Faculté des Sciences et Ingénierie, Sorbonne Université

Auto Regressive Models on Transport Data

Author: Ahmed Tidiane BALDE
Supervisors: Vincent GUIGUE, Nicolas BASKIOTIS

June 4, 2019

Summary

1. Intro
2. Data Analysis and Preprocessing
3. Models
4. Metrics
5. Results
6. Model of Models
7. Conclusion

Introduction

Goal

Predicting accurately the affluence of stations.

To achieve so, we will inspect and use Auto Regressive Models and its variants like Auto Regressive Moving Average Models.

Used Tables

While there are many tables, we are only interested in a few of them. The ones which contain features on **validations count** :

station
id
network_code
transporter_code
station_code
name
x
y
pricing_zone
validations_count

validation
user_id
station_id
mode
operation_date
time
nature
trip_portion_id

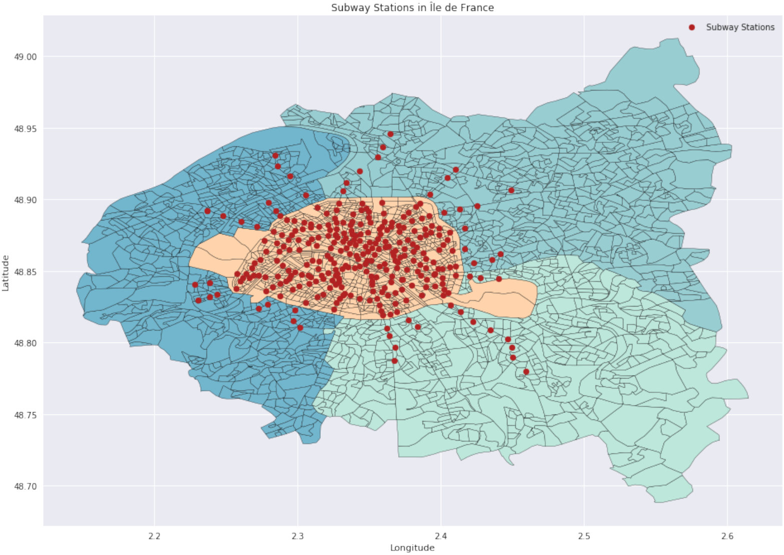
Below an exemple of rows in the *station* table.

id	nc	tc	sc	name	x	y	pricing_zone	val_count
2	12	112	266	Mairie	634777.498849	6860011.06668	4	554

As for *validation* table, here is an exemple :

user_id	station_id	mode	operation_date	time	nature	tp_id
8978714527637135004	53389	3	2015-10-01	19:46:12	031	2

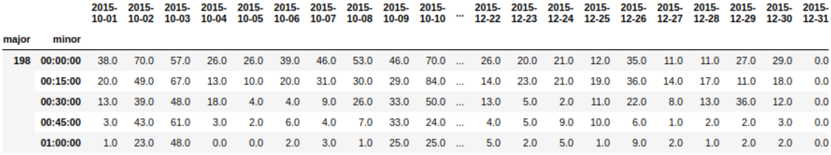
Station map



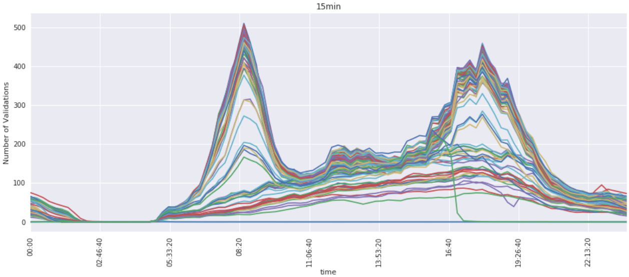
Subway Stations in Île de France
● Subway Stations
Latitude
49.00
48.95
48.90
48.85
48.80
48.75
48.70
Longitude
2.2
2.3
2.4
2.5
2.6

Discretization of validations table

By using validations table, we applied a *15* minutes discretization.

