US Flight Data Analysis

Forecasting Methods - Report

André Viana* Gonçalo Duarte† José Cabral‡

May 2021

Abstract

Nowadays, as easy as it is for any of us to travel from one city to another, one country to another, the flight industry is expected to be extremely profitable and resourceful. With hundreads of thousands of registered flights per day, there is a big volume of data possible to be analized. The year of 2020 brought change to our lives. Being hit by a pandemic changes the way all of us think and look at the world around us. Do we feel safe to travel in the next months? Can we even travel? Some questions arose last year and for many of them there are no answers yet. What we do know is that the way we travel has changed. That impacts numbers of the aviation industry. Despite being one single country, there are more than five thousand public airports across the USA. Some of these airports have tremendous traffic - which provides a huge amount of data possible to be analyzed. Develop analysis on flight data means providing - or trying to provide - answers to many questions, such as: how much did the number of flights decreased from 2019 to 2020 due to COVID 19? Is there a seasonality regarding the number of flights? Do cancellation rates follow a trend? How can we predict the number of flights on the next months based on past data? We hope to explore this questions and be able to provide meaningful answers.

*l54543, andreviana@aln.iseg.ulisboa.pt
†l48505, goncaloduarte@aln.iseg.ulisboa.pt
‡l54997, l54997@aln.iseg.ulisboa.pt

Contents

List of Tables

List of Figures

Introduction

This project aims at analyzing US flights – available at the Bureau of Transportation Statistics of United States Department of Transportations database – explore its evolution over the years concerning number of flights and flights cancellations Living through a pandemic is an unprecedented situation (in the most recent decades) that surely took a tool on the travel industry. Further on the project, it will be possible to visualize the evolution of the number of flights by month and year. An intense decrease is expected when comparing 2020 to previous years. We will focus our attention on four different airports:

- Anchorage, AK: Ted Stevens Anchorage International
- Chicago, IL: Chicago O'Hare International
- Los Angeles, CA: Los Angeles International
- New York, NY: LaGuardia

since these are some of the airports with the most traffic within their states. As well as the number of flights, this project also analysis patterns regarding the cancellation rates.

Disclaimer: The initial goal of the project was to develop analysis on flight data from New York airports. After an initial look at the series, we considered it was best for our project to work with airports from different states, since there is more heterogeneity.

For the purpose of this project, the analysis is developed on data until the end of 2019 (due to the out of ordinary changes that happened in 2020). A comparison with 2020 data is displayed on the end of the project.

1. Time Series Plots

For this project we decided to use the airports with more flights from the following states : Illinois, Alaska, California and New York.

On the plots below, it is possible to look at the behavior of both the number of flights (Figure 1) and the cancellation rates (Figure 2) for the four airports from January 2010 until December 2019.

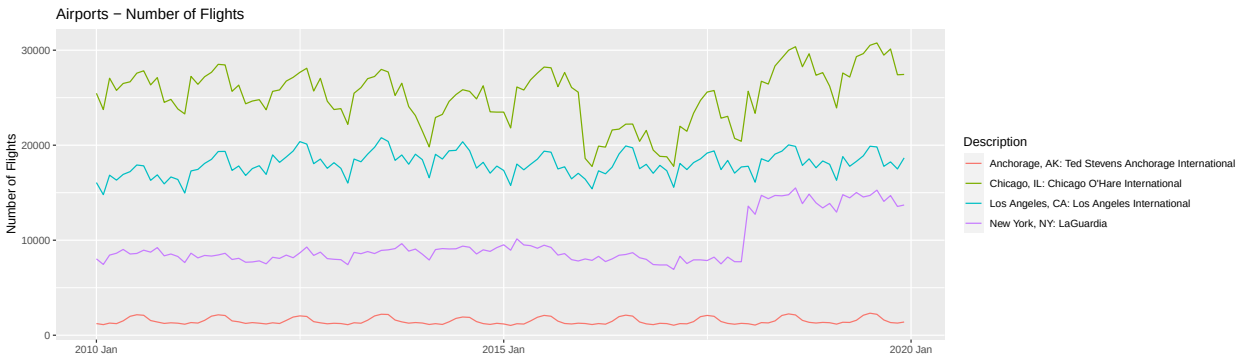

Figure 1: Airports - Number of Flights

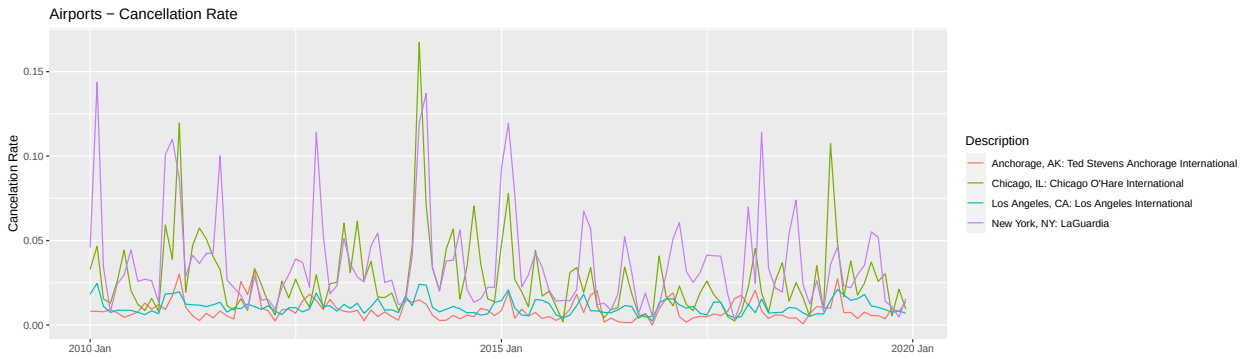

Figure 2: Airports - Cancellation Rate

A total of 8 time series integrates our project:

- 4 on number of flights (one per airport);
- 4 on rate of cancellations (one per airport).

After visualizing the series, we will now focus on seasonal behaviors. For that, two sets of plots will be displayed: **seasonal plots** and **seasonal subseries plots**:

Seasonal Plots

On the plot below, once again, two plots are displayed (1st regarding total number of flights, 2nd regarding the cancellation rate). On each of the plots, there are individual analysis for each airport that will allow us to understand how numbers change from month to month on each year. (one line per year)

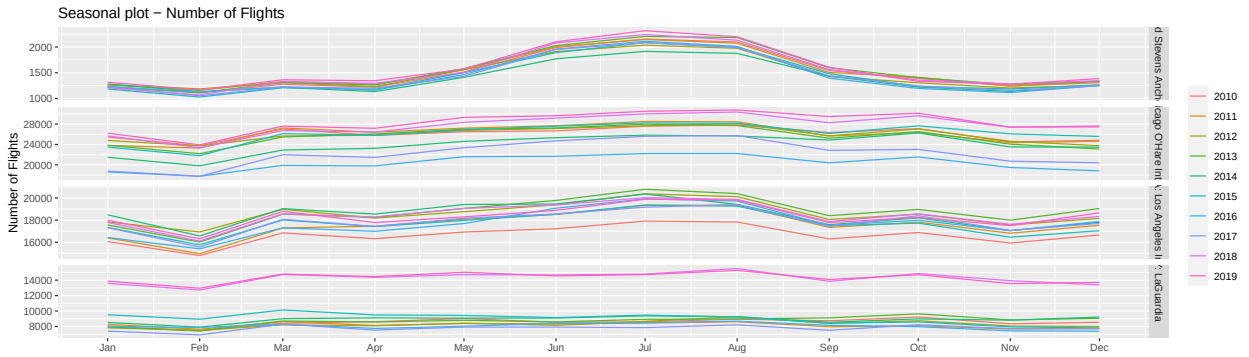

Figure 3: Seasonal plot - Number of Flights

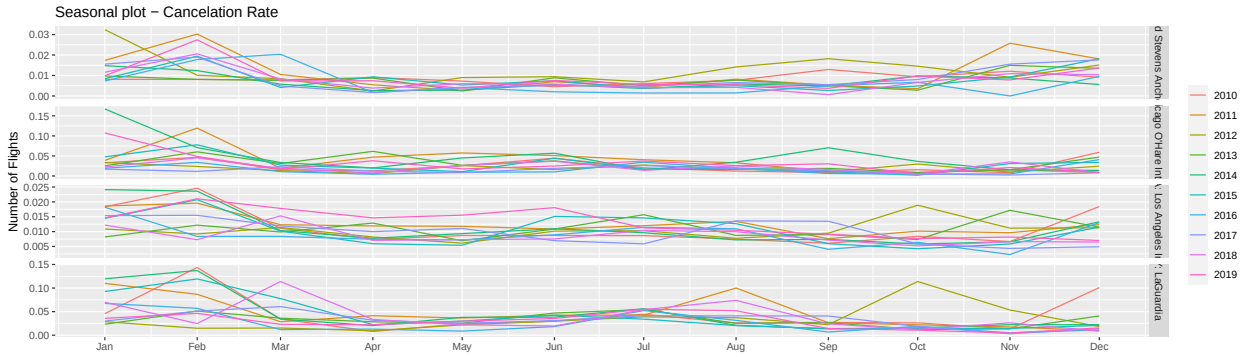

Figure 4: Seasonal plot - Cancelation Rate

Taking in consideration the previous visuals, note that, regarding **number of flights**:

- Ted Stevens Anchorage International: increase between June and August;

- Chicago O'Hare International: registered an increase between June and August;
- Los Angeles International: also has an increase between June and August;
- La Guardia: Not so strong seasonal effect but has a huge increase in 2018.

When it comes to the **cancellation rate**, all four airports show some instability in the beginning of the year, across all years. From March to August, cancellation rates tend to decrease a bit and then increase towards the end of the year. A special attention to La Guardia, that saw an uncharacteristic increase on October 2012, comparing to the same month of the remaining years.

Seasonal Subseries Plots

One different way to analyze and compare months from year to year is by plotting subseries - where it is possible to visualize, for each month, how the series behaves each year. Once again, two plots corresponding to the two features studied on this project:

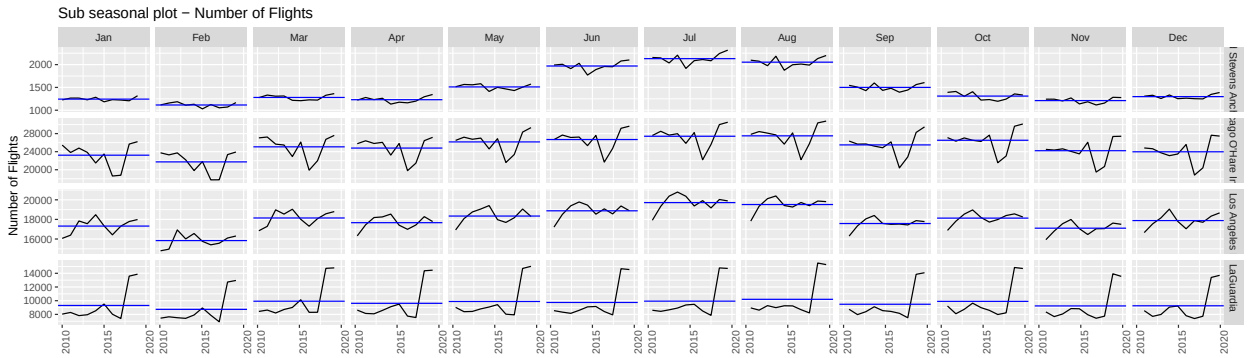

Figure 5: Sub seasonal plot - Number of Flights

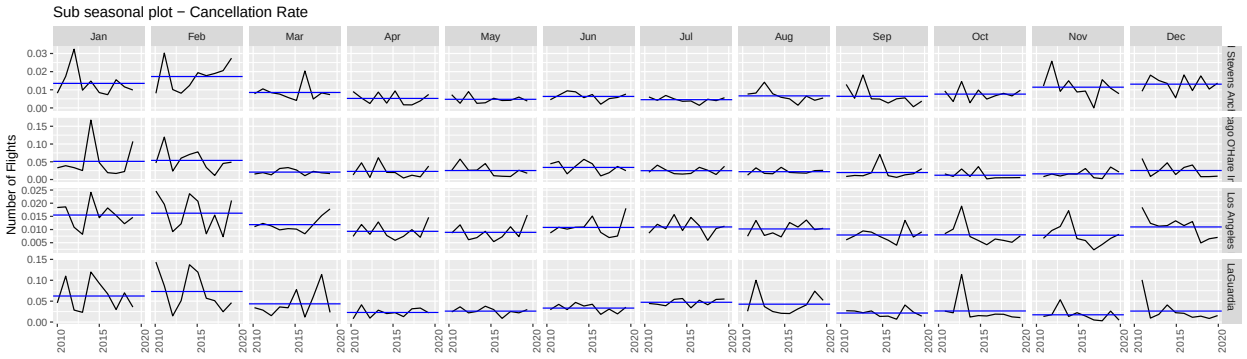

Figure 6: Sub seasonal plot - Cancellation Rate

To complement previous comments, note that La Guardia has a fairly constant average number of flights monthly across all months of the year (*blue line represents the average*), having suffered an increase on most recent years.

