JAM OR NO JAM ?
PREDICTING ROAD TRAFFIC IN PARIS

JULIEN LAKS

1/ PROBLEM STATEMENT

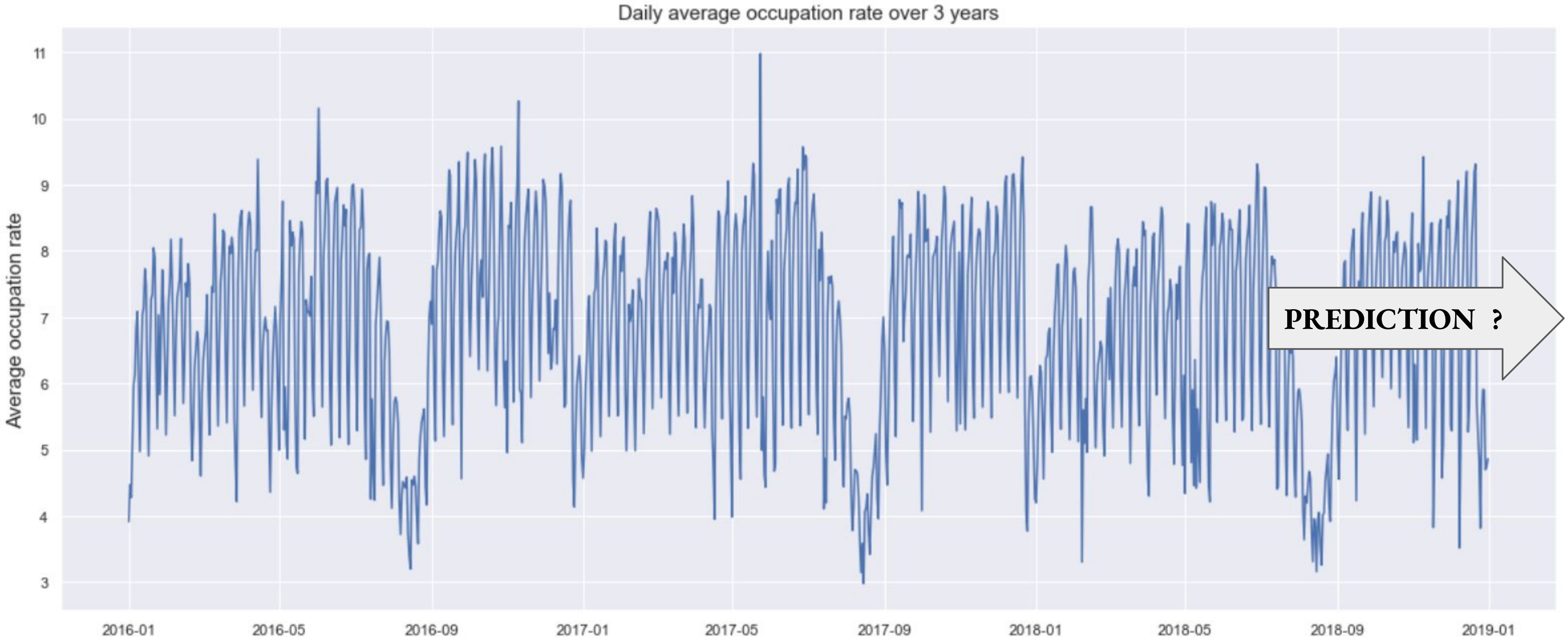
- Analyse global and local traffic trends **:** predict overall traffic intensity + predict **how** much cars will be at a certain place, a certain time, and a certain day of the year in Paris

- Traffic occupation rate **K** = % of time cars have occupied a road segment / hour

$0\% \leq K < 15\%$	Fluide
$15\% \leq K < 30\%$	Pré-saturé
$30\% \leq K < 50\%$	Saturé
$50\% \leq K$	Bloqué