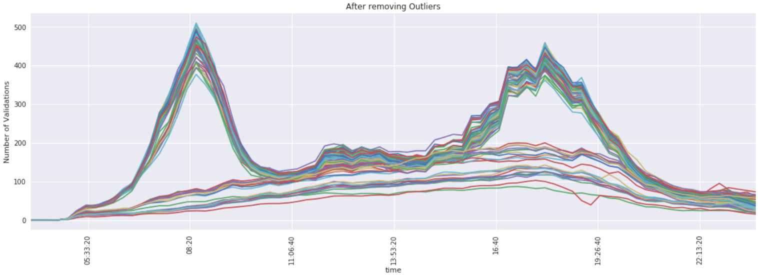


		2015-10-01	2015-10-02	2015-10-03	2015-10-04	2015-10-05	2015-10-06	2015-10-07	2015-10-08	2015-10-09	2015-10-10	...	2015-12-22	2015-12-23	2015-12-24	2015-12-25	2015-12-26	2015-12-27	2015-12-28	2015-12-29	2015-12-30	2015-12-31
major	**minor**																					
198	**00:00:00**	38.0	70.0	57.0	26.0	26.0	39.0	46.0	53.0	46.0	70.0	...	26.0	20.0	21.0	12.0	35.0	11.0	11.0	27.0	29.0	0.0
	00:15:00	20.0	49.0	67.0	13.0	10.0	20.0	31.0	30.0	29.0	84.0	...	14.0	23.0	21.0	19.0	36.0	14.0	17.0	11.0	18.0	0.0
	00:30:00	13.0	39.0	48.0	18.0	4.0	4.0	9.0	26.0	33.0	50.0	...	13.0	5.0	2.0	11.0	22.0	8.0	13.0	36.0	12.0	0.0
	00:45:00	3.0	43.0	61.0	3.0	2.0	6.0	4.0	7.0	33.0	24.0	...	4.0	5.0	9.0	10.0	6.0	1.0	2.0	2.0	3.0	0.0
	01:00:00	1.0	23.0	48.0	0.0	0.0	2.0	3.0	1.0	25.0	25.0	...	5.0	2.0	5.0	1.0	9.0	2.0	1.0	2.0	2.0	0.0



Preprocessed Data

Final preprocessed Data



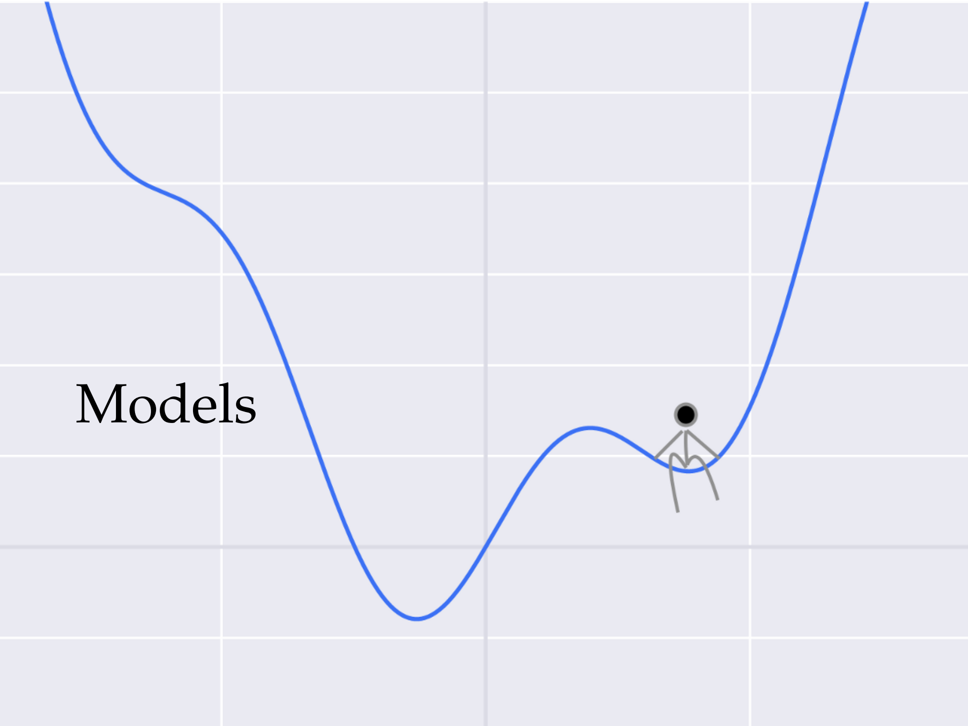
Normalization

Normalization between -1 and 1 according to each station

Data dimensions : **78 x 301 x 80**

Train : **57 x 301 x 80**

Test : **21 x 301 x 80**



Models

Auto Regressive model *(AR)*

$$X_t = c + \sum_{i=1}^{p} \theta_i X_{t-i} + \epsilon_t \tag{1}$$

where p is the auto-regressive parameter, θ_1, ..., θ_p are the model weights, c is a constant and ϵ_t the white noise,
wiki : *Autoregressive_model*.

