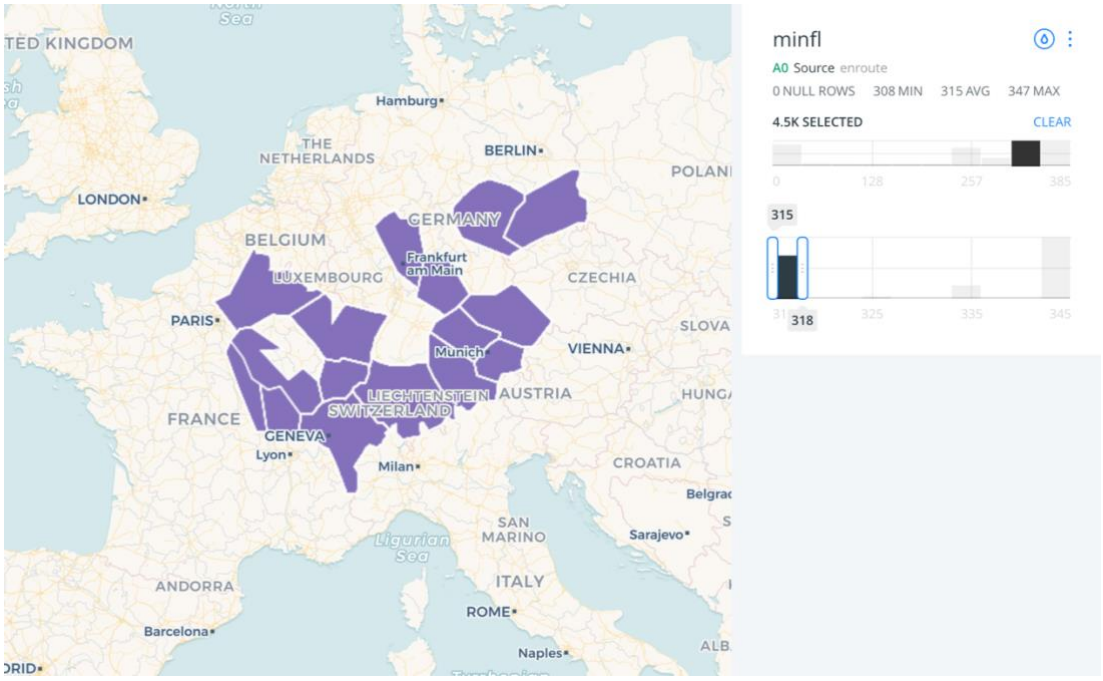


Fig.4.14. Minimum Flight Level Widget in Carto

Therefore, by plotting the centroids for each sector as well as the ARP for each airport, a live representation of regulations at any moment is feasible. There is an animation type for aggregating data that constructs a time-lapse with a heatmap style, which gives a quick idea of the status by regions. A screenshot of the animation is shown in Figure 4.15.