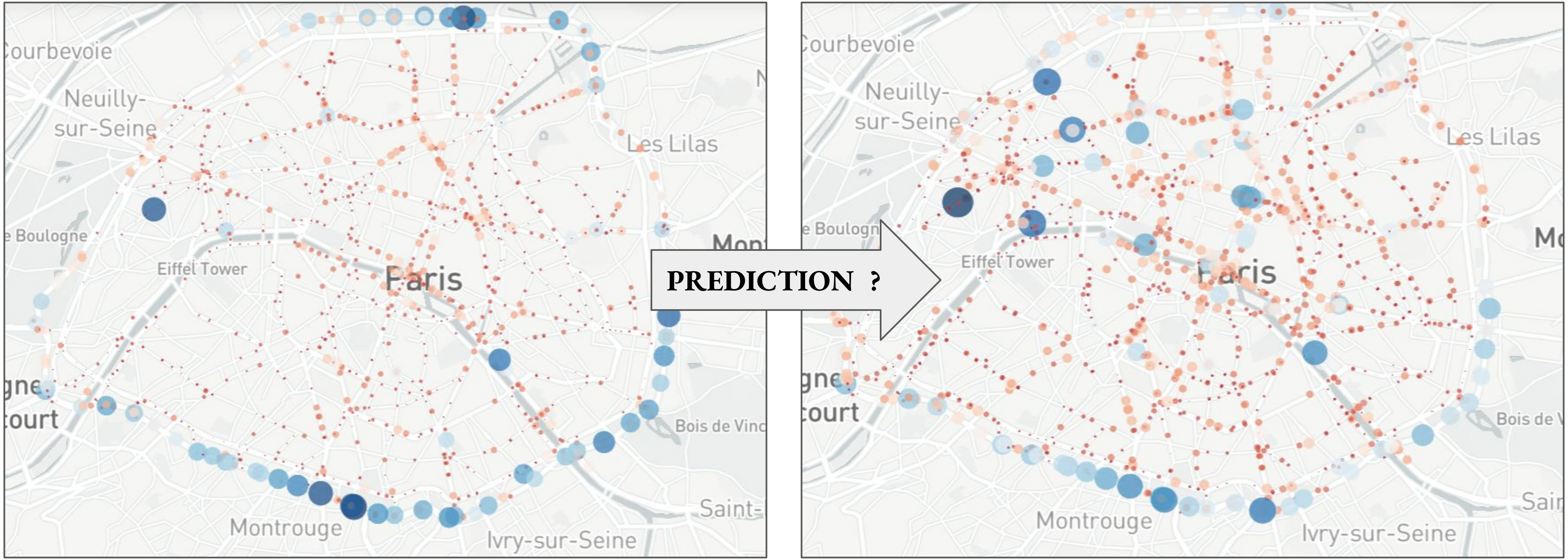
- TECHNICAL PRACTICE : Data Wrangling, Data Viz **(Plotly and Dash) ,** time series manipulation and analysis **(ARIMA, RNN),**

GOAL 1 : PREDICT WEEKLY AND DAILY AVERAGE TRAFFIC RATES



GOAL 2 : PREDICT HOURLY LOCAL TRAFFIC RATES AND JAMS

9 AM **12 PM**



2/ DATA ACQUISITION AND EDA

- Data collected from OPENDATA.PARIS website

- files : **geographical information,** of about **1700 recording stations**
  traffic measures, 150 datasets that needed to be concatenated

TIME

	timestamp	date	year	week	weekday	hour
0	2016-01-01 01:00:00	2016-01-01	2016	53	4	1
1	2016-01-01 02:00:00	2016-01-01	2016	53	4	2
2	2016-01-01 03:00:00	2016-01-01	2016	53	4	3
3	2016-01-01 04:00:00	2016-01-01	2016	53	4	4
4	2016-01-01 05:00:00	2016-01-01	2016	53	4	5

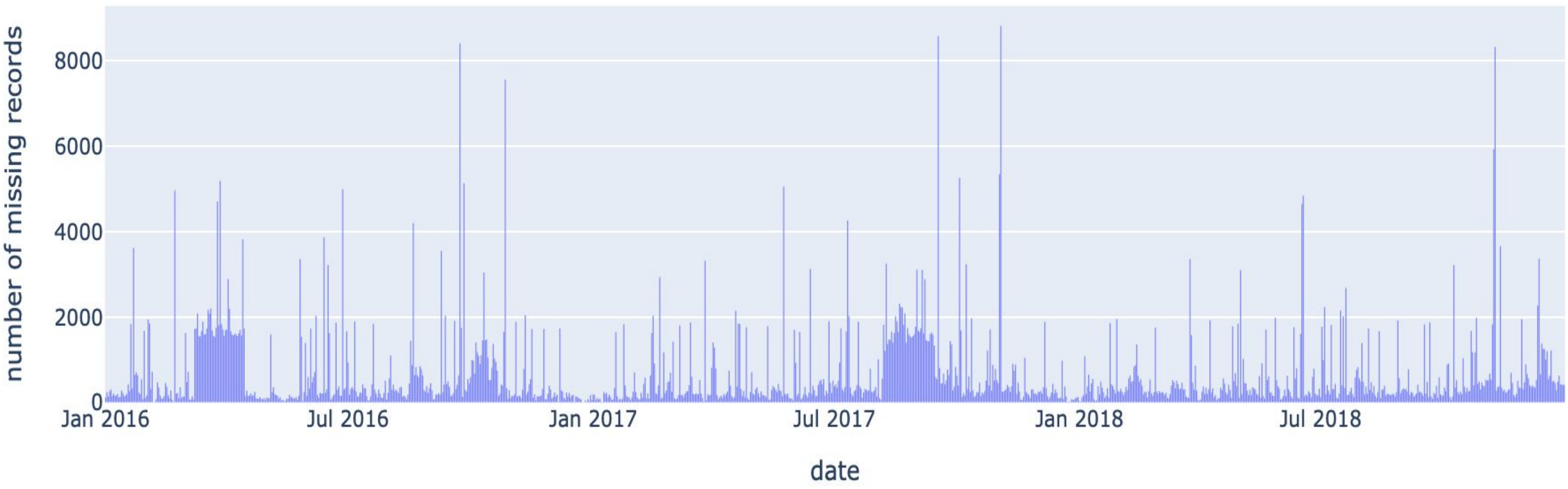
GEOGRAPHY

location_ID	road_ID	road_name	latitude	longitude
1	781	Quai_du_Louvre	48.859838	2.334242
2	781	Quai_du_Louvre	48.859375	2.336451
3	781	Quai_du_Louvre	48.859134	2.338776
4	781	Quai_du_Louvre	48.858747	2.341134
5	776	Quai_de_la_Megisserie	48.858214	2.343447