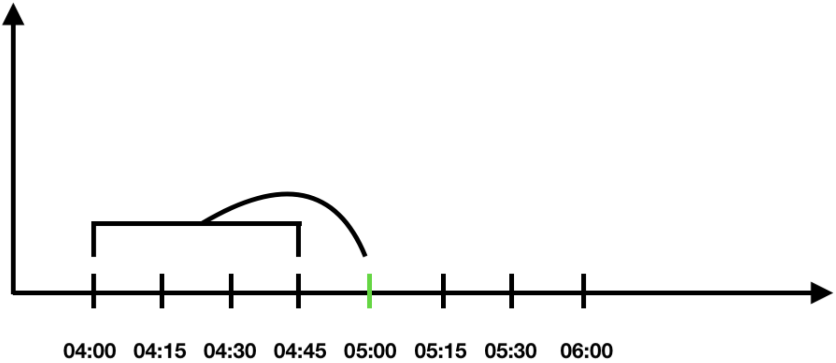
- Linear Regression
- XGBoost

Auto Regressive Moving Average *(ARMA)*

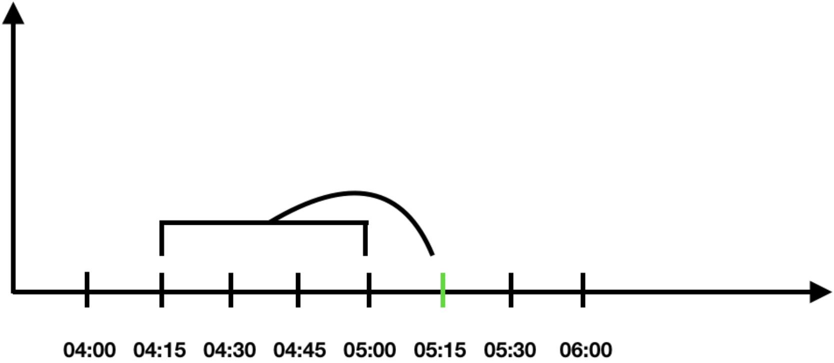
$$X_t = c + \sum_{i=1}^{p} \theta_i X_{t-i} + \epsilon_t \; + \sum_{i=1}^{q} \beta_i \epsilon_{t-i} \tag{2}$$

where p again is the auto-regressive parameter, q the *Moving Average(MA)* parameter, ϵ_1, ..., ϵ_q, are the errors,
wiki : *Autoregressive_moving_average_model*.

How does it work ?

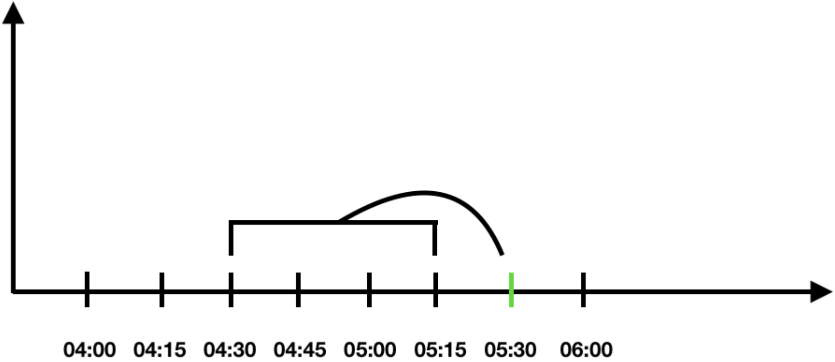


How does it work ?



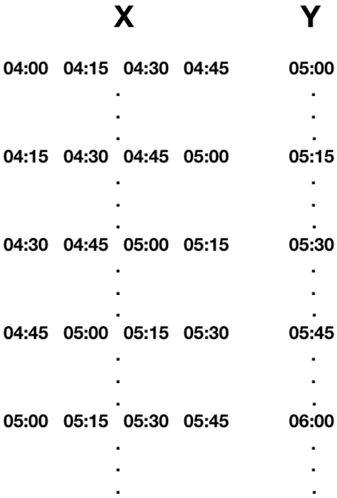
Forecasting
AR(4), forecasting at T+1

How does it work ?



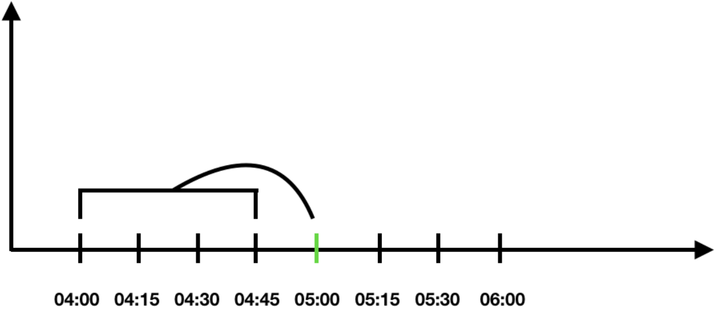
Forecasting
AR(4), forecasting at T+1

How does it work ?



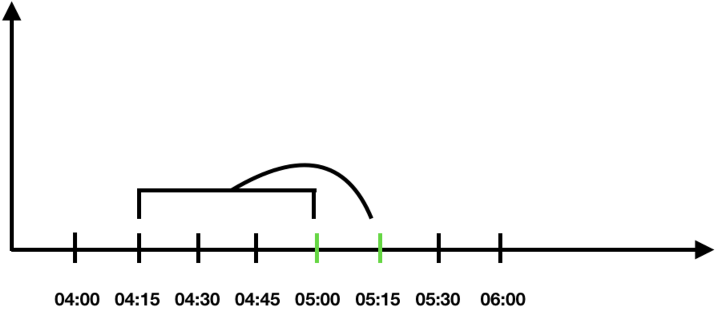
X				Y
04:00	04:15	04:30	04:45	05:00
		⋮		⋮
04:15	04:30	04:45	05:00	05:15
		⋮		⋮
04:30	04:45	05:00	05:15	05:30
		⋮		⋮
04:45	05:00	05:15	05:30	05:45
		⋮		⋮
05:00	05:15	05:30	05:45	06:00
		⋮		⋮

How does it work ?



