**Data description :**

Files :

• the coordinates of the Air quality stations (to predict) (files) :

* Beijing AirQuality Stations (.xlsx -> convert to .csv with Excel) (id/name\_aq, longitude, latitude)
* London AirQuality Stations (id, longitude, latitude)

• the coordinates of the grid weather points (files) :

* Beijing grid weather station (name\_grid, latitude, longitude)
* London grid weather station (station name, longitude, latitude)

• beijing 17 18 meo.csv (id/name\_meo, time (1/2), longitude, latitude, external parameters)

• beijing 17 18 aq.csv (id/name\_aq, time (1/2), PM SO…)

• beijing 201802 201803 aq.csv (id/name\_aq, time (2/2), PM SO…)

• beijing 201802 201803 me.csv (id/name\_meo, time (2/2), external parameters )

• Beijing historical meo grid.csv (name\_grid, longitude, latitude, time parameters(-weather))

• London historical aqi forecast stations.csv (PM measurements for each station id (*CD1, LW2,…),* time)

• London historical aqi other stations.csv (PI for each station id, for extra stations, not to predict)

• London historical meo grid.csv (station\_Name (*London\_grid\_XYZ*), longitude, latitude, external parameters, time)

*External parameters = temperature,pressure,humidity,wind\_direction,wind\_speed,weather(for Beijing)*

Import .csv in python :

Import csv

csv.reader(csvfile, dialect='excel', \*\*fmtparams)

**Before starting your analysis you have to clean/prepare your data (merge diﬀerent ﬁles, check/ handle missing variables, etc.) !!! :**

Merge datas :

Separate datas for Beijing and London.

**London :**

1. Use ” London\_historical\_aqi\_forecast\_stations.csv” in order to predict “London AirQuality Stations”.
2. Then 2 options : ( to test !! which one gives better result on validation test)

* Also add ” London\_historical\_aqi\_forecast\_stations.csv” : we have more training and tests datas. -> London\_historical\_aqi\_all\_stations.csv
* Do not add the datas, not linked to the stations we want to predict : could bring wrong values.

1. Add external parameters to predict :

London AirQuality Stations -> (longitude, latitude) -> London historical meo grid.csv

SOUCIS : pas mêmes arrondis (arrondir les London AirQuality Stations ?) (ils ont vraiment foutu des datas de merde ces cons) (même pour fusionner les colonnes sont pas dans le bon sens)

* Merge in one file : last one but with the id of stations to predict

**Beijing :**

*// TODO*

Missing variables solution :

* Interpolation( avec les valeurs autour, mais il y a des « sauts » entre les valeurs d’une donnée à la suivante)
* Delete (l’assistante a dit moyen mais Je vois pas trop d’autre solution, on en discute !)

Training and validation :

the data until 20 of March 2018

Split the data set properly in training and validation sets taking into account that the order matters.

Tests :

the 21th and 22th of March