Traffic Jam in the Air: Visualizing and Predicting US Airline Traffic Delays with Tableau and Exploratory

Mohamed Amir Omezzine
George Mason University
Volgenau School of Engineering
Fairfax, Virigina
Email: momezzin@gmu.edu

Wei Cao
George Mason University
Volgenau School of Engineering
Fairfax, Virigina
Email: Wcao2@gmu.edu

ZhiYu Tian
George Mason University
Volgenau School of Engineering
Fairfax, Virigina
Email: ztian5@gmu.edu

***Abstract*—Flight delay is a problem that occurs everyday in US airports. In 2007, nearly one in four airline flights arrived at its destination over 15 minutes late (Ball et al, 1). A flight is considered delayed when it arrived 15 or more minutes than the schedule. The air congestion is increasing daily which reached more than 87,000 flight per day across the United state in one day (Miller,1).Increasing travel delays varies from one airport to another in the US. Different causes such as security, weather and technical problems, are the main reasons for such delays. This paper analyzes a variety of delays causes for more than 30 airlines in the United States. In addition, an analysis of different type of data like departure delay time,arrival delay time,origin and number of cancellation due to weather, will answer different questions for travellers around the United States.Finally, time series modelling for 3 airlines was done to predict delays.**

1. Introduction : Problematic

The airline industry is characterized by the complexity of its operations. So when a delay occurs on a flight, the impact of it can become very important, even to create an effect snowball on all operational activities and resources involved. A delay can be caused by mechanical breakage ,bad conditions weather, or security reasons. its effect is not negligible and must be absorbed as soon as possible in order to limit the consequences. The options, in order to restore the situation when a delay occurs, are very restricted. So, the first is to catch up by accelerating the affected flight and the second is to take no action and therefore, to assume all the costs that result. According to Airlines.org, In 2017, the average cost of aircraft block (taxi plus airborne) time for U.S. passenger airlines was $68.48 per minute, 7.4 percent more than in 2016. In addition, a traveller today faces a lot of delays especially during summer. According to Wilson, more than half (%52) of airports have more summer than winter delays, although both seasons averaged an on-time rate of %77.1 for the airports we reviewed. By then, no detailed guide is provided to the traveller to give him knowledge about the best time to travel. At the same time, airports, airlines, and even the economy of the United States is suffering. The Schumer report estimated that in 2007, airport delays cost about 40.7 billion dollars to the US economy. The objective of the project that is presented is to develop a set of tools for analysis and assistance to the decision of a traveller and airports can make better decisions in operational and financial aspects of flight management with delays. In addition, this report analyzes a variety of cost components caused by flight delays, including cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The first motivation of this project is related to personal experiences. As the members of this project are international students, they faced the real problem more than once during their flights and most importantly in the busiest airports in the United States such as Ohare Chicago and Dulles Washington DC airports. Since flying schedule is so important to any traveller, a flight delay may cause the traveller to reschedule or even worse, miss a transaction flight. For that cause, its important to have an idea of how much time was wasted on flight delays, which airline have a reputation of the most delayed flight or which time of the year is better to avoid flying and find alternatives as a mean of transportation. All these questions will be answered according to a detailed analysis of the data of all US airlines arrival times, departure, reasons of delays etc. All this data is provided by the BTS (Bureau of Transportation Statistics) and after different and multiple research about the projects problem, its discovered that flight delays is a serious issue where different articles and papers analyze it.

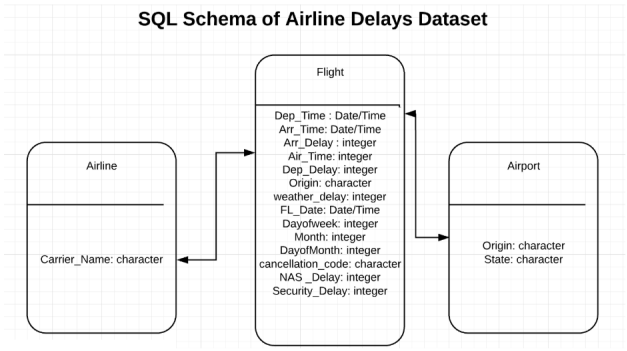
2. Dataset Description

The dataset consists of 29 variables with multiple types like numerical, categorical, and boolean. The data describes flight arrival and departure details for all commercial flights within the USA from 1987 to 2008. For this project, only year 2008 will be picked.There are more than 7 million records in total. Data mining analysis was applied on the dataset and a subset of the actual dataset was done. The new dataset consists of 17 variables will be explained further in this project.