Once again, just like on the previous analysis, regarding cancellations rates, they tend to be heigher (on average) and more inconsistent on the beggining of the year, decreasing after the first trimester. If we take a look, on October for La Guardia (regarding cancellation rates) there is a spike on 2012, which corresponds to the comments made on the previous seasonal plots.

Note that these two sets of plots (seasonal and seasonal subseries) allow us to interpret the same data from different "angles".

ACF Plots

Autocorrelation Function is a measure of the relationships between lagged values of a time series. By taking a look of correlograms (plot of the ACF) it is possible to identify possible seasonal and trend components.

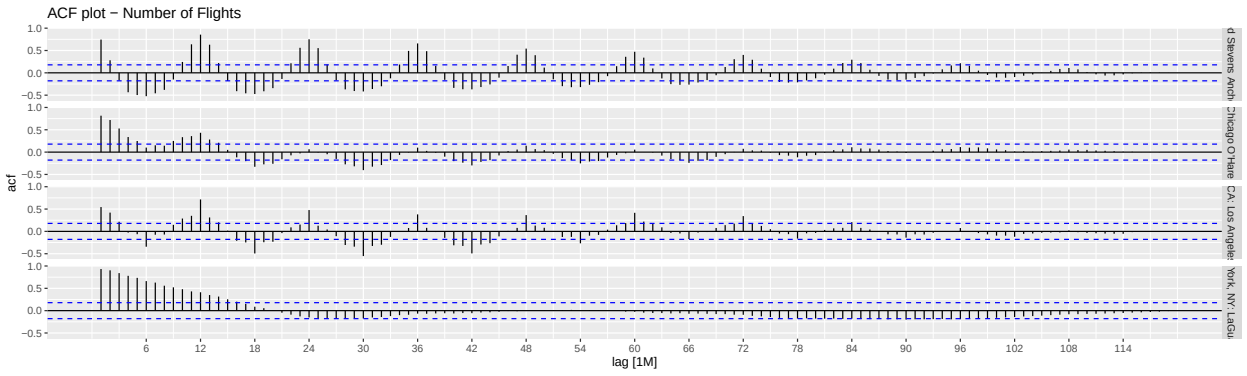

Figure 7: ACF plot - Number of Flights

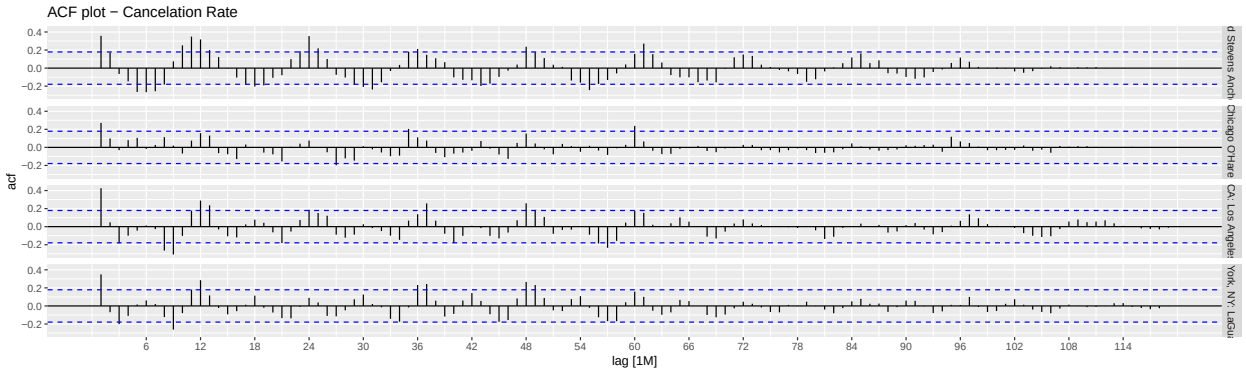

Figure 8: ACF plot - Cancelation Rate

To simplify our analysis, let's take a look at the first correlogram, regarding Ted Stevens Anchorage International on the number of flights: a clear seasonal pattern is displayed since the peaks tend to be 12 months apart and the troughs tend to be 6 months apart. Keeping in mind that we working with monthly flight data, is perfectly reasonable to think so (holidays seasons are expected to register an increase on flight numbers).

2. Decomposition and transformations

After taking a look at the data and the multiples series we will work with, there is one needed transformation to be developed:

- Cancellation rate: transform it so that all values are between 0 and 1.

Note that, on the following code, a function is created to replace any 0 value by 0.00001 on the cancellation rate. This is needed so that it is possible to apply a log transformation on the rates.

```
bound_transform <- function(x, lower = 0,upper = 1){
  x[x == 0] <- 0.00001
  log((x - lower) / (upper - x))
}

inv_bound_transform <- function(x, lower = 0, upper = 1){
  (upper - lower) * exp(x)/(1 + exp(x)) + lower
}

my_bound_transformation <- new_transformation(bound_transform, inv_bound_transform)
```

3. Forecaster Toolbox

The next step it o fit the most basic models on our time series. The models used are the following:

- Random Walk with Drift - forecasts equal to last value plus average change, controlling the growth rate of our time series;
- Snaive - forecasts equal to last value from same season;
- Naive - forecasts equal to last observed value;
- Trend

- Mean - forecasts equal to mean of historical data.

First, let's develop these analysis for the number of flights of each airport:

Number of flights

Table 1: Basic models accuracy - Number of FLights

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	drift	-442.46608	616.94116	538.71386
Anchorage, AK: Ted Stevens Anchorage International	mean	209.56140	467.60333	354.00000
Anchorage, AK: Ted Stevens Anchorage International	naive	-415.33333	589.26960	519.33333
Anchorage, AK: Ted Stevens Anchorage International	snaive	32.33333	48.27007	42.33333
Anchorage, AK: Ted Stevens Anchorage International	trend	220.93906	472.55498	353.71556
Chicago, IL: Chicago O'Hare International	drift	-479.43805	1508.27858	1220.81268
Chicago, IL: Chicago O'Hare International	mean	4271.24561	4486.87572	4271.24561
Chicago, IL: Chicago O'Hare International	naive	-350.00000	1418.10249	1177.66667
Chicago, IL: Chicago O'Hare International	snaive	416.33333	607.20562	480.33333
Chicago, IL: Chicago O'Hare International	trend	4887.46983	5072.85143	4887.46983
Los Angeles, CA: Los Angeles International	drift	-322.34513	1007.96836	943.72271
Los Angeles, CA: Los Angeles International	mean	671.41228	1143.95430	897.13743
Los Angeles, CA: Los Angeles International	naive	-235.00000	955.54278	881.33333
Los Angeles, CA: Los Angeles International	snaive	-69.50000	207.29407	178.16667
Los Angeles, CA: Los Angeles International	trend	236.19995	963.85838	811.96010
New York, NY: LaGuardia	drift	-413.66814	801.58823	650.04867
New York, NY: LaGuardia	mean	4983.26316	5020.24339	4983.26316
New York, NY: LaGuardia	naive	-212.00000	644.10869	556.66667
New York, NY: LaGuardia	snaive	-48.00000	249.06626	228.66667
New York, NY: LaGuardia	trend	2647.26235	2728.44000	2647.26235

After training and testing our models, the criterion to choose the most adequate amongst all possibilities is based on the ones that have the lowest RMSE on the test set.

Table 2: Best basic models - Number of Flights

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	snaive	32.33333	48.27007	42.33333
Chicago, IL: Chicago O'Hare International	snaive	416.33333	607.20562	480.33333
Los Angeles, CA: Los Angeles International	snaive	-69.50000	207.29407	178.16667
New York, NY: LaGuardia	snaive	-48.00000	249.06626	228.66667

As you can see on the table above, Seasonal Naive models were the best across all airports for the number of flights. These will now be the models used on this sector of the project.

Before analyzing in detail the residuals, first we will visualize the chosen models as well as the data for the four airports:

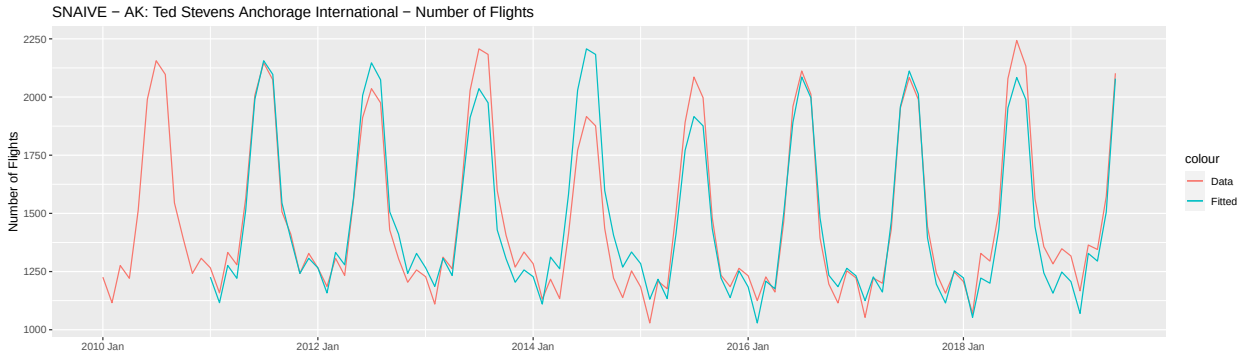

Figure 9: SNAIVE - AK: Ted Stevens Anchorage International - Number of Flights

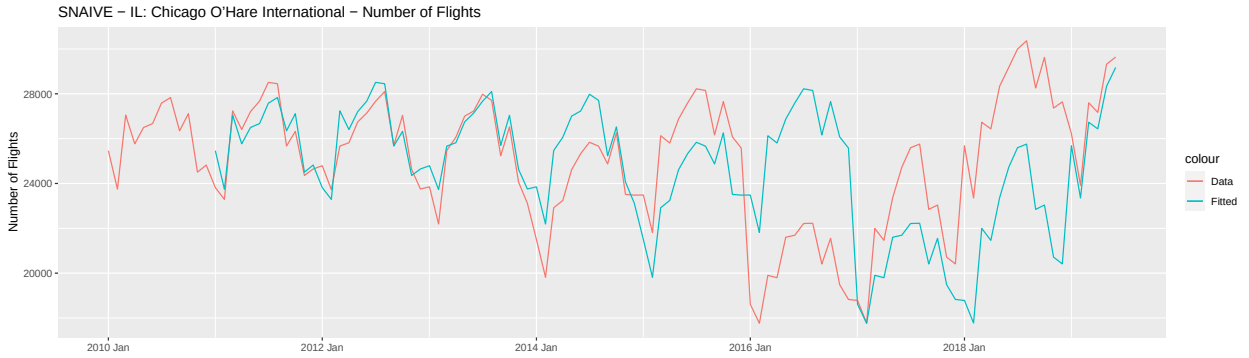

Figure 10: SNAIVE - IL: Chicago O'Hare International - Number of Flights

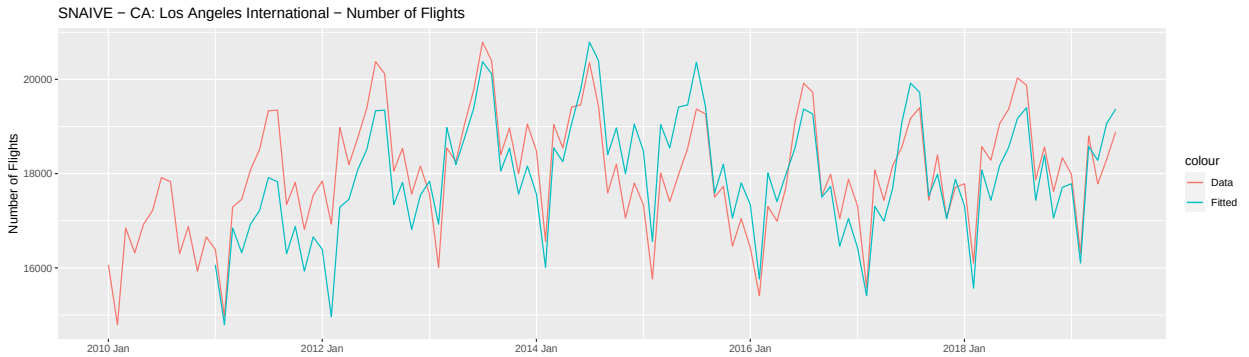

Figure 11: SNAIVE - CA: Los Angeles International - Number of Flights

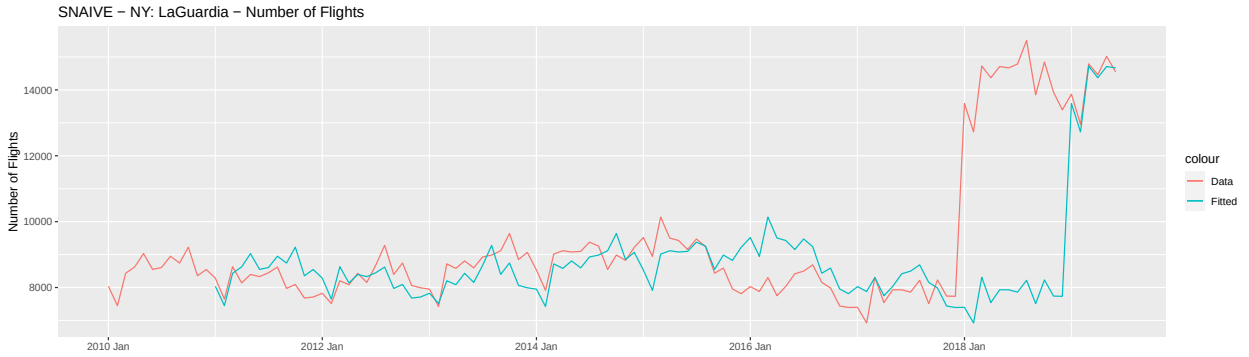

Figure 12: SNAIVE - NY: LaGuardia - Number of Flights

To verify our models, we will now focus on the residuals. In order to visually check residuals correlations, the next four figures present the Scatter, Histogram and ACF plot of the residuals. Note that having spikes out of the region limited in blue, means errors are correlated, which is unwanted.

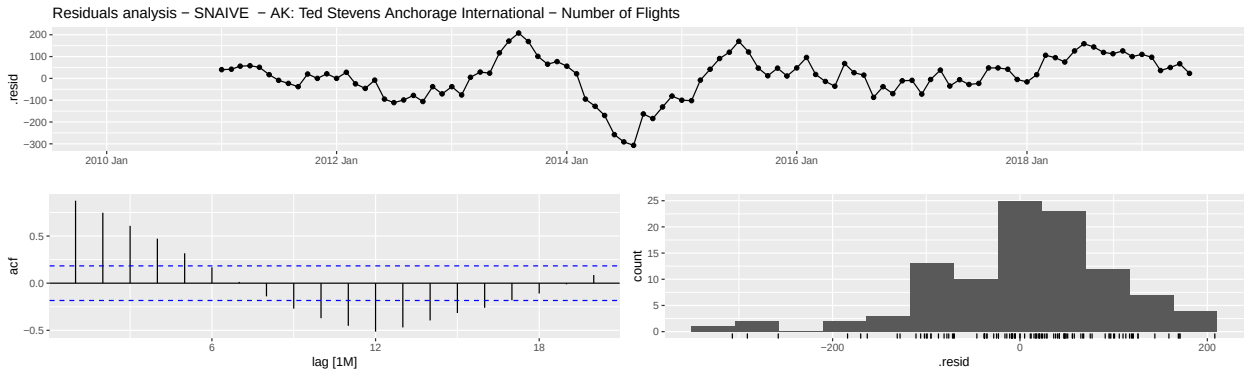

Figure 13: Residuals analysis - SNAIVE - AK: Ted Stevens Anchorage International - Number of Flights

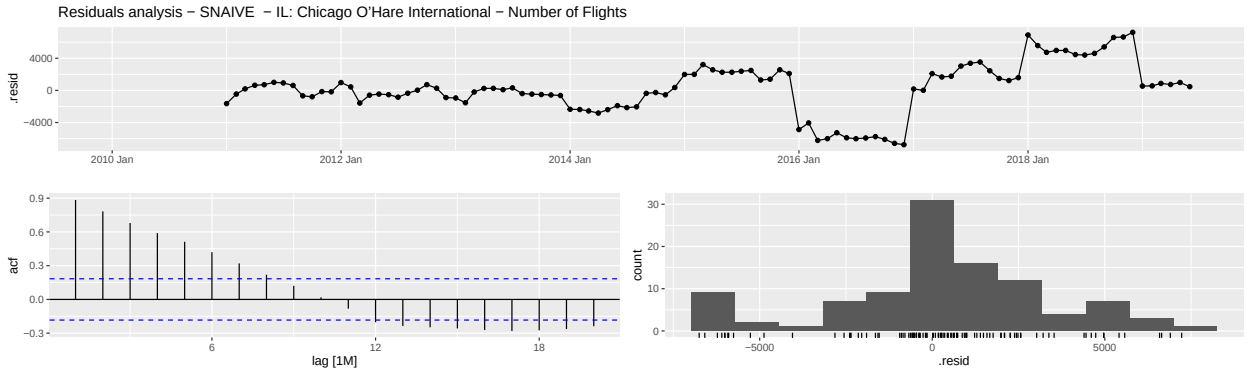

Figure 14: Residuals analysis - SNAIVE - IL: Chicago O'Hare International - Number of Flights

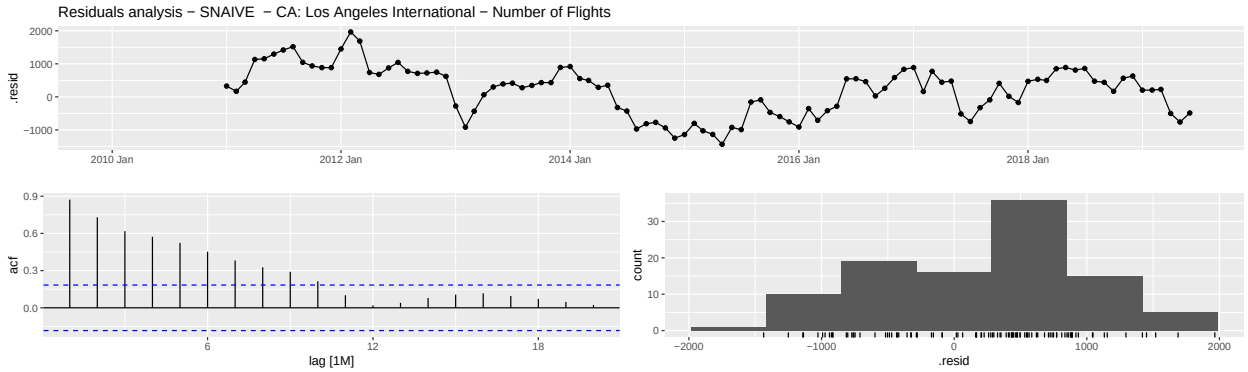

Figure 15: Residuals analysis - SNAIVE - CA: Los Angeles International - Number of Flights

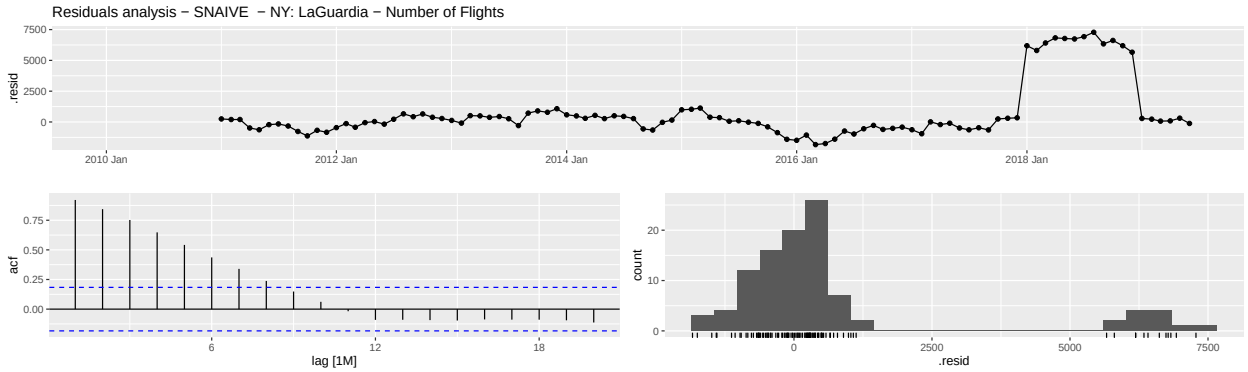

Figure 16: Residuals analysis - SNAIVE - NY: LaGuardia - Number of Flights

To verify our models, furthermore, we will now focus on the residuals through the Ljung box test – which allows us to test if the residuals are correlated. Note that: * H0: the residuals are autocorrelated; * H1: The residuals do not have autocorrelation.

The goal is to have errors as white noise processes, and so reject H0 and for that, p-values have to be > 0.05 (95% significancy).

Table 3: Ljung box test basic models - Number of Flights

Description	.model	lb_stat	lb_pvalue
Anchorage, AK: Ted Stevens Anchorage International	snaive	394.1981	0
Chicago, IL: Chicago O'Hare International	snaive	399.9217	0
Los Angeles, CA: Los Angeles International	snaive	316.5426	0
New York, NY: LaGuardia	snaive	371.0084	0

As we can see on table 3, none of the models have p-value > 0.05, which indicates us that these more basic models struggle to extract all the information from the data. We then need to test more complex models.

Cancellation Rate

The exact same work is now developed for the series regarding Cancellation Rates.

Start by training and testing the models.

