Analyzing Passenger Loads in The Valle de Aburrá Transportation Network

Área Metropolitana Del Valle De Aburrá

Project Team No. 21
Felipe Garcia, Sebastian Marin, Boris Martinez, Sandra Ruiz, Jorge Saavedra, Christian Velez, Carlos Taimal.

Abstract

Big cities around the world face the problem of optimizing their public transportation systems in order to provide a better service that fulfills the passenger's needs. Public entities have to deal with a complex situation where it converges diverse variables like the growth of the city, number of routes, number of passengers, number of buses, number of stops, schedules, etc. in order to balance demand and supply efficiently. Nowadays, thanks to the capacity to register huge amounts of data about the transport services in realtime, and the increasing computational capacity to process big chunks of data we can use Data Science methodologies that allow us to understand, analyze, and visualize these complex systems and therefore improve their planning and management.

Thus, and under the framework of the Data Science 4 All course, it is proposed through different kinds of methodologies to provide the Área Metropolitana del Valle de Aburrá (Metropolitan Area of the Aburrá Valley and hereinafter, AMVA) with tools that allow making decisions based on evidence and figures so that they can provide a more optimal service that adapts to the changing needs of the passengers who use it daily.

Team 21, has proposed through visual methods in the first instance, to be able to show the entity how the demand for the services they provide to the citizens of the AMVA behaves, by route, by an hour and day of the week as well as by road sections; besides, tools are provided to visualize the road corridors where the demand is greatest to take action accordingly.

Next, the data collection process, cleaning, the allocation of demand to the corresponding road sections, and as requested by the entity, the different problems that arose during the process, and some descriptions of interest are detailed. Subsequently, some methodological proposals are

shown to forecast the demand for transportation under the Time series framework and under the Machine Learning one using Random Forests and some problems that could be solved in later works are exposed.

1. Problem Statement and Background

Problem background

The planning of public transportation services is one of the most important areas of city planning. This allows us to generate an accurate offer to passengers without increasing operational costs nor generate traffic jams in important avenues. With the information on the number of passengers on each route and each point in the city, transportation planning policies can be designed to benefit the community in terms of saving travel time or optimizing transportation routes to expand the coverage of the transportation network.

As of 2017, on a daily basis, the transportation network of the AMVA mobilized 613.200 passengers such that 42% of them used the public transportation system, that is, 255.200 passengers of the total for whom the principal destination was Medellín. Of these travels, 66% were of mandatory nature (work and study), 22% corresponded to running errands and 14% of these were because of healthcare reasons, recreational nature, among others (Área Metropolitana del Valle de Aburrá, 2020a).

Additionally, it has been found that the maximum demand occurs around 6 - 7 AM such that 11% of the travels made in the day happened in this interval. Between the factors that influence the demand, there are the disponibility, time, and fees while factors such as the driver's attention and the user service did not influence it as much to get from one place to another (Área Metropolitana del Valle de Aburrá, 2020a).

In the area in question, there is the Sistema Integrado de Transporte del Valle de Aburrá, SITVA, which agglomerates among its services EnCicla (Public Bicycle System), which has 1.680 bicycles and as of today there are 100.893 passengers registered, the Metro (which moves 800.000 people a day on average along 31.3 km), the Cables (11.9 km of cables through which the city slopes are accessed by mobilizing approximately 41.000 passengers), the Tranvía (4.3 km line that mobilizes approximately 45.000 passengers), the Metroplus that daily mobilizes 125.000 passengers in a 26 k.m. network with a fleet of 30 articulated buses and 47 standard buses, the Buses Alimentadores and, Rutas Integradas which transport throughout the area approximately 110.000 people on 35 routes made up of 302 buses of 40 passengers and 65 of 19 that connect 1.033 stops in the AMVA. Some of these services are summarized below (Área Metropolitana del Valle de Aburrá, 2020b):

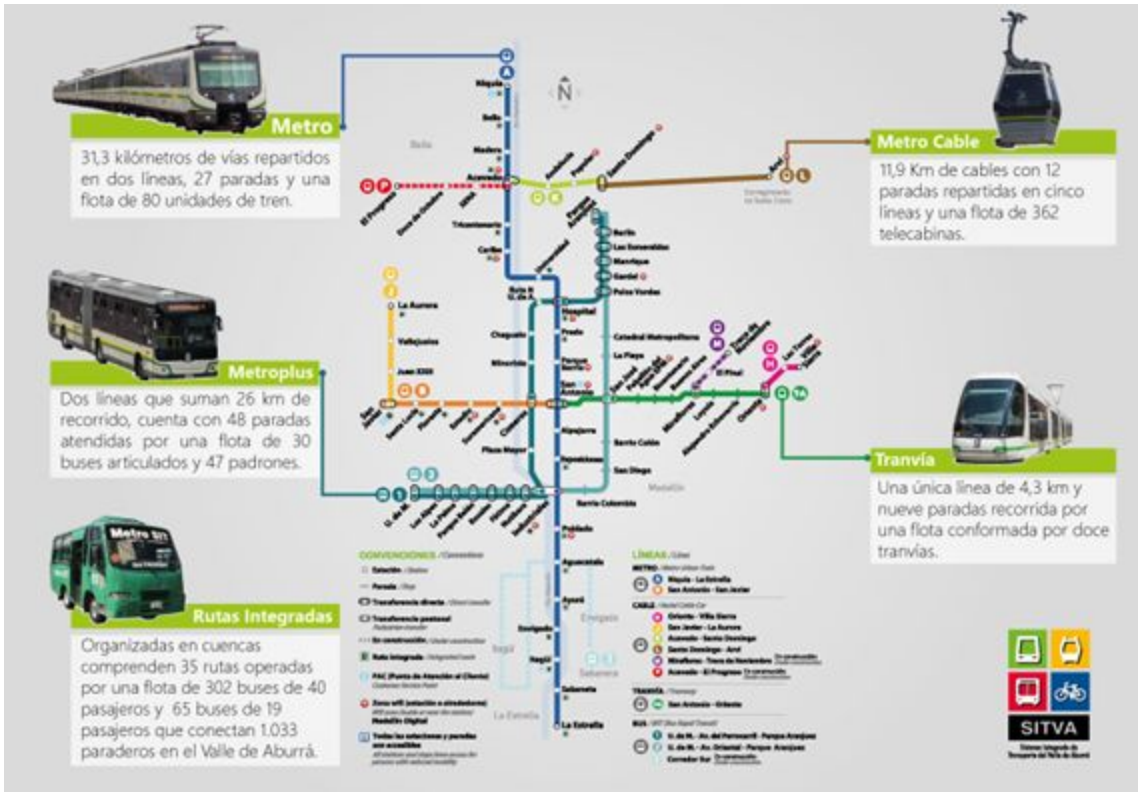


Fig.1 SITVA System. From: Área Metropolitana del Valle de Aburrá (2020b)

Lastly, another service that makes up the SITVA is the Sistema Integrado de Transporte Público Colectivo (TPC) that operates both at the municipal and metropolitan level and is made up of routes that run between one and two municipalities of the AMVA; the latter's services are provided by different transport companies and it is divided into 9 service areas or basins which are made up of specific sectors of the AMVA. These are:

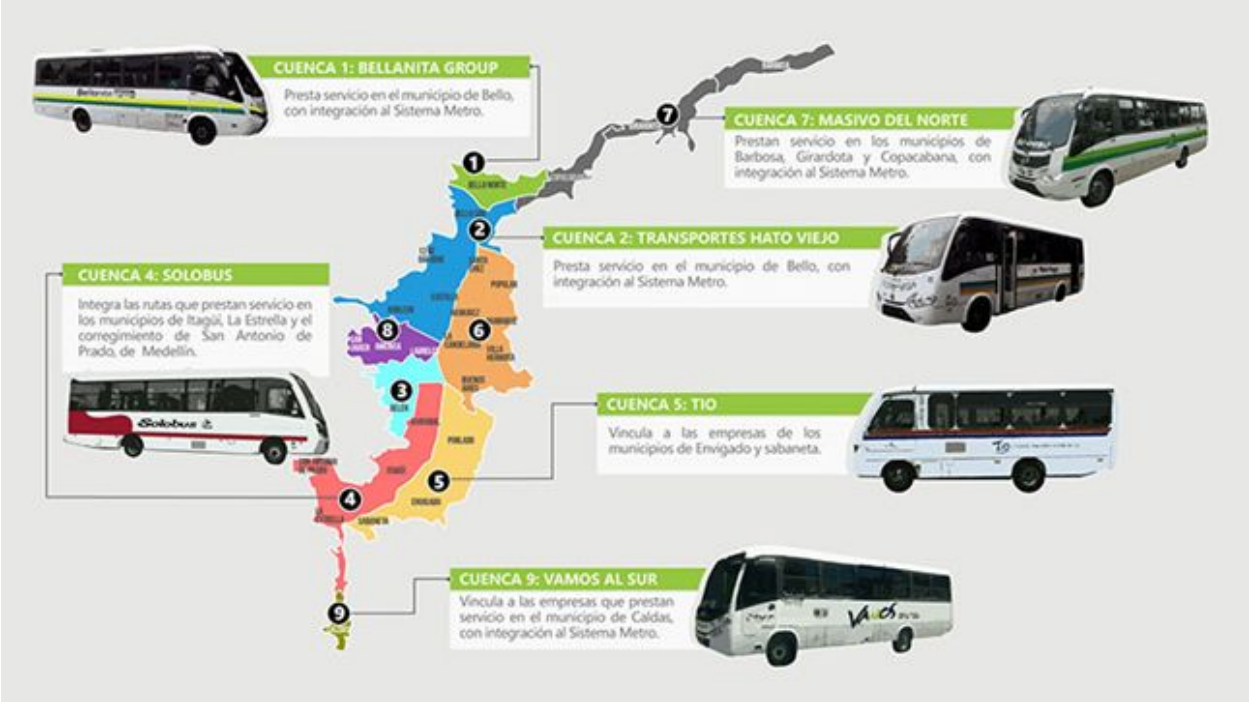


Fig.2 TPC operators. From: Área Metropolitana del Valle de Aburrá (2020b)

The here proposed solution is focused on the transport demand of this last service, the TPC, for which it is of interest to determine the distribution of passenger demand along the road corridors that make up the AMVA throughout the day and the routes for which the travels are given, this in a dashboard that allows visualizing the demand for each of the mentioned elements, either, a specific time or spatial intervals.

Problem statement

The Área Metropolitana del Valle de Aburrá area is formed by ten cities of Antioquia department: Barbosa, Bello, Caldas, Copacabana, Girardota, Envigado, Itagui, Sabaneta, Medellín and La Estrella and has a population of 3.9 million inhabitants and 1.157 km². As the second biggest urban agglomeration of Colombia, they face great challenges in urban mobility matters and currently they seek to identify high demand patterns in order to implement strategies that improve the competitiveness of public transportation (Wikipedia, n.d.).

The government entity wants to measure the passenger demand in the transportation network of Valle de Aburrá Region, particularly, the passenger load in each road network's arc. The expected results must be able to be filtered by date (time of the day, day of the week, or in

general, a time interval) and by route, so that way it is easier to observe the demand levels for the public transportation service through space and time parameters.

The methods to reach the result will be georeferenced data analysis, finding through heat maps where the highest passenger density occurs, and eventually, with the proper algorithm use the stations' data to find the nearest arc and then associate that demand in order to properly allocate the public transportation resources.

Thus, the main problem exposed by the entity is the adequate aggregation of the demand data to the respective sections that make up the TPC network. This corresponds to a problem of calculating distances between events from data supplied by the entity; both those of demand and the geographical ones that describe the stops and the length of the routes. Additionally and in the context of the course, prediction models based on the temporal correlation structure that determines the demand are proposed, as well as elements of a geographical nature, this as a plus to facilitate the planning of the demand by the entity.

2. Data wrangling and cleaning

Vehicles circulating in the transport network are equipped with passenger counter sensors that transmit information to a central by each one of the operators and logs these events: bus speed 0 Km/h, distance traveled 240 meters, the time that has passed 30 seconds, or change in the state of the doors. At each event, it is recorded the number of passengers boarding, alighting and the latitude and longitude of the stop. Then, data is transmitted online or in batch to the GTPC system which validates the information received for future control and reporting, as depicted in figure 3:

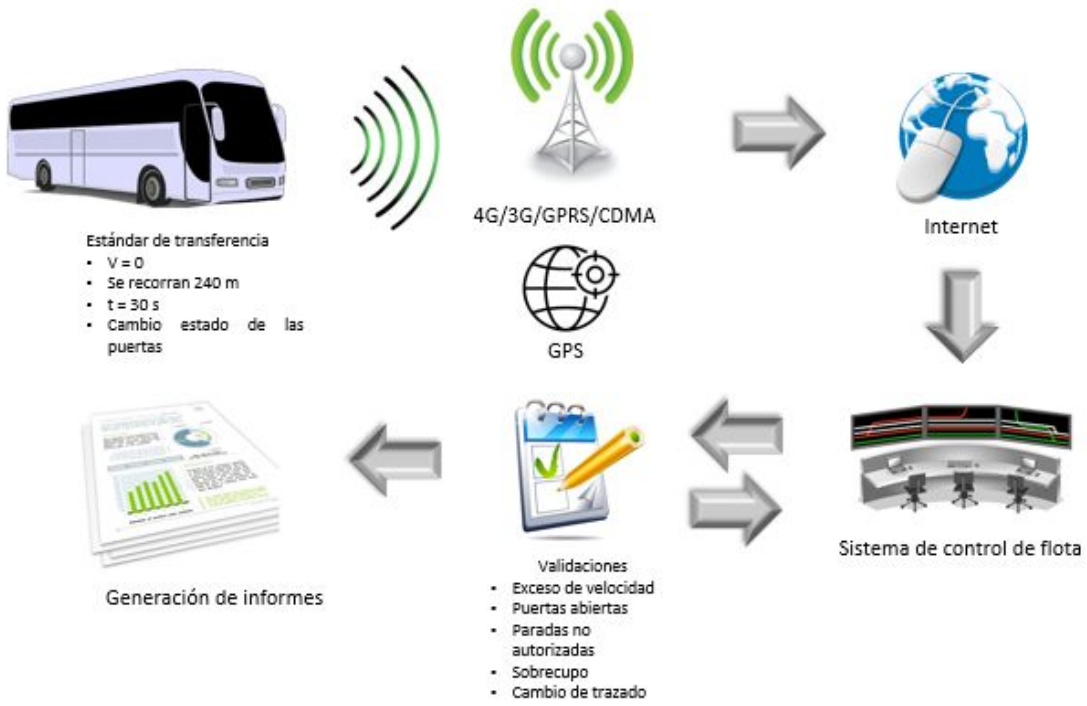


Fig.3 Data gathering from sensors

The system stores huge volumes of information every day (6 million records on average per day) and sometimes data about the location of the vehicle or the number of passengers recorded at each door are empty, imprecise or wronged. For this study, we took six months of operations from November 1st of 2019 until May 10th of 2020 which corresponds to 930 million records.

For the preprocessing of data, we proceed as follows (figure 4):

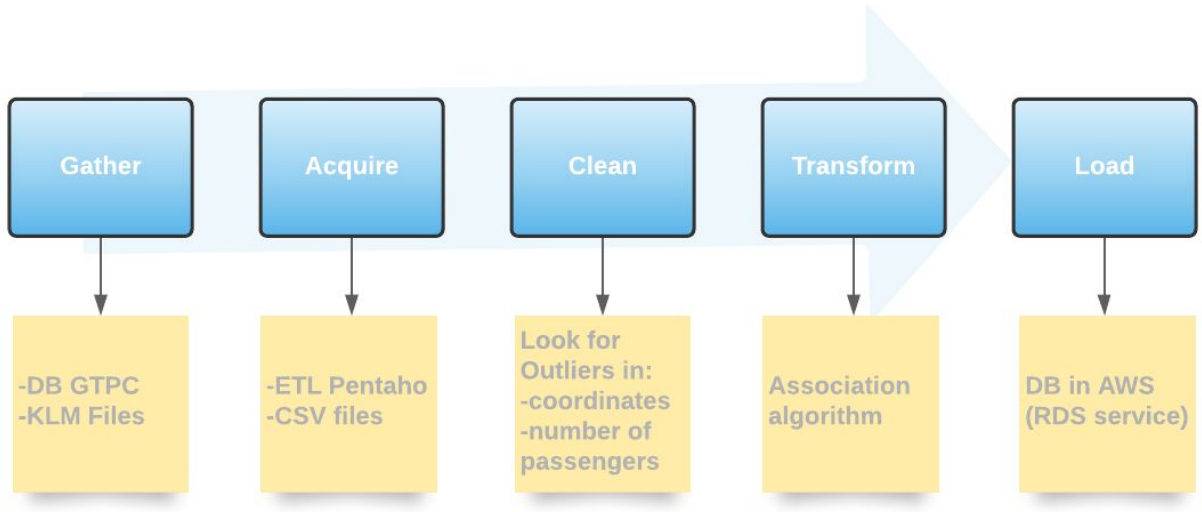


Fig.4 Data wrangling and cleaning stages

Gather: at this stage, we could access the source database (Oracle 12c) where all information has been collected. The model corresponds to 292 tables that store information about operators, files, capacities, speed, etc. but in our case, we only took data from tables about events, travels, defined routes, transport operators, and vehicles as described in this diagram:

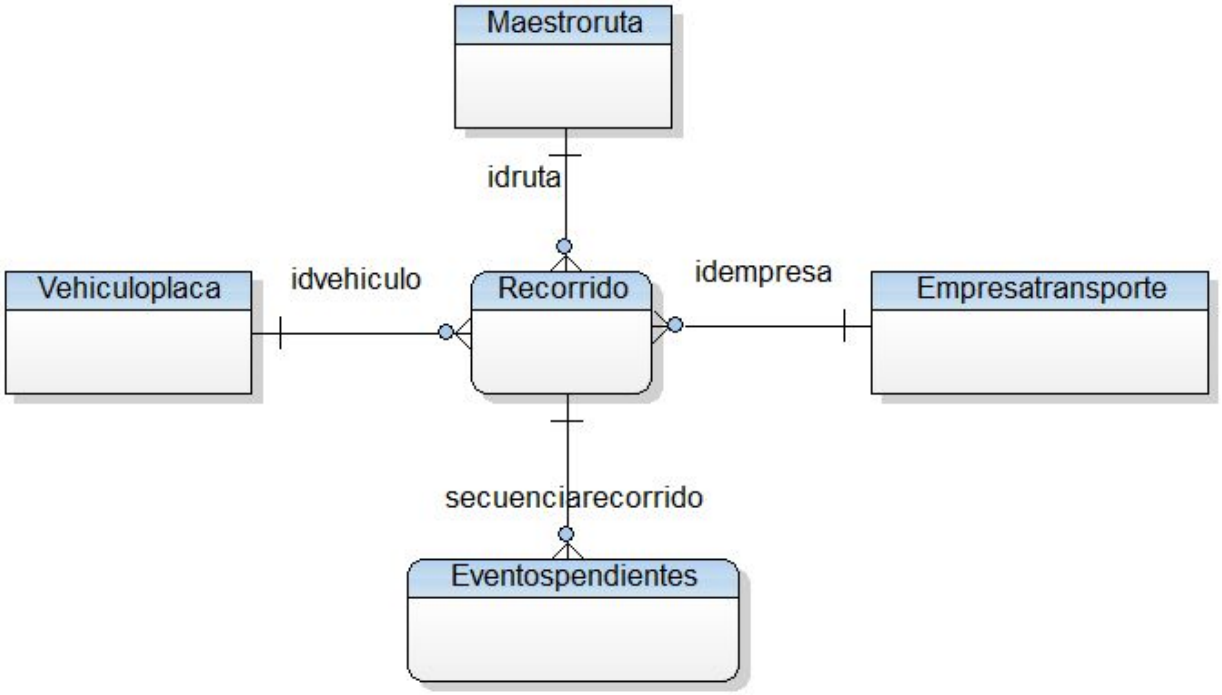


Fig.5 ER source model

On the other hand, we were given the KML (Keyhole Markup Language) files to display the geographic data about the routes on the maps.

Acquire:

For ETL (Extract, Transform, Load) steps, we used the Pentaho Data Integration tool to extract in CSV files only records that had values greater than zero on at least one of the columns of passengers boarding and alighting. Hence, we reduce data to 73 million records.

The data was stored by each day of the year such that each file has the following structure,

A	B	C	D	E	F	G	H	I	J	K
SECUENCIARE	RECORRIDOFIN	IDVEHICULO	CODIGORUTA	FECHAREGISTI	LATITUD	LONGITUD	SUBENDELANTERA	SUBENTRASERA	BAJANDELANTERA	BAJANTRASERA
196256842	S	6126	1011	2020-05-10 6:03	6.341.093	-75.545.803	0	0	0	2
196256842	S	6126	1011	2020-05-10 5:36	6.257.029	-75.572.562	0	0	0	1
196256842	S	6126	1011	2020-05-10 5:41	6.249.944	-7.557.494.200.(	1	0	0	0
196256842	S	6126	1011	2020-05-10 5:44	6.263.747	-75.574.894	0	0	0	1
196256842	S	6126	1011	2020-05-10 5:45	6.271.945	-75.573.454	0	1	0	0

Fig.6 Initial dataset example

Where, for each travel made (SECUENCIARECORRIDO) for a given vehicle (IDVEHICULO), and route (CODIGORUTA) in a given day, the information regarding where (LATITUD and LONGITUD) and when (FECHAREGISTRO) an event happened was captured by the censors and stored as specified above such that an event constitutes a boarding (SUBENDELANTERA and SUBENTRASERA) or alighting (BAJANDELANTERA and BAJANTRASERA) from the vehicle at issue. The data comprised 192 days or a bit more than 6 months of observations with some cases where there were observations for all around the day and in some other just for a few hours out of 24 that compose a day.

For the analysis of transport network data, it was also necessary to visualize the configuration of the road network on which the vehicles operate. Using files about the nodes and segments in gis format, we could visualize the data of the segments that make up the network.