2.1. Data Source

The dataset is extracted originally from two sources. The first source is the Bureau of Transportation Statistics and the other source is US Department of Transportation. Both datasets were published in Kaggle.com website. According to BTS, The Bureau of Transportation Statistics (BTS) is a politically objective supplier of trusted and statistically sound baseline, contextual, and trend information used to shape transportation policy, investments, and research across the U.S. and abroad(BTS webpage). BTS is a relevant source for aviation data, transportation economics, and motorways. The US Department of Transportation is a federal Cabinet department of the U.S. government concerned with transportation. The top priorities of DOT are are to keep the traveling public safe and secure, increase their mobility, and have our transportation system contribute to the nation's economic growth (DOT Website). One of the activities DOT focuses on is transportation statistics. Multiple datasets about airlines, airports, infrastructures, and passenger travel are organized and analyzed by DOT with correlation with BTS. The final dataset after the processing and mining consists of multiple metadata. According to Rouse, Metadata is data that describes other data(Rouse,1). In this project, the different types of data are summarized in an SQL ER diagram (Diagram 1). The first part of the schema is about flight info. There are 14 columns: Dates: consists of day of week, month, date, and year Arrival and departure time in hours, minutes, and seconds Time of delay in minutes: Departure delay and arrival delay Type of Delay: Weather delay, security delay and NAS delay Cancellation code: consists of a code of A,B,C,and D for 4 types of cancellation. (A = carrier, B = weather, C = NAS, D = security) Origin and destination The second part is the airport. There are 2 columns:
Origin: The airport name. In this project, the most important and busiest airports will be focused on. Example: O'hare airport Illinois(ORD), Los Angeles airport (LAX), Hartsfield-Jackson Atlanta International Airport (ATL) and others. In this dataset there are 304 airports. Their performance will be visualized in some graphs. Then the focus will switch on the busiest airports depending on the number of flights per year and AirTime. State: The state of which the airport exists. The final part is Airline: CarrierName: it consists of the US airlines existed in the datasets. Some examples are American, Aloha Air,United, US Air, Southwest etc.
There is a code for each airline. The codes are:
-9E Endeavor Air Inc.
-AA American Airlines Inc.
-AS Alaska Airlines Inc.
-B6 JetBlue Airways
-DL Delta Air Lines Inc.
-EV ExpressJet Airlines Inc.
-F9 Frontier Airlines Inc.
-G4 Allegiant Air
-HA Hawaiian Airlines Inc.
-MQ Envoy Air
-NK Spirit Airlines
-OH PSA Airlines Inc.
-OO SkyWest Airlines Inc.
-UA United Air Lines Inc.
-VX Virgin America
-WN Southwest Airlines Co.
-YV Mesa Airlines Inc.
-YX Republic Airlines

Diagram 1



3. Literature Review

3.1. Literature on Delay Analysis

The airline industry in the United States plays an important role in the US economy. The increase of delays in major airports in United States has been the subject of studies in recent years. The literature on delay analysis and its potential remedies extends back over several decades. According to Martin et al., Levine discussed in 1969 about this subject saying that pricing is a better means of allocating scarce airport capacity to meet the demand than other mechanisms being considered at the time, such as slot allocation (Martin et al , from Levine 1969). The Federal Aviation Administration (FAA) and the Bureau of Transportation Statistics describes the increase in delays and cancellation flights from 2003 to 2018. During that period, the data provided by the BTS showed that more than 20% of all flights between 2003 and 2018 were delayed. The BTS classified the causes of delays. There are 4 causes: Air carrier : the cause of the cancellation or delay was due to circumstances within the airline's control (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.). Extreme weather: Significant meteorological conditions (actual or forecasted) that, in the judgment of the carrier, delays or prevents the operation of a flight such as tornado, blizzard or hurricane. National Aviation System (NAS): Delays and cancellations attributable to the national aviation system that refer to a broad set of conditions, such as non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control. Late-arriving aircraft: A previous flight with same aircraft arrived late, causing the present flight to depart late. Security: Delays or cancellations caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening

