# Rapport de projet, Sparse wavelet

Louis Thiry, Alexandre Saint-Dizier

23 février 2017

## Introduction

Pendant la présentation des projets par les entreprises, nous avons été séduits par le projet de prédiction de la qualité de l'air à l'échelle de la rue de Plume Labs. En effet, le problème en question est un problème de physique, et l'on peut espérer s'inspirer de son intuition pour proposer des solutions. De plus, l'application, la prédiction de la pollution à l'echelle de la rue, est intéressante et utile.

Cependant, quand nous avons commencé à étudier les données, nous nous sommmes rendus compte qu'un bon nombre de paramètres relevant de la physique du problème étaient exclus dans les données fournies, ce qui empêchait d'avoir une approche physique et qui remettait sérieusement en cause l'ambition affichée : prédire la qualité de l'air à l'échelle de la rue. Nous avons contacté Plume Labs pour en savoir plus, et ils nous ont répondu que cela était volontaire, le but réel du projet étant de "prédire la pollution à des points précis à partir d'informations statiques sur ces points et d'informations structurelles extraites sur les séries temporelles".

C'est donc à ce problème que nous nous sommes attaqués, problème très intéressant, mais aussi très général pour lequel on voit mal comment on peut obtenir des prédictions précises.

## 1 Les données à notre disposition

Les données du problème concernent les contentrations de trois types de polluants ($NO_2$, $PM_{10}$, $PM_{2,5}$) sur une période de temps T donnée dans six villes inconnues, numérotées de 0 à 5. Chaque ville comporte 4 ou 5 stations (29 au total sur les six villes, numérotées de 1 à 29).

### 1.1 Variables explicatives

Pour chacune des stations, nous avons les variables explicatives suivantes :

— 18 variables statiques, qui nous renseignent sur l'entourage du point où les mesures sont faites :
  — la surface cumulée de zones résidentielles à faible densité dans un rayon de 25/50/150/250/500 mètres autour du point
  — la surface cumulée de zones résidentielles à haute densité dans un rayon de 25/50/150/250/500 mètres autour du point
  — la surface cumulée de zones industrielles dans un rayon de 500 mètres autour du point
  — la surface cumulée de zones portuaires dans un rayon de 2500 mètres autour du point
  — la surface cumulée d'espaces verts dans un rayon de 2500 mètres autour du point
  — la distance cumulée de routes dans un rayon de 50/150/250/500 mètres autour du point
  — l'inverse de la distance à la route la plus proche
— 9 variables dynamiques, dont on a les valeurs sur pratiquement toute la periode T, à intervalles de temps d'une heure :
  — La température
  — La vitesse du vent
  — l'orientation du vent, via la valeur du cosinus et du sinus de l'angle par rapport à une référence inconnue
  — l'ennuagement
  — l'intensité des précipisations
  — la probabilité de précipitations
  — la pression
  — une variable booléenne qui indique si le jour est calme ou non

### 1.2 Données d'entrainement

Comme données d'entrainement, nous avons les concentrations en polluants à intervalle de temps d'une heure sur toute la période pour 2 ou 3 stations par ville (17 sur 29 au total). Pour les 12 autre stations (2 par ville), nous n'avons **aucune** valeur de concentration en polluants.

### 1.3 Problème posé

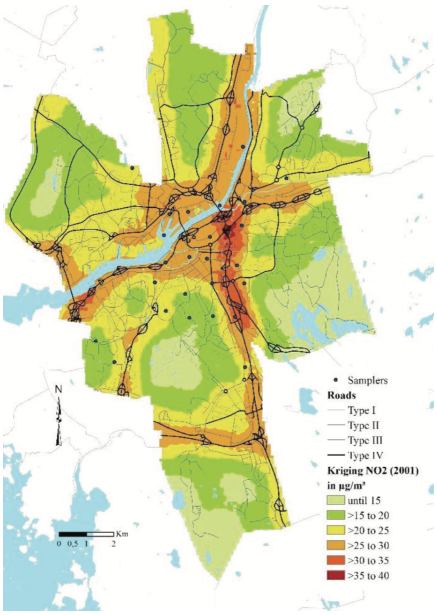
Le but est de prédire la concentration en polluant sur ces 12 stations sur sur tout la période T donnée.