Configuration parameters of the shapefile of the road network are given below:

Name: WGS 84 / UTM zone 18N Axis Info [cartesian]: - E[east]: Easting (metre) - N[north]: Northing (metre) Area of Use: - name:World -Nhemisphere - 78°W to 72°W - by country - bounds: (-78.0, 0.0, -72.0, 84.0) Coordinate Operation: - name: UTM zone 18N - method: Transverse Mercator Datum: World Geodetic System 1984 - Ellipsoid: WGS 84 - Prime Meridian: Greenwich

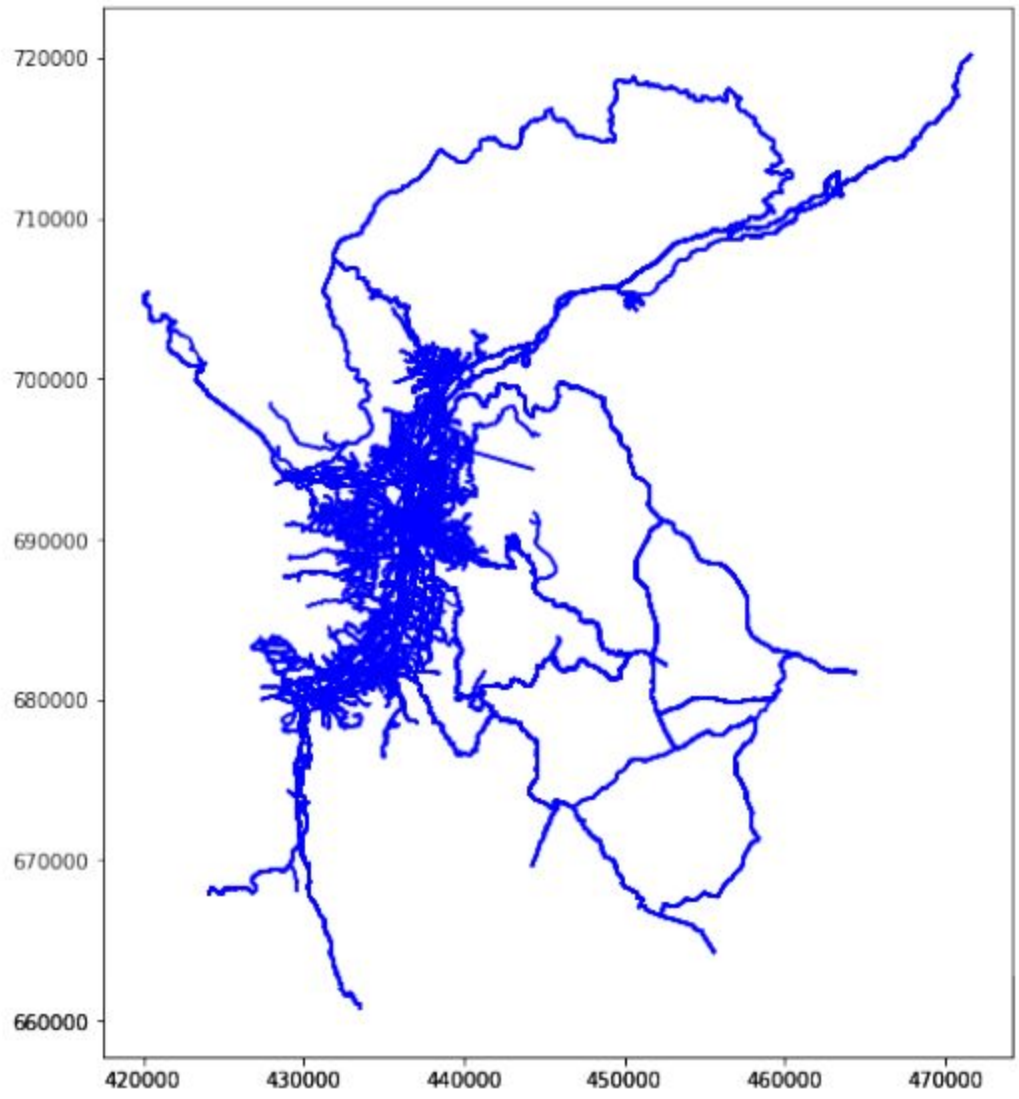


Fig. 7 Road network

Each link or route segment from the network connects two stops. These stops are called nodes so in order to assign the events to each link, we had to find the closest node to where each event happened and then, the next closest which is connected to the first one. This process is detailed in the subsequent section.

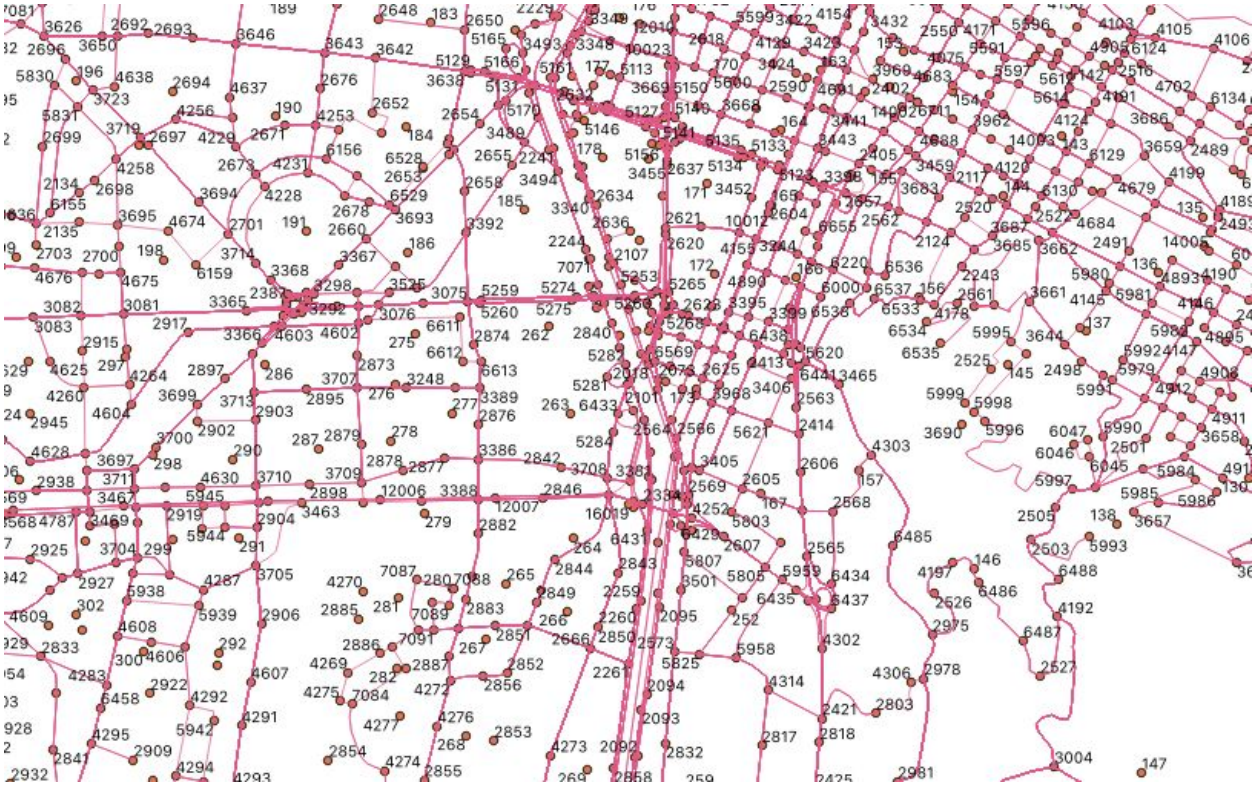


Fig.8 Example of nodes connecting a portion of the AMVA road transportation network

Events Assignation to the corresponding link

After gathering all the data from the entity, it was stored at Google Drive such that the computations were easier to implement in the Google Colab workspace; this because

1. The amount of data we had was such that at a local machine the process would not be possible to implement without the adequate software

2. We could collaborate in real-time and access the data after transforming it

So after loading the data into Google Drive, the first step consisted of cleaning it; the column regarding the date where the event happened was converted to an integer sequence and the hour was assigned to a different column. So then, we had two new columns, “DATEKEY” and “HOUR”. After this, both variables associated with a boarding or alighting event were summed, creating then the columns PAXUP for the boardings and PAXDW for the alightings.

Then for each day file, the DATEKEY, HOUR and PAXUP and PAXDW columns were added to the original data and saved creating a file per day with DATEKEY as name with a Comma separated Value extension (csv). For example, the events for the last day where there was available data, May 10th of 2020, we had the file 20200510 similar to the one in Figure 6 but with the additional columns stated above

DATEKEY	HOUR	PAXUP	PAXDW
20200510	6	0	2
20200510	5	0	1
20200510	5	1	0
20200510	5	0	1
20200510	5	1	0
20200510	5	1	0
20200510	5	1	0
20200510	5	1	0

Fig.9. Example of a DATEKEY file with the extra columns used to assign the link to each event.

The next step consisted of computing the distance from each event to each node and choosing the closest one. This process used as a distance measure the Haversine formula which is a special case of the Haversines Law used in spherical trigonometry for navigation and that takes into account the sphericity of the earth (Díaz, 2012). The events that were outside the AMVA were discarded.

After a node was associated with each event accounting for the distance, the other node closest to the first one we found was assigned such that the link comprising these two nodes was associated with each event.

This process was implemented for each day file and saved for later computation purposes.

Finally for each day file we aggregated the events by link, date and hour where these were given by the DATEKEY, HOUR and LINK variables. The final events dataset had the following structure:

DATEKEY	CODIGORUTA	HOUR	LINK	PAXUP	PAXDW
20200510	1001	5	10023-2641	2	0
20200510	1001	5	2080-2310	2	1
20200510	1001	5	2112-7010	2	1
20200510	1001	5	2304-2331	0	1
20200510	1001	5	2312-2309	0	4
20200510	1001	5	2317-2360	1	3
20200510	1001	5	2318-76	0	1
20200510	1001	5	2331-5204	0	1
20200510	1001	5	2348-7027	2	1

Fig.10. Passengers events dataset example with link assigned

Data Load:

After such aggregation and link assignment, the data was loaded to a PostgreSQL instance using AWS RDS service along with additional data from the links, routes, vehicles, and transportation service companies in case we would need them so it could be accessed later.

Cleaning procedure:

The columns of the number of passengers boarding and alighting had values between 0 and 99. These 99 values were considered outliers since they do not correspond to a normal bus stop. On the other hand, there were some coordinates (latitude and longitude) that do not correspond to the geographic area of Aburrá valley.

Detail of inconsistencies in the dataset

Some inconsistencies arose in the calculation of the passenger load per day, hour, and link because the passenger movement was sometimes wrong recorded and these measurement errors accumulated throughout the entire route of a vehicle that covers some route. A route consists of the succession of links (pairs of nodes) from the beginning to the end. Therefore, if there is a measurement error in any vehicle trip in the first links of the route, any calculation of a load of passengers on a later link will be affected.

We also observed the recording of events at unconventional hours and trips that transmit intermittent or incomplete information along a route and this is a pervasive problem in the data set. Analyzing complete trips, in conventional hours, they report passenger alightings at the beginning of the route, this means that the calculated passenger load is negative in the initial links. If we add all the trips that pass through these links and if the majority report a negative load calculation, we will obtain that the total volume of passengers is negative.

Negative passengers load

We observed measurement errors (more passengers getting off than boarding) accumulated during a period of time throughout an entire route. Therefore, successive negative loads will result in excessive negative loads in the end. If we add these results with the results of other trips on other routes that pass through the same links for a given day and time, the negative amounts will cancel out with the positive ones or the negative amounts will grow even more.

In general, an extreme value in both the number of passengers who get on and those who get off seriously distort the calculation of the load of passengers in a link and in those that follow along the route of the vehicle. Again, if we add these results to approximate the volume of passengers in any day and hour of operation in a certain link, the result obtained will be uninformative. It should be noted that values greater than 50 in the columns of boarding and alighting represent less than 1% of the distribution of these variables; however, the accumulated negative effect they produce in the calculations is large. In this figure, we depict the outliers found:

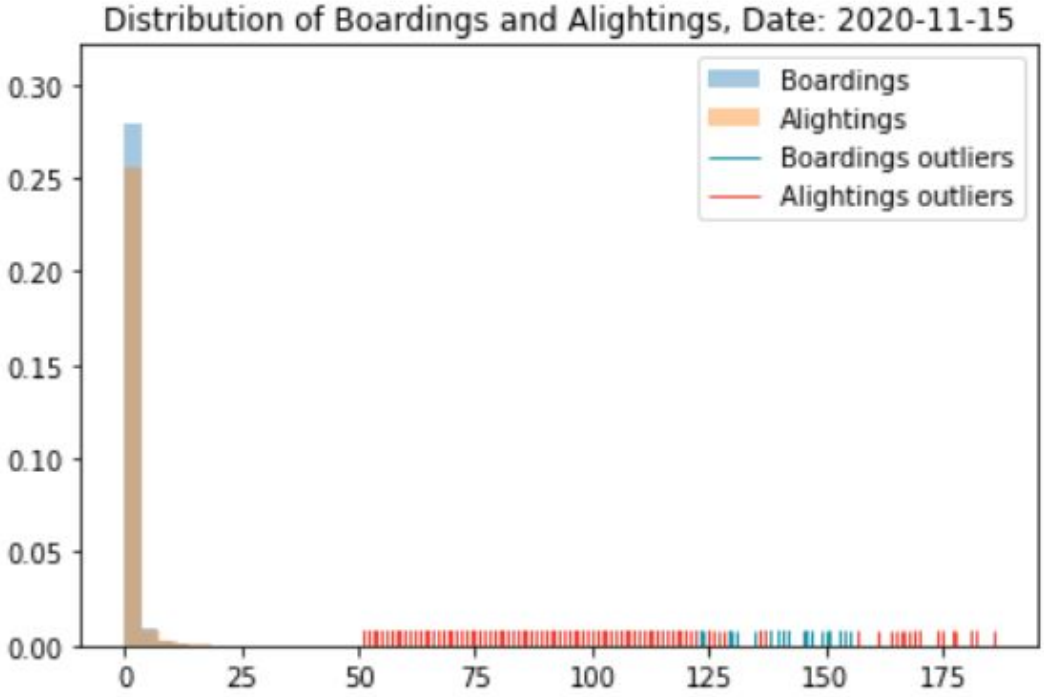


Fig.11 Distribution of boardings and alightings for a day

The difficulty with extreme values is that it is not possible to distinguish a false extreme value from a true extreme value (that is, many people get on or off), in addition, it is likely that the vehicles carry a passenger load greater than their maximum capacity, at least in some sections of the route. Below we show the distribution of the maximum reported capacity of the vehicles in the Valle de Aburrá transport network.

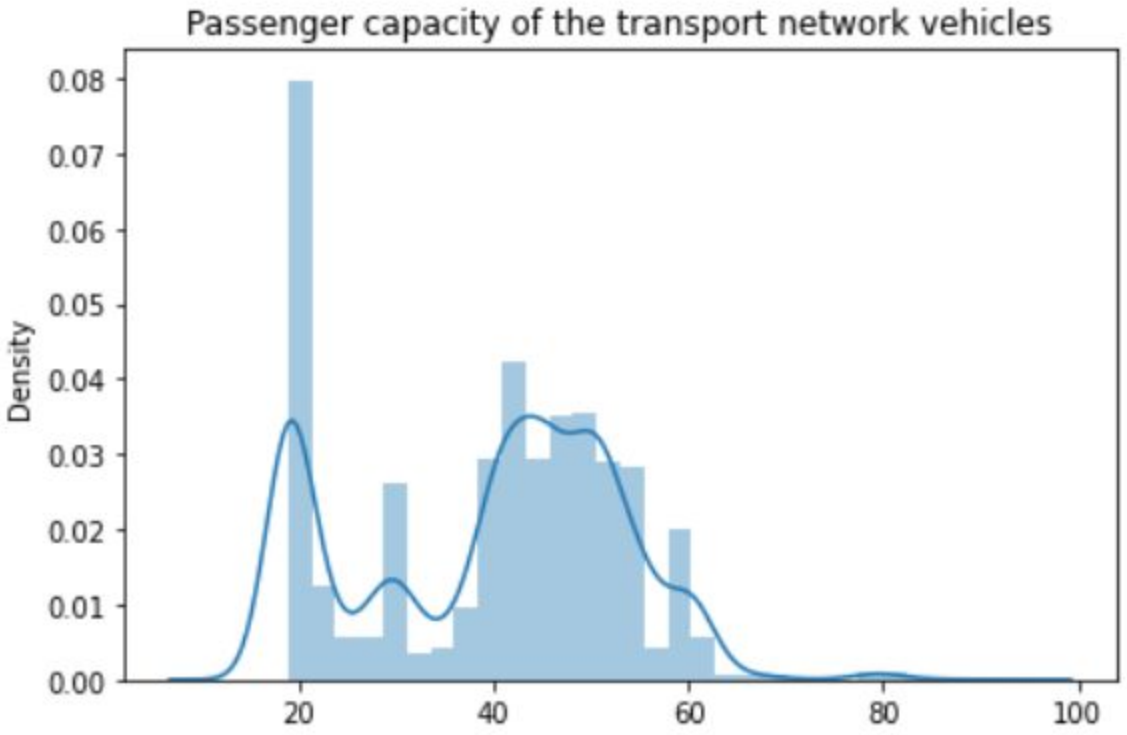


Fig.12 Passenger capacity of vehicles

A proposal to calculate the load and transform the data:

The central axis for the entity requirement was the passenger load computation associated with each one of the links which composes the transportation network of the AMVA region. As seen above, the data capturing process produces negative values of the load when aggregating so, therefore, adding by hour and not accounting for the links we will explore the passenger's load and will present some descriptive statistics to see how the demand behaves after removing also, the possible outliers or values that are over a defined threshold. These results are presented in order to analyze the transportation demand accounting for the passengers that alight from the vehicle analyzing possible trends behaviors that will be taking into account in modeling purposes.

The passenger load is understood as the net passenger quantity which occupies a specific vehicle in a given time interval; that is, the subtraction of the total of passengers that alight the vehicle from the total of passengers that board it.

As manifested by the entity, the variable must be broken down at a minimum hour level and as a maximum at a daily level, thus, it is needed to have the load data for each one of the 24 hours of the day for each day as it is required and also, for each link in the AMVA transportation network and each route which composes it.

After an initial transformation that consisted of dropping the events for the coordinates that were far from the AMVA periphery, assigning the links between two stops to each event, and aggregating by route, hour and day we came to the dataset that we proceeded to compute the passenger load.

Such dataset has the following structure:

	DATEKEY	CODIGORUTA	HOUR	LINK	PAXUP	PAXDW
0	20191101	1001	0	12012-2609	2.0	1.0
1	20191101	1001	0	12013-2603	2.0	0.0
2	20191101	1001	0	16008-3257	0.0	1.0
3	20191101	1001	0	2084-2081	0.0	1.0
4	20191101	1001	0	2112-7010	1.0	1.0

Fig.13 Passengers boarding and alighting by link

Where the first column corresponds to the day the event happened, the second one corresponds to the event route at issue and similarly, HOUR is associated to the time where this event happened which also, is the minimum level of disaggregation. The LINK column it's the link that connects two bus stops where an event of alighting and boarding happened in such an hour and for such a route so that these events are recorded in the PAXUP and PAXDW columns respectively. Then, it is of interest compute the difference between the total of passengers that board the vehicle and the total that alight from it, such that the passenger load is given by the next formula:

$$LOAD = PAXUP - PAXDW$$

However, some inconveniences can surge that are associated with the data recollection procedure. For example, there are records where the total of passengers that alight from the

vehicle is lower than the total of those who aboard it, there are observations for hours where the service is not provided and there are as well, events for that their behavior deviates from the "normality", i.e., load records that seem unlikely according to the variable distribution.

So then, in order to avoid incongruences for the variable computation it has proceeded the following way:

Be each record from the table with i = 1, 2, ..., all_records

Do:
IF PAXUP[i] > PAXDW[i]
LOAD[i] ← PAXUP[i] – PAXDW[i]
IF PAXUP[i] < PAXDW[i],
LOAD[i] ← PAXDW[i] – PAXUP[i]
IF HOUR[i] >0 y HOUR[i] <= 3
LOAD[i] ← 0

After computing the load regardless of the formerly mentioned conditions, this variable had the following behavior according to its distribution:

mean	0.20
std	10.21
min	-3028.00
25%	-1.00
50%	0.00
75%	1.00
max	2612.00
c.var	5078.74%

Adding Hourly Restrictions

Accounting for the hours restrictions where there can not be routes operating between the hours 00:00 and 02:59, the assignation LOAD[i] ← 0 was made for each one of those records.

Computing the position statistics, we had:

mean	0.20
std	10.21
min	-3028.00
25%	-1.00
50%	0.00
75%	1.00

```
max       2612.00
c.var     5071.17%
```

Even if there was not a significant change in the data distribution according to the statistics, there is quite a small reduction in the Coefficient of Variation of 0.14%.

Amending the negative loads

Given the problem's context where it does not make sense that there are more people who alight than those who aboard a vehicle, this values' order was reversed such that the load for each case was the difference between PAXDW and PAXUP; with this new restriction there was a significant change in the variable distribution which could be considered significant,

```
mean      3.25
std       9.67
min       0.00
25%       1.00
50%       1.00
75%       3.00
max       3028.00
c.var     297.24%
```

It makes more sense the values here obtained with a minimum of passengers load of 0 instead of a negative one, an average larger than 0 and a percentile 75 of 3 passengers. Nevertheless, after correcting the negative passenger load problem arises another one which is the presence of load values such that their behavior deviates from the rest of the load values observed.

Possible Outliers

Analyzing the percentiles, we had that 99% of the records had a load of 34 passengers or less. Using this value as a threshold we decided to assess the load values larger than this one; the minimum of passenger load for those records larger than the threshold is 35 passengers and as maximum, there are 3028 passengers such that the 75% reaches a load of 80 passengers or less. Still, for this subset, the present variation was such that the coefficient of variation took a value of 81.11%. This can be summed up next:

```
mean        74.89
std         64.87
min         36.00
```

25%	44.00
50%	57.00
75%	83.00
max	3028.00
c.var	81.11%

To deal with these values there were considered two scenarios; the first one consisted of replacing the load values with the passenger capacity of 99% of the vehicles and the second one, consisted of replacing them by the median of such capacity.

Using the first approach, we replaced by the percentile 99 of the vehicle's capacity for those in the transportation network of the AMVA; after proceeding with such approach we had:

mean	3.16
std	6.96
min	0.00
25%	1.00
50%	1.00
75%	3.00
max	62.00
c.var	220.33%

Where there was a reduction of 25.87% in the coefficient of variation for the variable at issue.