RECORDINGS

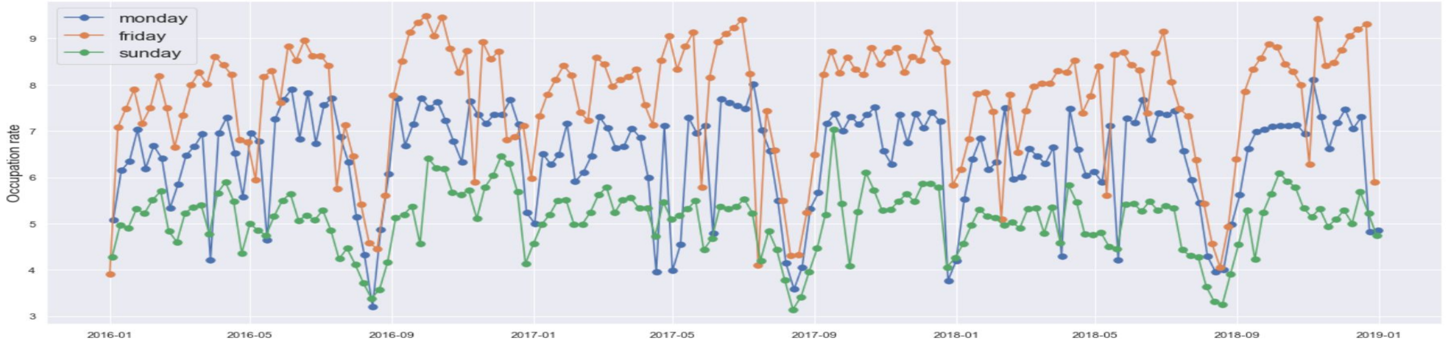
iu_nd_aval	libelle_nd_aval	t_1h	q	k
459	Bd_Kellermann-Damesme	2016-03-03 01:00:00	NaN	0.28278
459	Bd_Kellermann-Damesme	2016-03-03 02:00:00	NaN	0.12556

MISSING RECORDS

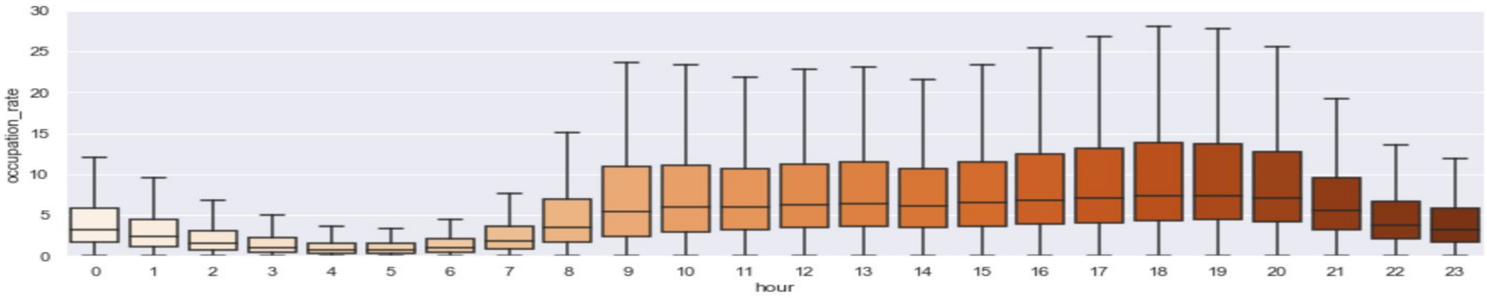


- Deleted recording stations with more than 5% of missing records
- Used **time interpolation method** to fill reamining missing values