Table 4: Basic models accuracy - Cancelation Rate

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	drift	-0.0182994	0.0195306	0.0182994
Anchorage, AK: Ted Stevens Anchorage International	mean	-0.0019948	0.0038531	0.0033592
Anchorage, AK: Ted Stevens Anchorage International	naive	-0.0176770	0.0187905	0.0176770
Anchorage, AK: Ted Stevens Anchorage International	snaive	-0.0032339	0.0053680	0.0040956
Anchorage, AK: Ted Stevens Anchorage International	trend	0.0004155	0.0033669	0.0027941
Chicago, IL: Chicago O'Hare International	drift	-0.0400055	0.0487283	0.0407324
Chicago, IL: Chicago O'Hare International	mean	-0.0056152	0.0125056	0.0100374
Chicago, IL: Chicago O'Hare International	naive	-0.0392042	0.0476933	0.0399329
Chicago, IL: Chicago O'Hare International	snaive	-0.0038149	0.0148863	0.0114190
Chicago, IL: Chicago O'Hare International	trend	0.0006189	0.0110764	0.0095455
Los Angeles, CA: Los Angeles International	drift	-0.0157273	0.0164313	0.0157273
Los Angeles, CA: Los Angeles International	mean	-0.0018746	0.0024083	0.0020242
Los Angeles, CA: Los Angeles International	naive	-0.0154941	0.0161553	0.0154941
Los Angeles, CA: Los Angeles International	snaive	0.0004919	0.0010060	0.0008096
Los Angeles, CA: Los Angeles International	trend	-0.0008210	0.0016974	0.0015292
New York, NY: LaGuardia	drift	-0.0494139	0.0623134	0.0523115
New York, NY: LaGuardia	mean	-0.0117019	0.0233953	0.0227086
New York, NY: LaGuardia	naive	-0.0485571	0.0612196	0.0514530
New York, NY: LaGuardia	snaive	-0.0188700	0.0245325	0.0200859
New York, NY: LaGuardia	trend	-0.0055682	0.0208880	0.0205447

Choosing the models with the lowest RMSE on the test set, we now obtain the following models as the best ones:

Table 5: Best basic models - Cancellation Rate

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	trend	0.0004155	0.0033669	0.0027941
Chicago, IL: Chicago O'Hare International	trend	0.0006189	0.0110764	0.0095455
Los Angeles, CA: Los Angeles International	snaive	0.0004919	0.0010060	0.0008096
New York, NY: LaGuardia	trend	-0.0055682	0.0208880	0.0205447

Once again, before looking at the residuals, we will plot the best models alongside with the data for the four airports:

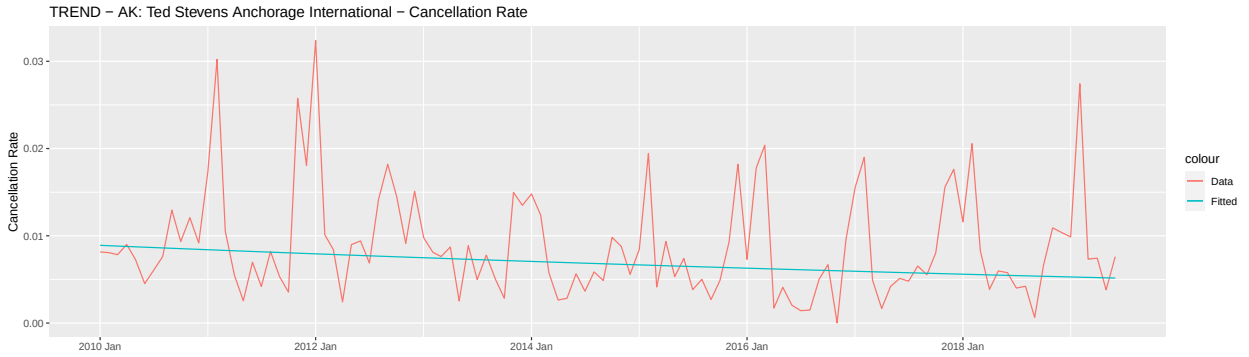

Figure 17: TREND - AK: Ted Stevens Anchorage International - Cancellation Rate

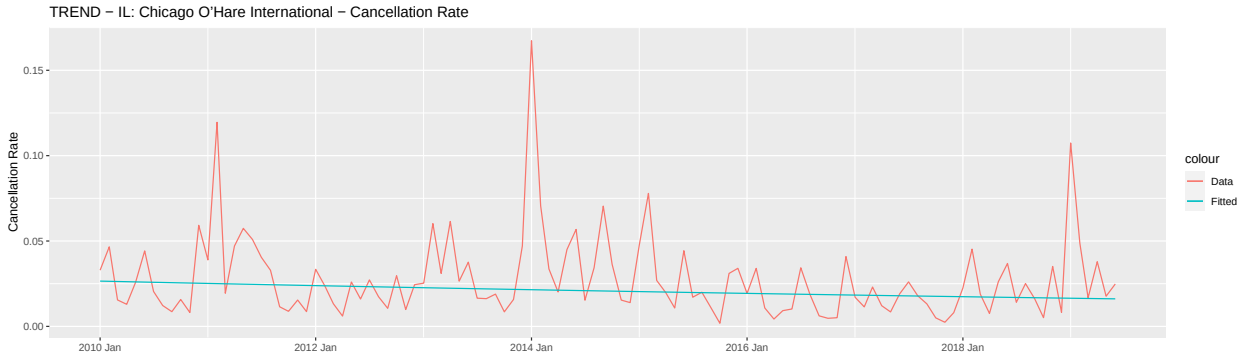

Figure 18: TREND - IL: Chicago O'Hare International - Cancellation Rate

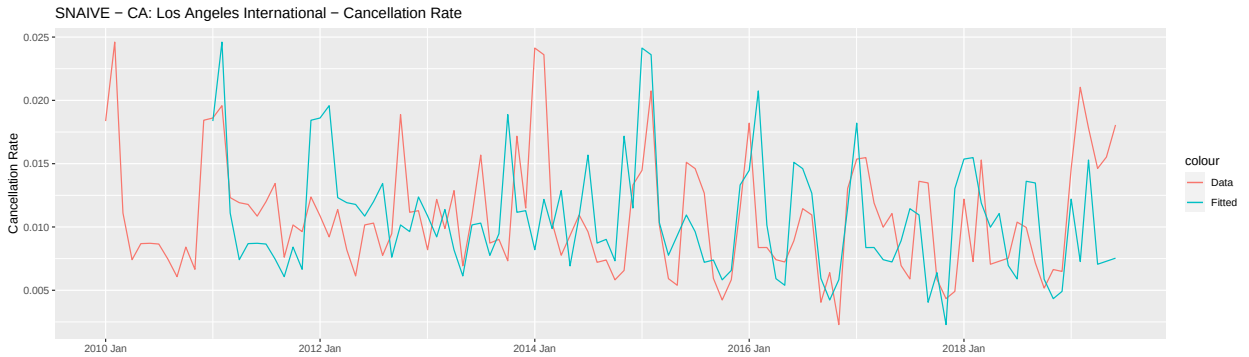

Figure 19: SNAIVE - CA: Los Angeles International - Cancellation Rate

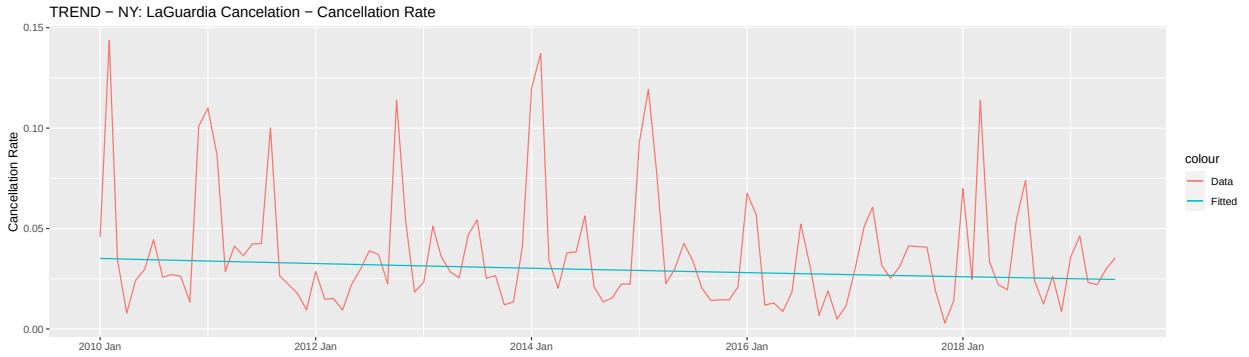

Figure 20: TREND - NY: LaGuardia Cancelation - Cancellation Rate

As for the Number of flights, the next step is to analyze the residuals and test autocorrelation. Let's first take a look at the Scatter, Histogram and ACF plot of the residuals, in order to check if the autocorrelation is significant or not.

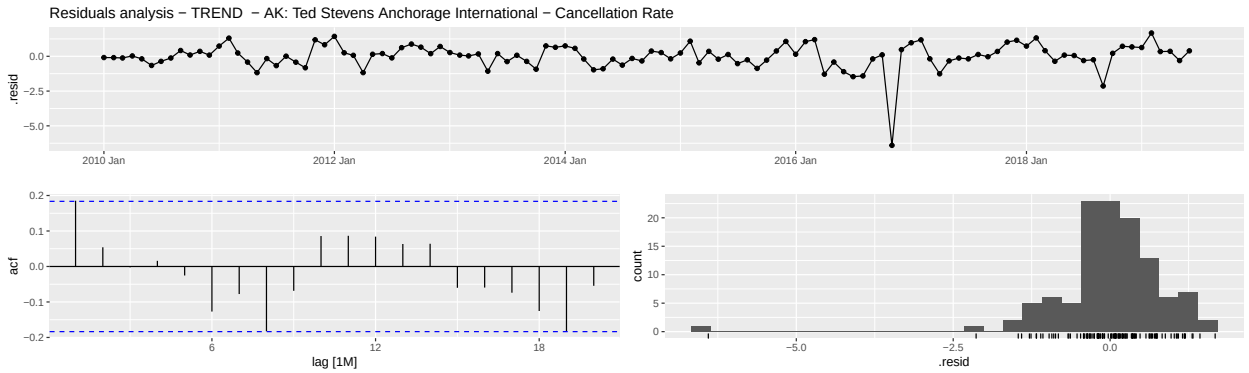

Figure 21: Residuals analysis - TREND - AK: Ted Stevens Anchorage International - Cancellation Rate

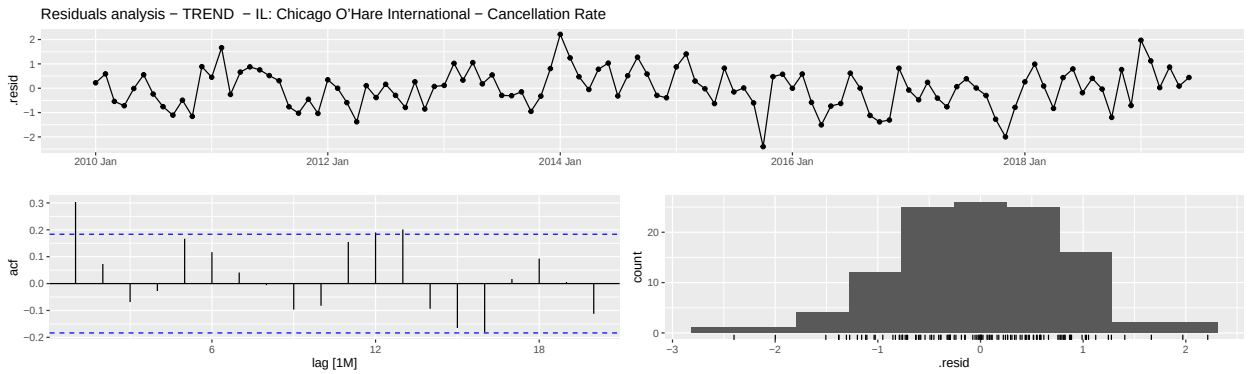

Figure 22: Residuals analysis - TREND - IL: Chicago O'Hare International - Cancellation Rate