Now, when we replaced such subset by the median of the capacity for the vehicles in the transportation network of the AMVA, the behavior was the following one:

mean	2.97
std	5.48
min	0.00
25%	1.00
50%	1.00
75%	3.00
max	43.00
c.var	184.49%

Then, it was chosen as a replacement for these values, the median of the capacity of the vehicles of the AMVA transportation network which added the least noise to the data and helped to decrease as low as possible the variable variation. While still, the variability could be considered high, until now, there has not been taken into account other features such temporal variations like the day of the week, the hour of the day where the events happened among others that could help to explain the overall variation. Finally, and regarding the initial load values, there was a reduction in the coefficient of variation that went from around 5000% to 184% approximately.

3. Descriptive analysis

Passengers Load Exploratory Data Analysis

Variable	Passengers Load
Unit	Total of People in the bus at a given time
First Date of Observation	2019-11-01
Last Date of Observation	2020-05-10
Frequence	Hourly
Minimum	0.00
Mean	18135.03
Median	17515.67
Maximum	75881.00
Coefficient of Variation	103.53%

Time Series Behavior

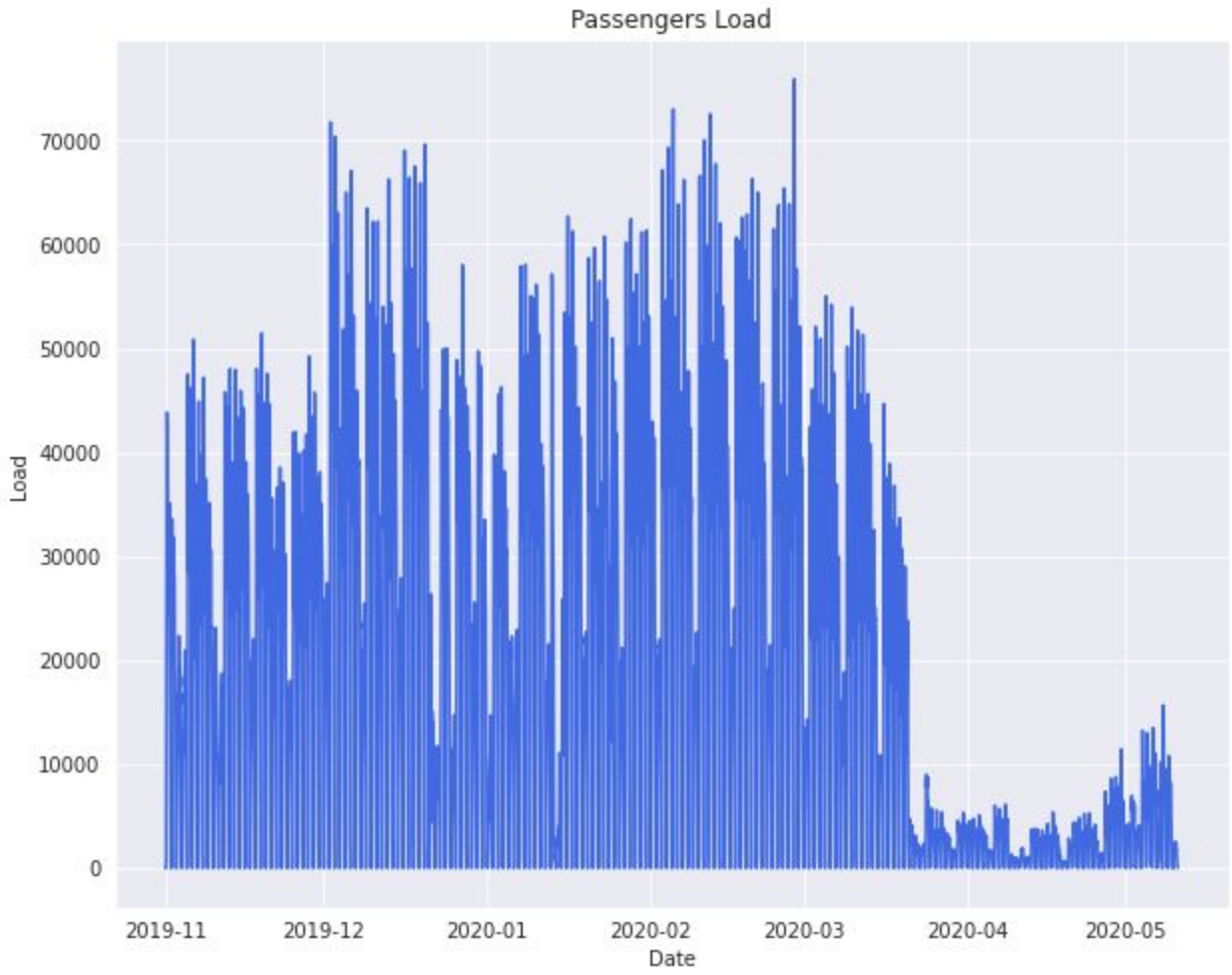


Fig. 14 Time series of passengers load

It can be appreciated that throughout the observation window there is a stational pattern that repeats as weeks go by; this corresponds to the hourly observations for each one of the days such that we advance week by week the transportation demand will vary accordingly.

On the other hand, there are some significant decreases at the beginning of the year, perhaps due to the festivities at the end and beginning of the year and more markedly, in mid-March, there is a decrease in demand as a result of the lockdown product of the pandemic generated by COVID-19.

Next, it will be proceeded to analyze in detail how each temporal and/or seasonal factor influences the demand and then, it will be proceeded to define features that will help to correctly model the variable at issue.

Passengers Load Behavior: Monthly

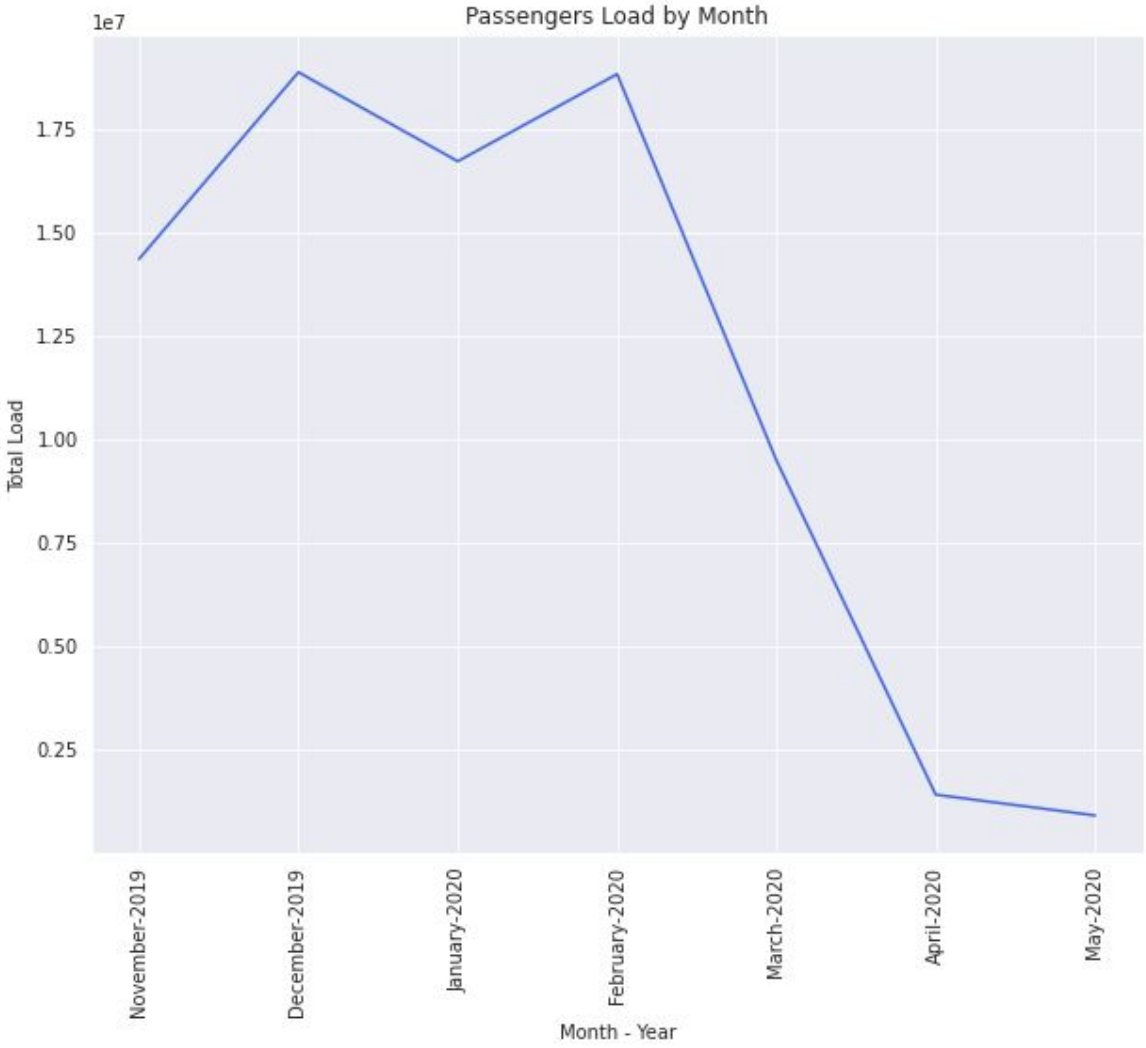


Fig. 15 Passengers load behavior

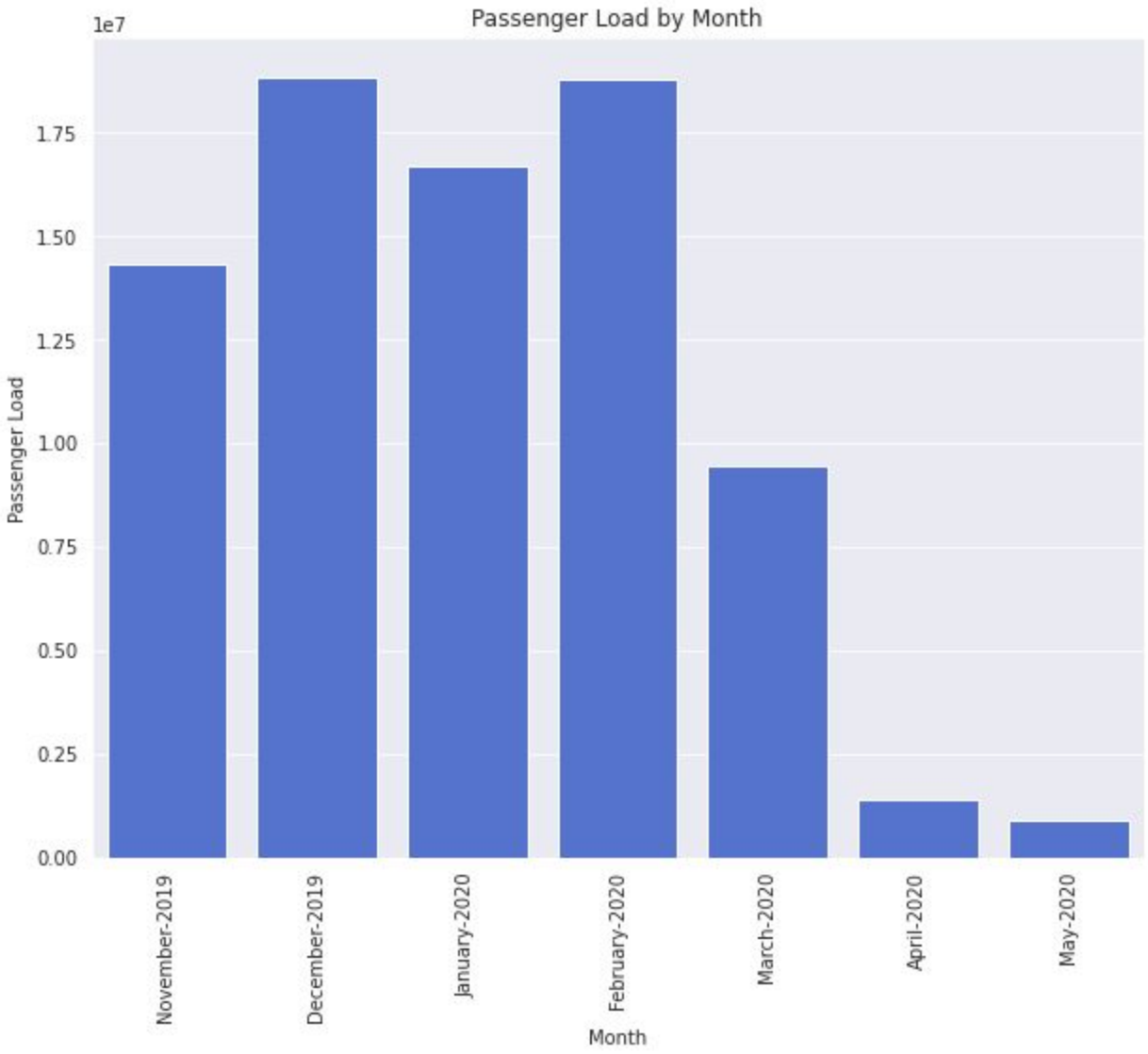


Fig. 16 Passengers load by month

Aggregating the load by month it can be appreciated that the maximum demand was reached around February and the minimums are reached after the lockdown as a result of the COVID-19 pandemic with fewer people traveling in the city to their workplaces, schools, and in general, outside their homes, affecting this, the demand for transportation service. There is also a low peak beginning the year as a result of people traveling for vacations at the end of the year and new year festivities.

Passengers Load Behavior: Weekly

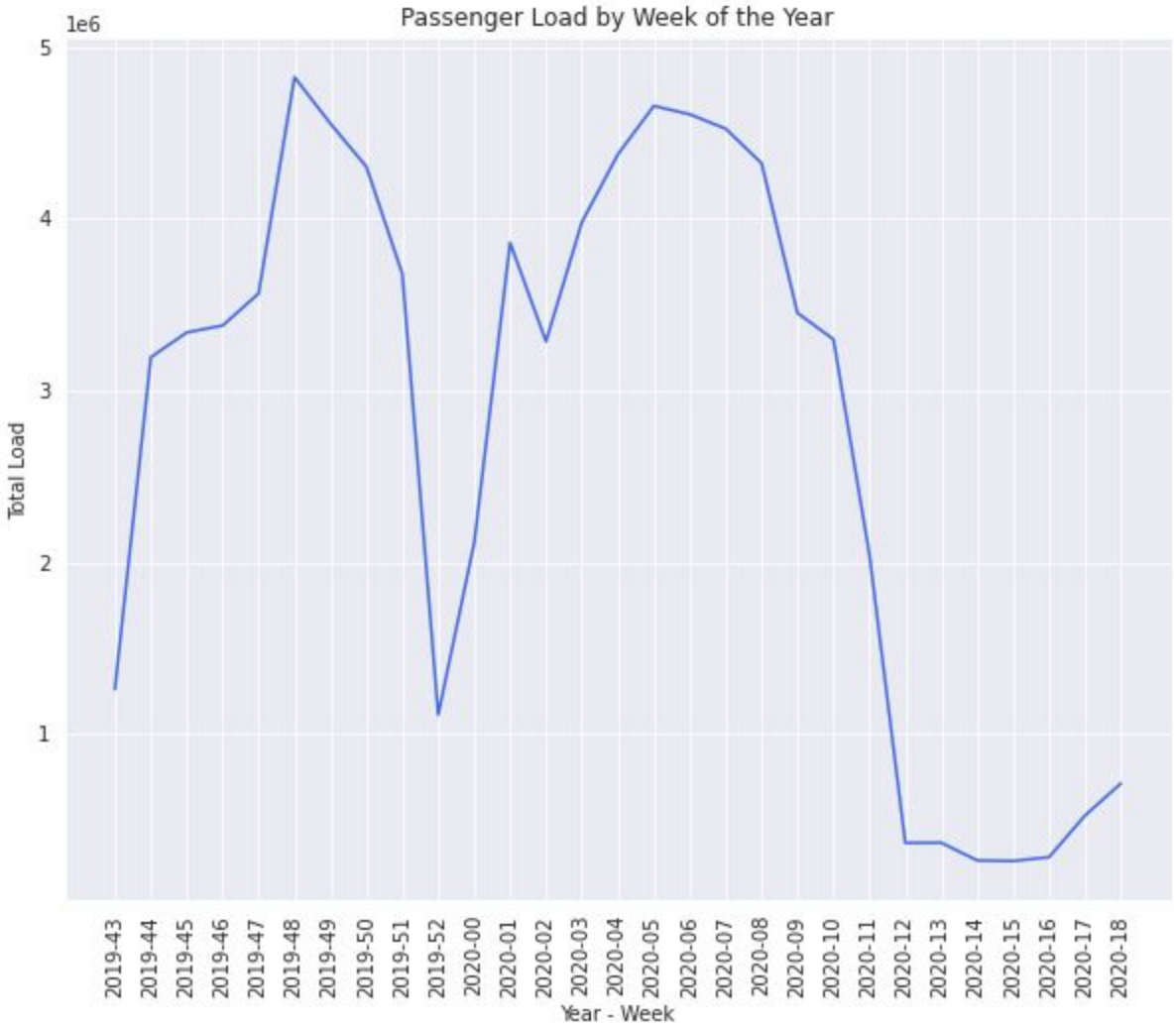


Fig. 17 Passengers load by week of the year

Regarding the public transportation demand variation, the highest peak is reached in the 48th week for 2019 and this corresponds to that which starts the 25th of November and ends the 1st of December undergoing from there a decrease in the demand that culminates in the lowest weekly peak for 2019 for the 52nd week, that is, the last seven days of the year which originates as stated before from trips out of town in the wake of the End of Year and New Year holidays.

The transportation demand reactivates again around the 5th week of 2020 which is the one where February begins and where it is expected for the services to fully stabilize after the end of vacations. After this maximum, begins a descending trend reaching a minimum in terms of weekly public transportation demand for the 12th week of the current year, again, this event originates in the lockdown after the COVID-19 pandemic lockdown as explained before; this

weeks go from March 16th and 22th (Valora Analitik, 2020) and later will be useful to model the effect of the lockdown in the transportation demand in the AMVA.

Passengers Load Behavior: Day of the Week

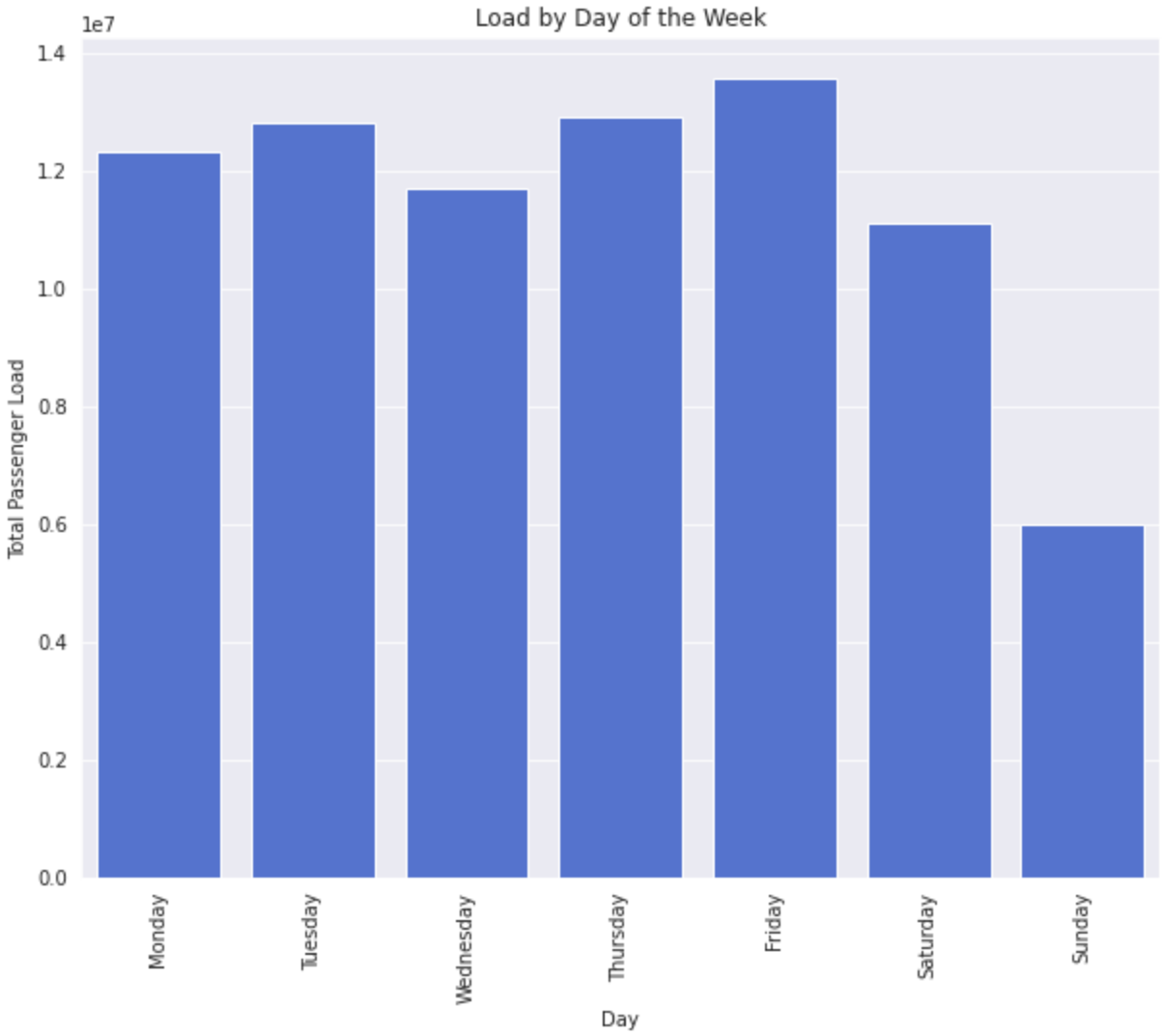


Fig. 18 Load by day of the week

Now, on a daily basis, those where the most passengers load or transportation demand occurs are Fridays, Thursdays, and Tuesdays and the lowest happens on Mondays, Wednesdays, Saturdays, and Sundays in descending order. For the first 6 days of the week, the demand does not vary much between each one of them and the sixth one, Sunday is the only one where happens a significative deviation from the general behavior; nevertheless, as for the first 6 days, the

demand remained almost constant the coefficient of variation for the variable at issue was of 22.25%, this for a daily aggregation.

