MATHÉMATIQUES
VISION
APPRENTISSAGE

Challenge: Prediction of spatiotemporal PM^{10} concentration

OREISTEIN Pierre, ZHAO Tong

April 3, 2019

Outline

Introduction

- **Main Task**:
  Prediction of the PM^{10} reading at one station given its urban features and the PM^{10} readings of its 10 nearest stations

- **Why?**
  - Relieve the lack of climate monitoring stations all over the world
  - Analyse important factors which influence more the air pollution

Related Methods

• The distribution of air pollution involves a physico-chemical complex process depending on a number of factors. Both the choice of features and the choice of models are controversial.

• **Two mainstream methods**:

- Deterministic model [VMD14] [RBM+12]
- Statistic model [CWY+12]

• Machine Learning methods, especially Deep learning methods [FLZ+15] [DR17] [DGS+19]

Dataset

We have:

- 695255 readings from 85 stations in the training dataset.
- 247473 readings from 31 stations in the test dataset.

	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standard	Standa
1	ID	station_id	hdres_100	hdres_500	ldres_100	ldres_500	industry_100	industry_500	urbgreen_100	urbgreen_500	roads
2	0	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
3	1	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
4	2	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
5	3	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
6	4	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
7	5	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
8	6	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
9	7	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
10	8	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
11	9	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
12	10	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712
13	11	105	0.000	0.000	1.000	0.929	0.000	0.033	0.000	0.022	0.712

Features Available: Static Features

- **Land Use Features**:
  Surrounding environment of a station. Counts of:
  - high-density residential area
  - low-density residential area
  - industrial area
  - green area

  at 100 meters and 500 meters

- **Road Features**:
  Length of the roads and major roads around the station to 25 meters, 100 meters and 500 meters.

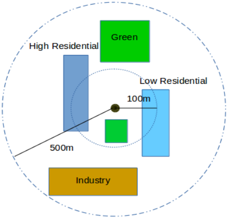


Figure: Land-used Features

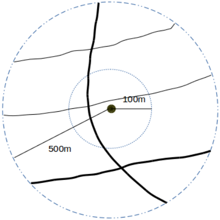


Figure: Road Features

Features Available: Dynamic Features & Missing Values

- **Nearby Readings**:
  Current readings of the 10 nearest stations and the corresponding distances.

  $\Longrightarrow$ Real Time information

  $\Longrightarrow$ In average 5% of the data are missing.

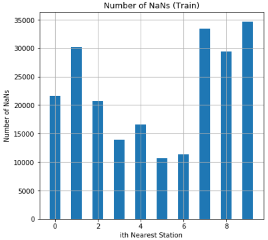


Figure: NaNs in Train Set

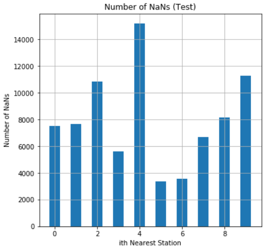


Figure: NaNs in Test Set

Data Analysis: Extrem Values = Outliers ?

Distribution of PM^{10} readings

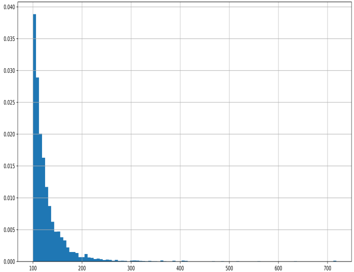


Figure: Distribution of the PM 10 over all stations in the training set

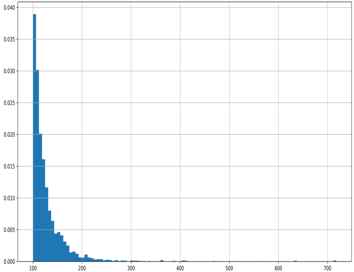


Figure: Distribution of the PM 10 over all stations in the testing set

- The pollution is concentrated over a small range of values
- Few extreme values: up to 750. 0.46% (and 0.15%) superior to 100 in the training set (in the testing set respectively)

Data Analysis: Outliers ?