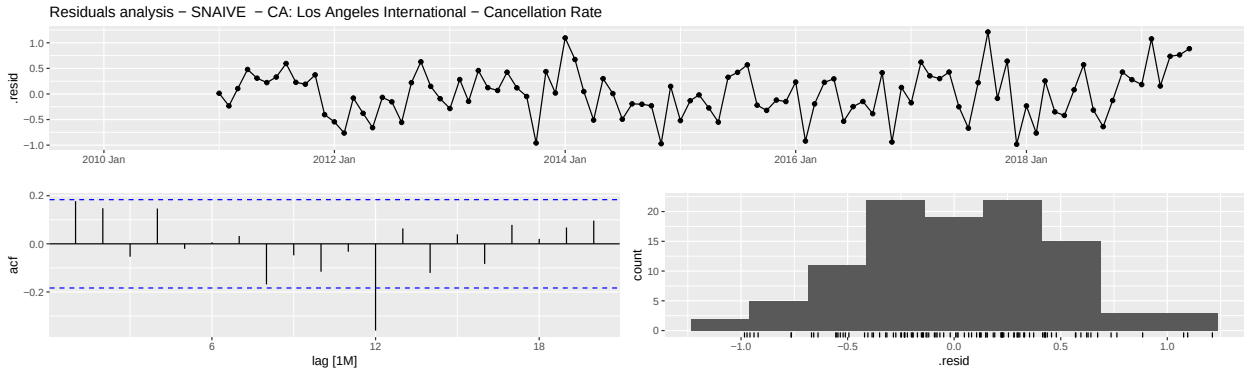

Figure 23: Residuals analysis - SNAIVE - CA: Los Angeles International - Cancellation Rate

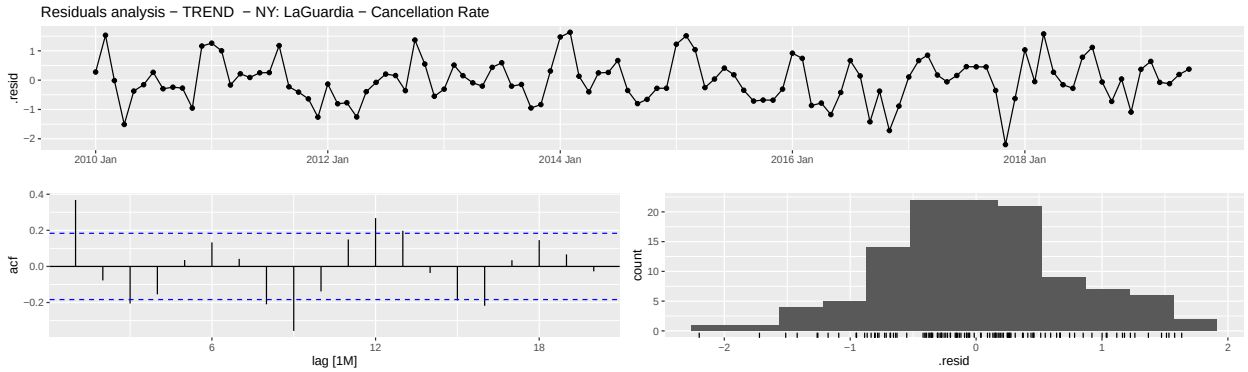

Figure 24: Residuals analysis - TREND - NY: LaGuardia - Cancellation Rate

Note that, by looking at the ACF plots, it seems that the residuals of the model for AK airport series are uncorrelated (which is desired). We can also do a Ljung box test to prove this, as it follows:

Table 6: Ljung box test basic models - Cancellation Rate

Description	.model	lb_stat	lb_pvalue
Anchorage, AK: Ted Stevens Anchorage International	trend	12.88298	0.2302886
Chicago, IL: Chicago O'Hare International	trend	19.35375	0.0359929
Los Angeles, CA: Los Angeles International	snaive	36.51107	0.0489248
New York, NY: LaGuardia	trend	51.01873	0.0000002

In fact, p-value for the residuals of the AK series is > 0.05, which shows us that the errors are uncorrelated.

So, using 6 months predictions to test our models, the best are the following:

Number of flights: Alaska: snaive Illinois: snaive Chicago: snaive New York: snaive

Cancellation Rate: Alaska: trend Illinois: trend Chicago: snaive New York: trend

4. Exponencial Smoothing for the airport of Chicago

To simplify our analysis, we will focus our attention on the number of flights of the Chicago airport and cancellation rate of the New York airport to develop exponential smoothing models.

Exponential smoothing models (ETS) are "divided" in three components:

```
Error - which can be either additive or multiplicative;
Trend - which can be non-existing, additive, multiplicative or damped;
Seasonality - which can be none existing, additive or multiplicative.
```

We will use a STL decomposition to choose the category for each component

Number of flights

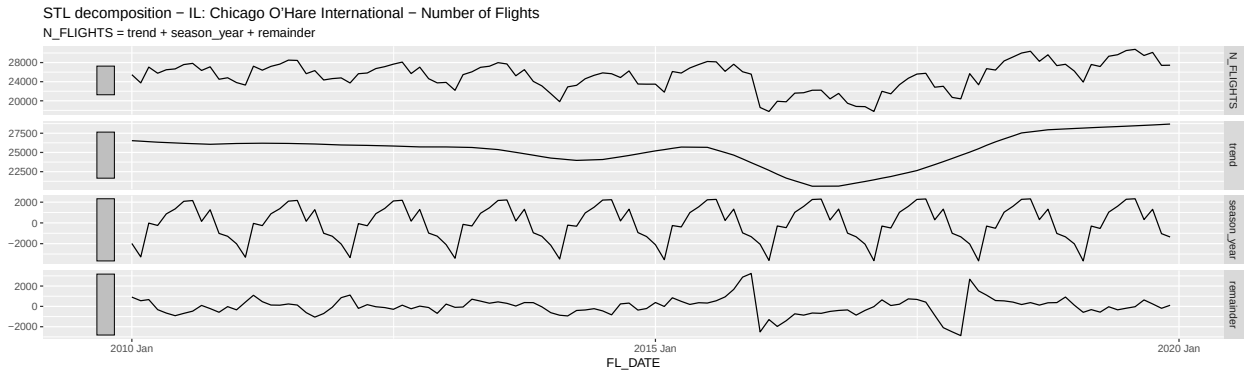

Figure 25: STL decomposition - IL: Chicago O'Hare International - Number of Flights

So, after analyzing the decomposition, we considered:

Error = A: The errors' variance stays constant, so we considered this parameter as Additive.
Trend = N: The trend component stays fairly constant, meaning we fixed the parameter as None.
Season = A: As the variance of the seasonality component also stays constant, the parameter was considered to be Additive.

We also tested the auto model, which looks to find the "best" possible combination of the parameters for the given data:

Table 7: ETS models - IL: Chicago O'Hare International - Number of Flights

Description	ana	auto
Chicago, IL: Chicago O'Hare International	<ETS(A,N,A)>	<ETS(A,N,A)>

As described on the table above, the automatic model is the same as the one we considered. This will result on the same results, when looking at performance criterions.

Table 8: ETS metrics - IL: Chicago O'Hare International - Number of Flights

Description	.model	sigma2	log_lik	AIC	AICc	BIC	MSE	AMSE	MAE
Chicago, IL: Chicago O'Hare International	ana	1044139	-1052.441	2134.882	2139.78	2175.925	915911.4	1707554	499.6087
Chicago, IL: Chicago O'Hare International	auto	1044139	-1052.441	2134.882	2139.78	2175.925	915911.4	1707554	499.6087

On the following figure, we can now look at the ETS model as well as the data.

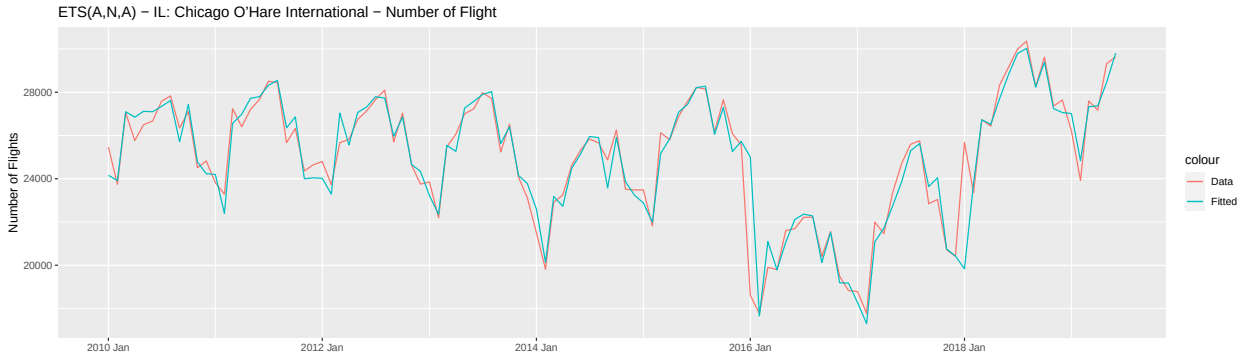

Figure 26: ETS(A,N,A) - IL: Chicago O'Hare International - Number of Flights

Cancellation Rate

Just as for the most basic models, the same procedure is now applied to the Cancellation Rate series, starting with visualizing the STL decomposition in order to decide on the Error, Trend and Seasonal parameters.

Once again, to simplify, the detailed analysis will only be developed step by step for one airport: this time for La Guardia, New York.

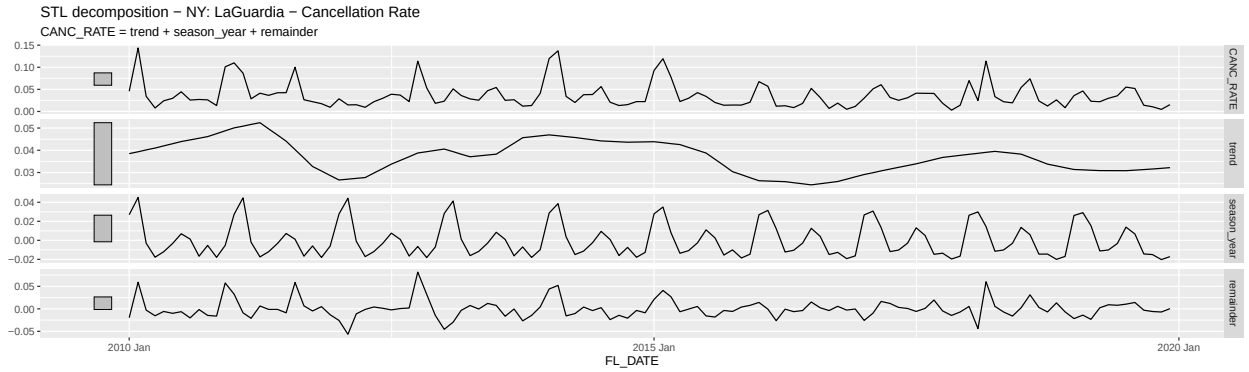

Figure 27: STL decomposition - NY: LaGuardia - Cancellation Rate

The following was considered regarding the parameters:

Error = A: The errors' variance remains constant, so it was assigned Additive.
Trend = N: The trend also remains fairly constant, having been considered as None.
Season = M: As the variance of the seasonality component grows throughout time, this parameter was considered Multiplicative.