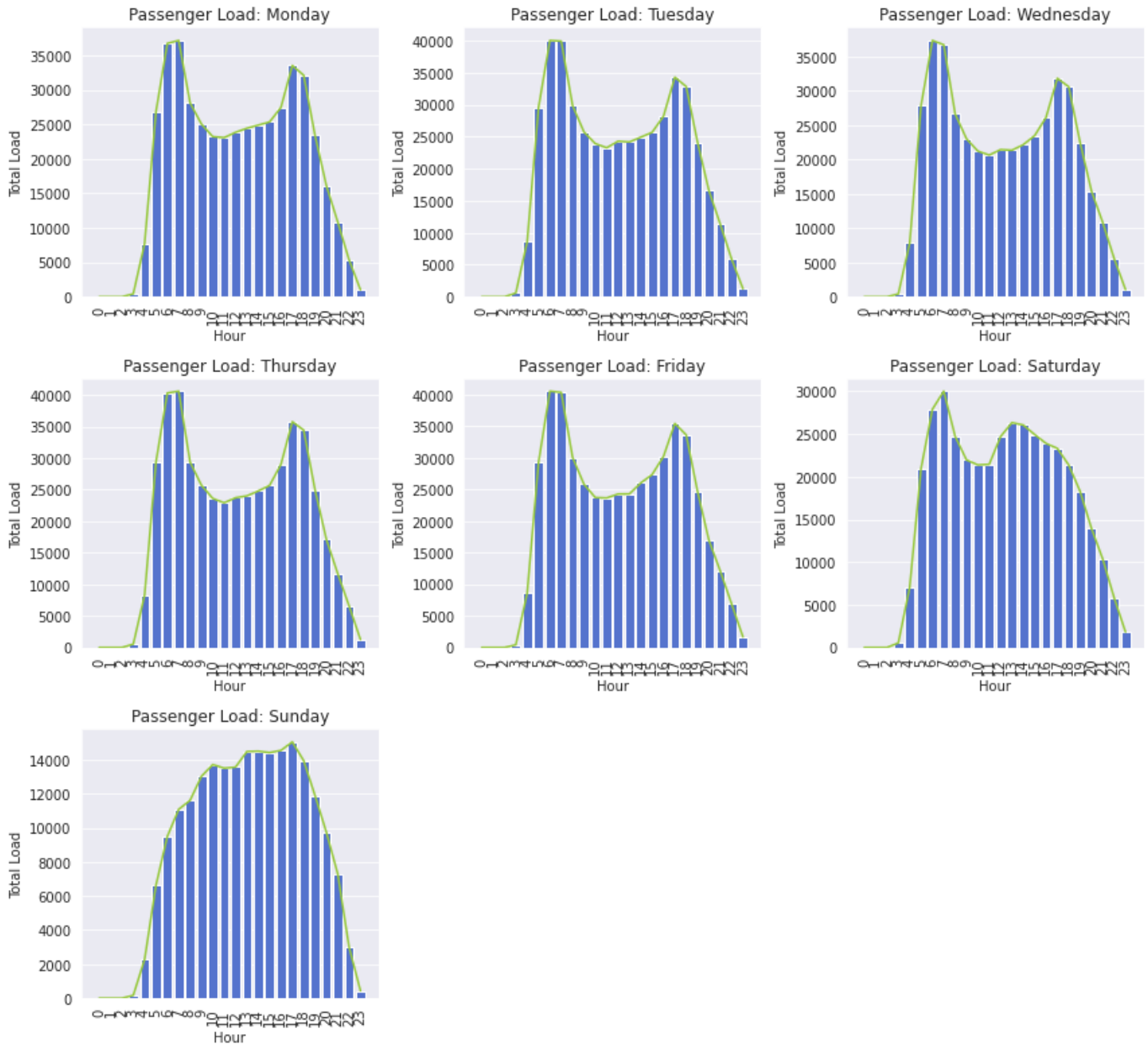


Fig. 19 Load by day of the week

Disaggregating the behavior previously observed according to each of the hours of the day of the week is found a repeating pattern for the first six; this corresponds to the beginning of the workday around 6 - 7 hours and then the start of a decrease in demand that reaches a valley in the daily distribution around the 9 - 10 hours; analogous to the previous peak, there is one around

17-18 hours that corresponds to the return to the home of people at the end of the working day at issue after which, the demand begins to decline as the day progresses until it reaches a minimum at 23 hours where public transport routes end their trips.

A similar behavior occurs on Saturdays, with a peak around 6-8 hours, but unlike the weekdays, where there is a valley that associated with the minimum around is broader, this is more pronounced because in most of the cases the end of the Saturday working day, occurs around 12-13 hours, when the second peak or maximum demand for transport occurs for this day. From this value, a decrease in demand begins until the end of the day as in the case of the others previously observed.

Regarding Sundays, the behavior is approximately unimodal so that since the operation of routes begins, the number of passengers who demand the service begins to increase until reaching a maximum around 12 hours and then a little more pronounced peak at 17 - 18 hours; this behavior can be originated in the realization of outdoors activities as this day is of rest for most of the population in the region and the country and these days are destined to visit museums, parks, biking, lunching outside home among others (Uber, 2020) .

Passengers Load Behavior: Holidays

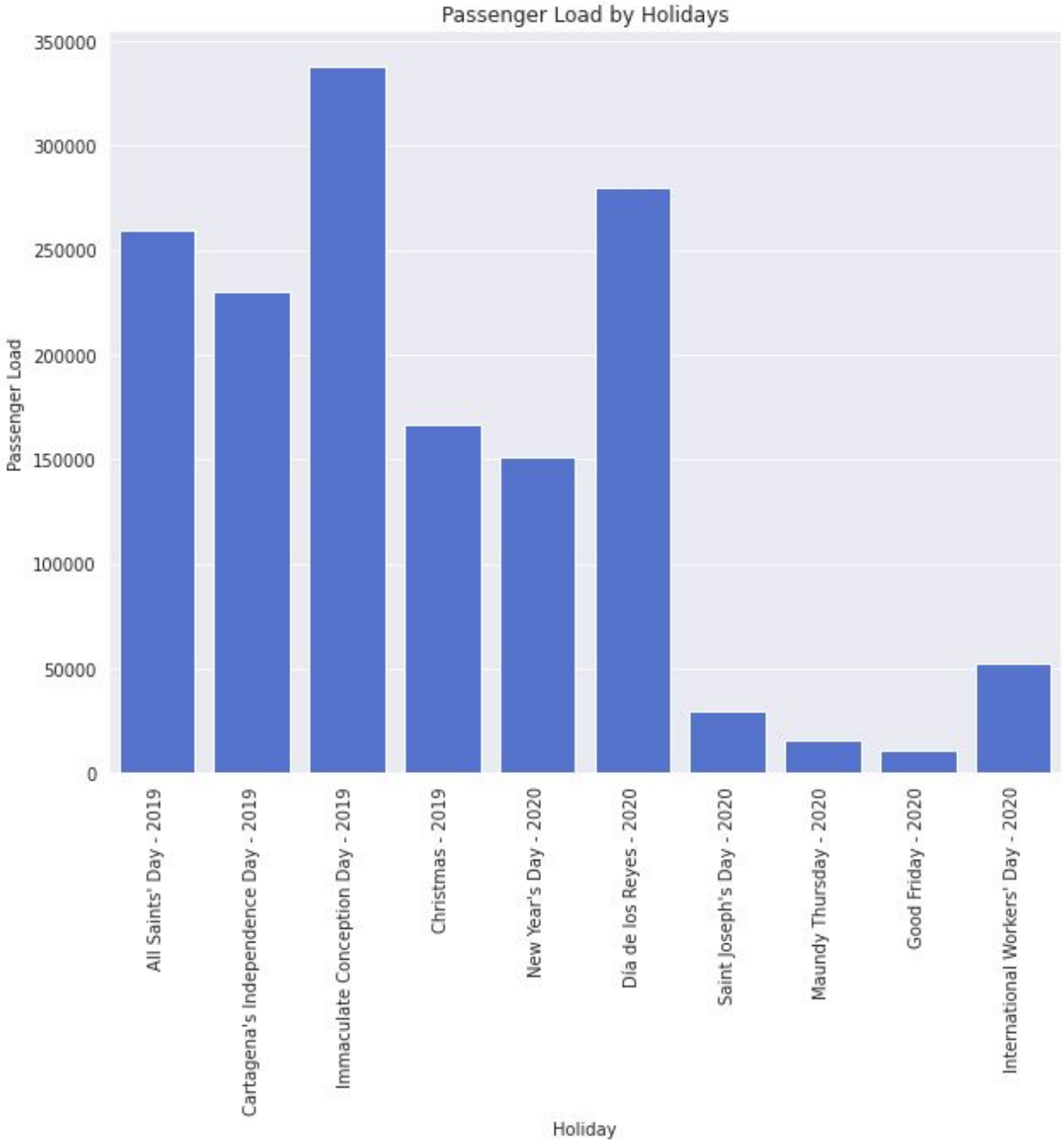


Fig. 20 Passengers load by holidays

The Immaculate Conception Day or Día de las Velitas is the one for which there's a higher public transportation demand; this could happen as this is the first of the end of year holidays and as this is a highly celebrated holiday in Medellín (Bolaños, 2017) where Christmas lighting is one of the main tourist attractions of the city during December (Turismo en Medellín, 2020).

The next day where there's a high demand for transportation is the Día de los Reyes corresponding to the 6th of January; as in this day the Christmas lights are turned off (EPM,2020) also it could begin a return to the everyday activities in the region at issue. For the next holidays there is a decrease in the demand for public transportation, again, originated in the lockdown previously stated that began around the second week of March being this consequent with the rest of the days of this period for the observation window. Nevertheless, there's a slight increase for the 1st of May, i.e., International Workers Day where for Medellín some demonstrations were held by Working Centrals and Syndicates (Venegas, 2020).

Passengers Load Behavior: Heatmap

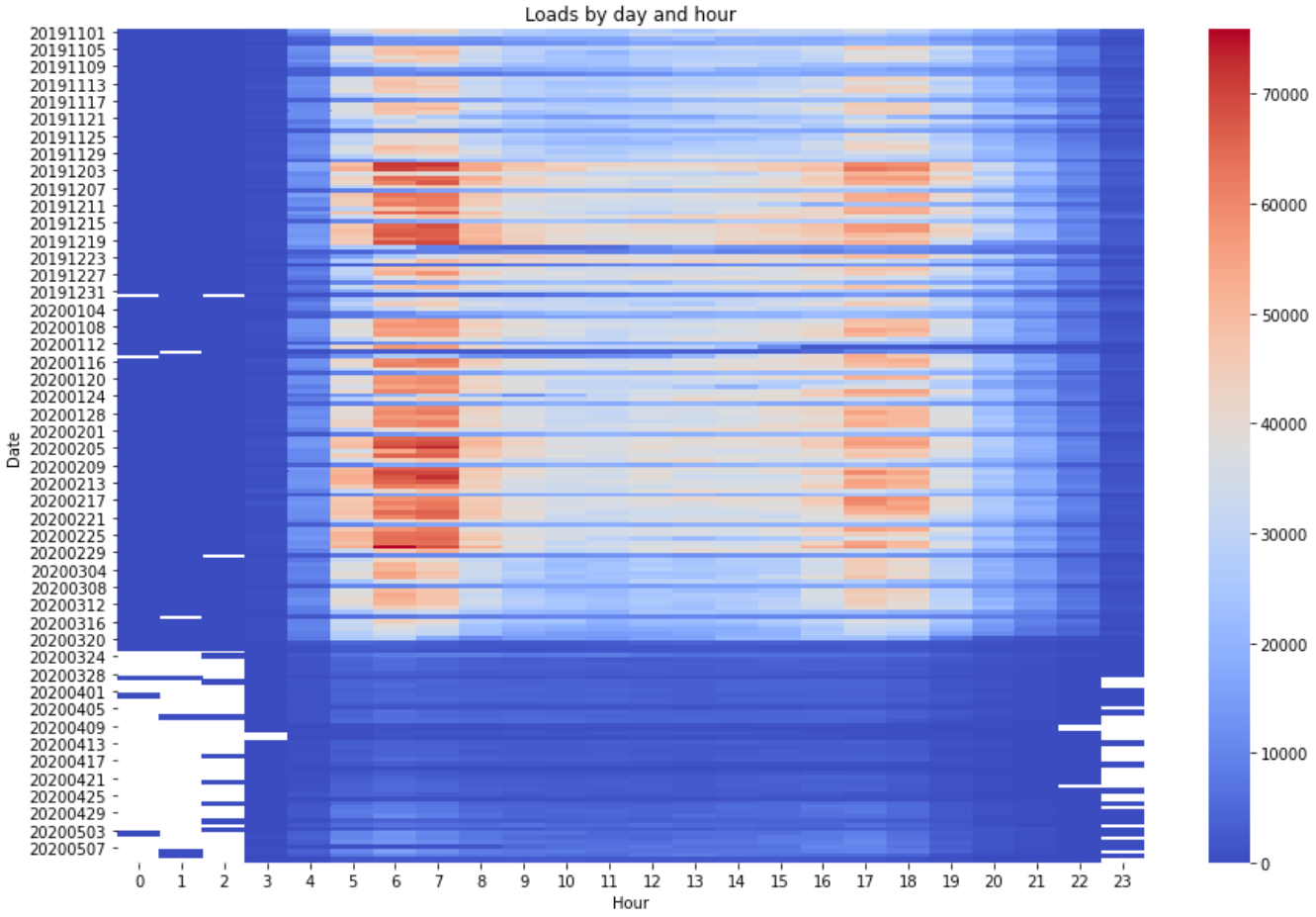


Fig. 21 Loads by day and hour

Now, to conclude, the overall demand behavior is larger around the 6 - 7 hours and the 17 - 18 hours such that the first interval corresponds to 11% of the trips made along the day in the AMVA (Área Metropolitana del Valle de Aburrá, 2020) and the second one is associated to the end of the working day at issue. This behavior is present for all the days of the week for all the observed months with a lower intensity after the second week of March product of the mentioned lockdown. Overall, around the 3 hours begins an increase in the demand for public transportation while before this, the travel data are 0 and null or non-existent in some cases. Finally, after the 18 hours begins a decrease in the demand until it reaches levels of 0 around 23 hours.

4. Model and statistical analysis

Since the calculation of load of passengers showed negative values in many links, we focused our predictive models only on passengers boarding in order to predict future boarding at any hour of a day and the total number of passengers boarding by a particular day of the month.

For predictions by hour we defined a model based on Random Forest Algorithm and for predictions of the day of the month, we used time series analysis using ARIMA and SARIMA models, which are described in detail in the next sections.

Background cases about forecast of transportation demand

As Chang (2019) points out, the transportation demand is not a direct one but a derivative which depends on what is transported, whether it is passengers or merchandise. In his case, the author implemented a maritime transportation cargo demand model for the port of Callao in Peru using a SARIMAX model (Seasonal ARIMA with eXogenous variables) for three representative terminals of the Port of Callao: APMTC, DPWC, and TC projecting forward 4 periods, 2019 - 2023. Among the exogenous variables used are some macroeconomic ones obtained from the Central Reserve Bank of Peru composed of Commercial, Confidence, and Copper Price indicators. In most of the cases, it is obtained that seasonal autoregressive models, SAR (p), are the ones that perform the best along with some AR (p), and a few models of moving average, MA (q) and GARCH (p, q). An adjusted Coefficient of Determination or R-squared of 48.49% is obtained in the best of cases and in the worst, one of 47.02%; MAPE had a value of 6.33% in the best case and 9.4% in the worst, indicating a good capacity of the models to generalize the results to future observations.

On the other hand, Xue et al. (2015) point out the little attention that models of this type have received in terms of conducting studies and propose an approach to forecasting bus demand in

the city of Shenzhen, China for a specific route with data ranging from August to November 2013; this approach consists of combining three models of time series for different time scales (weekly, daily and every 15 minutes) in order to produce a short-term forecast (15 minutes), which is called Iterative Multiplicative Model (IMM) this with the aim of "not only assist in boosting bus operation efficiency, but also minimize the operation cost and improve service quality and reliability". Good measurements were obtained using ARIMA-type models (p, d, q) that were improved using a GARCH (p, q) and subsequently combined; the final results in terms of MAPE indicate that the weekly model had a value of 23.20%, the daily one of 12.42% and the adjusted one for every 15 minutes of 14.93% in such a way that when they were combined, the MAPE was 9.08%

Li & Axhausen (2019) studied different types of models for the forecast of transportation demand; specifically, the demand for Taxis in New York, USA, and Shanghai, China. In the first case, there are data that go from January 1, 2016, to June 30 of the same year and in the second one, these were collected for April 2018; there is data on the trajectory of 12095 taxis, the GPS timestamp, the longitude and latitude of the stop and the status of the passengers (which definition is not specified), all this, with a data recollection frequency of 3 seconds. Thus, there are a total of 7571012 trips such that after cleaning the data remained 6506551; from Shanghai, only the data from the metropolitan area were used, that is, 2346714 records.

The models compared were time series (Simple Moving Average, Weighted Moving Average, Seasonal Moving Average (Day), Seasonal Moving Average (Week), Exponential Smoothing and ARIMA), Random Forests, XGBoost, Multilayer Perceptron Neural Networks, Long short-term memory (LSTM) Neural Networks (NN), LSTM-NN with Onehot-Encoding and LSTM-NN with embedding. The effectiveness of the models was measured using the SMAPE or Symmetric Mean Absolute Percentage Error and the RMSE (Root-Mean-Square Error). It is found that in terms of RMSE, the LSTM-NN model with embedding was the best in both cities, while in terms of SMAPE, Neural Networks, Random Forest and XGBoost were the best compared to the rest. Thus, it is concluded, on one hand, that time series models have the worst performance compared to machine learning or deep learning models and that using embedding to treat categorical variables is better than using One-hot Encoding although measuring the errors in terms of temporal variation, the authors do not present conclusive results. Regarding the spatial behavior, neural networks presented the best performance by area units for both cities, but compared to the previous result, LSTM-NN with embedding or with One-Hot Encoding did not present good results.

Liu et al. (2020) propose another approach by which, instead of forecasting demand based on historical data, they propose to carry out such work by identifying hotspots or the points where the transportation demand is higher; the problem was oriented towards the demand for taxis in the city of Xi'an in China using Random Forests, Ridge Regression and a Forecast Model that

combines the late two. There were around 40 million observations as a result of measurements every 5 seconds for 30 days regarding the taxi demand. Meteorological factors such as air quality, the concentration of different types of pollutants, the type of climate (Sunny, Cloudy, Rainy or Hazy), wind speed, temperature, relative humidity, and precipitation were used as covariates as well as variables referring to the temporal frequency of the event, such as the time, the day of the week, if the day is a working day and if the day is a holiday or not. It is found that although the climatic variables contributed to the explanation of the demand, the temporary ones are the ones that contribute the most and it is also found that the combined model presents the best results.

Modeling using Random Forest Algorithm

Random forest is one of the most popular tree-based supervised learning algorithms that can be used to solve both classification and regression problems. Random forest tends to combine hundreds of decision trees and then trains each decision tree on a different sample of the observations. The final predictions of the random forest are made by averaging the predictions of each individual tree.

The algorithm has the following steps:

1. Resample the dataset creating B subsamples
2. Adjust a decision tree i for each subset with $i = 1, 2, ..B$ obtaining in each one, a prediction or classification of the variable accordingly.
3. Voting will be performed for every predicted result.
4. Select the most voted prediction result as the final prediction.

One first issue applying this method to the dataset is its temporal correlation and a possible spatial influence in addition to the labels of cities. For this study, we adjusted the model using cross-validation applied to time series but instead of sampling randomly we selected datasets in an orderly manner.

For a dataset with n observations, and a subset for training where $i = 1, 2, ..., j$,dataset for testing must be found in the dataset where $i = j, j+1, .., n$ with $i < i+1 < i+2 < \cdots < j-1 < j$ and similarly $j+1 < j+2 < \cdots < n-1 < n$. This resampling procedure can be observed as follows:

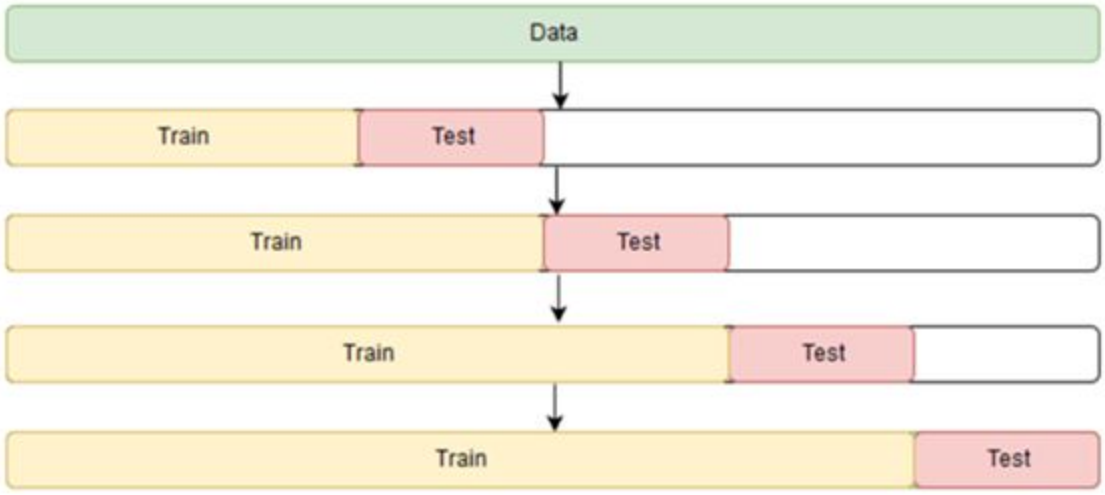


Fig.22 Cross validation in time series. From Shrivastava (2020)

It is observed that the extension of the series expands as we select a new testing dataset. It starts with a small subset for training, forecast for the later data points, and then checking the accuracy. The same forecasted data points are included as part of the next training dataset and subsequent data points are forecasted.