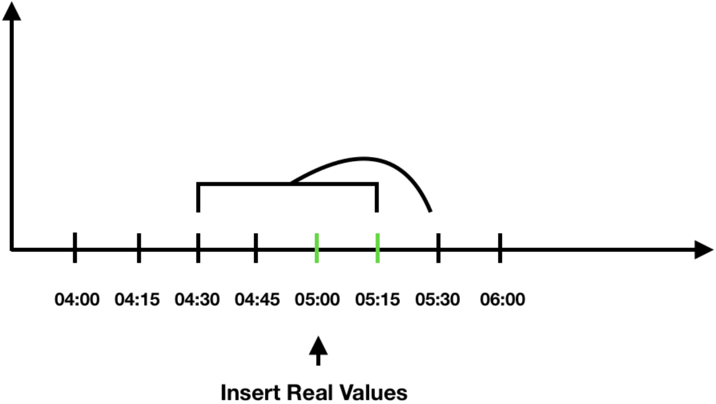
Forecasting
AR(4), forecasting at T+2

How does it work ?

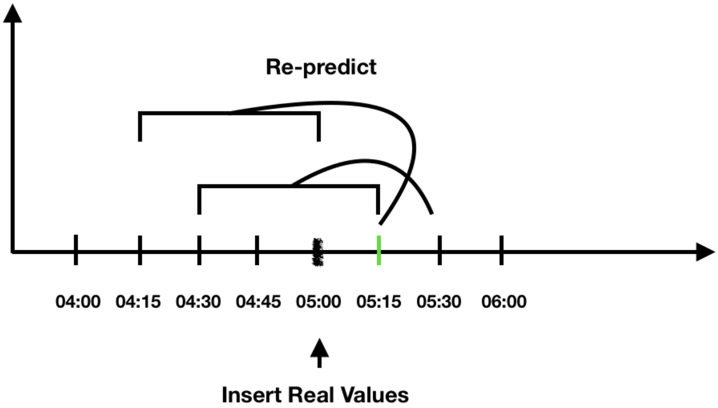


AR(4), forecasting at T+2

How does it work ?



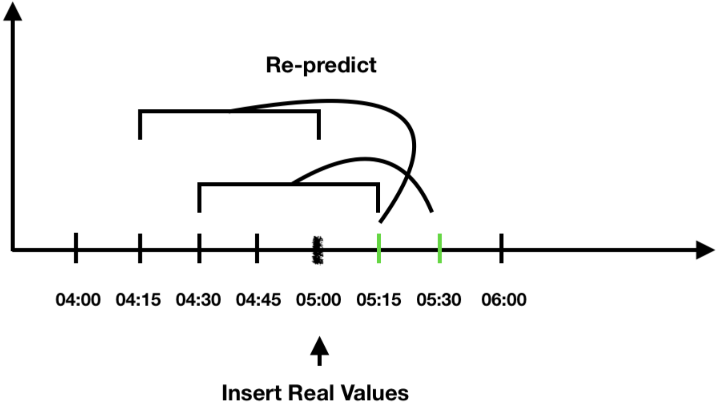
How does it work ?



Forecasting

AR(4), forecasting at T+2

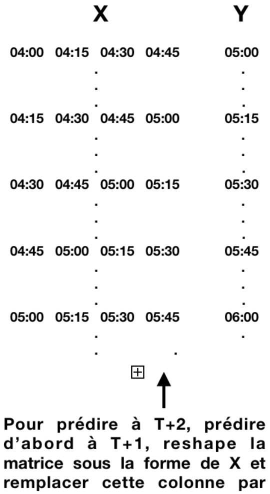
How does it work ?



Forecasting

AR(4), forecasting at T+2

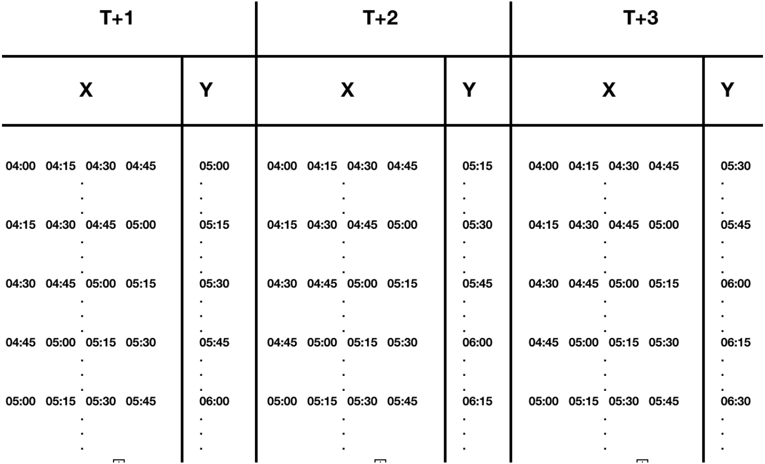
How does it work ?