Once again, the automatic model was also trained, this time being different from ours, having Additive seasonal component:

Table 9: ETS models - NY: LaGuardia - Cancellation Rate

Description	anm	auto
New York, NY: LaGuardia	<ETS(A,N,M)>	<ETS(A,N,A)>

Analyzing the AICc of both models, the automatic one seems better than the one we proposed.

Table 10: ETS metrics - NY: LaGuardia - Cancellation Rate

Description	.model	sigma2	log_lik	AIC	AICc	BIC	MSE	AMSE	MAE
New York, NY: LaGuardia	anm	0.4142817	-212.2658	454.5316	459.4295	495.5745	0.3634050	0.3734387	0.4537052
New York, NY: LaGuardia	auto	0.4105022	-211.7434	453.4868	458.3847	494.5297	0.3600896	0.3700140	0.4486999

Contrary to what happened for the number of flights, for the cancellation rate, our model was different than the one provided by the automatic function, which leads us to three different lines on the plot – one for each model and a third for the data itself.

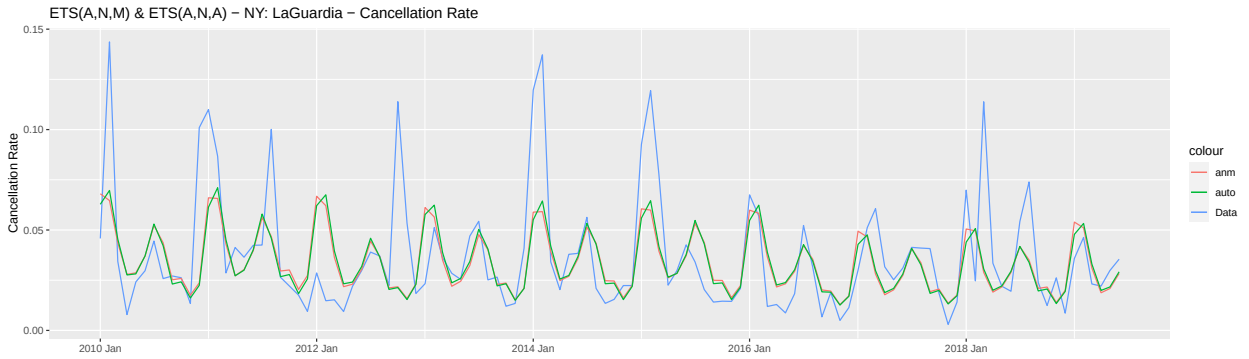

Figure 28: ETS(A,N,M) & ETS(A,N,A) - NY: LaGuardia - Cancellation Rate

5. ARIMA

Number of Flights of IL: Chicago O'Hare International

After considering the models previously described, it is now time to develop ARIMA models – models that take into account temporal dependency. Note that it is required to have stationary time series. For a series to be stationary both the mean and variance have to be constant over time. Plus, the covariance between observations also has to be constant (at most depending on the lag).

So, first, let's look at the data and check if there is the need to apply first differences and/or seasonal differences. (Working for the Chicago O'Hare airport)

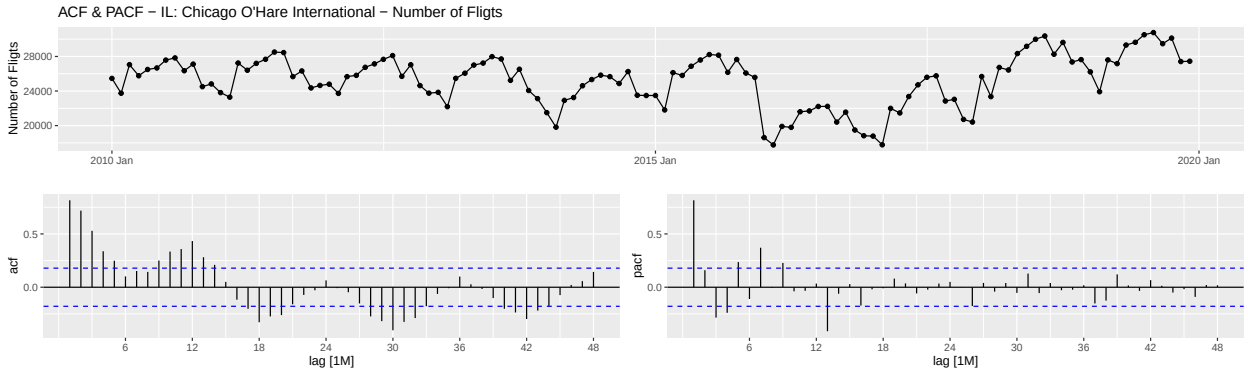

Figure 29: ACF & PACF - IL: Chicago O'Hare International - Number of Fligts

Looking at the data, we will apply both first differences and seasonal differences in order to get a stationary series.

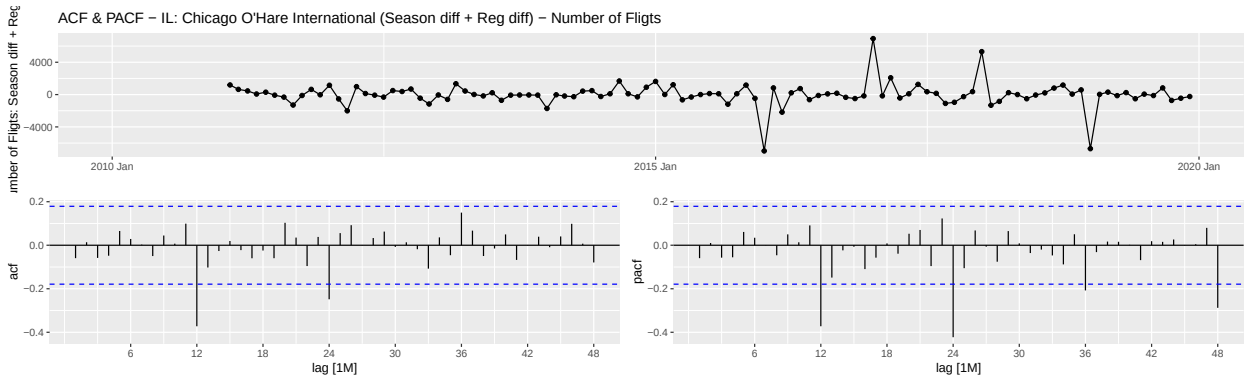

Figure 30: ACF & PACF - IL: Chicago O'Hare International (Season diff + Reg diff) - Number of Fligts

After getting a stationary series and looking at both ACF and PACF to decide on the Moving Average part and Autoregressive part, respectively, several were proposed, as well as the auto arima (developed by Hyndman). Note that, for each model, the second parameter is set to 1 which corresponds to applying first differences.

- ARIMA(0,1,0)(0,1,1): The acf shows a spike at lag 12, which suggest a SMA(1).
- ARIMA(0,1,0)(0,1,2): The acf shows a spike at lag 12 and 24, which suggest a SMA(2).
- ARIMA(0,1,0)(1,1,2): The acf and pacf shows a spike at lag 12, which suggest a SAR(1) and SMA(1).
- ARIMA(0,1,0)(1,1,2): The acf and pacf shows a spike at lag 12, which suggest a SAR(1) and SMA(2).
- ARIMA(0,1,0)(2,1,2): The acf and pacf shows a spike at lag 12 and 24, which suggest a SAR(2) and a SMA(2).
- Auto: ARIMA(0,1,0)(2,1,0).

Table 11: Ljung box test ARIMA models - IL: Chicago O'Hare International - Number of Flights

.model	lb_stat	lb_pvalue
arima010011	22.92230	0.4653161
arima010012	22.91494	0.4065461
arima010111	22.91853	0.4063421
arima010112	18.44886	0.6204491
arima010212	13.40262	0.8594488
auto	18.96057	0.6477618

Looking at the Ljung box test we can see that all models reject the H0 and this means that the errors are uncorrelated.

Table 12: ARIMA metrics - IL: Chicago O'Hare International - Number of Flights

Description	.model	sigma2	log_lik	AIC	AICc	BIC
Chicago, IL: Chicago O'Hare International	arima010212	724887.3	-849.3178	1708.636	1709.267	1721.711
Chicago, IL: Chicago O'Hare International	arima010011	1029758.5	-855.4147	1714.829	1714.952	1720.060
Chicago, IL: Chicago O'Hare International	arima010112	1007019.4	-853.7840	1715.568	1715.985	1726.028
Chicago, IL: Chicago O'Hare International	arima010012	1041154.6	-855.4142	1716.828	1717.076	1724.674
Chicago, IL: Chicago O'Hare International	arima010111	1040631.5	-855.4144	1716.829	1717.076	1724.674
Chicago, IL: Chicago O'Hare International	auto	1363416.7	-859.8703	1725.741	1725.988	1733.586

By analyzing the AICc of the models, note that an ARIMA(0,1,0)(2,1,2) has the best performance. On the next plot the fitted model is presented on top of the data:

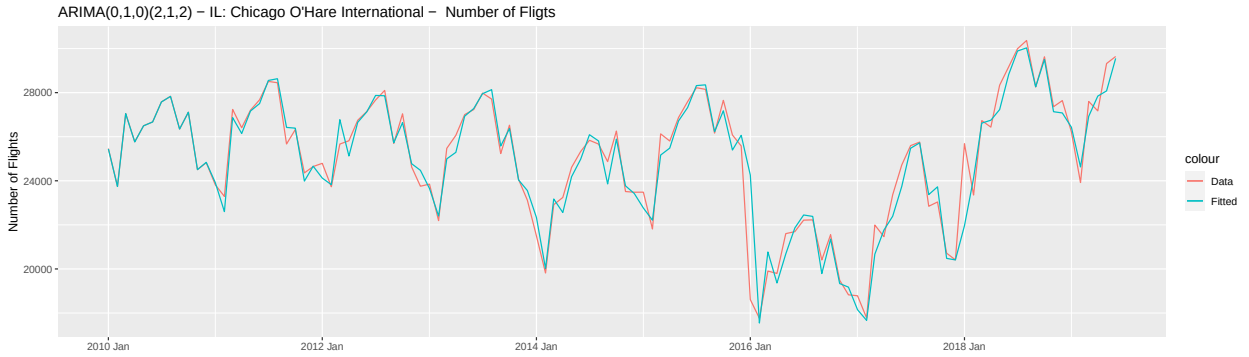

Figure 31: ARIMA(0,1,0)(2,1,2) - IL: Chicago O'Hare International - Number of Fligts

Cancelation Rate of New York, NY: LaGuardia

Similarly to the ETS models, for the Cancellation Rate, we will focus on the La Guardia airport, New York. The first step is, once again, to plot the data and check for the need of differences.