For measure the accuracy of the model, we used Mean Absolute Percentage Error (MAPE) defined as:

$$MAPE = \frac{1}{n_{test}} \sum_{t=1}^{n_{test}} \frac{|y_t - \hat{y}_t|}{y_t} \times 100$$

Where y_t is the value observed and $\hat{y}_t$ denotes its forecast, and the mean is taken over. The difference between y_t and $\hat{y}_t$ is divided by the actual value y_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying by 100% makes it a percentage error.

Scaled errors a given by:

$$q_j = \frac{e_j}{\frac{1}{T-1}\sum_{t=2}^{T}|y_t - y_{t-1}|} /$$

Where e_j is the difference between the observed value and the forecast one $e_j = y_t - \widehat{y_{t-1}}$. Thus, it is sought that the error is less than the one obtained when predicting it as the previous value. Then, MASE is the average mean of the values:

$$\text{MASE} = \text{mean}(|q_j|)$$

It is ideal then, that the MASE is less than 1 indicating that the error of the model when predicting each value is less than the naive model (Hyndman & Athanasopoulos, 2018).

For our dataset we took the last days of the measurement, that is the first 10 days of May 2020 for testing the model to forecast the values based on the data points previous since 01/11/2019 and aggregating one day in each iteration until 10 of May 2020, as shown in this figure:

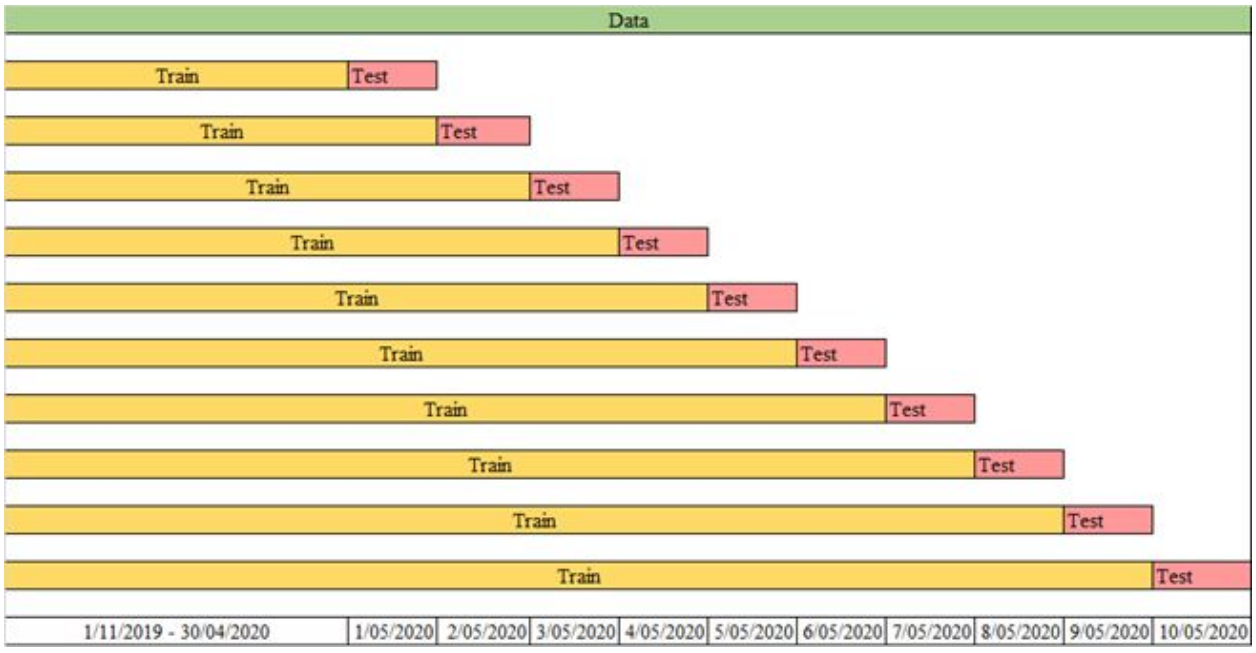


Fig.23 Dataset for training and testing

Each day a random forest model was adjusted with 1000 trees and the other parameters like the max number of levels in each decision tree or the number of data points placed in a node before the node is split were defined as default.

The final model for the last day returned a MASE of 0.18, i.e. error model of random forest is better at 82% than naive model.

Fig.24 Random forest model forecasting

It can be observed that the model captures the variance associated with the observations during the day.

This behavior is better observed by city,

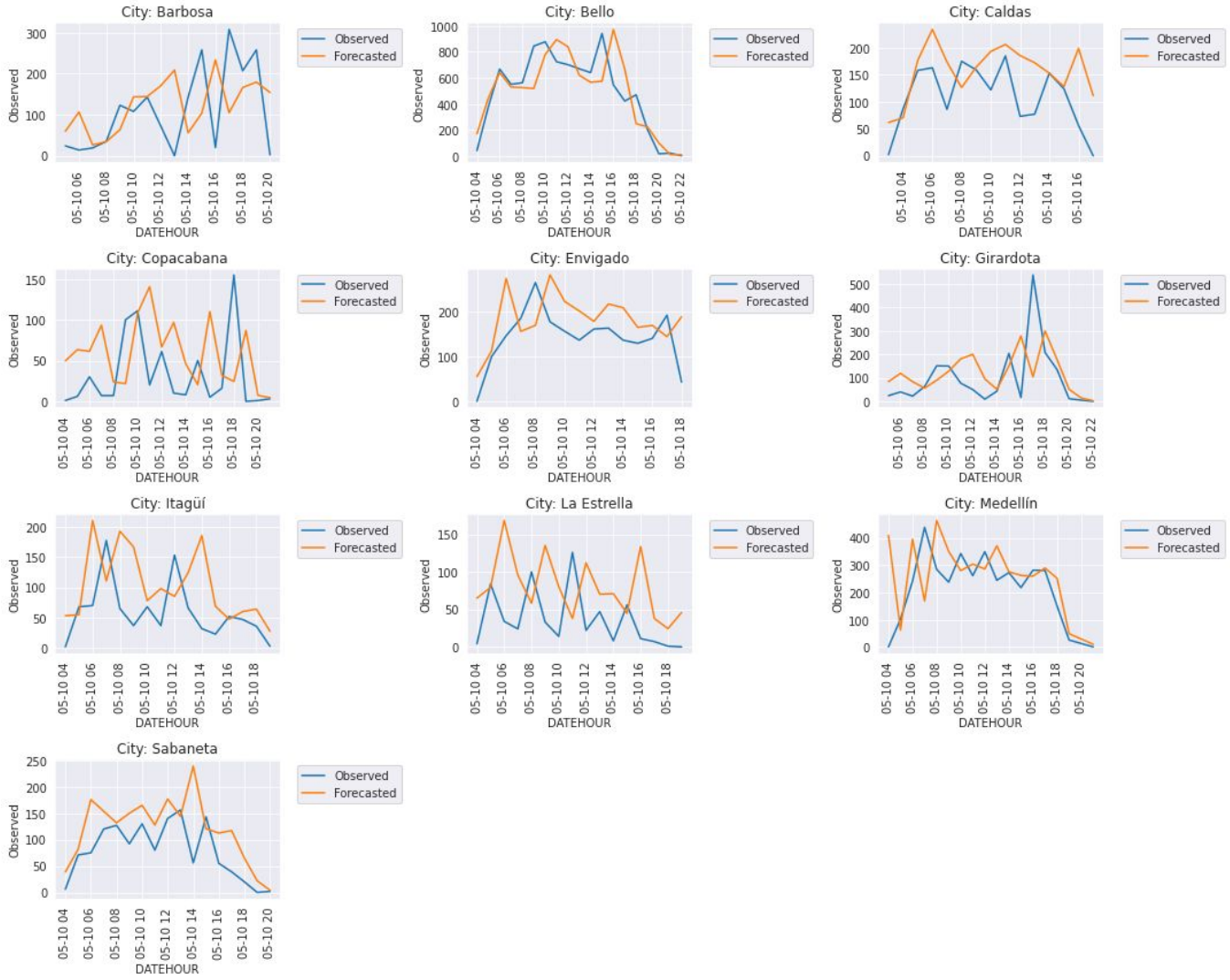


Fig.25 Observed values vs. Forecasted values by City

The relationship between the predicted and the true values still have some variation so that the predicted results do not completely fit a straight line as expected but still there's some linear relationship that could be improved using more features or tuning the model.

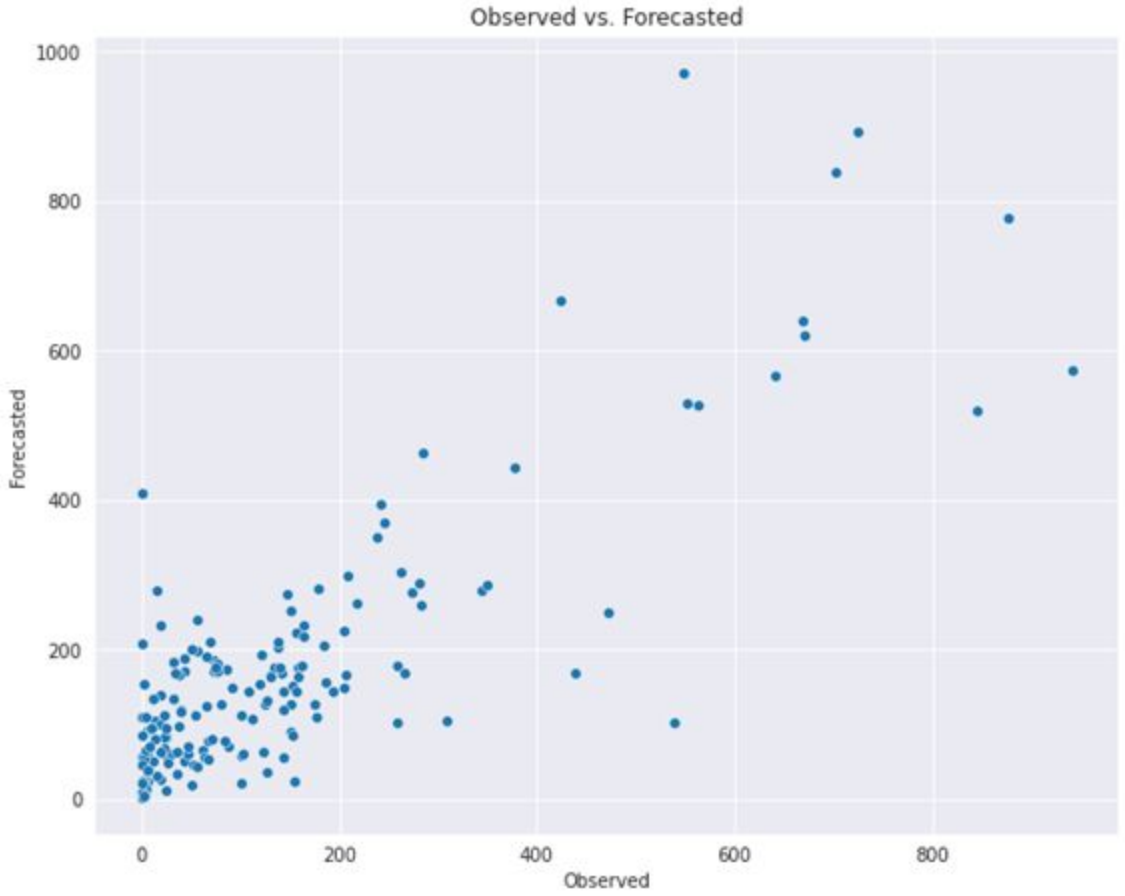


Fig.26 Observed values vs. Forecasted values

On the other hand, and taking into account how much demand varies between one day and another, the prediction intervals are very wide, as can be seen below:

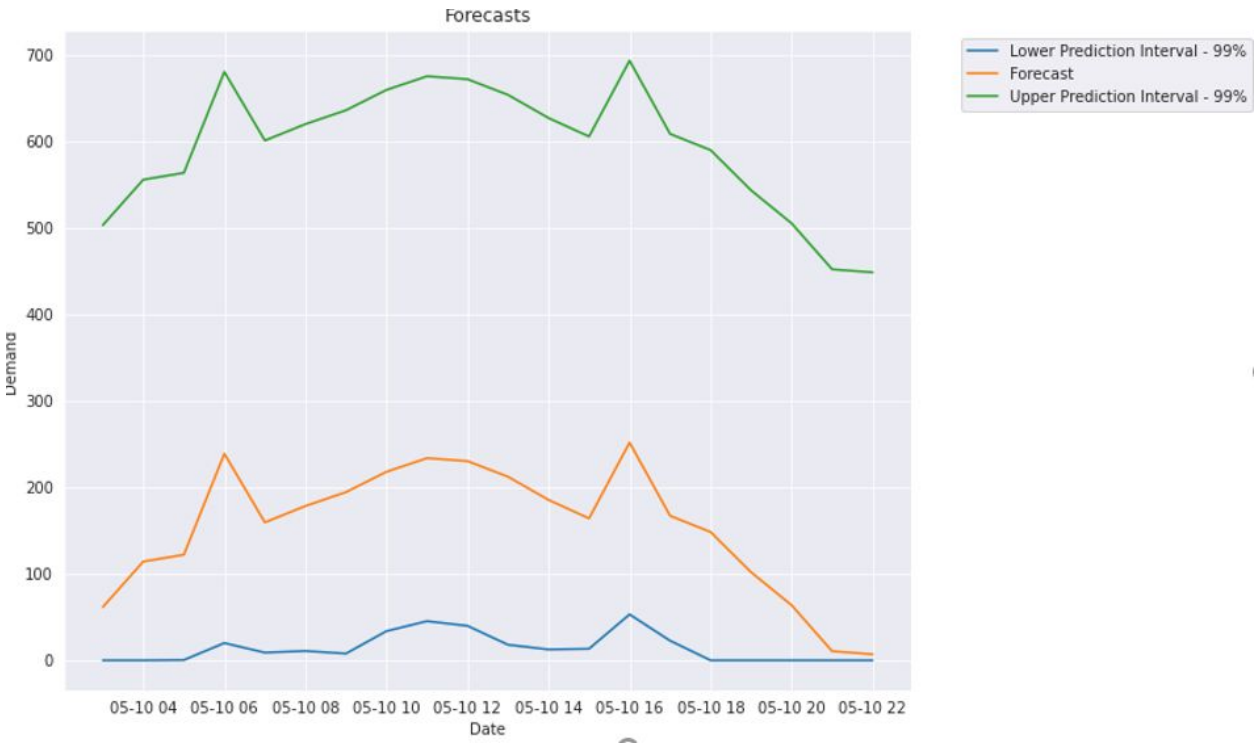


Fig.27 Prediction intervals

However, and based on the MASE, this model can be explored later to generate forecasts with a more detailed adjustment on both temporal and spatial factors as well as the hyperparameters of the model.

On the other side, it was found that the best variables regarding the explanation of the passengers boardings were the previous two records and the hour of the day where the event happened:

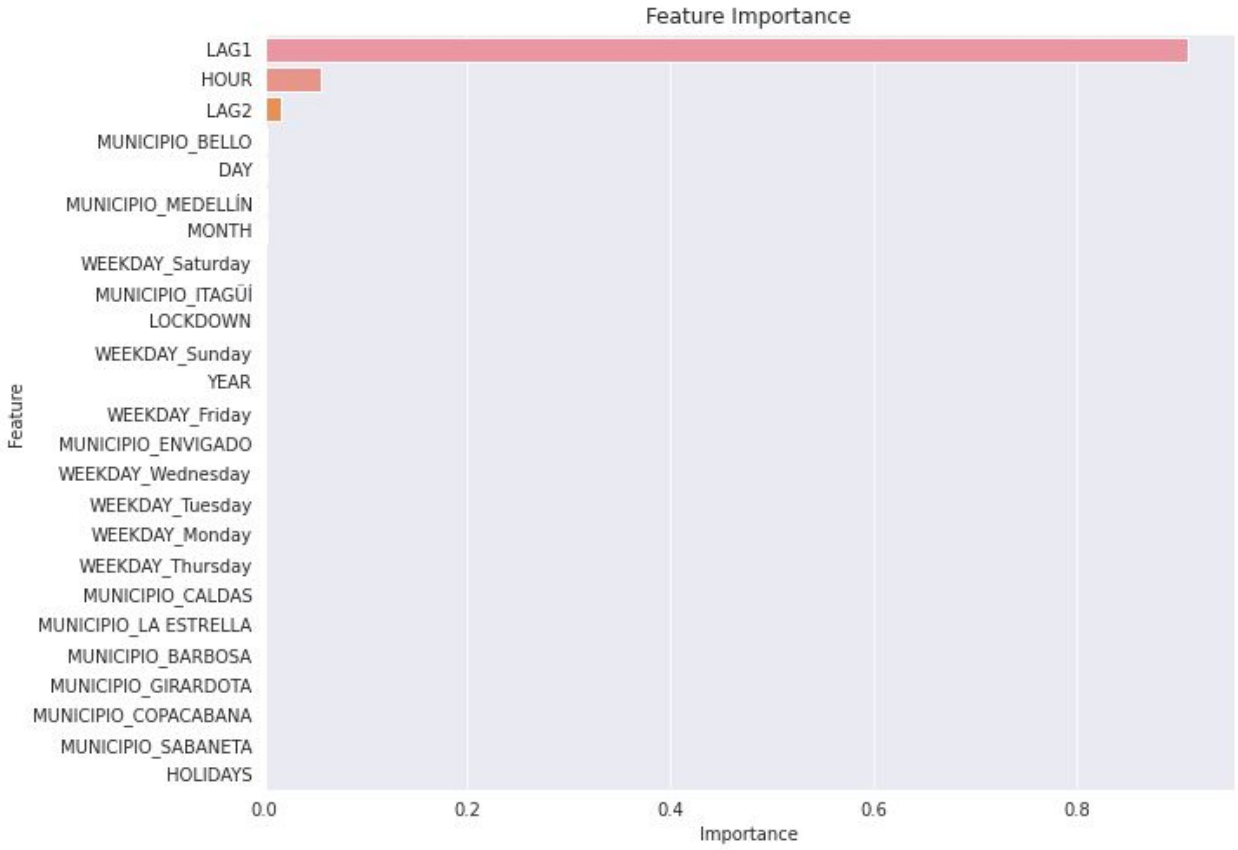


Fig.28. Feature Importance

Time series analysis (ARIMA/SARIMA models)

For time series forecasting there are several models that allow making predictions about future observations based on historical data.

Time series can be decompose into three separate time series:

- Trend: linear increasing or decreasing behavior over time
- Seasonality: repeating patterns related to calendar movements
- Noise (residuals): non-systematic component that corresponds to a random variation in the series

As described in this diagram:

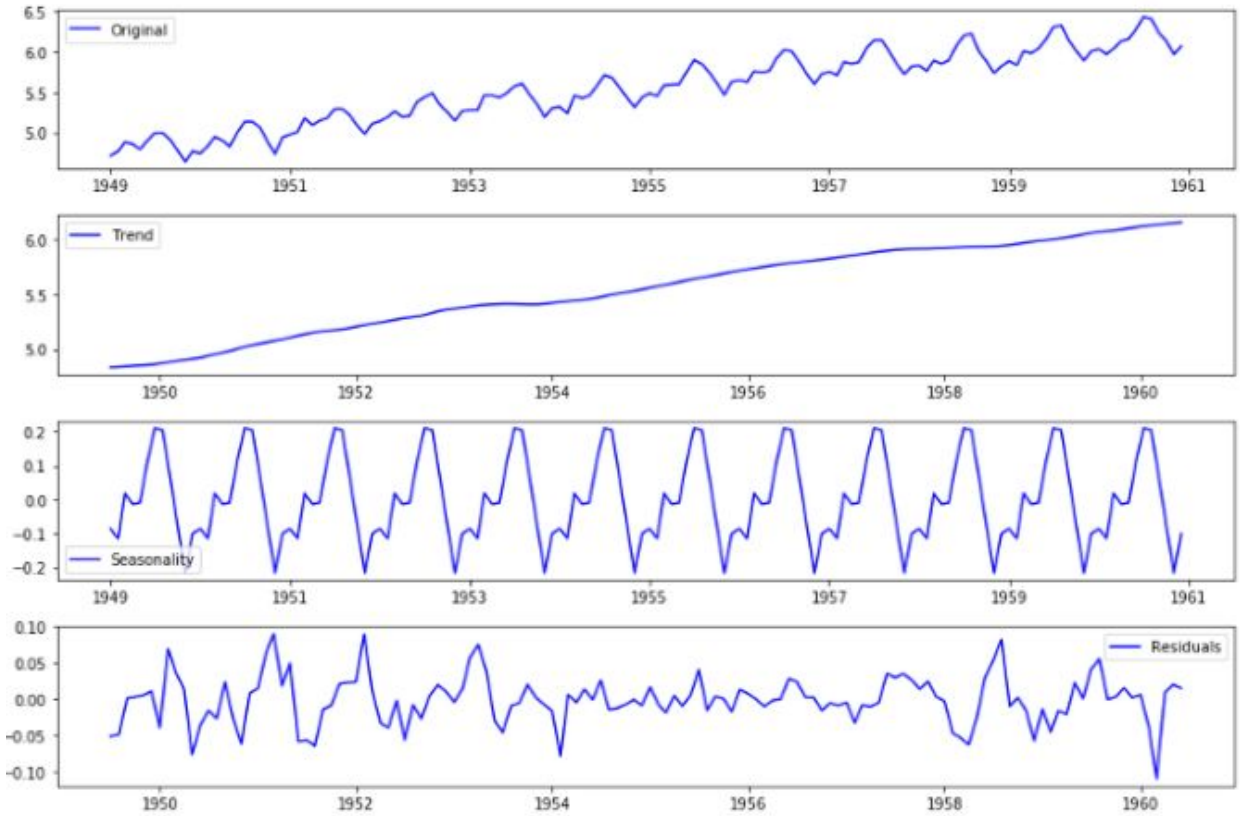


Fig.29 Time series decomposition

One of the most widely used models is Autoregressive Integrated Moving Average, (**ARIMA**) for time series data forecasting. Its acronym describe the the key components of the method:

AR (autoregression): the model uses observations from previous time steps to predict the value at the next step

I (integrated): uses differencing of raw observations in order to make time series stationary.