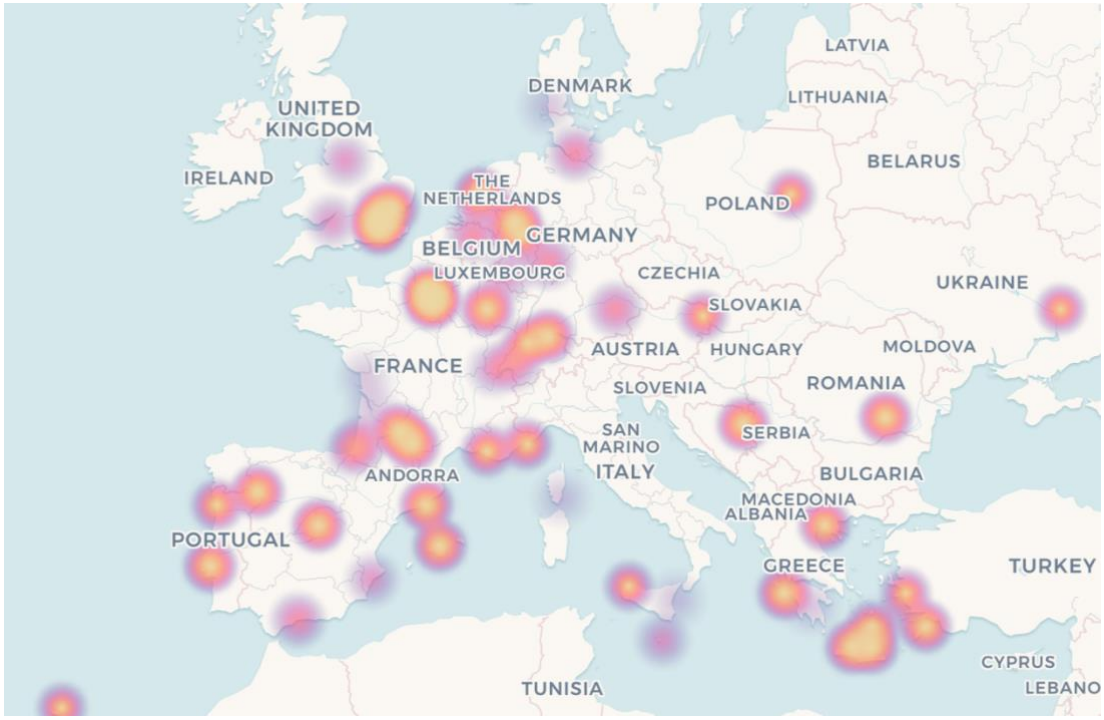


Fig.4.15. Regulations Heatmap in Carto

To make it more interactive and easier to retrieve the information, Carto contains a pop-up generator. With it, all the data for a specific regulation may be consulted on screen just by clicking on it. Figure 4.16 shows an example of a regulation at Adolfo Suarez Madrid-Barajas airport with all information associated to it.

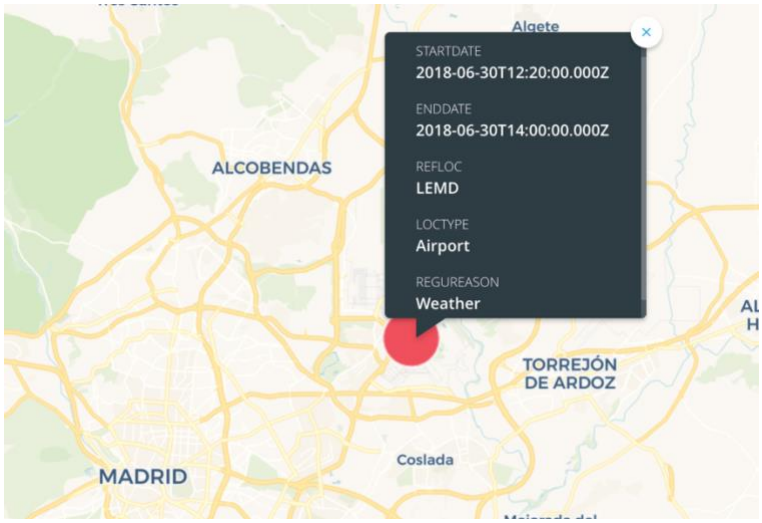


Fig.4.16. Pop-Up Information in Carto

5. Conclusions

5.1. Accomplishments

As stated in the objectives, the complete project follows a Big Data Analytics workflow. The main steps trailed have been the ones that were established, while adding some subtasks as data normalization and Exploratory Data Analysis. It has been shown that following the indicated timeline is crucial as each task depends on the previous ones, i.e. Exploratory Data Analysis has no real value if the graphs have not been normalized beforehand. Therefore, after the completion of the thesis, a full guide for the stages of a Big Data Analytics work has been successfully built.

In terms of data acquisition, taking information from different sources, a Data Inventory for ATFM Regulations is completed. Looking for open data only, a reasonable number of features have been obtained directly from the sources while others needed to be extracted with web scraping. With all the tasks of transforming and integrating data, not only the temporal values that were used but also geographical information about reference locations as well as number of daily flights are put together. This gives a varied and very rich set that can be used for diverse researches in the ATM field.