equipment and/or long lines in excess of 29 minutes at screening areas. According to FAA, extreme weather is the main cause of flight delays. 62% of delays are caused by weather during the period of 2003 and 2018. Allan et al. examined weather delays at New York major airports from September 1998 through August 2000. During that period the FFA implemented an Integrated Terminal Weather System (ITWS) which provides improved integration of weather data into timely, accurate aviation weather information.
This is an essential component in reducing delays and improving National Airspace System (NAS) capacity use while enhancing aviation safety (FFA).The methodology used in the study has considered major causes of delays (convective weather inside and well outside the terminal area, and high winds) that have generally been ignored in previous studies of capacity constrained airports such as Newark International Airport (EWR). Allan et al. research found that the FAA air traffic personnel at the NY TRACON and towers using ITWS/TCWF is providing primary delay reductions of over 27,000 hours, with downstream passenger savings of an additional 22,000 hours, resulting in a total monetary savings (using standard FAA values for airline direct operating costs and passenger time costs) of over $150,000,000 per year. In addition, the research have found that the usual paradigm of assessing delays only in terms of Instrument Meteorological Conditions (IMC) and Visual Meteorological Conditions (VMC) conditions and the associated airport capacities is far too simplistic as a tool for determining which air traffic management investments will best reduce the avoidable delays.

3.2. Review on Methodology of Flight Delay Analysis

Suzuki (2000) proposed a a new method of modeling the relationship between on-time performance and market share in the airline industry. The idea behind the method is that the passengers decision to remain (use same airline) or switch (use other airlines) at time t depends on whether they have experienced flight delays at time t1 or not. Sternberg et al. present a thorough literature review of approaches used to build ight delay prediction models from the Data Science perspective. The research is focused on proposing a taxonomy and summarizing the initiatives used to address the ight delay prediction problem, according to scope, data, and computational methods, giving particular attention to an increased usage of machine learning methods. Ball et al. focused on the cost of delays. The research analyzes a variety of cost components caused by flight delays, including cost to airlines, cost to passengers, cost of lost demand, as well as the indirect impact of delay on the US economy. The research team employs a statistical cost estimation methodology to estimate how delays affect airline costs. In addition, the research focuses on decision making when a traveller is choosing an airport for a connecting flight. Avoiding the major hubs by using smaller airports will help you to avoid flight delays. They claim that secondary airports are mostly less congested and therefore they are less prone to flight delays.

3.2.1. Conclusion of Review. Data and statistical analytics models and simulation models are used to analyze the issue of flight delays. Since flights data size is huge, its a challenge to perform a perfect analysis. Using different resources and several published work, this paper will focus on analyzing the flight delays data of the 2008 year and a time series model will be build through a 10 year period. A comparison will be established with the existing data of the year of 2018. In addition, this paper will develop a clear conclusion to the traveller concerning flight delays and answering some major questions to avoid it. This paper will detect the pattern of delay from the airport level in which delays occur, give basic statistics on their magnitudes and frequencies. Besides, major airports in the United States will be under the microscope regarding the delay time, the major causes of delays, and the worst time of delays in comparison the month of the year/season. The data used will be described in the proposed approach section.

3.3. Proposed Approach

The methodology of this project is based on analyze and visualize. The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. The source of the dataset is is the Bureau of Transportation Statistics and the other source is US Department of Transportation. For this project we will analyze only the data from the year 2008 and through the project work the team will develop a statistical model for the data (potentially a time series regression model). Since the dataset is large (nearly 120 million records in total for total size of 1.6 gigabytes) the team will use Tableau, Exploratory and Rstudio for analysis work and visualization. The different types of data are summarized in an SQL ER diagram (Diagram 1). The first approach done is data cleaning and preparation for descriptive analysis. The dataset contains some attributes and fields that the team think are unnecessary such as tail number for airplanes and flight number. These columns don't add any perspective to this project and only important columns were selected in the project such as CarrierName, Type of Delay, Arrival and departure time in hours, minutes, and seconds ,Time of delay in minutes etc. During the process of data cleaning, Tableau and Rstudio were used as tools. In the second steps, the team is including other databases such as weather database for certain cities since weather is a main cause of delays and airports database which describes the locations of US airports, with the fields :iata which is the international airport abbreviation code,name of the airport,city and country in which airport is located, and the latitude and longitude of the airport. Since the list of airports is big, the team will focus on the top 10 airports in the United States in different graphs. Detailed graphs will be shown and explained in the final report. The team will develop graphs to understand which airline has the less delays by extracting the time

of departure and arrival of multiple flights of each airline. Besides, since the data provides the causes of delays for each airline, the team will develop graphs for the main causes of delays for each airline and which airline to avoid and airport too. Finally, the team will focus on a model to develop for the project concerning flight delays. This model will be based on the data of 2008 flight delays to predict certain attributes such as delay time for 10 year period. This model is still under consideration and in the next steps the team will analyze evaluate and discuss such model and its effectiveness.