MA (Moving Average): The model uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

ARIMA is a class of models that explains a given time series based on its own past values, that is, its own lags and the lagged forecast errors, so that the equation can be used to forecast future values. Although the method can handle data with a trend, it does not support time series with a seasonal component.

An extension to ARIMA that supports the direct modeling of the seasonal component of the series is called **SARIMA** (Seasonal Autoregressive Integrated Moving Average).

SARIMA is an extension of ARIMA that explicitly supports univariate time series data with a seasonal component. It adds three new hyperparameters to specify the autoregression (AR), differencing (I) and moving average (MA) for the seasonal component of the series, as well as an additional parameter for the period of the seasonality.

SARIMA configuration requires selecting hyperparameters for both the trend and seasonal elements of the series.

Trend Elements
They are the same as the ARIMA model:

- p: Trend autoregression order.
- d: Trend difference order.
- q: Trend moving average order.

Seasonal Elements
There are four seasonal elements that are not part of ARIMA that must be configured:

- P: Seasonal autoregressive order.
- D: Seasonal difference order: seasonal differencing is similar to regular differencing, but, instead of subtracting consecutive terms, it subtracts the value from previous season.
- Q: Seasonal moving average order.
- m: The number of time steps for a single seasonal period.

For our model, we focus the analysis on the behavior of passenger boardings on a route on any given day. Some routes recorded less than 6 hours in several days, mainly during the quarantine periods from March. For this reason, we set a daily period for the time series and take the maximum of the boardings.

We used the maximum hourly boardings for each day because the mean and median approximate the typical number of boardings each day, but they do not capture the magnitude of the passenger volume on an entire day. On the other hand, the total sum of boardings per day approximates the volume of transport demand, however, several hours were not observed in some days, therefore the demand could be underestimated. The maximum number of boardings per day seems to be a more balanced statistic.

Analyzing particular routes like the ones depicted below, we observed that there are missing values on some days, which can affect the prediction model for the time series.

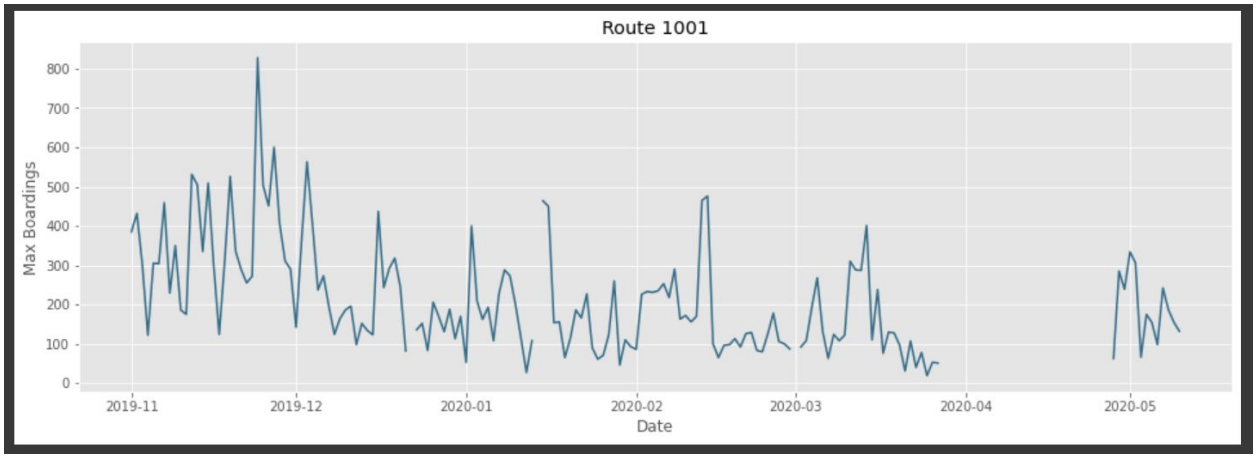


Fig.30 Time series for route 1001

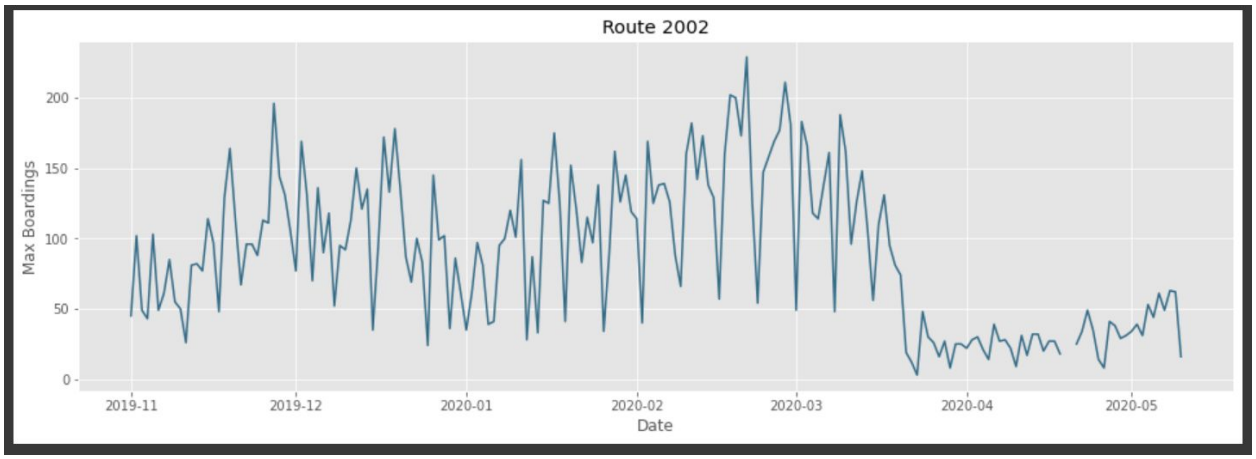


Fig.31 Time series for route 2002

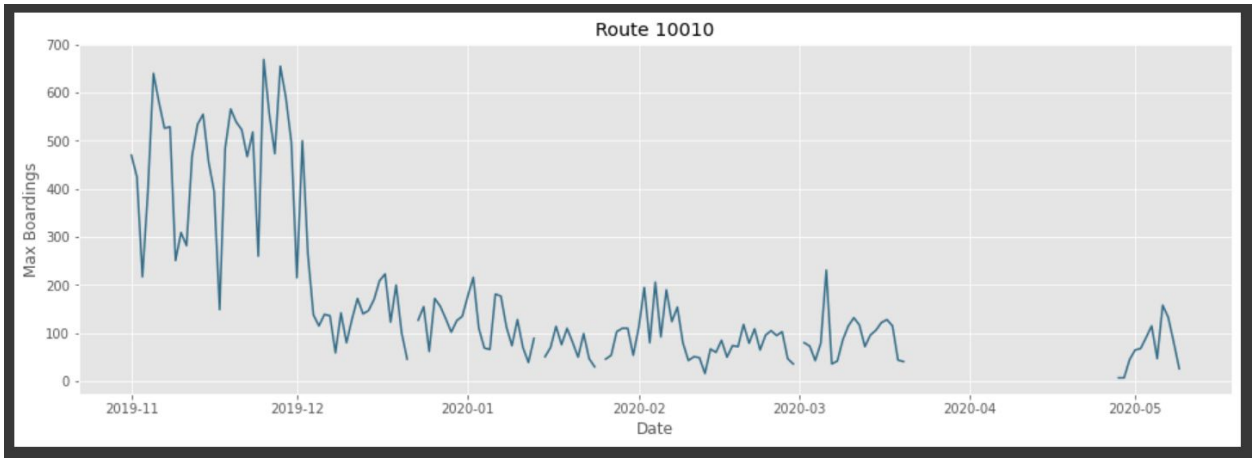


Fig. 32Time series for route 10010

For that reason, we truncated those series with more than 20 missing values in that time interval. Then, we approximated the number of boardings on the remaining missing values days through an imputation method like the seasonal mean.

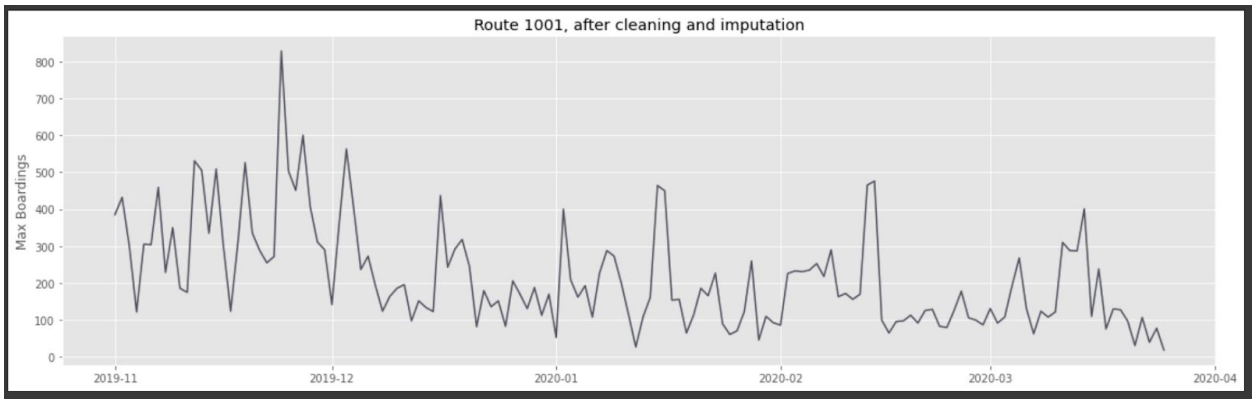


Fig. 33 Time series for route 1001 after cleaning and imputation

We decomposed the time series of routes into its principal components: trend, seasonal index, and residuals. For example, we took route 1001 and decomposed it using multiplicative and additive models:

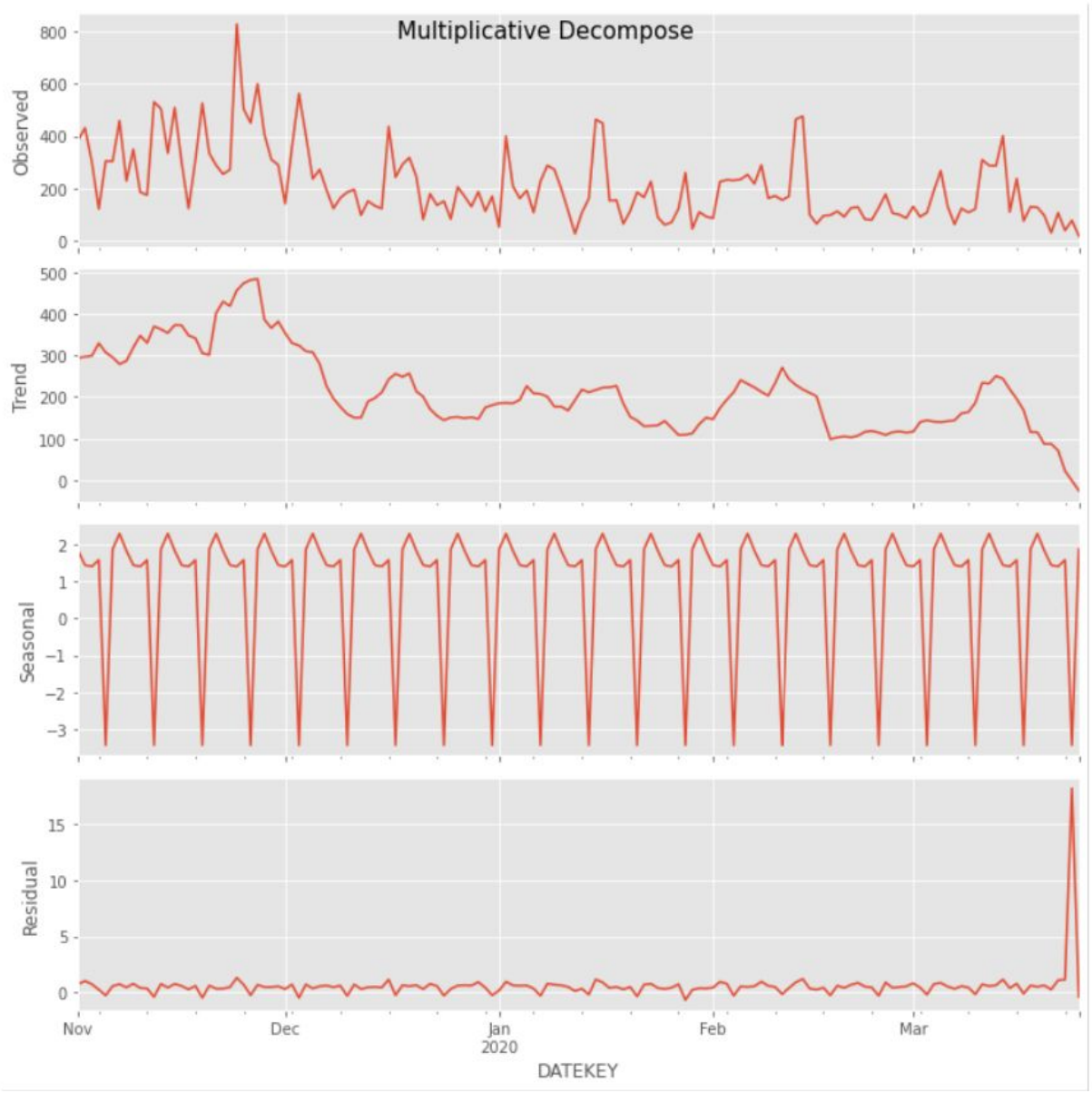


Fig. 34 Multiplicative decompose model

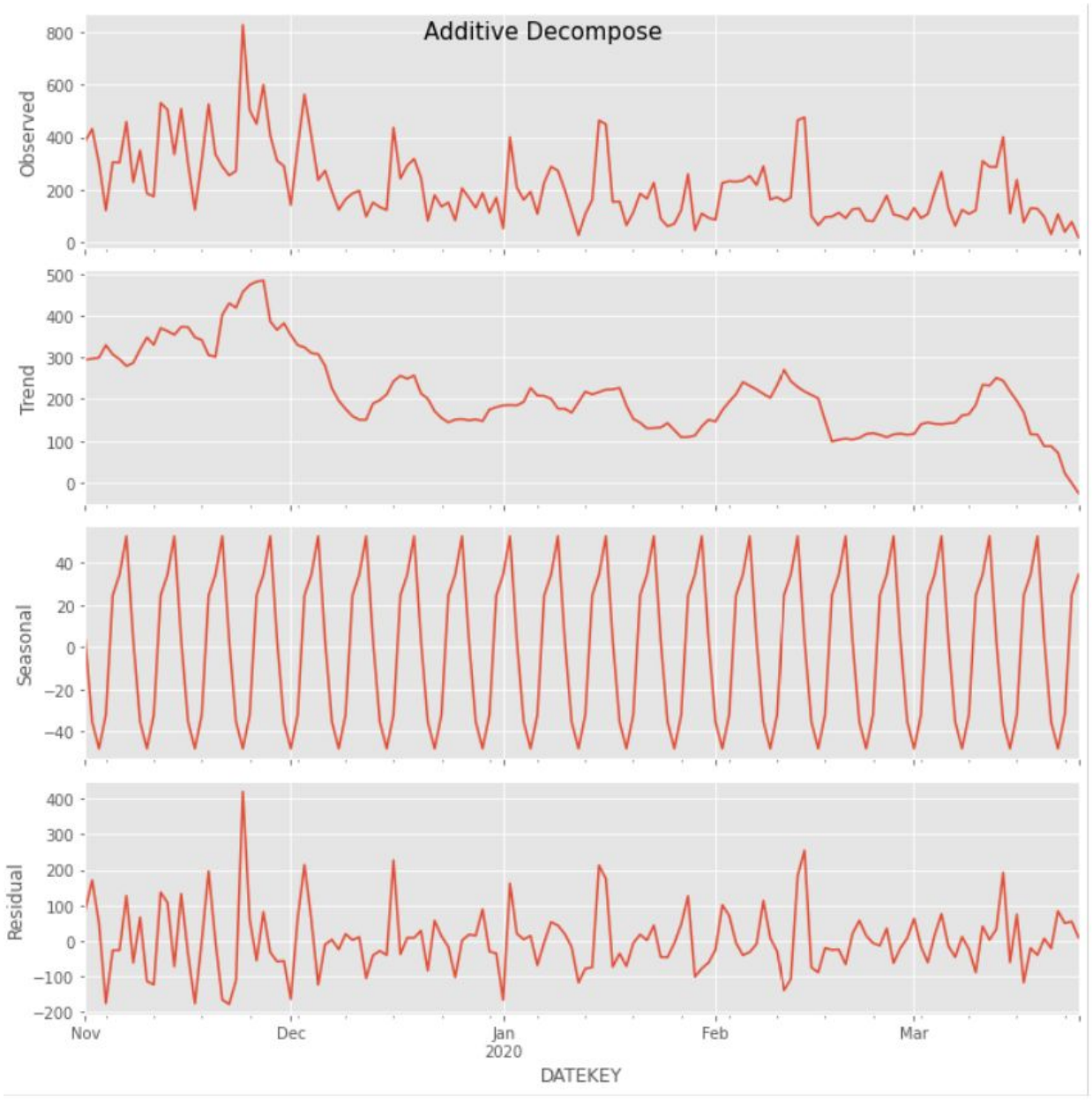


Fig. 35Additive decompose model

Time series for the previous route appears to have a nonconstant mean, variance, and a seasonal component. The diagnosis is probably similar to the other routes. Also, time series appears to come from an additive process.

In order to determine if the series is stationary or not, we implemented the unit root test Augmented Dickey-Fuller (ADF). Under the null hypothesis that the series presents a unit root,

we reject the non-stationarity hypothesis if the test statistic is less than the 5% significance level. These are the results of the test for some routes:

```
Route: 1001
 ADF Statistic: -2.14, p-value: 0.23, to 5% significance level: Non-Stationary
Route: 1002
 ADF Statistic: -9.60, p-value: 0.00, to 5% significance level: Stationary
Route: 1003
 ADF Statistic: -7.87, p-value: 0.00, to 5% significance level: Stationary
Route: 1004
 ADF Statistic: -4.36, p-value: 0.00, to 5% significance level: Stationary
Route: 1005
 ADF Statistic: -4.28, p-value: 0.00, to 5% significance level: Stationary
Route: 1006
 ADF Statistic: -4.34, p-value: 0.00, to 5% significance level: Stationary
Route: 1009
 ADF Statistic: -6.17, p-value: 0.00, to 5% significance level: Stationary
Route: 1010
 ADF Statistic: -5.89, p-value: 0.00, to 5% significance level: Stationary
Route: 1011
 ADF Statistic: -3.89, p-value: 0.00, to 5% significance level: Stationary
Route: 1015
 ADF Statistic: -4.04, p-value: 0.00, to 5% significance level: Stationary
Route: 1018
 ADF Statistic: -4.27, p-value: 0.00, to 5% significance level: Stationary
```

Fig.36 Stationarity test

For the non-stationary series we applied the ADF test, as it can be observed in the following graphs:

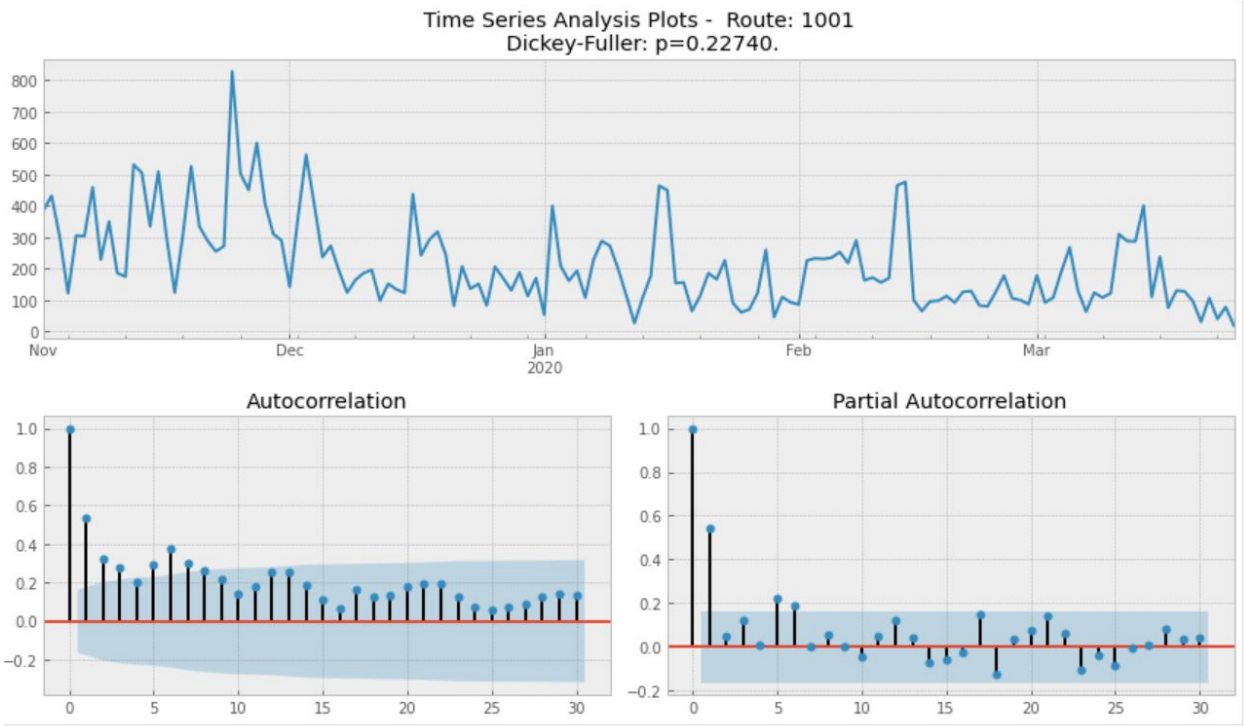


Fig.37 ADF test for route 1001

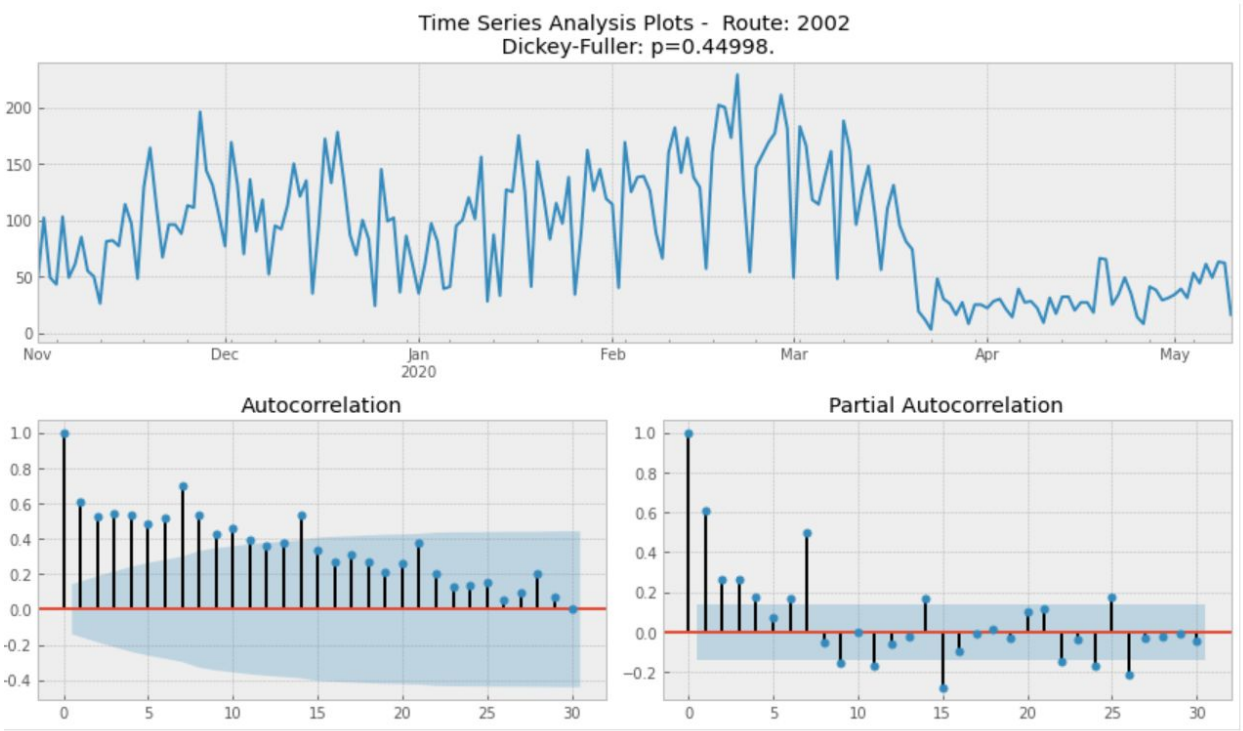


Fig.38 ADF test for route 2002

The series showed a certain downward trend and high variability at the beginning or end of the time horizon. For that reason, we eliminated the trend component and took the first difference of each series.