True Value	Value_0	Value_1	Value_2	Value_3	Value_4	Value_5	Value_6	Value_7	Value_8	Value_9
718.00	16.0	3.00	8.00	28.00	8.0	21.30	20.70	23.60	32.50	30.50
717.00	11.0	5.00	10.00	26.00	10.0	22.10	25.50	20.80	29.90	30.60
635.60	26.3	10.20	12.17	16.20	7.4	24.60	25.93	15.70	21.70	22.70
562.00	72.0	19.37	58.00	34.00	79.0	64.00	41.00	49.00	10.00	9.00
516.00	13.0	19.00	26.00	27.00	10.0	34.00	22.00	23.00	13.00	14.00
487.00	71.0	19.37	52.00	24.00	73.0	44.00	38.00	49.00	13.00	7.00
470.00	70.0	57.00	49.00	56.00	62.0	112.20	65.00	62.00	38.00	50.00
437.70	15.4	16.90	13.46	17.16	18.2	23.90	19.70	31.54	10.50	12.40
409.22	24.0	21.00	37.00	27.00	37.0	42.00	22.00	25.00	15.00	39.00
408.00	388.0	273.00	123.00	75.00	178.0	152.83	130.00	112.00	28.01	31.27

Data Analysis: Static Features

Static Features

Estimation of the mean PM^{10} readings for each station using all training data. Then computation of the Pearson correlation between each features and the mean reading.

Feature	Correlation	Feature	Correlation	Feature	Correlation
hdres_100	0.262098	ldres_100	-0.106761	industry_100	-0.067775
hdres_500	0.270868	ldres_500	-0.115931	industry_500	-0.042046
urbgreen_100	-0.035763	roads_length_25	0.132217	major_roads_length_25	0.175380
urbgreen_500	-0.078163	roads_length_100	0.119255	major_roads_length_100	0.133029
-	-	roads_length_500	0.166722	major_roads_length_500	0.217183

$\Longrightarrow$ High residential areas and major roads play an important role.

Data Analysis: Dynamic Features

Dynamic Features

Computation of the Pearson correlation between each pair of current reading at the center station and the one at a nearby station.

Feature	Correlation	Feature	Correlation
value_0	0.705016	value_5	0.656450
value_1	0.705132	value_6	0.559369
value_2	0.694476	value_7	0.558495
value_3	0.705147	value_8	0.506016
value_4	0.722658	value_9	0.504334

$\Longrightarrow$ All of them have high correlation but there is much likely high correlation between the features itself.

Metric

The metric is the mean squared logarithmic error (MSLE).

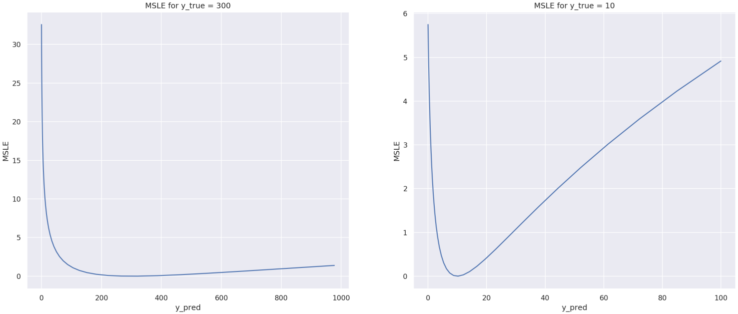
$$\text{MSLE}(I, T) = \frac{1}{N}\sum_{i=1}^{N}\Big(\log(T_i + 1) - \log(I_i + 1)\Big)^2$$

where:

- T is the target series
- I is the predict series

Metric Effect

$\Longrightarrow$ The choice of the log makes sense for avoiding a too important penalty for the extreme values.



Data Filling

The missing values correspond of cases where sensors did not transmit any messages.

- **0-filling**: NaN values are replaced by 0.
- **Benchmark-based Filling**: NaN values are replaced by the Benchmark

$$\hat{y} = \frac{\sum_{i=0}^{9} v_i/d_i}{\sum_{i=0}^{9} 1/d_i}$$

- **Interpolation-based Filling**: It assumes that the data are ordered in time and interpolate the missing value with the previous and next records.

$$v_i(t) = \frac{v_i(t^+) - v_i(t^-)}{t^+ - t^-} * \frac{t}{2} + v_i(t^-)$$

Features Expansion

There is a limited number of features. To enhance the embedding of the data, it added some combinations of the original features.

- **Benchmark**: add the benchmark as a guideline for the method of regression.
- **Inverse of the distance**: the inverse of the distance play a role of normalisation of the value of the nearby stations.
- **Clipping**: As only few readings are higher than 100, it clips the values between 0 and 100 to avoid a behaviour of outliers.