X				Y
04:00	04:15	04:30	04:45	05:00
04:15	04:30	04:45	05:00	05:15
04:30	04:45	05:00	05:15	05:30
04:45	05:00	05:15	05:30	05:45
05:00	05:15	05:30	05:45	06:00

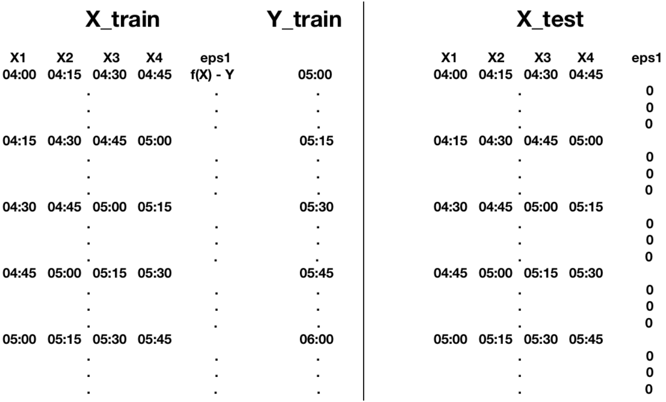
Pour prédire à T+2, prédire d'abord à T+1, reshape la matrice sous la forme de X et remplacer cette colonne par

Forecasting
AR(4), forecasting at T+2



T+1		T+2		T+3	
X	Y	X	Y	X	Y
04:00 04:15 04:30 04:45	05:00	04:00 04:15 04:30 04:45	05:15	04:00 04:15 04:30 04:45	05:30
⋮	⋮	⋮	⋮	⋮	⋮
04:15 04:30 04:45 05:00	05:15	04:15 04:30 04:45 05:00	05:30	04:15 04:30 04:45 05:00	05:45
⋮	⋮	⋮	⋮	⋮	⋮
04:30 04:45 05:00 05:15	05:30	04:30 04:45 05:00 05:15	05:45	04:30 04:45 05:00 05:15	06:00
⋮	⋮	⋮	⋮	⋮	⋮
04:45 05:00 05:15 05:30	05:45	04:45 05:00 05:15 05:30	06:00	04:45 05:00 05:15 05:30	06:15
⋮	⋮	⋮	⋮	⋮	⋮
05:00 05:15 05:30 05:45	06:00	05:00 05:15 05:30 05:45	06:15	05:00 05:15 05:30 05:45	06:30
⋮	⋮	⋮	⋮	⋮	⋮

ARMA(4), forecasting at T+2



X_train					Y_train
X1	X2	X3	X4	eps1	
04:00	04:15	04:30	04:45	f(X) - Y	05:00
		.		.	.
		.		.	.
		.		.	.
04:15	04:30	04:45	05:00		05:15
		.		.	.
		.		.	.
		.		.	.
04:30	04:45	05:00	05:15		05:30
		.		.	.
		.		.	.
		.		.	.
04:45	05:00	05:15	05:30		05:45
		.		.	.
		.		.	.
		.		.	.
05:00	05:15	05:30	05:45		06:00
		.		.	.
		.		.	.
		.		.	.

X_test				
X1	X2	X3	X4	eps1
04:00	04:15	04:30	04:45	
		.		0
		.		0
		.		0
04:15	04:30	04:45	05:00	
		.		0
		.		0
		.		0
04:30	04:45	05:00	05:15	
		.		0
		.		0
		.		0
04:45	05:00	05:15	05:30	
		.		0
		.		0
		.		0
05:00	05:15	05:30	05:45	
		.		0
		.		0
		.		0

Baseline

Different cases for each auto-regressive models :

- General Baseline
- Baseline per day
- Baseline per station
- Baseline per station per day

The goal is to do better than the baselines. Specially the last one which is quite strong.

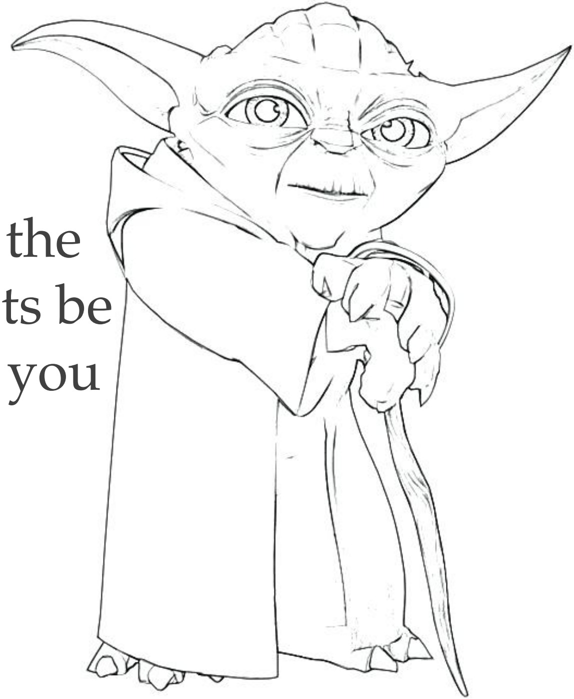
METRICS
INDICATOR GOALS
SALES REVIEW POTENTIAL OBJECTIVES
METRICS
DATA
IMPLEMENT
GOAL
GROWTH
KEY
GOAL
PLAN
BUSINESS STUDY VALUE
KEY
DASHBOARD
COMMUNICATION
INDICATOR
SALES
GOAL
IMPLEMENT
METRICS
TARGET
DASHBOARD SUCCESS
IMPROVEMENT
DATA
KEY
IMPROVEMENT
COMMUNICATION
GOAL
TARGET
PROGRESS
BUSINESS STUDY VALUE DATA
PLAN
IMPLEMENT
DATA
METRICS
DASHBOARD
BUSINESS STUDY
GROWTH
KEY DASHBOARD SUCCESS STRATEGIC SALES REVIEW POTENTIAL OBJECTIVES
MANAGE IMPROVEMENT REVIEW