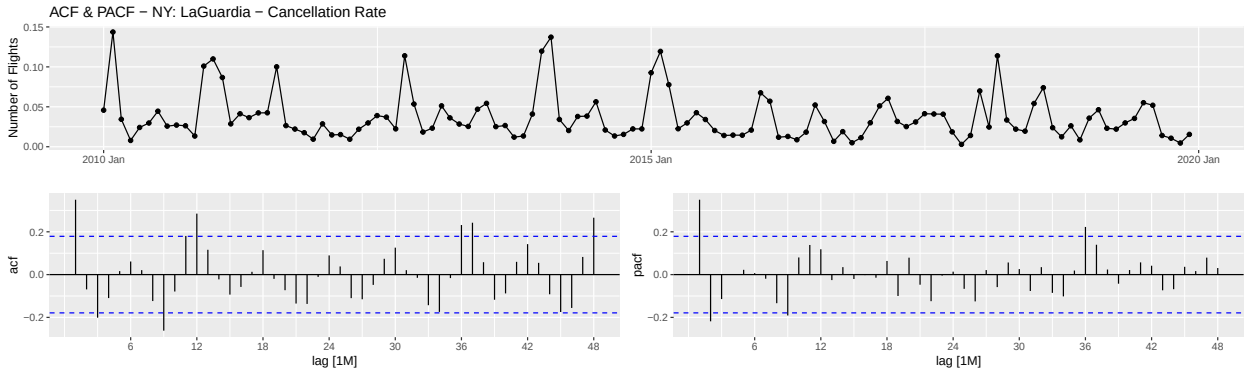

Figure 32: ACF & PACF - NY: LaGuardia - Cancellation Rate

We decided not to apply any differences, since the series seems fairly stationary. Focusing on the ACF and PACF, we thought the following were reasonable:

- ARIMA(0,0,1)(0,0,1): The ACF shows a spike at lag 1, which suggests a MA(1). Plus, there is also a spike at lag 12, suggesting a SMA(1). Regarding the PACF it is gradually decaying.
- ARIMA(0,0,1)(1,0,0): Once again, the spike at lag 1 from the ACF suggests a MA(1). Spotting seasonality on this time series leads us to try a SAR(1).
- ARIMA(0,0,3)(0,0,1): The ACF shows a spike at lag 3, which suggests a MA(3). Plus, there is also a spike at lag 12, suggesting a SMA(1). Regarding the PACF it is gradually decaying.
- ARIMA(0,0,3)(1,0,0): Once again, the spike at lag 3 from the ACF suggests a MA(3). Spotting seasonality on this time series leads us to try a SAR(1).
- Auto: ARIMA(0,0,1)(1,0,0).

Table 13: Ljung box test ARIMA models - NY: LaGuardia - Cancelation Rate

.model	lb_stat	lb_pvalue
arima001001	30.08039	0.1165236
arima001100	26.21983	0.2423635
arima003001	27.35514	0.1255649
arima003100	24.93715	0.2038461
auto	26.21983	0.2423635

Looking at the Ljung box test we can see that all models reject the H0 and this means that the errors are uncorrelated.

Table 14: ARIMA metrics - NY: LaGuardia - Cancellation Rate

Description	.model	sigma2	log_lik	AIC	AICc	BIC
New York, NY: LaGuardia	arima001100	0.4486709	-114.9205	237.8410	238.2079	248.7858
New York, NY: LaGuardia	auto	0.4486709	-114.9205	237.8410	238.2079	248.7858
New York, NY: LaGuardia	arima001001	0.4515071	-115.2253	238.4506	238.8176	249.3954
New York, NY: LaGuardia	arima003100	0.4559210	-114.7667	241.5334	242.3184	257.9506
New York, NY: LaGuardia	arima003001	0.4582364	-115.0104	242.0209	242.8059	258.4380

After analyzing the AICc criterion, the ARIMA(0,0,1)(1,0,0) proved to have the best results, being plotted as it follows:

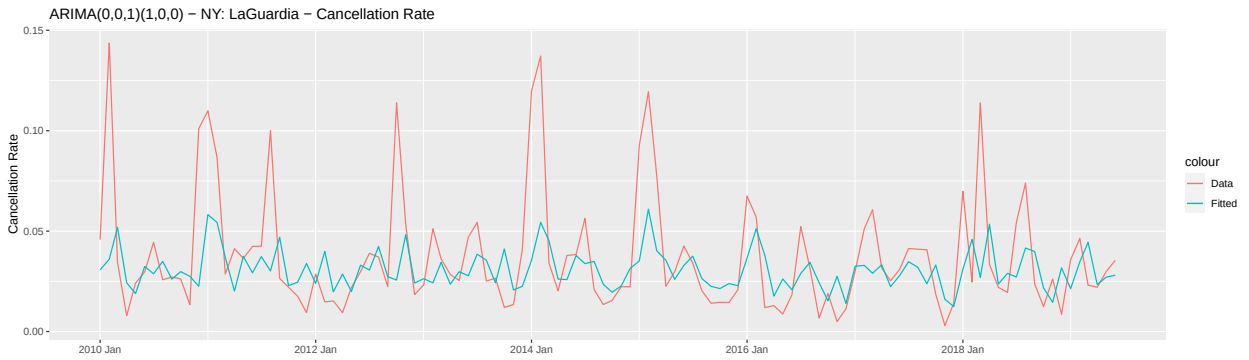

Figure 33: ARIMA(0,0,1)(1,0,0) - NY: LaGuardia - Cancellation Rate

6. Model comparision

On this stage of the project, we aim at comparing the models previously trained and tested.

Number of flights

The following table compiles the models detailed on the previous stages, for each airport, regarding the number of flights:

Table 15: Models accuracy - Number of flights

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	arima	36.98840	47.98864	42.63852
Anchorage, AK: Ted Stevens Anchorage International	drift	-442.46608	616.94116	538.71386
Anchorage, AK: Ted Stevens Anchorage International	ets	-12.70655	59.73660	51.93726

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	mean	209.56140	467.60333	354.00000
Anchorage, AK: Ted Stevens Anchorage International	naive	-415.33333	589.26960	519.33333
Anchorage, AK: Ted Stevens Anchorage International	snaive	32.33333	48.27007	42.33333
Anchorage, AK: Ted Stevens Anchorage International	trend	220.93906	472.55498	353.71556
Chicago, IL: Chicago O'Hare International	ana	624.36083	706.49524	624.36083
Chicago, IL: Chicago O'Hare International	arima010212	364.88950	571.30687	469.78959
Chicago, IL: Chicago O'Hare International	snaive	416.33333	607.20562	480.33333
Los Angeles, CA: Los Angeles International	arima	385.14365	442.95789	385.14365
Los Angeles, CA: Los Angeles International	drift	-322.34513	1007.96836	943.72271
Los Angeles, CA: Los Angeles International	ets	381.53583	472.51210	381.53583
Los Angeles, CA: Los Angeles International	mean	671.41228	1143.95430	897.13743
Los Angeles, CA: Los Angeles International	naive	-235.00000	955.54278	881.33333
Los Angeles, CA: Los Angeles International	snaive	-69.50000	207.29407	178.16667
Los Angeles, CA: Los Angeles International	trend	236.19995	963.85838	811.96010
New York, NY: LaGuardia	arima	-163.27554	424.33735	334.80863
New York, NY: LaGuardia	drift	-413.66814	801.58823	650.04867
New York, NY: LaGuardia	ets	81.92077	230.29342	172.55517
New York, NY: LaGuardia	mean	4983.26316	5020.24339	4983.26316
New York, NY: LaGuardia	naive	-212.00000	644.10869	556.66667
New York, NY: LaGuardia	snaive	-48.00000	249.06626	228.66667
New York, NY: LaGuardia	trend	2647.26235	2728.44000	2647.26235

"Competition" for the best models has the following winners, for each airport:

Table 16: Best models - Number of Flights

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	arima	36.98840	47.98864	42.63852
Chicago, IL: Chicago O'Hare International	arima010212	364.88950	571.30687	469.78959
Los Angeles, CA: Los Angeles International	snaive	-69.50000	207.29407	178.16667
New York, NY: LaGuardia	ets	81.92077	230.29342	172.55517

An important stage of any project of this nature is to produce and plot forecasts. To not turn the report too extensive, forecasts will be displayed only for the Chicago O'Hare airport – using the best model amongst all trained (which was the ARIMA(0,1,0)(2,1,2).

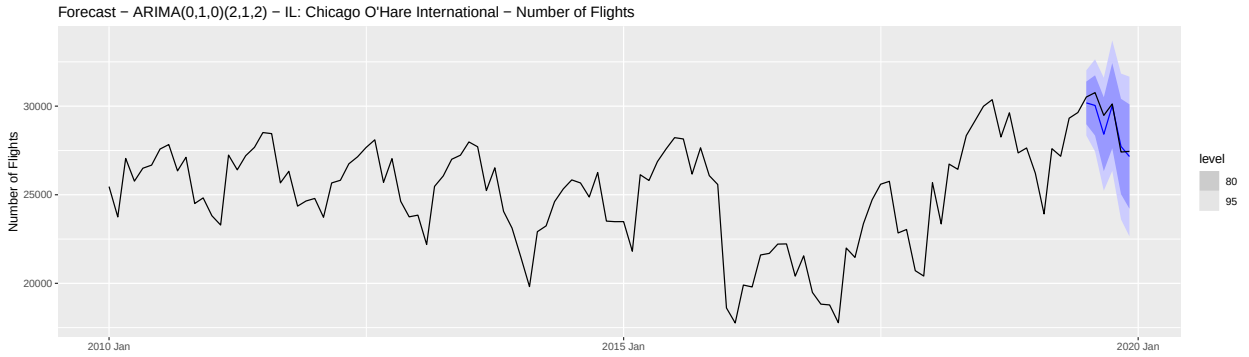

Figure 34: Forecast - ARIMA(0,1,0)(2,1,2) - IL: Chicago O'Hare International - Number of Flights

Cancelation rate

The same process is applied to the Cancellation Rate series:

Table 17: Models accuracy - Cancellation rate

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	arima	-0.0020263	0.0038845	0.0033866
Anchorage, AK: Ted Stevens Anchorage International	drift	-0.0182994	0.0195306	0.0182994
Anchorage, AK: Ted Stevens Anchorage International	ets	-0.0010520	0.0020690	0.0017419
Anchorage, AK: Ted Stevens Anchorage International	mean	-0.0019948	0.0038531	0.0033592
Anchorage, AK: Ted Stevens Anchorage International	naive	-0.0176770	0.0187905	0.0176770
Anchorage, AK: Ted Stevens Anchorage International	snaive	-0.0032339	0.0053680	0.0040956
Anchorage, AK: Ted Stevens Anchorage International	trend	0.0004155	0.0033669	0.0027941
Chicago, IL: Chicago O'Hare International	arima	-0.0040663	0.0106890	0.0094303
Chicago, IL: Chicago O'Hare International	drift	-0.0400055	0.0487283	0.0407324
Chicago, IL: Chicago O'Hare International	ets	0.0005653	0.0104574	0.0088523
Chicago, IL: Chicago O'Hare International	mean	-0.0056152	0.0125056	0.0100374
Chicago, IL: Chicago O'Hare International	naive	-0.0392042	0.0476933	0.0399329
Chicago, IL: Chicago O'Hare International	snaive	-0.0038149	0.0148863	0.0114190
Chicago, IL: Chicago O'Hare International	trend	0.0006189	0.0110764	0.0095455
Los Angeles, CA: Los Angeles International	arima	-0.0010447	0.0013969	0.0012405
Los Angeles, CA: Los Angeles International	drift	-0.0157273	0.0164313	0.0157273
Los Angeles, CA: Los Angeles International	ets	-0.0018221	0.0026307	0.0018221
Los Angeles, CA: Los Angeles International	mean	-0.0018746	0.0024083	0.0020242
Los Angeles, CA: Los Angeles International	naive	-0.0154941	0.0161553	0.0154941
Los Angeles, CA: Los Angeles International	snaive	0.0004919	0.0010060	0.0008096
Los Angeles, CA: Los Angeles International	trend	-0.0008210	0.0016974	0.0015292
New York, NY: LaGuardia	ana	-0.0038991	0.0097250	0.0092928
New York, NY: LaGuardia	arima001100	-0.0109552	0.0187838	0.0171114

Description	.model	ME	RMSE	MAE
New York, NY: LaGuardia	trend	-0.0055682	0.0208880	0.0205447

This time, winners are the next models:

Table 18: Best models - Cancellation Rate

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	ets	-0.0010520	0.0020690	0.0017419
Chicago, IL: Chicago O'Hare International	ets	0.0005653	0.0104574	0.0088523
Los Angeles, CA: Los Angeles International	snaive	0.0004919	0.0010060	0.0008096
New York, NY: LaGuardia	ana	-0.0038991	0.0097250	0.0092928