Model Requirements

Several factors need to be taken into account when it chooses the models.

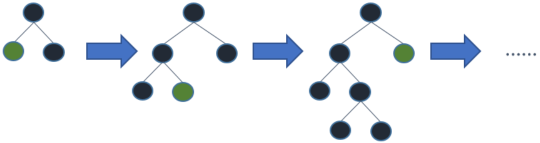
- **Model expressivity**: Ability of the model to understand the complex relationships among the three types of features.
  $\Longrightarrow$ Linear Regression are not sufficient.
- **Scalability**: The training set is composed of 700000 readings.
  $\Longrightarrow$ Model like SVM will not scale well.
- **Efficiency**: Two submissions per day at maximum. It needs to perform a cross-validation. Thus the model should be trained during a short period of time.

Given all the constraints, we chose two models:

- Light GBM
- Neural Network,

Light GBM

• Light GBM is a gradient boosting framework that uses tree based learning algorithm.



Leaf-wise tree growth

Advantages:

- **High speed and low memory**: Use histogram-based method.
- **Allow feature diversity**: No assumptions on the relationships of features.
- **Stability of results**: No sensitivity to the initialization thus good reproducibility.

Disadvantages:

- **Overfitting issue**: The first 24 features of each station remain the same for all its readings, which may cause a problem.
- **Inflexible evaluation metric**: Complicated to customize the metric. Using of the RMSE as a metric.

Results for Gradient Boosting Tree (1)

First Model: Naive Regressor:
A single regression model.

Results of the cross-validation:

Fold	1	2	3	4	5
Train MSLE	0.1120	0.1054	0.1312	0.0934	0.1002
Val MSLE	0.1760	0.1342	0.1798	0.1521	0.1111

Best parameters:

Parameter	**Choice**
max depth	5
number leaves	25
learning rate	0.01
feature fraction	0.9
bagging fraction	0.8
bagging freq	5

Then a model is trained on the whole training set:

- MSLE on training set is 0.1023.
- It achieves a score of 0.1391 on the public leaderboard.

Results for Gradient Boost Tree (1)

Visualization of the feature importance.

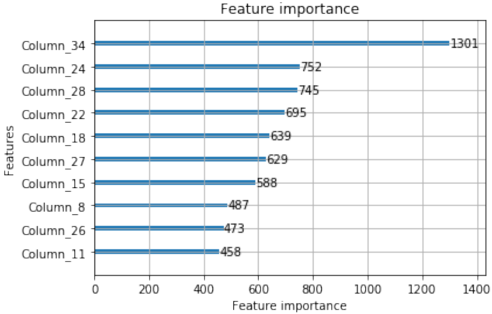


Figure: Feature Importance (Single Model + Benchmark Filling)

- Column 34: Benchmark
- Column 24: value_0
- Column 28: value_3
- Column 22: distance_8
- Column 27: value_2
- Column 15: distance_1
- Column 8: roads_length_25
- Column 26: value_2
- Column 11: major_roads_length_25

Results for Gradient Boosting Tree (2)

Second Model: Multi-regressors with K-means Clusters:
Use K-means to cluster all the stations having similar static features to one group. Then a tree model is trained for each group, respectively.

The final model achieves:

- MSLE on training set is 0.1058.
- It achieves a score of 0.1538 on the public leaderboard.

Best parameters:

Parameter	Choice
max depth	3
number leaves	18
learning rate	0.01
feature fraction	0.9
bagging fraction	0.8
bagging freq	5

Why?

- Each model has less data than before
- Cross-validation becomes unreliable
- In each group, the static features provide few information

Neural Networks

• Some promising works has recently been released [FLZ+15] using neural networks.

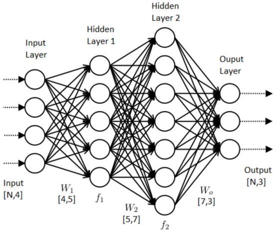


Figure: Example of fully connected neural networks with two hidden layers

Advantages:

- **Auto-embedding**: Learns an embedding.
- **Scalability**: Existence of efficient library.

Disadvantages:

- **Overfitting issue**: Neural networks can easily over-fit.
- **Stability**: Many hyper-parameters that influence a lot the over-fitting.