The Dickey-Fuller test indicated that all series were stationary after removing the trend. Therefore, there was no need to calculate differences for the data to become stationary (d=0). The next graph shows the stationarized series together with their Autocorrelation and Partial Autocorrelation functions.

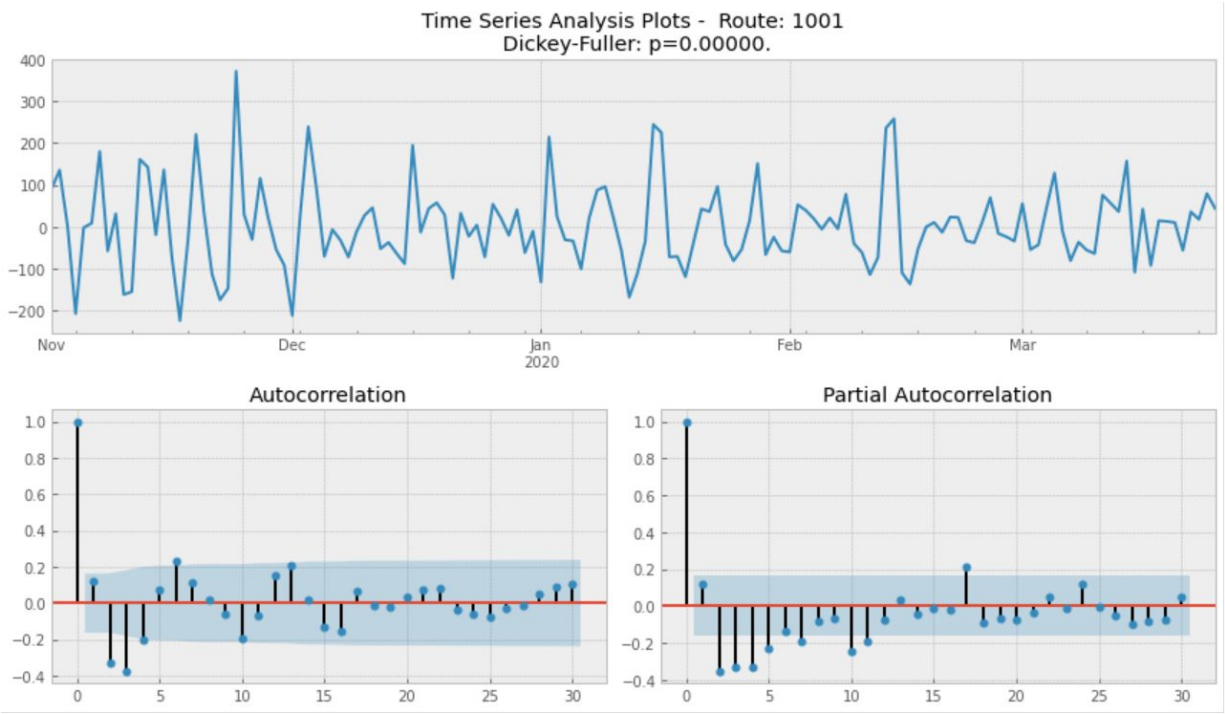


Fig.39 Stationarity test for route 1001

Thus, for prediction, we applied ARIMA model obtaining these results:

ARMA Model Results

Dep. Variable:	PAXUP	No. Observations:	146
Model:	ARMA(1, 1)	Log Likelihood	-901.825
Method:	css-mle	S.D. of innovations	116.356
Date:	Wed, 11 Nov 2020	AIC	1811.649
Time:	19:38:49	BIC	1823.584
Sample:	11-01-2019 - 03-25-2020	HQIC	1816.498

	coef	std err	z	P>|z|	[0.025	0.975]
const	215.6446	24.076	8.957	0.000	168.456	262.833
ar.L1.PAXUP	0.6854	0.155	4.409	0.000	0.381	0.990
ma.L1.PAXUP	-0.2031	0.228	-0.890	0.375	-0.650	0.244

Roots

	Real	Imaginary	Modulus	Frequency
AR.1	1.4590	+0.0000j	1.4590	0.0000
MA.1	4.9238	+0.0000j	4.9238	0.0000

Fig.40 Fitting Arima model

The summary of the fitted model showed that the coefficients were significant at 5%. However, we can test different combinations of the parameters (p, d, q) to minimize the AIC information criterion and fit the best model to the series.

According to the AIC criterion, the best model was **ARIMA (1,0,0)** and we tried the model prediction for the next 30 days:

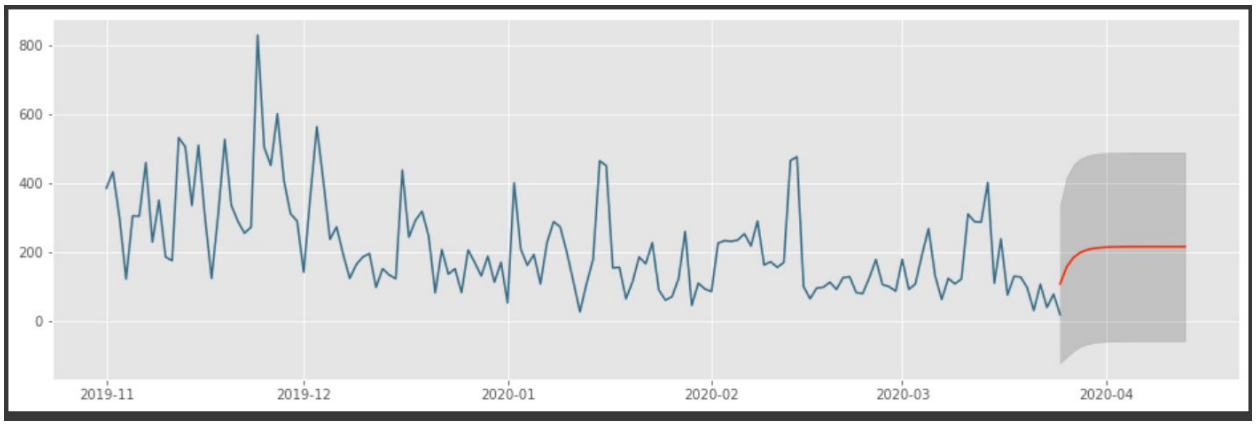


Fig.41 Prediction using ARIMA(1,0,0)

The prediction does not perform well because we are not considering the seasonality of the series. Previously, in the additive decomposition, we observed that the series seemed to repeat itself every seven days (1 week), that is, daily data with weekly seasonality, so we tried fitting a SARIMA model and look at the prediction for the next 30 days:

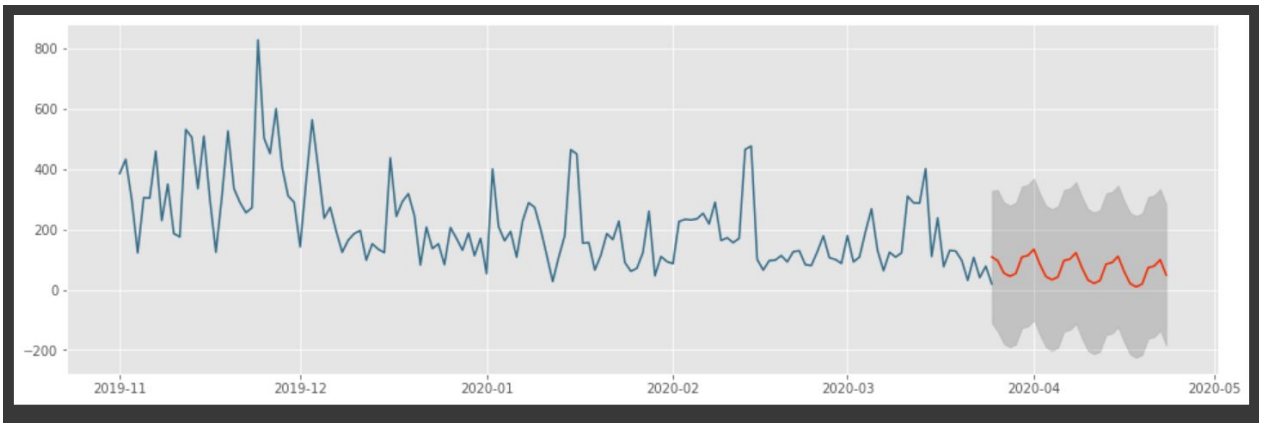
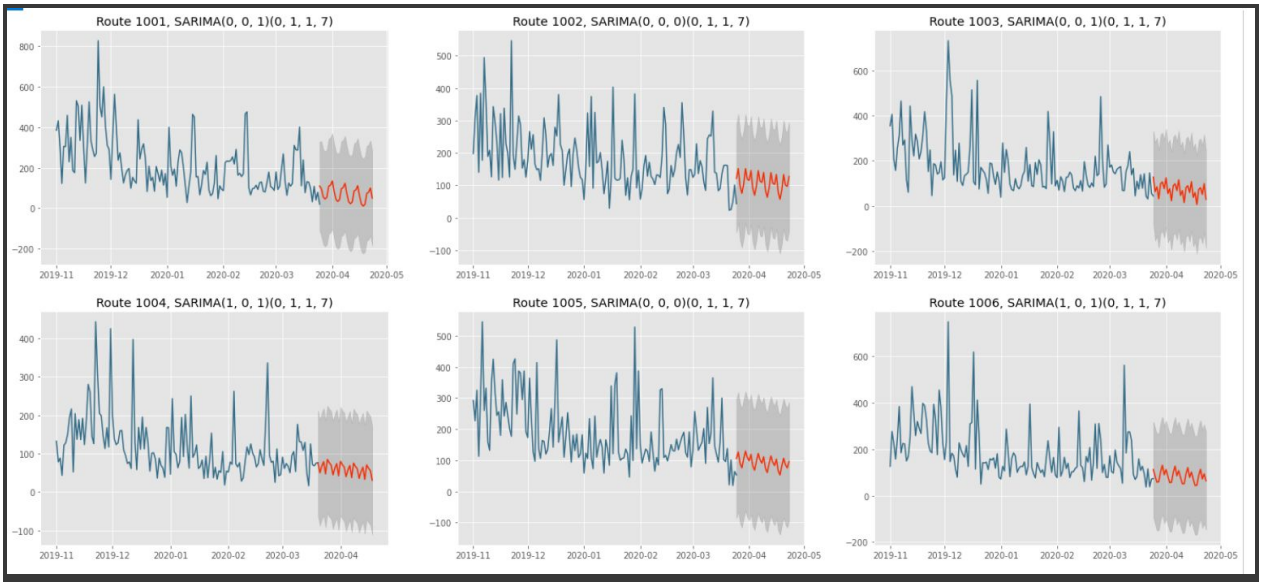
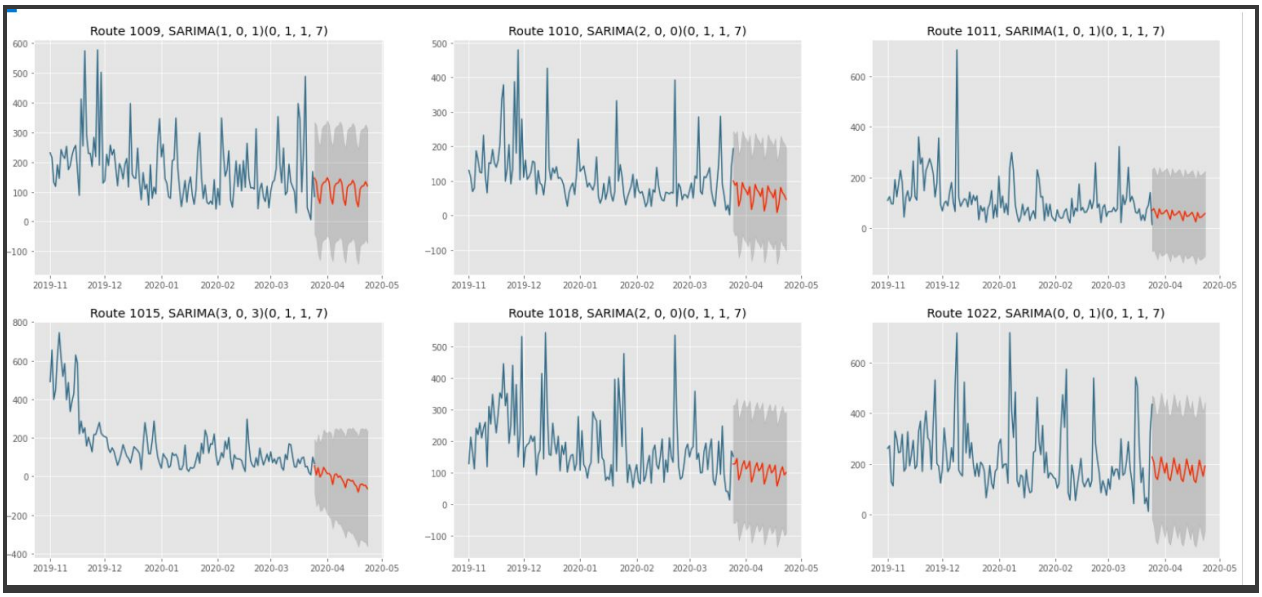
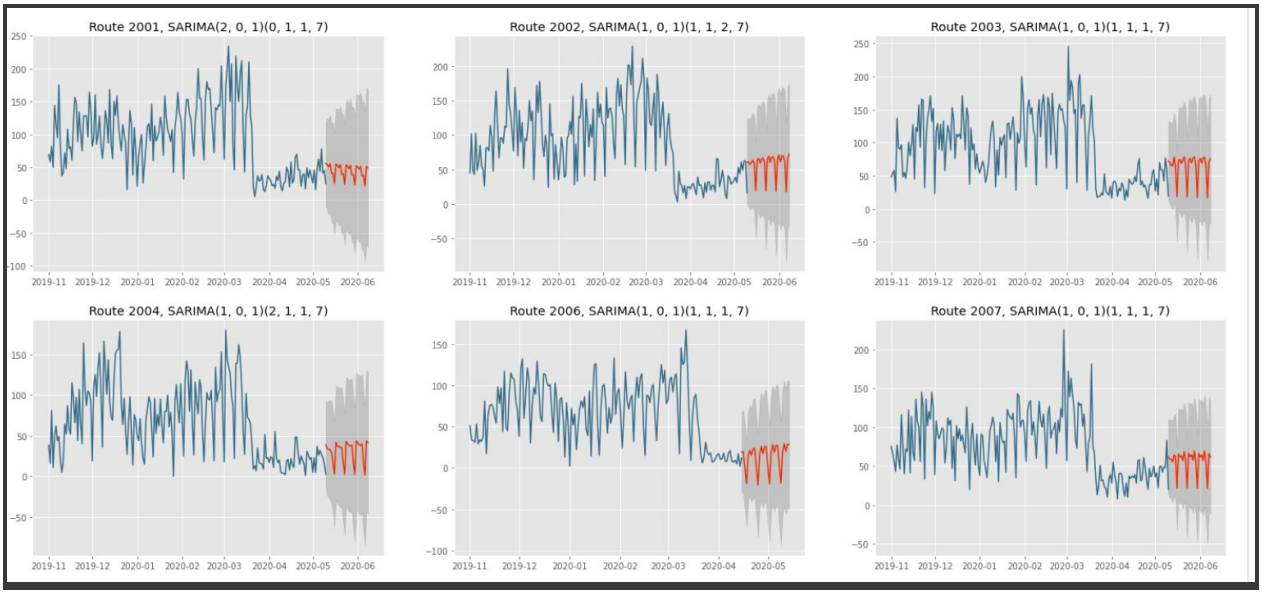


Fig.42 Prediction using SARIMA for one route

Finally, we built the SARIMA model for each route using pmdarima's auto_arima() function. For that, we set seasonal=True, frequency m=7, and let the algorithm define the best set of parameters according to the ADF test and the AIC information criteria. Once the models were adjusted, we forecasted the maximum daily boardings per route for the next 30 days and saved the results that can be observed as follows:




Route 1009, SARIMA(1, 0, 1)(0, 1, 1, 7)
Route 1010, SARIMA(2, 0, 0)(0, 1, 1, 7)
Route 1011, SARIMA(1, 0, 1)(0, 1, 1, 7)
Route 1015, SARIMA(3, 0, 3)(0, 1, 1, 7)
Route 1018, SARIMA(2, 0, 0)(0, 1, 1, 7)
Route 1022, SARIMA(0, 0, 1)(0, 1, 1, 7)


Route 2001, SARIMA(2, 0, 1)(0, 1, 1, 7)
Route 2002, SARIMA(1, 0, 1)(1, 1, 2, 7)
Route 2003, SARIMA(1, 0, 1)(1, 1, 1, 7)
Route 2004, SARIMA(1, 0, 1)(2, 1, 1, 7)
Route 2006, SARIMA(1, 0, 1)(1, 1, 1, 7)
Route 2007, SARIMA(1, 0, 1)(1, 1, 1, 7)

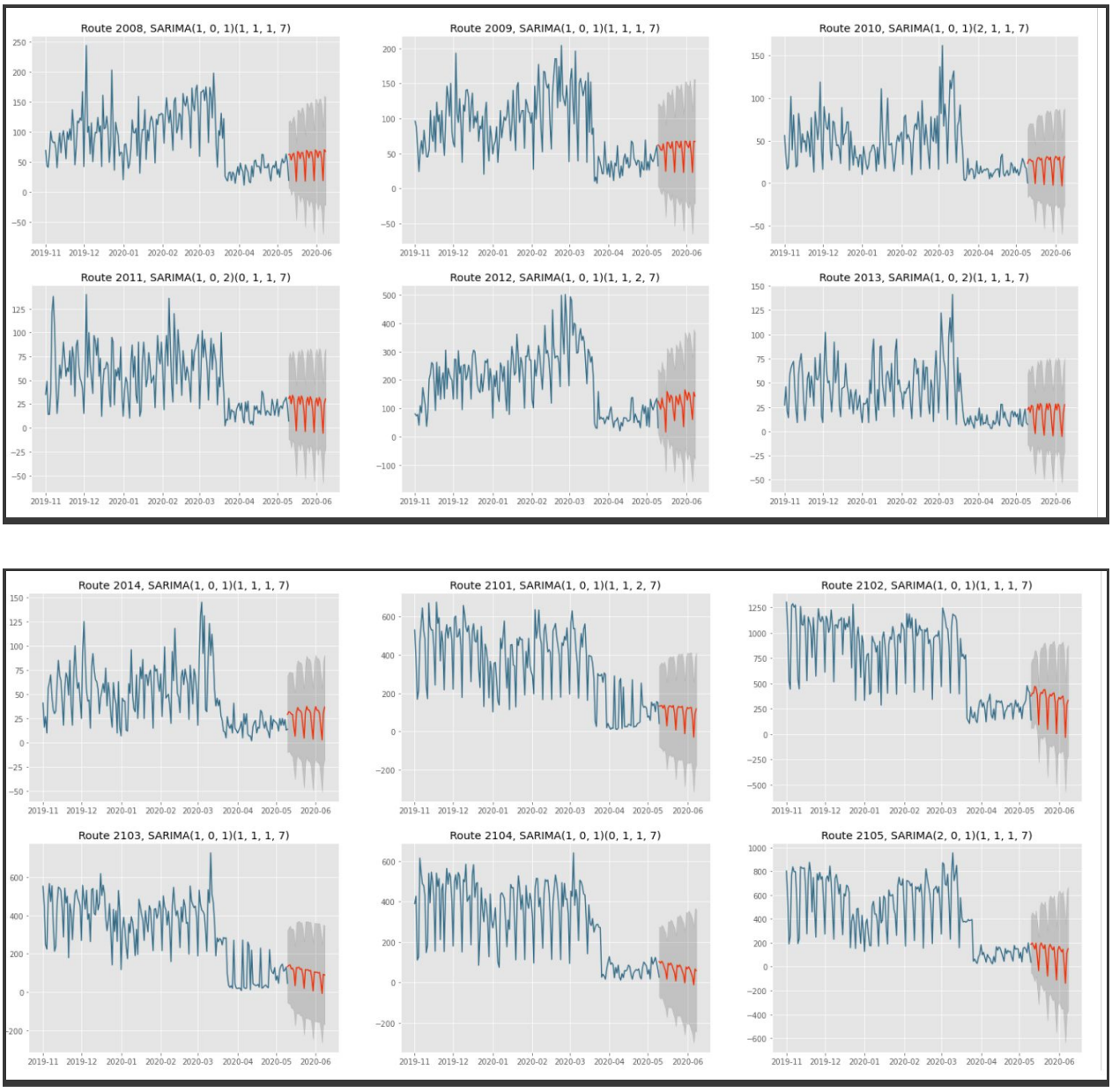


Fig.43 Prediction using SARIMA for the routes

5. Dashboard

After fitting the two models we proceeded to put it all together in a dashboard using Power BI that can be accessed using the following link https://tinyurl.com/y5ao296g.