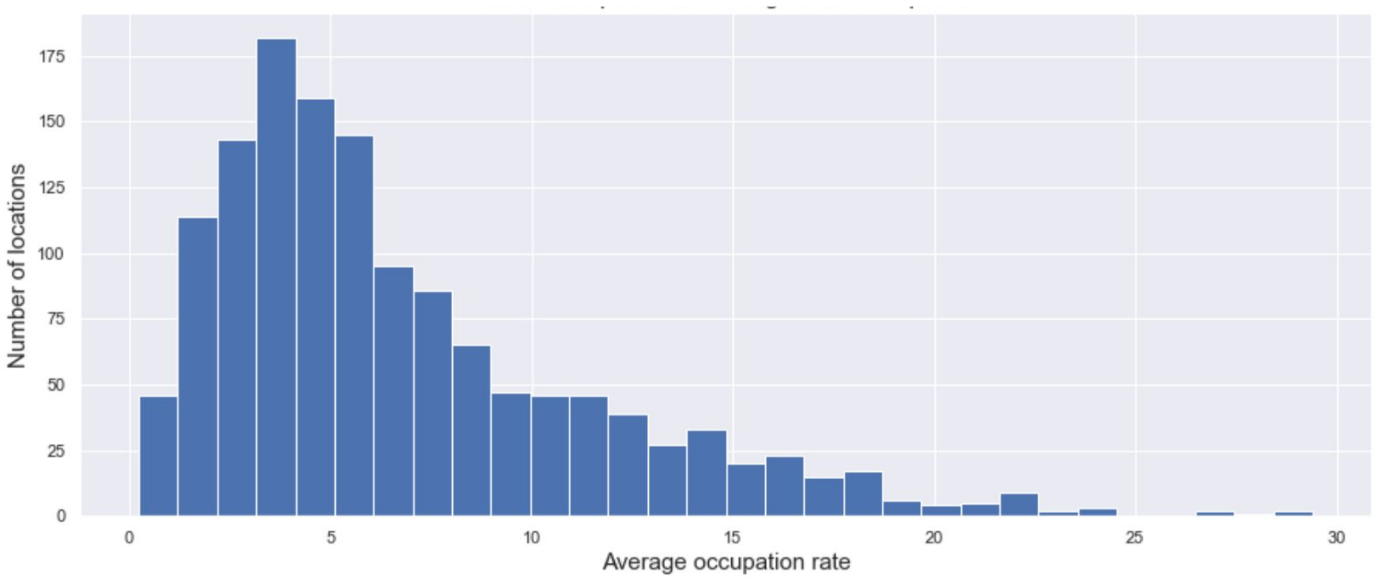
AVERAGE DAILY TRAFFIC RATES FOR 3 DIFFERENT WEEK DAYS



TYPICAL HOURLY TRAFFIC RATES

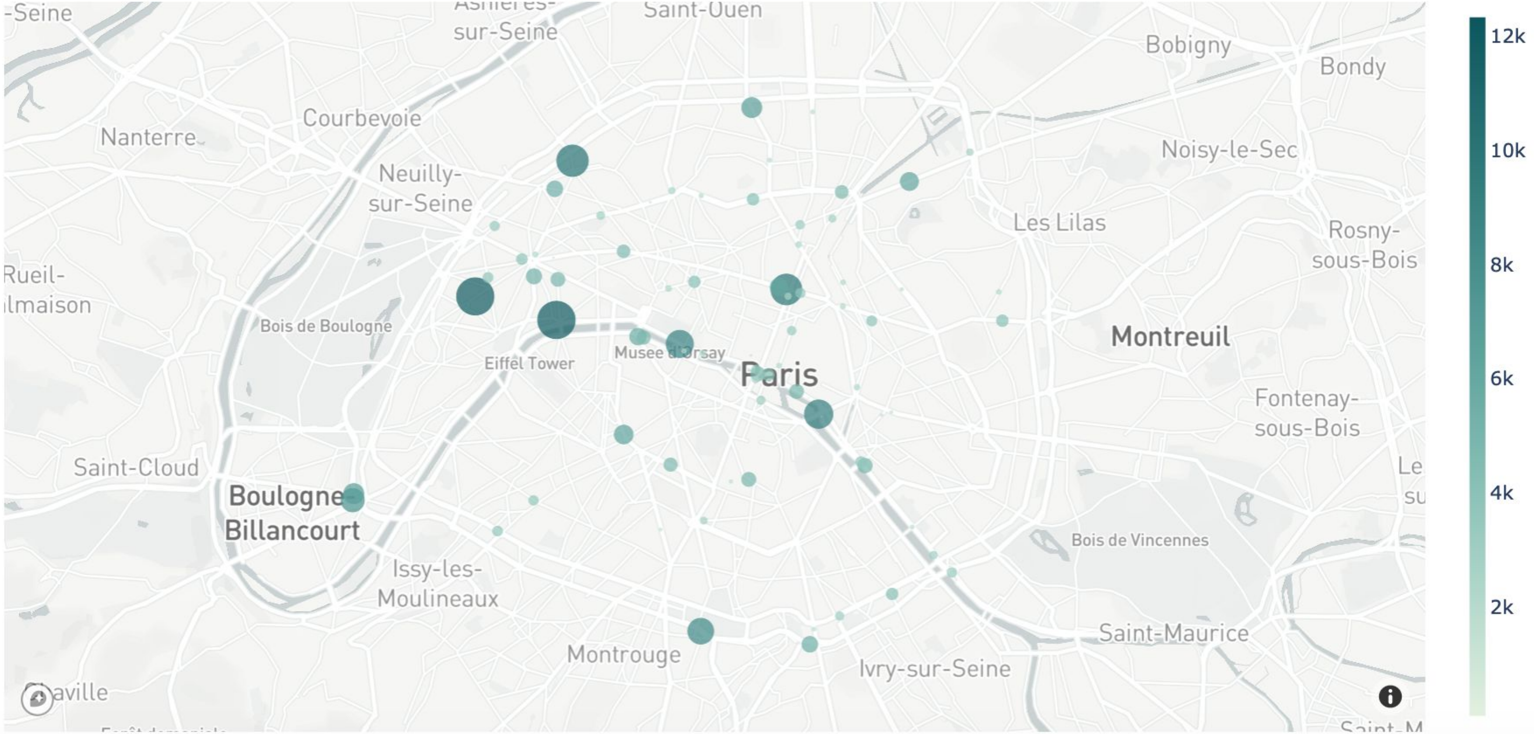


AVERAGE TRAFFIC RATES DISTRIBUTION

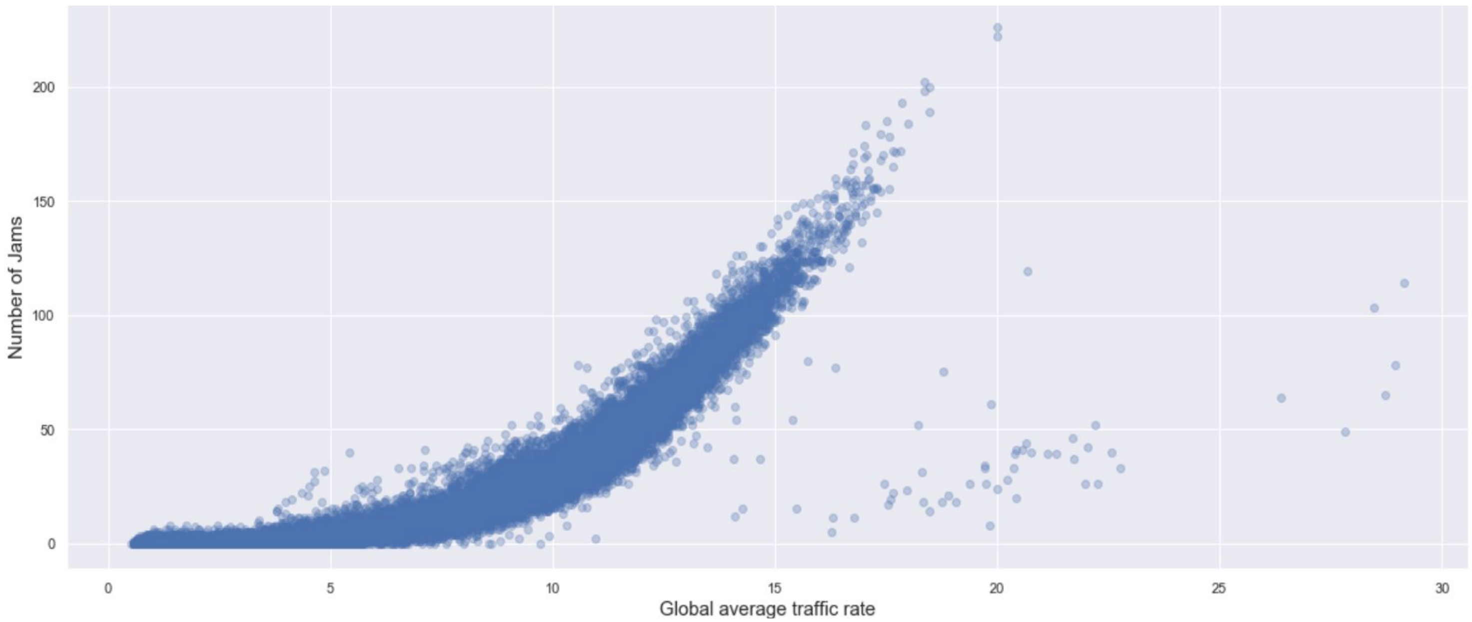


- The average traffic rate distribution is **skewed to the right** : The traffic is globally fluid in most locations, but tends to be very high and staturated at some specific places.

NUMBER OF JAMS AT EACH LOCATION OVER THREE YEARS (EXCLUDING OUTER HIGHWAY)