Results for the Neural Network

• **Results without grid search**: The choice of the hyper-parameters was quite empirical:

Parameter	**Choice**
fill NaN	Interpolated-based Filling
features added	Benchmark, Inverse Distances
Dropout for each layer	0, 0, 0
Batch Size	64
Epochs	10
Number of neurons	16
Batch Normalisation	False

This simple network achieved

- A score of 0.1322 on the public leader-board.
- A score of 0.1341 on the intermediate academic ranking.

$\Longrightarrow$ Thanks to this simple network we were ranked 1st until today (3 April 2019).

• **Results with Grid Search**: No meaningful results.

Analysis of the predictions

- **Top 10 Samples with the biggest errors of predictions**

Error of Prediction	True value of PM^{10}	Predicted value of PM^{10}
15.484452	717.0	13.033408
15.448437	718.0	13.117446
13.444888	635.6	15.271488
11.034476	361.0	12.063552
10.689042	437.7	15.683257
10.219151	283.0	10.614259
10.070142	361.0	14.154469
9.749145	361.6	14.973613
8.909923	516.0	25.130247
8.857376	307.0	14.704784

- Predictions far from what expected. $\Longrightarrow$ Link with the choice of the metric.
- No strong correlations with the value of the others stations when the air pollution is high (> 100).

Final Ranking

Ranking	Date	User(s)	Public score
1	March 21, 2019, 11:47 p.m.	pierreO & zt	0.1322
2	March 14, 2019, 2:04 p.m.	antoine	0.1339
3	March 31, 2019, 7:56 p.m.	cbilli44 & Mathieu78 & aurelien	0.1346
4	April 1, 2019, 8:30 p.m.	mathieuB & ThomBouchMVA	0.1365

Figure: Public Ranking Board (3rd April)

Ranking	Date	User(s)	Final score (date 2019-03-22)
1	March 21, 2019, 11:47 p.m.	pierreO & zt	0.1341
2	March 2, 2019, 2:32 a.m.	mhajabri	0.1431
3	March 2, 2019, 2:26 p.m.	bsreda & mhajabri	0.1434
4	March 14, 2019, 5:01 p.m.	Rom1P	0.1605

Figure: Intermediate Ranking Board (3rd April)

Conclusion

- **Motivation**: Interest for the challenge because its purpose and difficulty
- **Contribution**: A new method relying on other station to predict the air pollution.
- **Model Developed**: Neural Networks seem a promising field of research for further development.
- **Perspectives**: New investigations could be done:
  - Creating a mixed model with meteorological features
  - Managing high values: outliers ? Create another models for high values ? Using local meteorological features ? Change of metrics ?

References I

Chu-Chih Chen, Chang-Fu Wu, Hwa-Lung Yu, Chang-Chuan Chan, and Tsun-Jen Cheng.
Spatiotemporal modeling with temporal-invariant variogram subgroups to estimate fine particulate matter pm2. 5 concentrations.
Atmospheric environment, 54:1–8, 2012.

Mahmoud Reza Delavar, Amin Gholami, Gholam Reza Shiran, Yousef Rashidi, Gholam Reza Nakhaeizadeh, Kurt Fedra, and Smaeil Hatefi Afshar.
A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of tehran.
ISPRS International Journal of Geo-Information, 8(2):99, 2019.

Nadjet Djebbri and Mounira Rouainia.
Artificial neural networks based air pollution monitoring in industrial sites.
In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–5. IEEE, 2017.

References II

Xiao Feng, Qi Li, Yajie Zhu, Junxiong Hou, Lingyan Jin, and Jingjie Wang.
Artificial neural networks forecasting of pm2.5 pollution using air mass trajectory based geographic model and wavelet transformation.
Atmospheric Environment, 107:118–128, 2015.

Laura Ranzato, Alberto Barausse, Alice Mantovani, Alberto Pittarello, Maurizio Benzo, and Luca Palmeri.
A comparison of methods for the assessment of odor impacts on air quality: Field inspection (vdi 3940) and the air dispersion model calpuff.
Atmospheric environment, 61:570–579, 2012.

Laura E Venegas, Nicolás A Mazzeo, and Mariana C Dezzutti.
A simple model for calculating air pollution within street canyons.
Atmospheric environment, 87:77–86, 2014.