This app has visualizations to see how many passengers board the vehicles in the AMVA; the days where the demand is higher and the spatial behavior of this variable.

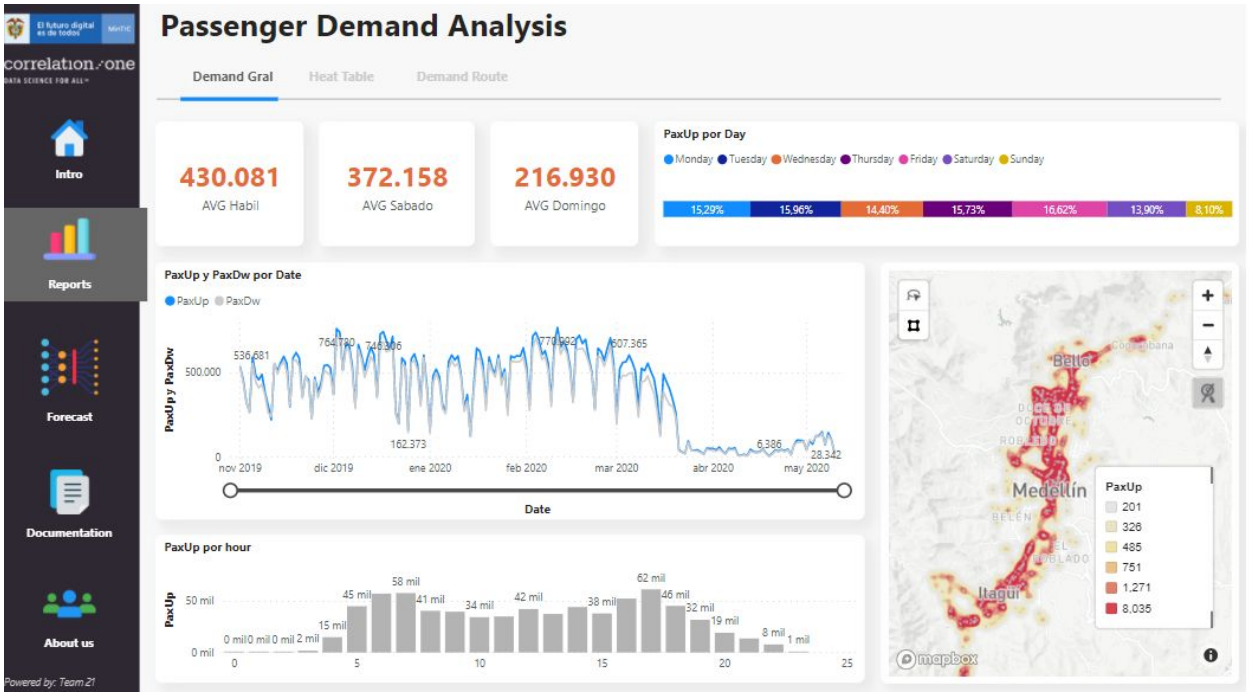


Fig.44 Dashboard: Vehicle demand

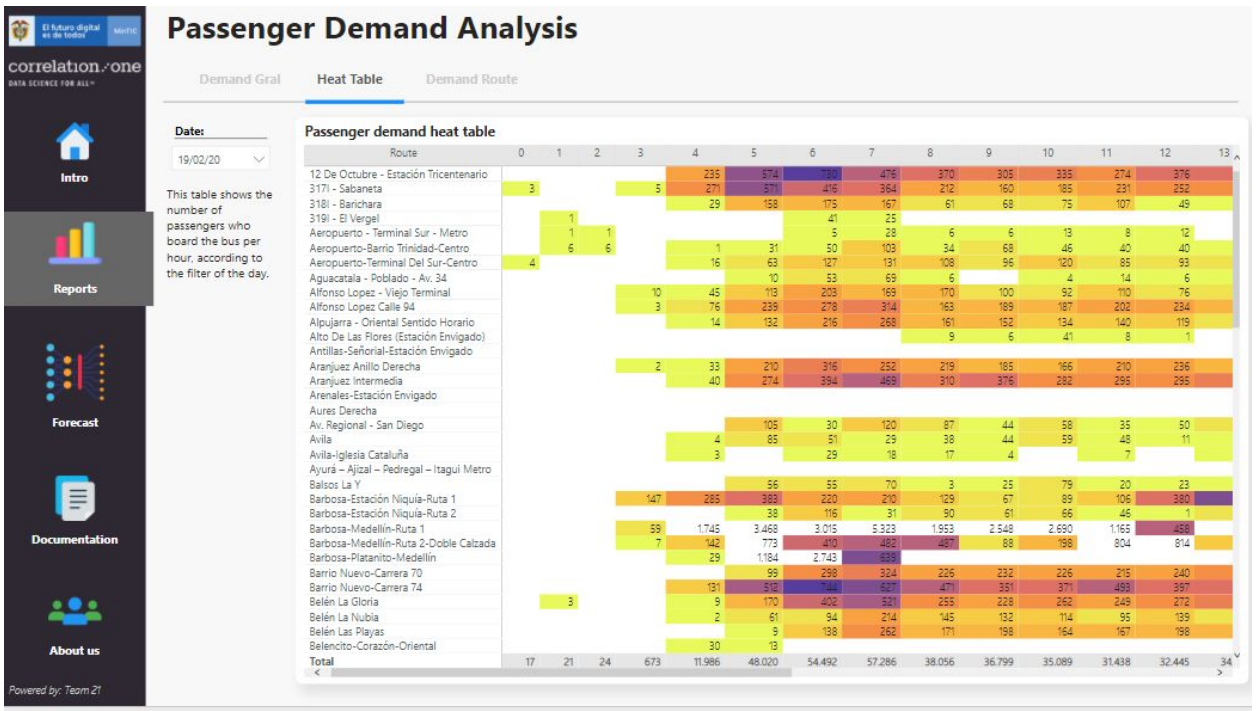


Fig.45 Dashboard: Demand by day and hour

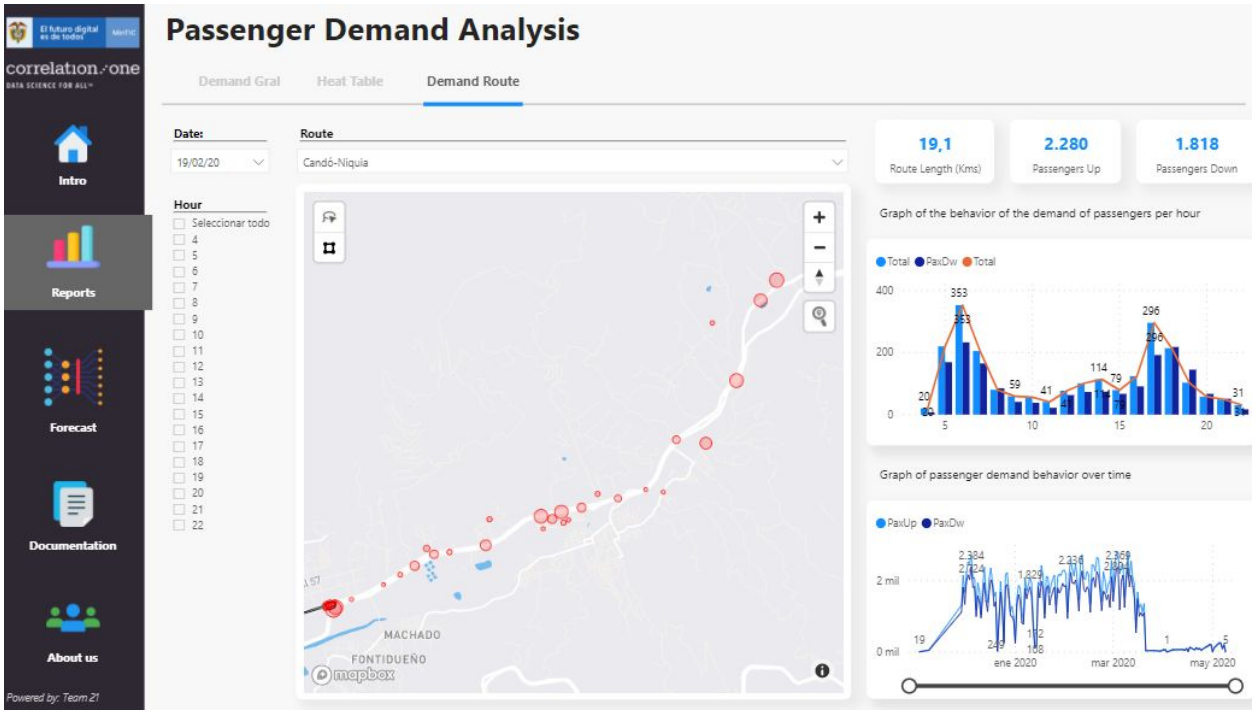


Fig.46 Dashboard: Demand spatially

As well, the app has plots regarding the forecast models fitted before,

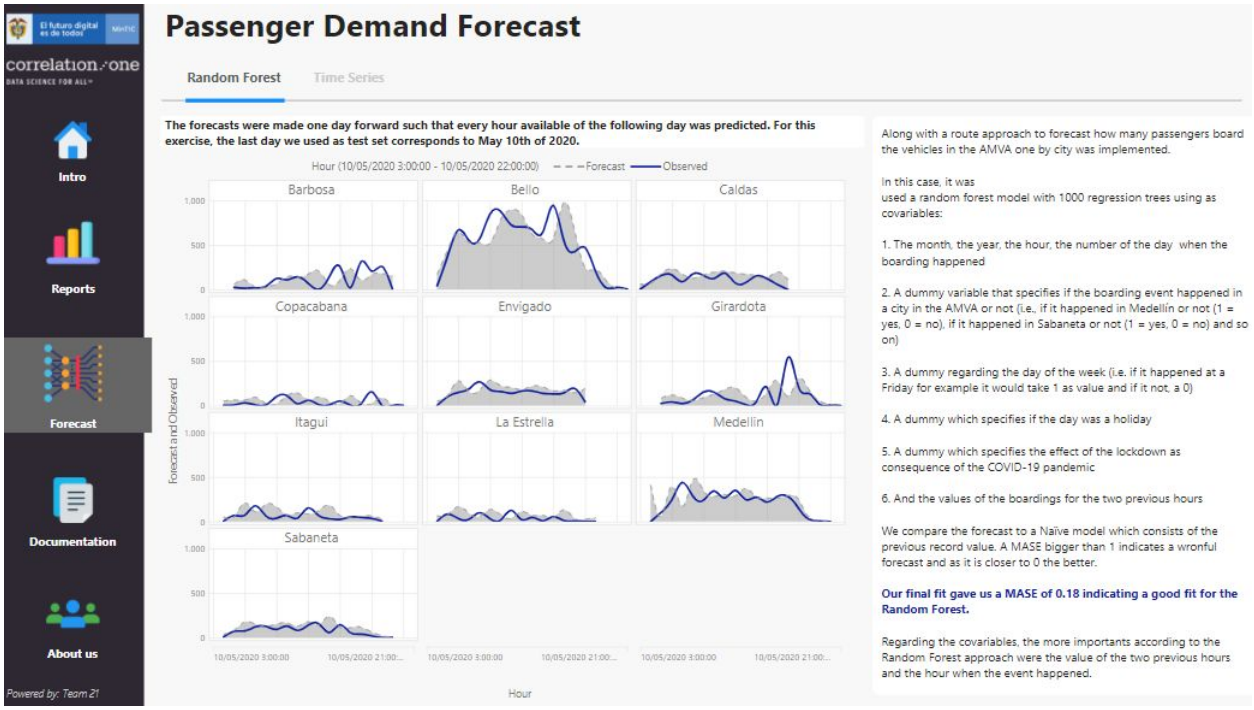


Fig.47 Dashboard: Forecast, Random Forest

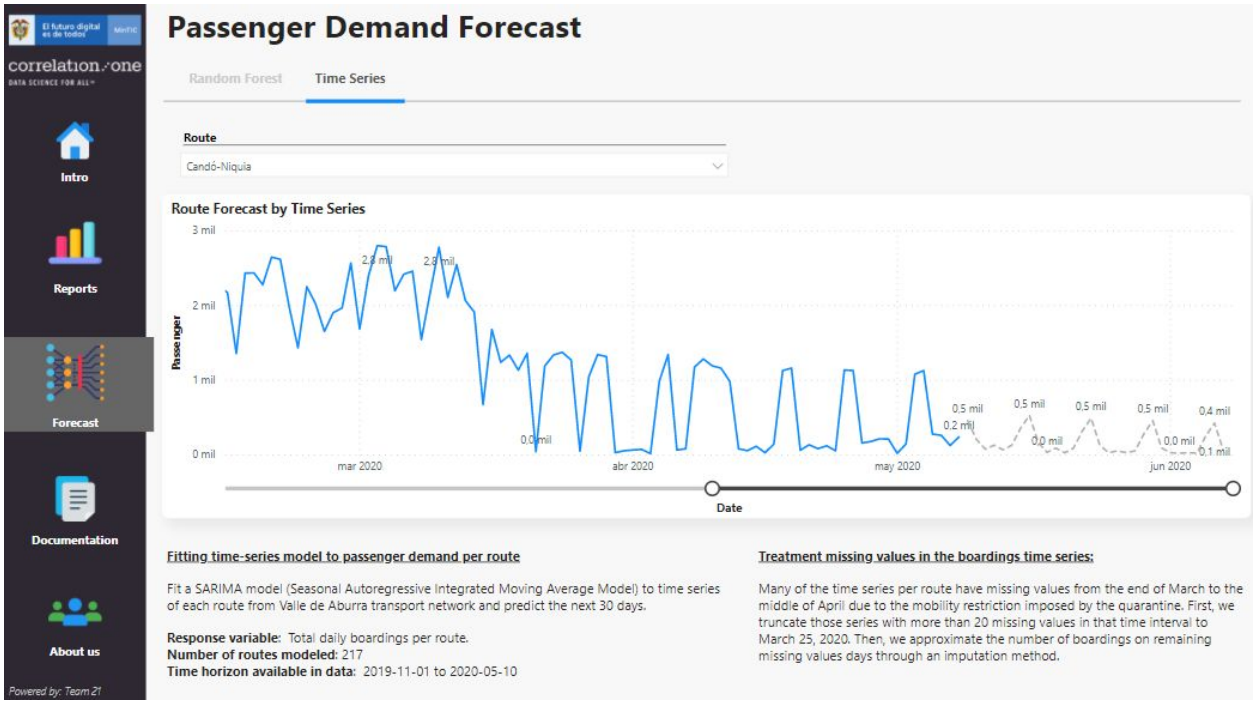


Fig.48 Dashboard: Forecast, Time Series

6.Conclusions and future work

Exploratory data analysis showed that despite the huge volume of information gathered daily from the vehicle's sensors, only around 20% of the data has values about boardings and alightings greater than zero and it was valuable for analysis. Therefore, it is recommended that transport companies that operate at the road network tune the devices that collect the data.

Inconsistencies in data like the recording of events at unconventional hours, passenger alightings at the beginning of the route, more passengers getting off than boarding, or extreme values in both the number of passengers who get on and those who get off, seriously distort the calculation of the load of passengers in a given link. For this reason, it is important to establish thresholds considered as normal movements of passengers.

In order to calculate the load of passengers for a given link, on an hourly basis, it was necessary to associate each observation at a specific time and longitude/latitude to the closest node of the road network using the Haversine Formula. This approach reduced greatly the volume of data for exploratory and predictive analysis and allowed us to visualize the loads of passengers on the map of Aburrá Valley.

The transportation demand modeling has been approached from different perspectives, being one of the most common, corresponding to the analysis of Time Series and Forecasts. Additionally, innovative techniques have also been used, such as some framed within Machine Learning and Deep Learning. Nevertheless, some authors reiterate the absence of this type of study in the literature such that this subject still could be explored in detail using some of the previously mentioned techniques in future works.

A dashboard was created so the entity could use it in order to understand how the passengers use their services and to forecast the demand on an hourly basis and on a daily one, nevertheless, there's still the need to improve the gathering of the information reiterating the importance of calibrating the sensors.

7. Bibliography

Área Metropolitana del Valle de Aburrá (2020a, November 5). Calidad y Cobertura. *Área Metropolitana del Valle de Aburrá*.

https://www.metropol.gov.co/la-movilidad/transporte-p%C3%BAblico/calidad-y-cobertura#:~:text=El%20estudio%20ha%20permitido%20observar,de%20transporte%20privado%20(26%25)

Área Metropolitana del Valle de Aburrá (2020b, November, 5). SITVA: Sistema Integrado de Transporte del Valle de Aburrá. *Área Metropolitana del Valle de Aburrá.* https://www.metropol.gov.co/movilidad/Paginas/transporte-publico/sitva.aspx

Bolaños, L.F (2017). Los 10 lugares recomendados para celebrar el Día de las Velitas. *La República*. https://www.larepublica.co/ocio/los-10-lugares-para-pasar-el-dia-de-velitas-2578765

Brownlee, J. (2018). A Gentle Introduction to SARIMA for Time Series Forecasting in Python. *Machine Learning Mastery*. https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/

Brownlee, J. (2017) Time Series Components. *Machine Learning Mastery*. https://machinelearningmastery.com/decompose-time-series-data-trend-seasonality/

Chang, V. (2019). Un análisis de series de tiempo mediante modelos SARIMAX para la proyección de demanda de carga en el puerto del Callao. Revista de Análisis Económico y Financiero, 1(3), 15 – 31. https://www.aulavirtualusmp.pe/ojs/index.php/raef/article/view/1694

Davis, D. (2020). Random Forest Classifier Tutorial: How to Use Tree-Based Algorithms for Machine Learning. *Freecamp.org*. https://www.freecodecamp.org/news/how-to-use-the-tree-based-algorithm-for-machine-learning/

Díaz, A. (2012). Trabajo de Cátedra: Calculadora Espacial. *Cátedra: Informática I, Universidad Tecnológica Nacional Facultad Regional Bahía Blanca*. http://41jaiio.sadio.org.ar/sites/default/files/41_EST_2012.pdf

EPM (2020). Alumbrados Navideños de Medellín se apagan este lunes festivo 6 de enero. *EPM*. https://www.epm.com.co/site/alumbrados-navidenos-de-medellin-se-apagan-este-lunes-festivo-6-de-enero

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, 2nd edition, OTexts: Melbourne, Australia. otexts.com/fpp2/. Accessed on November, 13, 2020.

Li, A. & Axhausen, K.W. (2019). Comparison of short-term traffic demand prediction methods for transport services. Arbeitsberichte Verkehrs- und Raumplanung, 1447. www.research-collection.ethz.ch/handle/20.500.11850/356143

Liu, Z., Chen, H., Li, Y. & Zhang, Q. (2020). Taxi Demand Prediction Based on a Combination Forecasting Model in Hotspots. Journal of Advanced Transportation. 2020. www.hindawi.com/journals/jat/2020/1302586/

Pemberti, G. (2020). Presentación Técnica Proyecto GTPC. *Información Retos AMVA*. drive.google.com/drive/folders/1ZO6c0xtNiF_WuATGMqmuJxXdHlJcKQAa

Prabhakaran, S. (2019). ARIMA Model – Complete Guide to Time Series Forecasting in Python. *Machine Learning Plus*. www.machinelearningplus.com/time-series/time-series-analysis-python/

Prabhakaran, S. (2019). Time Series Analysis in Python – A Comprehensive Guide with Examples. *Machine Learning Plus*. www.machinelearningplus.com/time-series/time-series-analysis-python/

Shrivastava, S. (2020, November 13). Cross Validation in Time Series. *Cross Validation in Time Series*. /medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4

Turismo en Medellín (2020, November 5). Tour Luces en Medellín. *Turismo en Medellín*. www.turismoenmedellin.com/turismed/vp5385/sp/tour-alumbrados-luces-medellin-colombia

Uber (2020, November 5). Qué hacer un domingo en Medellín para terminar bien el fin de semana. *Uber*. www.uber.com/es-CO/blog/que-hacer-un-domingo-en-medellin/

Valora Analitik (2020, November 5). Antioquia sí se sumará a los departamentos con aislamiento preventivo. *Valora Analitik*. www.valoraanalitik.com/2020/03/19/antioquia-se-sumaria-a-los-departamentos-con-aislamiento-preventivo/

Venegas (2020). Marchas del 1 de mayo en Medellín. *El Colombiano*. www.elcolombiano.com/multimedia/imagenes/galeriamarchas_del_1_de_mayo_en_medellin-KXEC_292866

Venos, N. (2019). Time Series Analysis in Python. *Time Series Analysis in Python*. medium.com/@nathanvenos/time-series-analysis-in-python-ab582dd803cd

Wikipedia. (n.d.). The Metropolitan Area of the Aburrá Valley. Retrieved November 3, 2020 from en.wikipedia.org/wiki/The_Metropolitan_Area_of_the_Aburr%C3%A1_Valley

Xue, R., Sun, D. & Chen, S. (2015). Short-Term Bus Passenger Demand Prediction Based on Time Series Model and Interactive Multiple Model Approach. Discrete Dynamics in Nature and Society, 2015. www.hindawi.com/journals/ddns/2015/682390/