4. Preliminary Results

According to Laerd Statistics, Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data(Descriptive and Inferential Statistics, 1). The first preliminary results showed in this milestone is descriptive statistics of the data. In this analysis, some data are included. The mean, median, maximum, minimum, and standard deviation are calculated for the 4 types of delays : Weather delay, NAS delay, Security delay, and Carrier delay. In addition, AirTime, Arrival Delay, and Departure Delay are variables included in the analysis. The team uses Rstudio and Exploratory for the descriptive statistics. The results are explained in Table 1. The unit for Weather delay, NAS delay, Security delay, and Carrier delay is delays. The unit for AirTime, Arrival Delay, and Departure Delay is minutes.

Table 1 : Descriptive Statistics of The Flight Delays Dataset Variables for 2008

	Mean	Median	Max	Min	STD
Weather Delay	3.7	2.0	1,148	0.0	20.1
NAS Delay	17.16	6	1,357	0	31.89
Security Delay	0.075	0	392	0	1.83
Carrier Delay	15.77	0	2,436	0	40.1
Air Time	112.9	94	1,350	0	72.3
Arrival Delay	56.82	37	2,461	15	57.16
Departure Delay	48.54	32	2,467	-61	59.11

Table 1 shows some major results about flight delays of the year 2008. During that year, 2461 minutes are the maximum arrival delays occurred. Weather delays and NAS delays are the highest causes of delays in 2008. The second result shows the average minutes of delays per flight in major airports in the United States. Figure 1 is a map developed by Tableau shows the maximum average of delays per minutes is occured in Illinois state with 70 minutes of delay in average and the lowest in Washington state with 52 minutes. New York has a high average minutes of delay too with 68 minutes. In third place comes New Jersey with 67 minutes. Those 3 states include airports with highest number of flights and passengers per day according to Bureau of Transportation Statistics.

The map shows the maximum and minimum of average minutes of delays in different states. The highest is in Illinois with 72 minutes and the lowest is in Washington with 52 minutes. In addition, the map shows small differences between states if we focus on the color. New

Figure 1: Avg. minutes of delays per flight in United States

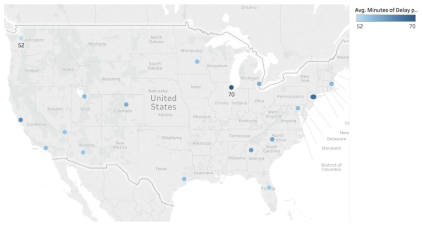


Figure 1.

Figure 2| Arrival delay vs Air Time per month ~ Top 10 airports in USA

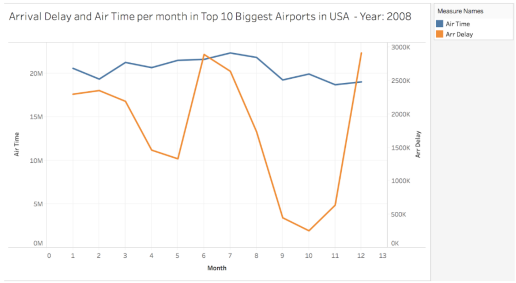


Figure 2.

York has a highest average minutes of delay too with 68 minutes.

In third place comes New Jersey with 67 minutes. Those 3 states include airports with highest number of flights and passengers per day. According to BTS data , O'Hare international airport had 419,429 flight operations in 2008. (Chicago, IL: Chicago O'Hare International). In 2014, the number is doubled to reach 881,933 flight operation which made OHare the busiest airport in the world(Mutzabaugh, 1). In this case, the air time, which consists of the total time in minutes airplane consumed during flights, is higher in busy airports such as John F Kennedy airport in New York or HartsfieldJackson Atlanta International Airport. The next visualization in Figure 2 shows a line chart for the arrival delay and air time per month in top 10 biggest airports in the United States in 2008.