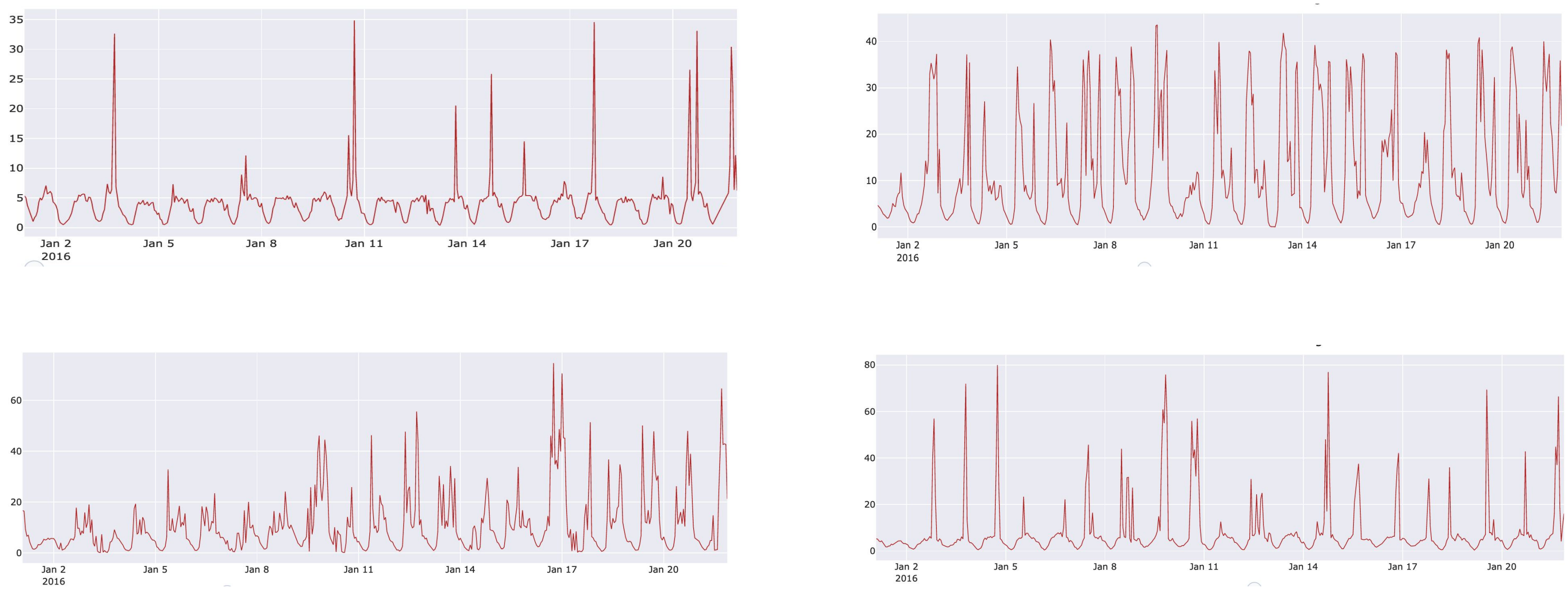


NUMBER OF JAMS VS GLOBAL TRAFFIC AVERAGE



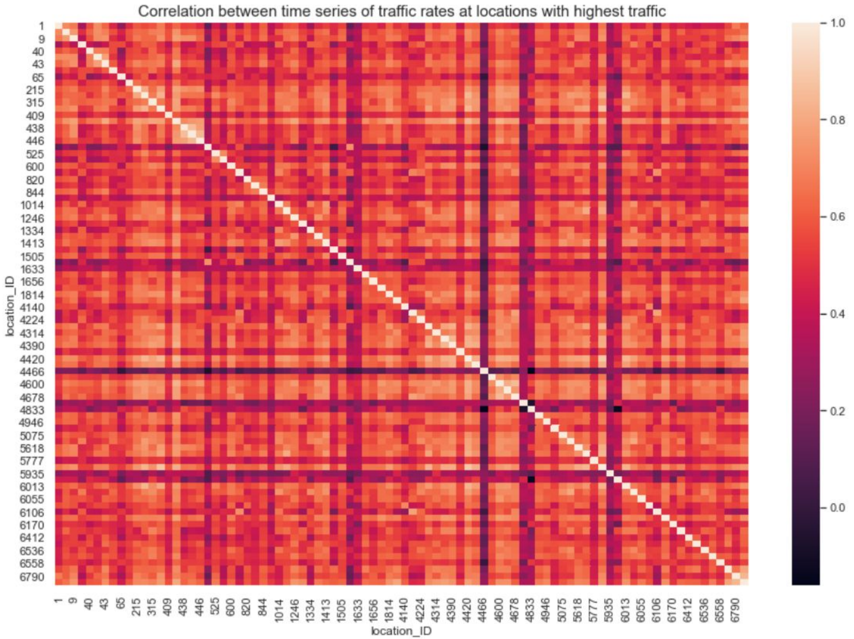
Bend in the curve at around **12%,** at which point the number of jams starts to increase more steeply. If we wanted to reduce the global number of traffic jams in Paris, this certainly be a critical value

HOURLY TRAFFIC RATES AT DIFFERENT LOCATIONS



- Although traffic rates at different locations tend to increase and decrease together, the above plots clearly suggest that there are also strong local particularities and variations in the traffic trends.

CORRELATIONS BETWEEN TRAFFIC RATES AT LOCATIONS WITH HIGHEST TRAFFIC



- Globally highly correlated
- Will information about other locations add to the predictive power of our forecasts for local traffic rates ?