Le problème est donc un problème de régression très général : étant données les variables explicatives et l'évolution temporelle des concentration en polluants à certain points sur une période T, prédire l'évolution temporelle des concentration en d'autre points à partir des données météorologiques et statiques. Pour ce faire, nous n'avons que très peu d'exemples (2 ou 3 par ville), bien que la période T soit longue. La tâche s'annonce donc compliquée.

### 1.4 Remarques sur les données

Quelques remarques sur les données :

— Nous n'avons aucune donnée géographique qui nous renseigne sur la position relative des différents points les uns par rapport aux autre, ni sur la position relative des éléments (par exemples les routes) qui sont dans l'entourage du point auquel est fait la mesure. Nous ne pouvons donc pas espérer avoir de résultats précis. Prenons par exemple la carte de pollution de la ville Gothenburg qui est prise en exemple dans l'article [] :




On constate que la pollution est phénomène spatial, et on peut sans aucun doute trouver des points qui auront les même valeurs statiques (surfaces cumulées) et pour lesquels le niveau de pollution est pourtant très différent. Notons au passage que les routes sont classés selon différents types qui semblent être déterminants pour la pollution en $NO_2$, or nous n'avons qu'un type de route à notre disposition

— A une heure donnée, les données météorologique (température, vent, pression, ennuagement) sont les mêmes pour toutes les stations d'une zone. Ce sont donc uniquement les données statiques qui différencient les stations les unes des autres.

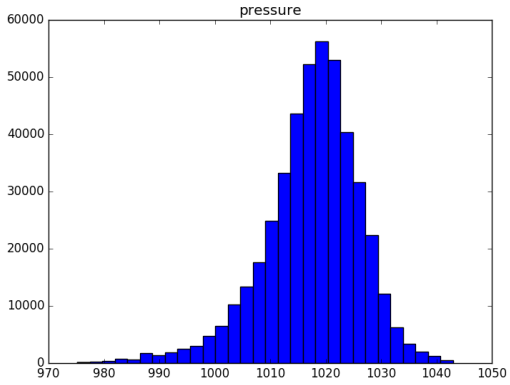
— L'angle d'orientation du vent nous est donnée par son cosinus et son sinus, dont la somme des carrés ne font pas 1 en général (moyenne = 0.995, écart type = 0.06 , maximum = 1.995 , minimum = $2.4 \cdot 10^{-11}$)

— Les données statiques nous sytématiquement répétées pour chaque valeur temporelle d'un polluant. Nous avons donc fait un script pour vérifier que ce sont effectivement toujours les mêmes pour une station donnée.

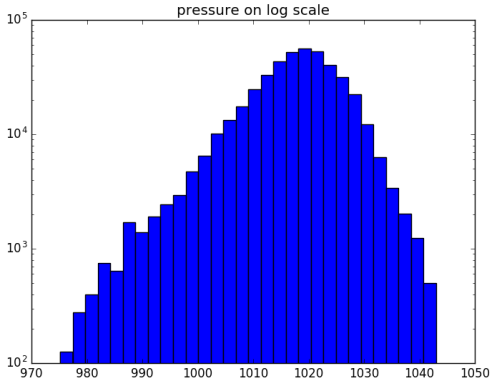
## 1.5 Exploration des valeurs

### 1.5.1 Les données dynamiques

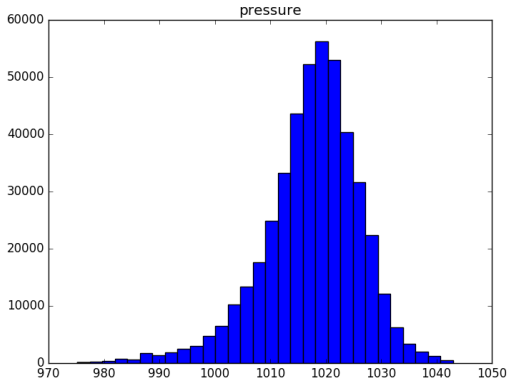
Pour se rendre compte de la répartition des valeurs des données dynamiques, on affiche les histogramme de ces valeurs sur une echelle logarithmique :