The forecasts are produced and displayed using the ARIMA(0,0,1)(1,0,0) for La Guardia airport, New York:

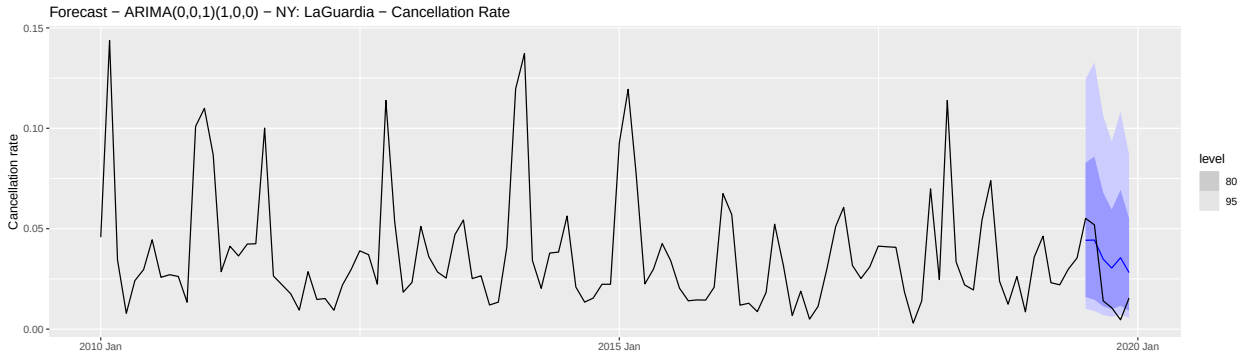

Figure 35: Forecast - ARIMA(0,0,1)(1,0,0) - NY: LaGuardia - Cancellation Rate

7. Impact of Covid-19 in our models accuracy

The final segment of this project covers the impact of covid19 on the numbers of the airports. To develop the previous models, 2020 was excluded since it was such an atypical year translated to fairly unstable flight data.

The process on this stage is to take the best models produced before, for each airport and for both number of flights and cancellation rates and re-train and test them, using as test date the 2020 period.

Number of flights

Before we dive into the analysis, data regarding 2020 number of flights is displayed:

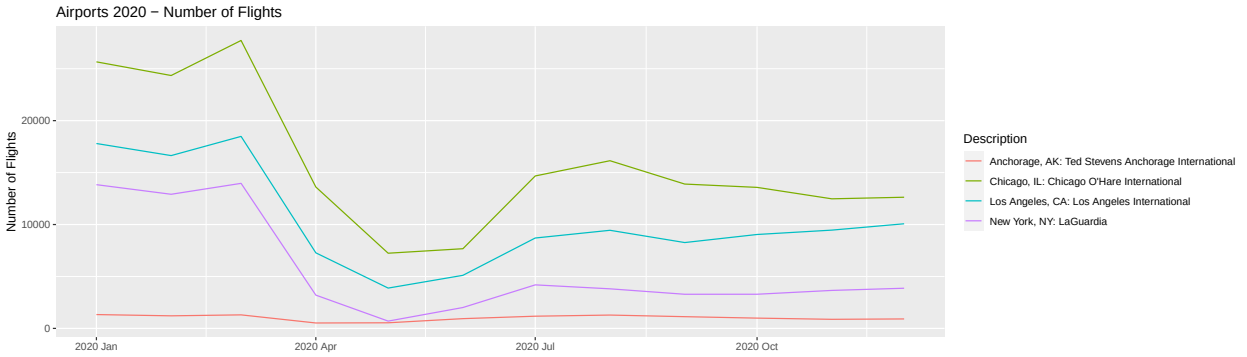

Figure 36: Airports 2020 - Number of Flights

As seen above, the number of flights suffered an intense decrease since around March 2020.

Next, we compare models' accuracy from post and pre covid:

-Post Covid

Table 19: Models accuracy post 2020 - Number of Flights

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	arima	-558.9625	693.1655	562.9291
Chicago, IL: Chicago O'Hare International	arima010212	-10071.8831	12470.4846	10960.9884
Los Angeles, CA: Los Angeles International	snaive	-7982.0000	9374.8774	8036.8333
New York, NY: LaGuardia	ets	-8422.7819	9640.4722	8422.7819

By considering 2020 to test our models, the RMSE is considerably higher than what is obtained by not considering that period (Table 18). This is expected by the instability described above.

By taking a look at the forecasts produced regarding the Chicago O'Hare airport, this difference is evident:

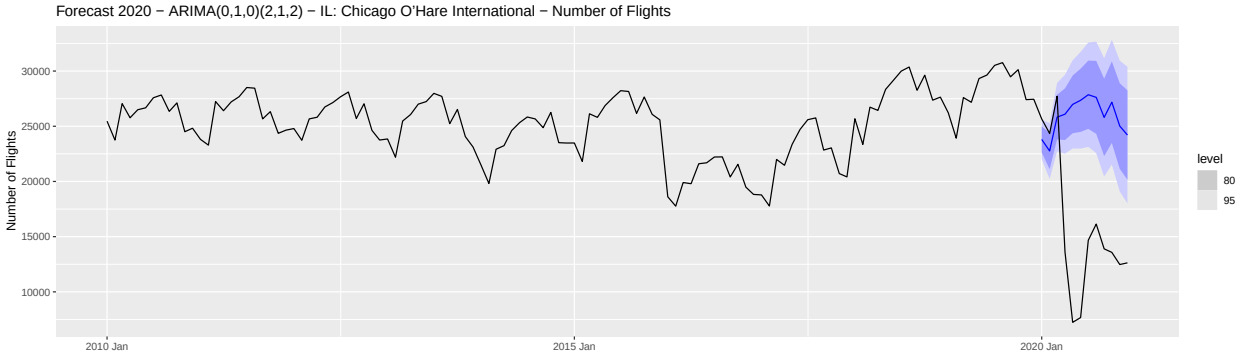

Figure 37: Forecast 2020 - ARIMA(0,1,0)(2,1,2) - IL: Chicago O'Hare International - Number of Flights

Cancelation rate

Just as for the number of flights, we will now analyze data concerning the cancellation rate taking 2020 into consideration. A significant increase is expected.

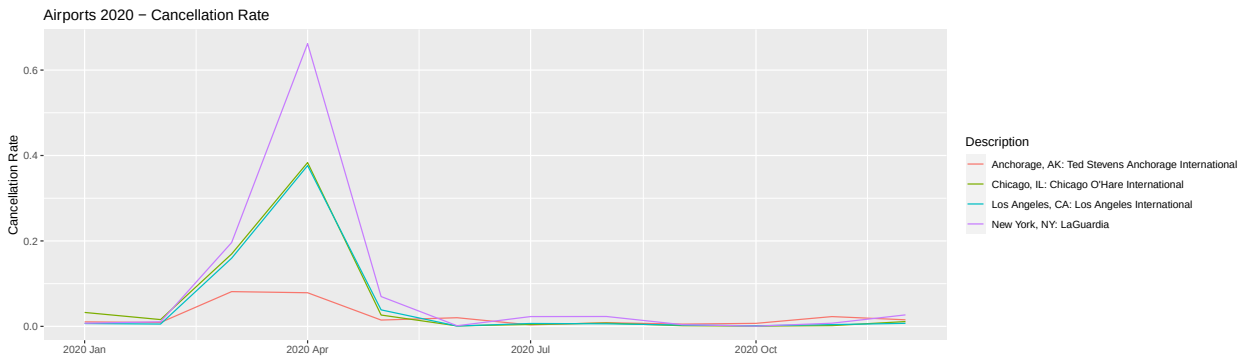

Figure 38: Airports 2020 - Cancellation Rate

Note that, as expected an increase of cancellation rates was seen. Comparing the four airports in study, Ted Stevens, AK, suffered the least increase whereas La Guardia, NY had the highest rate of cancellations at around April.

Once again the best models for each airport will now be trained using all data (until December 2019) and then tested on the critical period (2020).

-Post Covid

Table 20: Models accuracy post 2020 - Cancellation Rate

Description	.model	ME	RMSE	MAE
Anchorage, AK: Ted Stevens Anchorage International	ets	1.30746e-02	0.0302902	0.0171433
Chicago, IL: Chicago O'Hare International	ets	2.55999e-02	0.1144924	0.0588160
Los Angeles, CA: Los Angeles International	snaive	-7.98200e+03	9374.8774036	8036.8333333
New York, NY: LaGuardia	arima001100	5.12534e-02	0.1888443	0.0861535

As expected, if we take a look at the RMSE from our models, there is a significant increase when including 2020 data comparing with the pre 2020 results (Table 20), particularly for the snaive model (modelled Los Angeles International data).

This is clearly observable when plotting forecasts for 2020 and compare them with the real data. On the following plot, this comparison is displayed for La Guardia, NY. The model estimated much lower cancellation rates, not being able to respond to such unprecedented changes.

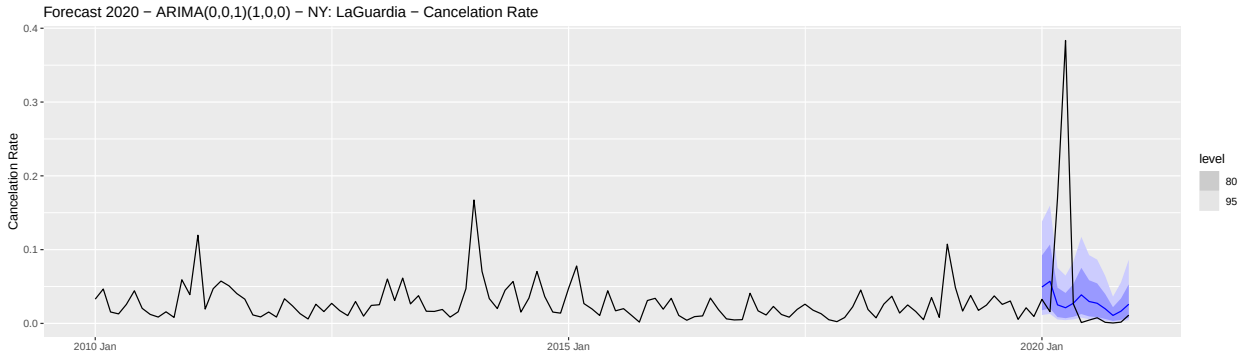

Figure 39: Forecast 2020 - ARIMA(0,0,1)(1,0,0) - NY: LaGuardia - Cancelation Rate

Conclusion

Reaching the end of this project, we now have a better understanding on how to go from historical data to being able to produce considerably accurate forecasts. In fact, one thing that stood our attention was the different ways each model "looks" at data, fitting it in its own way. As described through the project, we want to focus on different sets of models, analyze its properties and assumptions to then compare and come up with the best that we could produce. Note that for each airport, the most appropriate model was different when working with the number of flights and working with the cancellation rate. For example, for La Guardia, New York, on the time series for the number of flights, the best model was an ETS, whereas for the cancellation rate it was an ARIMA(0,0,1)(1,0,0). As was stated on the introduction, one other goal of this project was to compare 2020 data when using the models produced without considering this period. 2020 was a difficult year across all industries – including the aviation industry. Surely the number of flights suffered a fall at around March/April and the cancellation rates increased on a big scale. By comparing our previously trained and tested models excluding 2020, we then noticed an increase on the errors produced (as described on section 7 of this report). We surely hope numbers will go back to "normal" in the future and the industry is capable of recovering from such a tough period. To conclude, we are now more capable of analyzing time series, understating their features and singularities, and treat them in order to produce accurate forecasts.