With reference to the tools used for all the stages, the objective was to look for open source software that may be universally used. Working with Python in Jupyter Notebook has proven to be a very efficient manner to deal with big datasets due to Pandas gigantic set of functions. Moreover, with the 1-month free trial of Carto, it has been possible to manipulate geographical data as well as visualize it with many add-ons available. In terms of additional tools needed, AWK has been utilized for text transformation into table, and some JavaScript for the web scraping. Although it is not recommended for big amounts of data, Excel has also proven to be convenient when manipulating certain subsets. For the analysis part, WEKA has been a time-saving machine learning tool thanks to its simplicity and efficiency.

Concerning the main objective of the project, a prediction model has been generated. Producing the desired temporal fields (season, month, time of the day…), a Naive Bayes classifier has been implemented in WEKA to the most crowded reference locations. The results have shown that in average, the model predicts better than random guessing, so that it has some added value. Nevertheless, the recall for regulations did not prove a great efficiency of the predictor and further work needs to be done so results ameliorate.

About the visualization tool, Carto delivers a complete set of geographical functions that could derivate in becoming a useful interactive display for stakeholders. With the multiple functionalities, mixed with better results coming from the model, it could be a powerful tool for decision making for the FMP and the CFMU operators. Nevertheless, due to the complexity of the project, only historical data has been manipulated in Carto as a way of showing its capabilities.

5.2. Limitations

While the project aimed to encounter a prediction model for each reference location, it has not been possible to achieve so. First issue that was found had to do with the sources. Organizations as EUROCONTROL have realized that there is value in data and that may translate into monetary value. For that reason, even with the student access, the data repository of EUROCONTROL was very restricted in terms of downloading. Therefore, by taking data from public reports, a very small representation of regulations was acquired, only for 2017 and the first half of 2018.

Regarding to the same limitation, since small quantities of data were available, predictive model needed to be simple and robust. Naive Bayes is a widely used algorithm that performs great at tasks as text mining. However, in this specific application, since the temporal variables used were all categorical, no correlation was considered for training the model, thus, a very modest result was obtained. Not only that, but as regulations are not equally distributed among the different reference locations, some of them did not have enough instances to be evaluated and only the three more crowded ATC sectors have been analyzed. The results then are still not great but at least reasonable to validate the model.

5.3. Future Directions

Having in mind the opportunities and limitations discovered by this project, several subsequent projects may be proposed. To begin with, data accessibility would be a field to research on. As explained previously, there is data available in the web but with plenty of difficulties to access them. Thus, as a way of improving the existing dataset, broader data sources, with more years of information available, must be encountered so that predictive models can get more complex and reach a higher amount of reference locations.

Furthermore, new fields of research may bring more insight for ATFM regulations prediction. This could be the case of adding weather phenomena variables to the dataset. By doing so, maybe not all regulations are better predicted but some correlation with delay issues regarding weather may be found. Nevertheless, these variables tend to be very complex and they require a deep research and lengthy manipulation of the data.

5.4. Project Management

For the full development of the project, an estimation of the budget is presented in Table 5.1 as well as an indication of the methodology followed in Figure 5.1.

Item	Quantity	Unitary Price	Subtotal
Workforce Costs			
Junior Engineeer	150 hours	50 €/h	7.500 €
Senior Engineer Supervisor	20 hours	250 €/h	5.000 €
Software			
Carto	1	1 Month Free Trial	0 €
Microsoft Excel	1	Univ. License Soft.	0 €
Jupyter Notebook	1	Free	0 €
WEKA	1	Free	0 €
Hardware			
Computer Amortization	1	20 €/month	120 €
Communications (Wi-Fi, telephone)	1	5 €/month	60 €
Other Expenses			
Transportation	12 meetings	10 €	120 €
Others			80 €
TOTAL			**12.880 €**