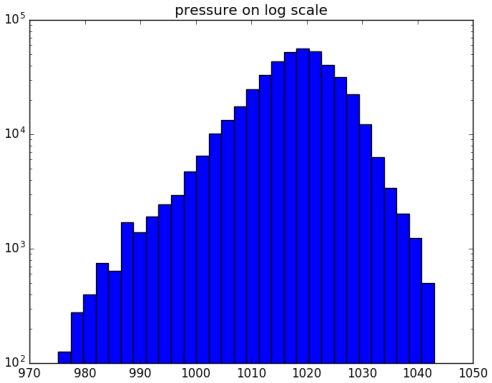

pression



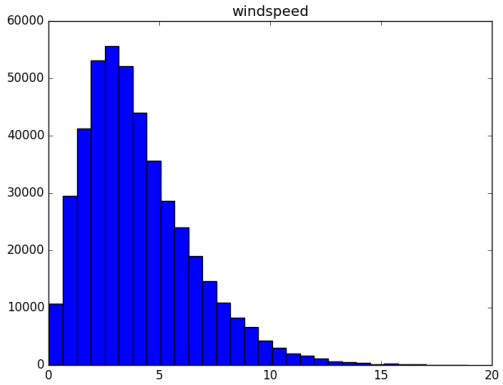

pression sur échelle log



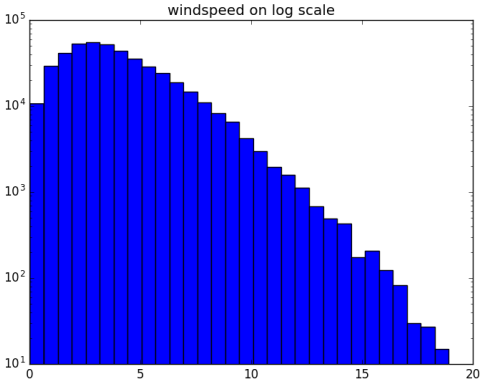

pression



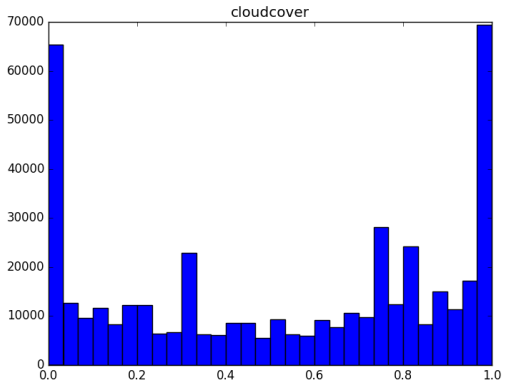

pression sur échelle log



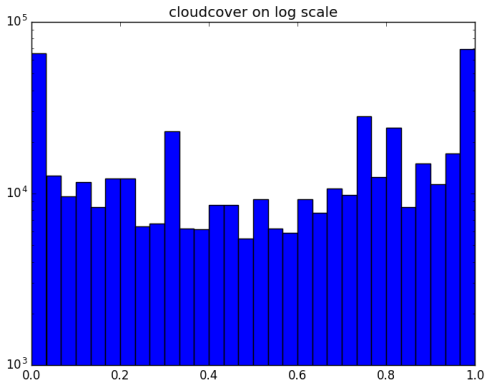

vent



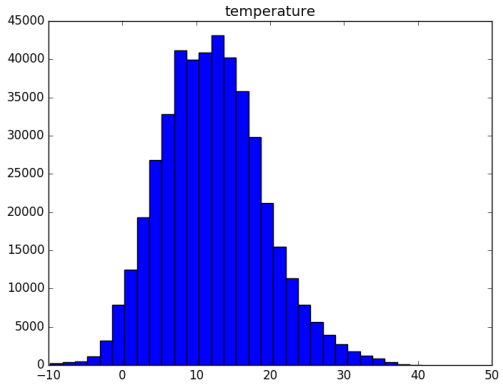

vent sur échelle log



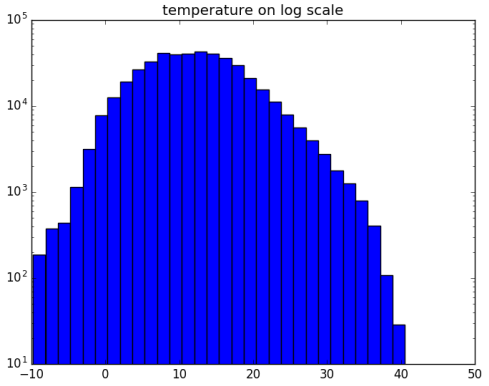

ennuagement




ennuagement sur échelle log




temperature




temperature sur échelle log

Il apparaît donc que les données sont plus étalées sur un échelles logarithmique. Dans la mesure ou nous cherchons à appliques des techniques de régression simples et avec une forte régularité (regression lineaire), le fait d'utiliser des échelles logarithmiques est une première transformation que nous pouvons appliquer à nos données avant d'utiliser des modèles de régression simples.

### 1.5.2 Les données statiques

On a seulement 29 jeux de données statiques (1 par station), on pourrait les afficher dans un tableau, mais cela manque d'intérêt.

# 2 Solutions

## 2.1 Plus proche voisins

On commence par implémenter une methode des K plus proches voisins sur les données brutes séparées polluant par polluant. On fait plusieurs essais et voici ce qu'on obtient :

| score | methode | K |
|---|---|---|
| 597.379 | par polluant | K = 5 |
| 617.229 | par polluant | K = 3 |
| 613.892 | par polluant | K = 4 |
| 613.892 | par polluant et par zone | K = 4 |

Etonnament, le fait de faire la méthode par polluant et par zone ne change pas le résultat. De plus, on n'est pas très loin du score du benchmark proposé par Plume Labs (501)