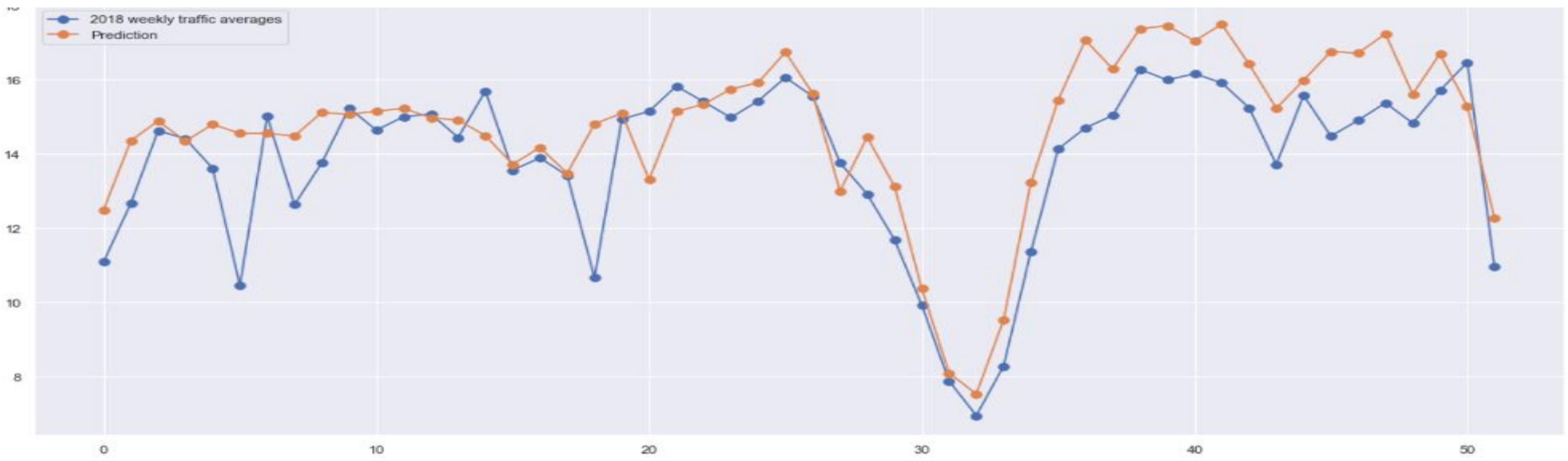
3/ FORECASTING

a/ Predicting weekly average traffic rates with simple averaging method

b/ Daily average traffic rates prediction with weekly seasonal arima

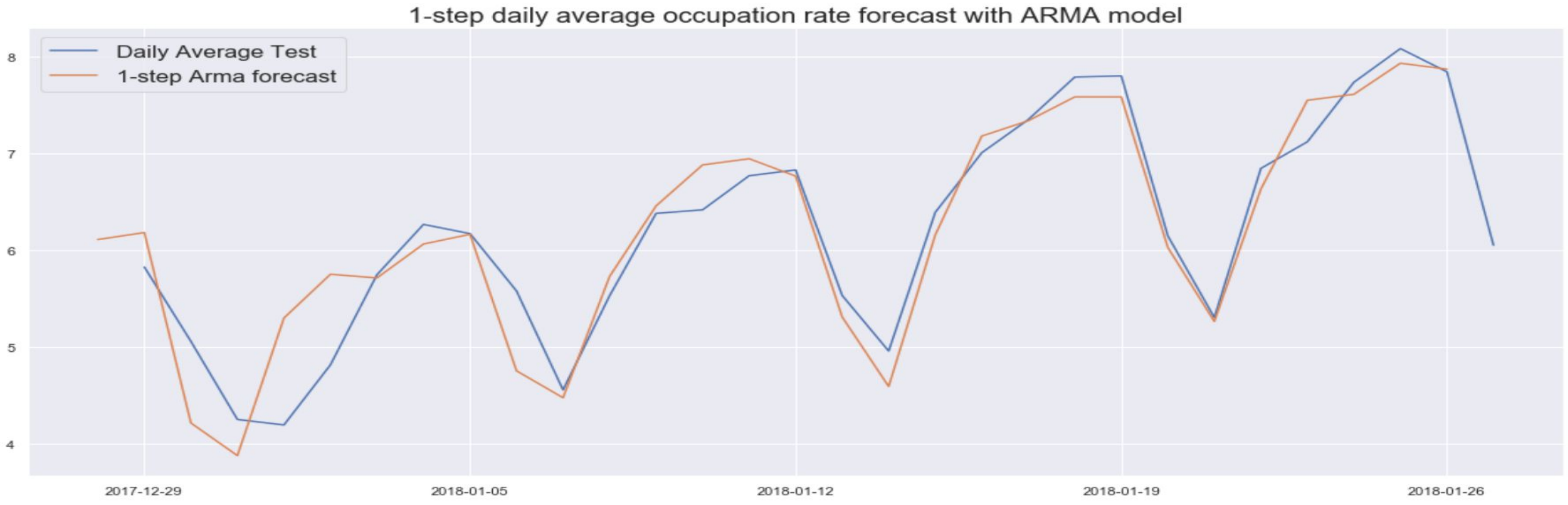
c/ Multi-Step forward hourly predictions at specific location (here, Boulevard des Invalides)

WEEKLY AVERAGE TRAFFIC PREDICTION OVER ONE YEAR WITH SIMPLE AVERAGING



- Historical data was insufficient to capture subtle dependencies : I simply used the average of the two preceding years to predict weekly average traffic rates in 2018
- Result is still quite satisfactory