The top 10 airports are chosen by the number of airtime (more or equal to 15 minutes of delay) and from an article by Melanie Renzulli in TripSavvy website that shows top 10 busiest airports in the US according to number of passengers and flights. The list contains:

-Hartsfield-Jackson Atlanta International Airport
-O'Hare International Airport
-Los Angeles International Airport
-Dallas/Fort Worth International Airport
-John F. Kennedy International Airport
-Denver International Airport
-San Francisco International Airport
-McCarran International Airport
-Phoenix Sky Harbor International Airport
-George Bush Intercontinental Airport

Figure 3 shows a significant result. During the month of

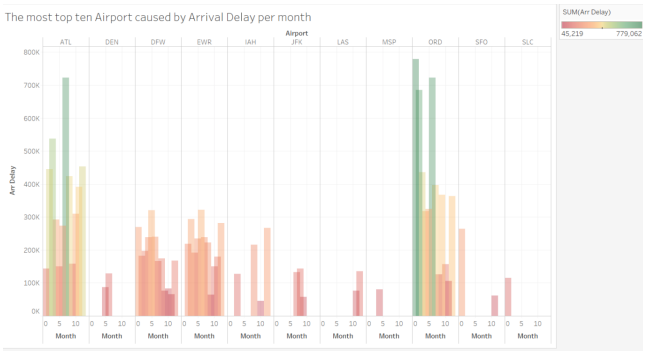


Figure 3. Delays per month in Top 10 airports in USA

Figure 4: Arrival delay in day of month of January 2008.

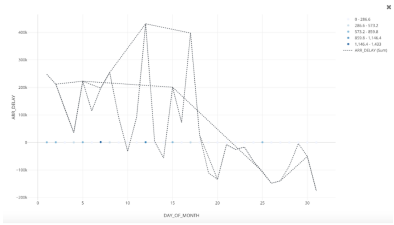


Figure 4. Arrival Delay in day for January month in USA 2008

February , April, and May, a low of more than 5 million minutes in delay occurred during that period. In the same time, the air time increased slightly. Between May and June, a high increase happened in delay that reached more than 20 million minutes but after the month of June until month of November the total minutes of delays in top 10 airports in US dropped to nearly 2.5 million minutes. In the same period, the air time dropped slightly under 20 million minutes. After the month of November, the total number of delays increase again to reach more than 20 million. For the next analysis, the focus will on the month of January. This month is chosen because some important holidays come through such as the new year and its the month after christmas holiday. Its expected to be a busy month for airport and for delays. The first analysis in Figure 4 is about the Arrival delays in the days of the month of January.

For Figure 3, we noticed that the number of delays is increasing during the first months of the year. For example, for Chicago ORD airport the number of delays during the month of January was more than 700,000 minutes. For Atlanta airport it was around 400,000 minutes of delays. In addition, those two airports have the highest arrival delays during the year

For the next analysis, the focus will on the month of January. This month is chosen because some important holidays come through such as the new year and its the month after christmas holiday. Its expected to be a busy month for airport and for delays. The first analysis in Figure 4 is about the Arrival delays in the days of the month of January.

Figure 5: Arrival delay per day of week in month of January in 2008

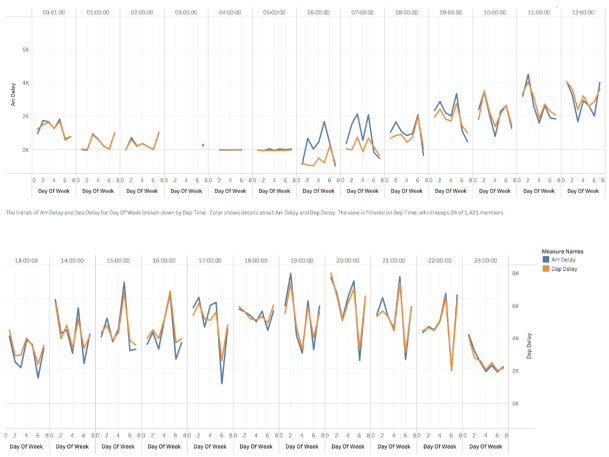


Figure 5.