In order to evaluate our models we used different metrics among which :

RMSE

$$RMSE = \sqrt{\frac{1}{N} \sum_{y \in Y \text{ and } \hat{y} \in Y_{pred}} (y - \hat{y})^2} \tag{3}$$

May the results be with you



General model

Without using the Baseline : AR, Xgboost

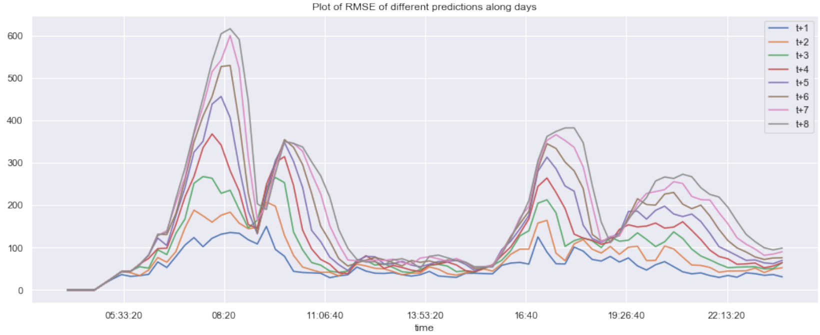
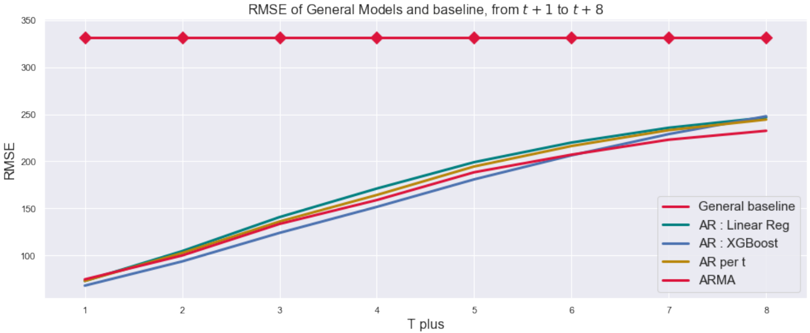


Figure: RMSE Score error for General Model with Baseline per station: XGBoost

General Model
Comparison



General model
With Baseline per station: AR, Xgboost

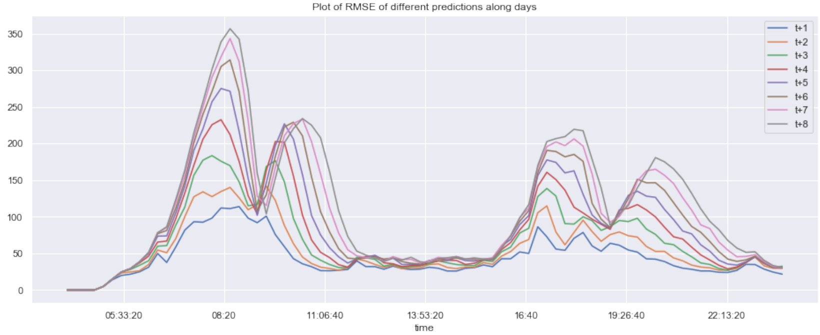
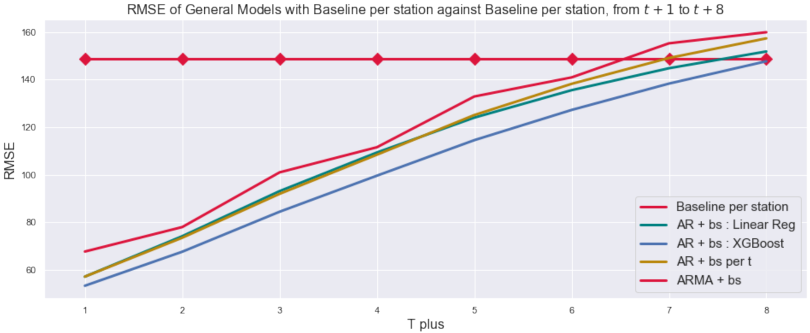