DAILY AVERAGE PREDICTION WITH WEEKLY SEASONAL SARIMA

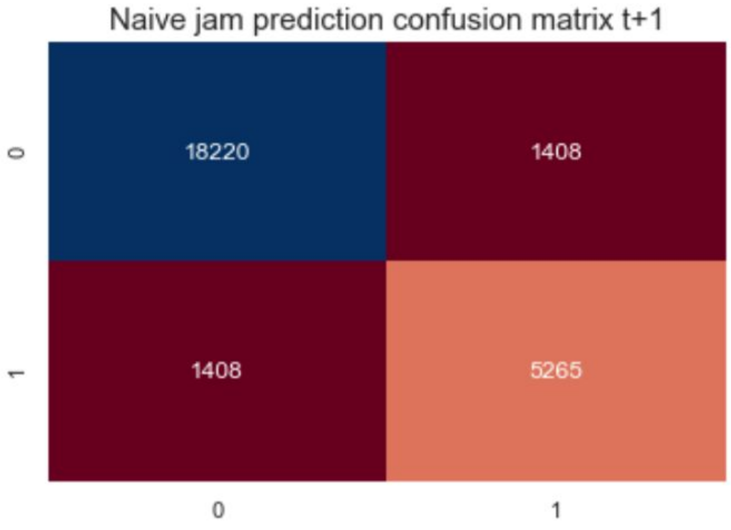


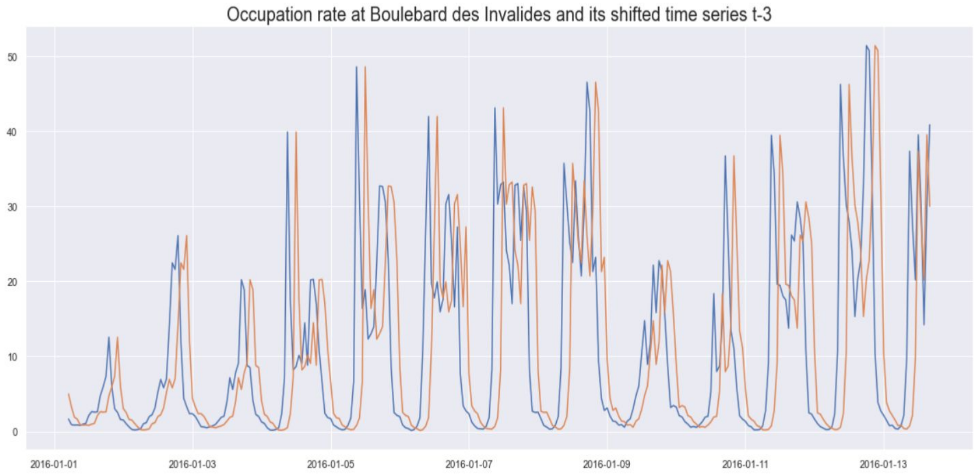

```
MAE for one step forward SARIMA on weekly averages :  0.5837899172164931
```

Next step : using exogenous SARIMA with weekday and holiday information

Baseline : persistent forecasting

$K(t+n)) = K(t)$



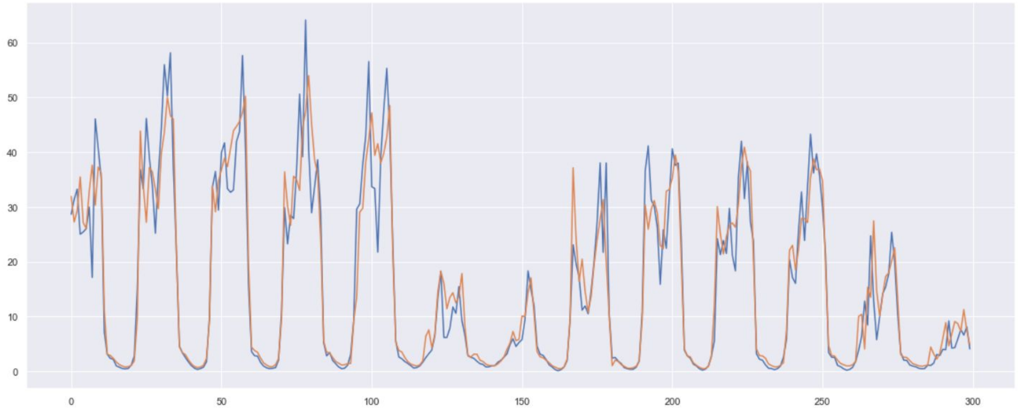


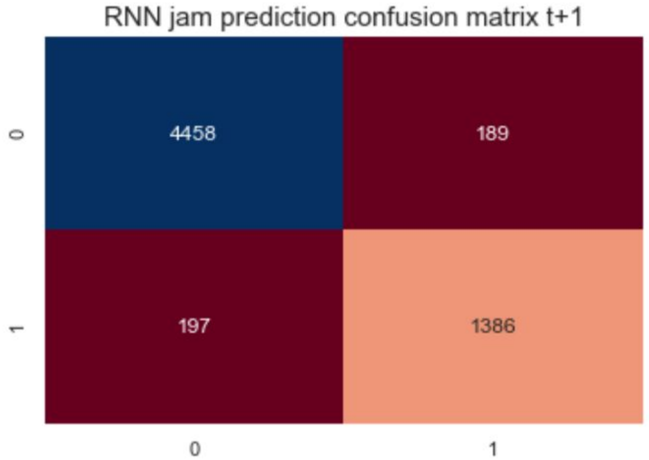

```
Accuracy: 0.8929318276871602
F1 score: 0.7890004495729057
Recall: 0.7890004495729057
Precision: 0.7890004495729057
```

```
absolute mean error for naive forecasting :  5.287193058819057
```

Best model : Multi-Step forward forecasting with LSTM

Number of units : 120
Feed in time series length : 336 (2 weeks)






```
Accuracy: 0.9380417335473515
F1 score: 0.8777707409753008
Recall: 0.8755527479469362
Precision: 0.88
```

```
MAE for 1-hour into the future LSTM  3.15375413695095
```

4/ CONCLUSION

-Successfully predicted daily traffic averages using SARIMA Model. Adding Exogenous variables like the day and holidays could certainly enhance performance.

-Was able to predict hourly traffic rates at specific locations with great precision using LSTM model.

-If we change last unit of LTSM to softmax, we can obtain a model that predict probability of Jam occuring, with great accuracy.