Table 5.1. Project Budget

TASK NAME	SUB TASK	DURATION	March-18	April-18	May-18	June-18	July-18	Aug-18	Sept-18
Documentation		120 days							
Objetives definition		60 days							
State of the Art study		30 days							
Data extraction		60 days							
Data Pre-processing									
	Data Cleansing	60 days							
	Data Integration	60 days							
	Data Transformation	30 days							
Data Analysis									
	Exploratory data analysis	30 days							
	Data mining	60 days							
	Data Visualitations	60 days							
Conclusions		30 days							
Complete project execution		210 days							

Fig.5.1. Project Gantt Chart

References

[1] "Pessimistic Sector Capacity Estimation," *EUROCONTROL*, Nov., 2003, pp. 3-5 [Online]. Available: https://www.eurocontrol.int/sites/default/files/library/026_Pessimistic_Sector_Capacity.pdf
[Accessed on: Jun. 20, 2018]

[2] "EUROCONTROL in SESAR", *EUROCONTROL*, Feb. 3, 2010 [Online]. Available: https://www.eurocontrol.int/publications/eurocontrol-sesar
[Accessed on: Jun. 20, 2018]

[3] Rodrigo Marcos *et al.*, "Visual Analytics and Machine Learning for Air Traffic Management Performance Modelling," presented at the 6th SESAR Innovation Days, TU Delft, Netherlands, Nov 8-10, 2016 [Online]. Available: https://www.sesarju.eu/sites/default/files/documents/sid/2016/SIDs_2016_paper_36.pdf
[Accessed on: Jun. 25, 2018]

[4] David Gianazza, "Forecasting workload and airspace configuration with neural networks and tree search methods," Artificial Intelligence, Elsevier, 2010, 174 (7-8), pp. 530-549 [Online]. Available: https://hal-enac.archives-ouvertes.fr/hal-01020725/document
[Accessed on: Jun. 25, 2018]

[5] Karthik Gopalakrishnan and Hamsa Balakrishnan, "A Comparative Analysis of Models for Predicting Delays in Air Traffic Networks," Department of Aeronautics and Astronautics, MIT, Cambridge, MA, USA, 2017 [Online]. Available: http://www.mit.edu/~hamsa/pubs/GopalakrishnanBalakrishnanATM2017.pdf
[Accessed on: Jun. 25, 2018]

[6] DDR2 Reference Manual for General Users, v.2.9.5, EUROCONTROL, Brussels, Belgium, 2018, pp. 7 / 89-91 [Online]. Available: https://ext.eurocontrol.int/ddr/files/documentation/ddr2_userguide_generic.pdf. [Accessed on: Jul. 5, 2018]

[7] "Post-operations performance adjustment process," *EUROCONTROL*, 2018. [Online]. Available: https://www.eurocontrol.int/publications/post-operations-performance-adjustment-process. [Accessed on: Jul. 5, 2018]

[8] "Traffic volume set - EUROCONTROL ATM Lexicon," *EUROCONTROL*, 2011. [Online]. Available: https://ext.eurocontrol.int/lexicon/index.php/Traffic_volume_set. [Accessed on: Jul. 5, 2018]

[9] "Traffic volume - EUROCONTROL ATM Lexicon," *EUROCONTROL*, 2014. [Online]. Available: https://ext.eurocontrol.int/lexicon/index.php/Traffic_volume. [Accessed on: Jul. 5, 2018]

[10] "Reference location - EUROCONTROL ATM Lexicon," *EUROCONTROL*, 2014. [Online]. Available: https://ext.eurocontrol.int/lexicon/index.php/Reference_location. [Accessed on: Jul. 18, 2018]

[11] ATFCM User Manual, 22nd ed. EUROCONTROL, Brussels, Belgium, 2018, pp. 48 / 93-94 [Online]. Available: https://www.eurocontrol.int/sites/default/files/content/documents/nm/network-operations/HANDBOOK/atfcm-users-manual-current.pdf. [Accessed on: Jul. 18, 2018]

[12] "Demand Data Repository (DDR2) - Historical Page," *EUROCONTROL*, 2018. [Online]. Available: https://ext.eurocontrol.int/ddr/historicaltraffic. [Accessed on: Jul. 29, 2018]