Figure: RMSE Score error for General Model with Baseline per station: XGBoost


RMSE of General Models with Baseline per station against Baseline per station, from $t+1$ to $t+8$
RMSE
160
140
120
100
80
60
1
2
3
4
5
6
7
8
T plus
Baseline per station
AR + bs : Linear Reg
AR + bs : XGBoost
AR + bs per t
ARMA + bs

General model

With Baseline per station per day: AR, Xgboost

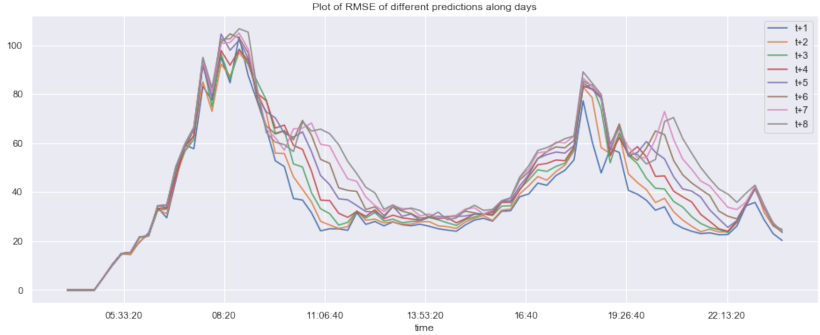
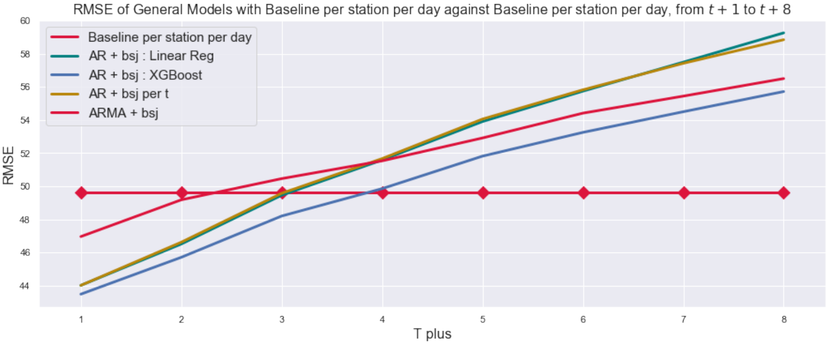


Figure: RMSE Score error for General Model with Baseline per station: XGBoost

General model

Comparison



Specific Models

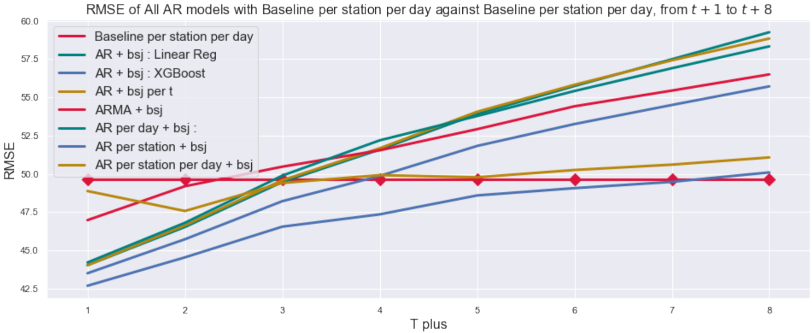
We then tried to test the AR models in a more accurate and specific way, for instance:

- One model per day
- One model per station
- One model per station and per day

We have seen that when we add the baseline per station and per day it works better so for each of those listed above, we did add the baseline ber station and per day.