The line graph summarize the total delays in the month of January during the 31 days. The highest rate of the delays occurred during the 10th day and the 17th day. During the middle of the month, the total number of delays passed 400,000 minutes. After the 17th day of January, a significant drop happened. Until the end of the month, airlines in the US are having a negative minutes of delays which means early arrivals increased in US airports. The question that arises here is whats the exact time of the day airlines delays happens the most ?

From Figure 5, we can see that there are variations between Arrival delay and Departure Delay in each time frame in each day of the week. During the first hours of the day, there are few delays between 2,000 minutes in delays to 0 minutes in 4AM. From 6AM until the end of the day at 22:00 PM, we can see that delays increases and decreases at the same level and time. The highest is during 19:00 PM and 21:00 PM. After 22:00 PM, the total delays in Arrival and Departure decreases. In addition, the busiest day of week is day 5 (Friday) with highest probability of delays. During the time 15:00 PM,16:00 PM, and 21:00 PM we can notice a peek in the line of Arrival and Departure delays. The total minutes of delays drops after day 5 in those times.

The conclusion from analyzing Figure 3,4,and 5 is concerning the best time for travelling. For the top 10 busiest airport in the US, the best time to fly is in April, September, and October. During summer, airports have the highest percentage of delays. During holidays like christmas , the number of delays increase too. During the month on January, which is after the holiday, delays are minimum in the last 10 days of the month. Its not consistent to say that this occurs in every month of the year. For the time of travelling, early in the morning are the best times where delays for departure and arrival are not hight. Evening flights have a high number of delays. Besides, travellers should avoid flying during weekends.

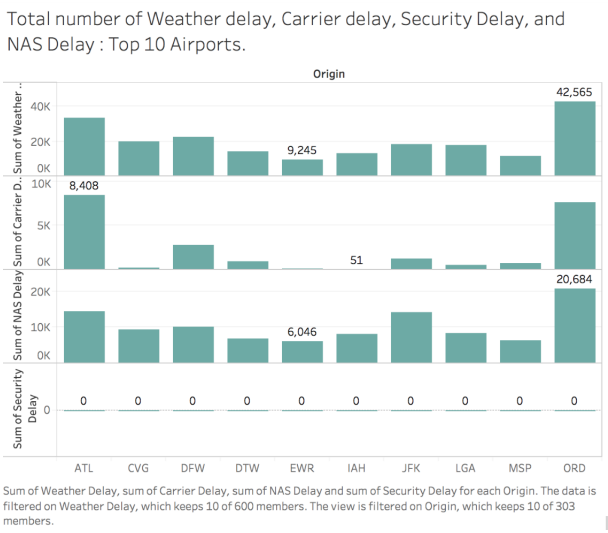


Figure 6.

The data shows an increase in delays during Fridays and Saturdays. This could be explained by the increase number of travellers during weekends which makes airport carrier management have difficulties in being on schedule.

5. Weather Delay Analysis

The next graphical analysis focuses on the reasons of delays. The 4 reasons are weather , security, carrier, and National Airspace System (NAS). According to FAA, delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc (Types of Delay, 2017). Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival (Types of Delay, 2017). Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas (Types of Delay, 2017). Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike etc (Types of Delay, 2017).
Figure 6 shows the US airports with total more than 100,000 min delays due to weather, NAS, security, and carrier. In addition, the graph shows the maximum and minimum value for each type of delay.

We can notice that in all airports there are no security delays in the year 2008. OHare international airport is in the lead for weather delays and NAS delays. Hartsfield-Jackson Atlanta International Airport has the maximum number of delays due to Carrier. Newark Liberty International Airport has the minimum number of delays due to weather and NAS. George Bush Intercontinental Airport has the minimum number of delays due to Carrier issues. In the next analysis, we will focus on one type of delay which is weather delay. With more than 42,000 delays in OHare international airport because of weather, its important to study the relation between month and number of delays due to weather.// In 2008, OHare international Airport in Chicago is in the lead with more than 40,000 delays due to weather conditions. In Figure 7 we can notice the distribution of the total number of delays due to weather per month in top 10 airports by weather delays. As ORD has the maximum delays, we can notice that in month of December, 12,000 delays occurred due to weather in the biggest airport in Illinois.

6. Time Series Model

This section is still underconstruction.

7. Conclusion

Underconstruction

Acknowledgments

The authors would like to thank...

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.