## 2.2 Méthode triviales

On essaie ensuite trois méthodes triviales :

— On met la valeur zéro pour toutes les prédictions : le score de 800.000 environ. Cela nous donne un ordre de grandeur sur les score : toute méthode donnant un score supérieur à 800.000 n'est vraiment pas adaptée.

— On met une valeur constante pour chaque polluant égale à la moyenne de toutes les valeurs pour ce polluant : on obtient une score de 350.000 environ. C'est nettement mieux que le benchmark de Plume Labs. Il est donc probable que leur solution overfitte les données d'entrainement.

— On calcule la valeur moyenne pour chaque polluant à chaque instant donné : on obtient un score de 440.000 environ. Cette méthode n'apporte rien par rapport à la valeur moyenne

## 2.3 Regression linéaire

On teste un modèle linéaire sur les données. Puisque le modèle linéaire contient les fonctions constantes, a priori, le score devrait être au pire de l'ordre de la valeur obtenue en mettant la moyenne, c'est à dire 350.000. Etonnament, on obtient un score de 430.000. Cela veut dire qu'avec une classe de fonction aussi simple que les fonctions linéaires, on overfitte déjà sur les données d'entrainement. Mais ce n'est pas tellement étonnant ; en effet, nous n'avons que 29 jeu de données statiques alors que ces données vivent dans un espace de dimension 18. Quelque soit la méthode employée, à moins d'avoir de la chance, on ne peut pas espérer obtenir une bonne prédiction.

## 2.4 Gradient boosting