One model per station per day

Comparison



Best model : AR per station
With Baseline per station per day

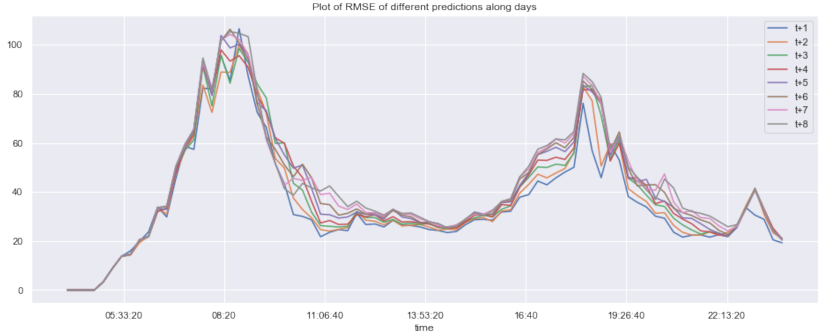


Figure: RMSE Score error for General Model with Baseline per station: XGBoost

Best model : AR per station
With Baseline per station per day

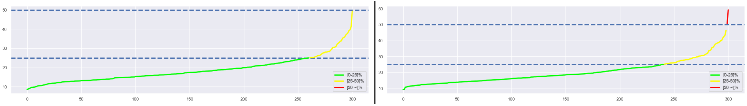


Figure: MAPE curves of stations

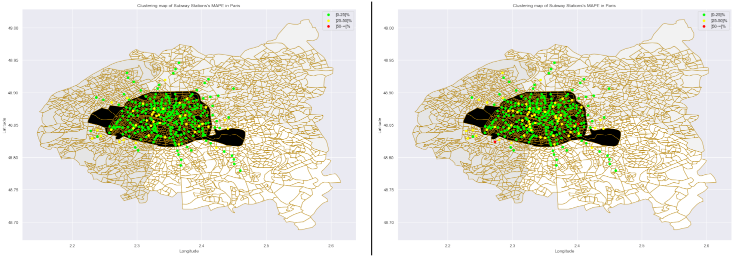
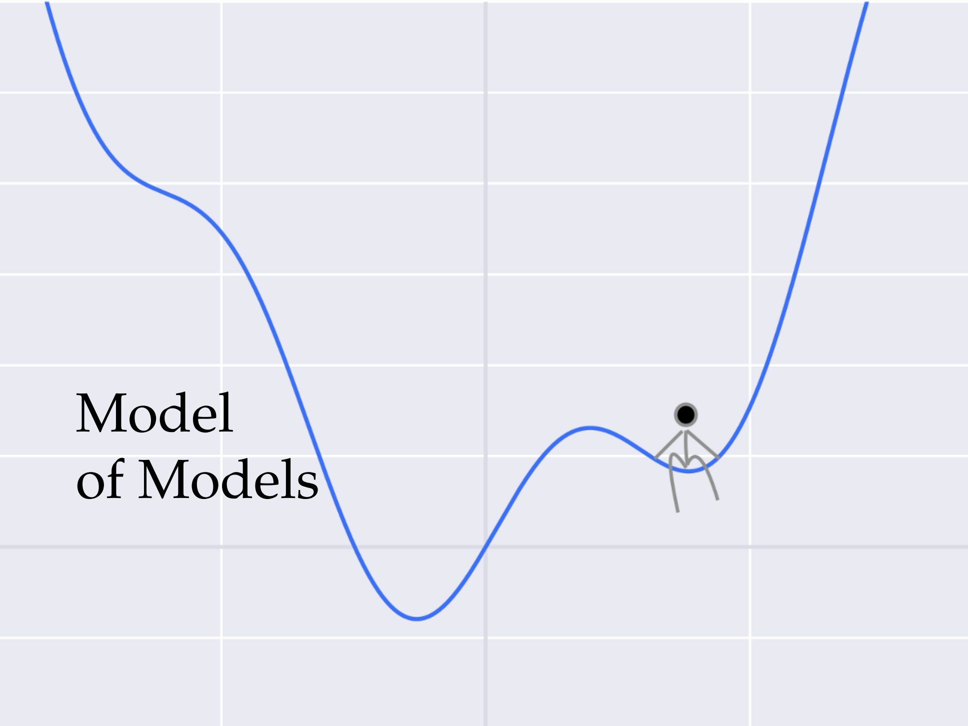
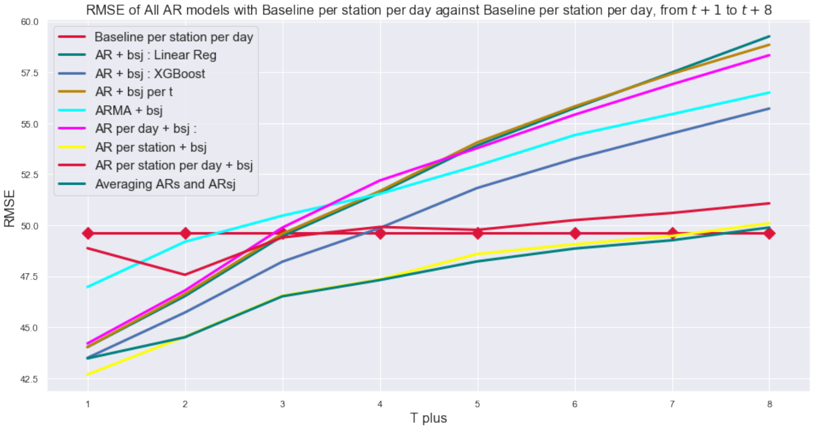


Figure: MAPE clusters map of stations

Model
of Models




RMSE of All AR models with Baseline per station per day against Baseline per station per day, from $t+1$ to $t+8$
Baseline per station per day
AR + bsj : Linear Reg
AR + bsj : XGBoost
AR + bsj per t
ARMA + bsj
AR per day + bsj :
AR per station + bsj
AR per station per day + bsj
Averaging ARs and ARsj
RMSE
60.0
57.5
55.0
52.5
50.0
47.5
45.0
42.5
T plus
1
2
3
4
5
6
7
8

Stacking

Model Stacking also called **Meta Ensembling** consists in combining information from mutiple predictive models (**base models**) to generate a new one.

Often times, **Stacking** provides better results.

Conclusion

Thank You !

Questions ?