[13] "Demand Data Repository (DDR2) - Dataset Page," *EUROCONTROL*, 2018. [Online]. Available:
https://ext.eurocontrol.int/ddr/datasets.
[Accessed on: Jul. 30, 2018]

[14] "Open data downloads," *OurAirports*, 2018. [Online]. Available:
http://ourairports.com/data/.
[Accessed on: Jul. 30, 2018]

[15] Vertica Documentation. HPE Vertica Analytics Platform. Vertica, 2018, pp. 1615-1616 [Online]. Available:
https://www.vertica.com/docs/8.0.x/PDF/Vertica_8.0.x_Complete_Documentation.pdf.
[Accessed on: Aug. 10, 2018]

[16] "AWK Workflow," *Tutorialspoint*, 2018. [Online]. Available:
https://www.tutorialspoint.com/awk/awk_workflow.htm.
[Accessed on: Aug. 10, 2018]

[17] Antonino Ingargiola, "What is the Jupyter Notebook?," *Jupyter/IPython Notebook Quick Start Guide*, 2015. [Online]. Available:
https://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html#.
[Accessed on: Aug. 10, 2018]

[18] N. Babar, "The Levenshtein Algorithm," *Cuelogic Blog*, Jan. 25, 2017. [Online]. Available:
https://www.cuelogic.com/blog/the-levenshtein-algorithm/.
[Accessed on: Aug. 21, 2018]

[19] SQRRL Team, "An Introduction to Machine Learning for Cybersecurity and Threat Hunting," *Threat Hunting Blog*, June 16, 2016. [Online]. Available:
https://sqrrl.com/an-introduction-to-machine-learning-for-cybersecurity-and-threat-hunting/.
[Accessed on: Aug. 21, 2018]

[20] "European Aviation Environmental Report 2016," *EEA, EASA and EUROCONTROL*, Rep. TO-01-15-323-EN-N, 2017, pp. 15 [Online]. Available: https://ec.europa.eu/transport/sites/transport/files/european-aviation-environmental-report-2016-72dpi.pdf.
[Accessed on: Aug. 27, 2018]

[21] "Aerodrome Traffic Zone (ATZ) - Definition", *Skybrary*, 2016. [Online]. Available: https://www.skybrary.aero/index.php/Aerodrome_Traffic_Zone_(ATZ).
[Accessed on: Aug. 27, 2018]

[22] Mark Leslie, "Intoduction to PostGIS - Section 8: Spatial Indexing," *OSGeo*, 2009 [Online]. Available: http://revenant.ca/www/postgis/workshop/indexing.html.
[Accessed on: Sep. 3, 2018]

[23] The Local, "Iberia staff announce Christmas strike," *The Local Europe AB*, Dec. 12, 2017 [Online]. Available: https://www.thelocal.es/20171212/iberia-staff-announce-christmas-strike.
[Accessed on: Sep. 3, 2018]

[24] Walber, "Precision and recall", *Wikimedia Commons*, 2014. [Online]. Available: https://commons.wikimedia.org/wiki/File:Precisionrecall.svg.
[Accessed on: Sep 6, 2018]

[25] B. Stecanella, "Machine learning. A practical explanation of a Naive Bayes classifier," *MonkeyLearn Blog*, May 25, 2017. [Online]. Available: https://monkeylearn.com/blog/practical-explanation-naive-bayes-classifier/.
[Accessed on: Sep. 6, 2018]

* Pandas: powerful Python data analysis toolkit, Release 0.23.4, Wes McKinney and PyData Development Team, Aug. 6, 2018 [Online]. Available: https://pandas.pydata.org/pandas-docs/stable/pandas.pdf. [Accessed on: Sep. 16, 2018]

* Joel Grus, *Data Science from Scratch*. 1st ed., Sebastopol, CA, USA: O'Reilly, 2015.

* MIT Critical Data, *Secondary Analysis of Electronic Health Records*. Cambridge, MA, USA: Spring, 2016 [Online] Available: https://link.springer.com/content/pdf/10.1007%2F978-3-319-43742-2.pdf [Accessed on: Sep. 20, 2018]