Prévision de la qualité de l'air dans les stations de métro parisiennes

Jérémy Guérin, Nathan Huet, Alex Westbrook

Mars 2021

Contents

1 Introduction

La qualité de l'air est un enjeu important pour les métropoles comme celle de Paris. L'importante pollution liée à l'activité humaine et aux déplacements a des effets néfastes sur la santé, c'est pourquoi ces villes tentent de la réduire par diverses mesures notamment la réduction de l'utilisation de la voiture. Il arrive que la ville se retrouve dans un pic de pollution, auquel cas les dangers pour la santé sont accrus. Pour sortir de cette situation, Paris prend des mesures temporaires comme l'interdiction des véhicules les plus polluants et la diminution de la limitation de vitesse.
Afin d'éviter les pics de pollution, il serait utile d'appliquer les mesures temporaires plus tôt. Pour cela, il est important de pouvoir prévoir la qualité de l'air et c'est l'objectif de cette étude. Plusieurs quantités jouent un rôle majeur dans la mesure de la qualité de l'air : les particules fines et les oxydes d'azotes. Ces polluants proviennent principalement des combustions fossiles, pour le chauffage ou le trafic routier. Nous nous intéresserons dans ce projet à la quantité d'oxydes d'azote dans l'air dans plusieurs stations de métro de Paris, il ne s'agit donc pas de pollution extérieure mais intérieure dans un souterrain fermé. Le facteur météorologique, qui est usuellement l'un des facteurs principaux pour prédire la qualité de l'air, ne jouera sans doute pas un rôle très important dans notre analyse prévisionnelle. Dans notre cas la diffusion d'oxydes d'azote proviendra surtout des moteurs des trains. Nous allons plus précisément étudier la concentration moyenne quotidienne en oxydes d'azote. Dans l'idée de prévoir des pics de pollution, l'idéal serait de prédire les maxima sur chaque jour mais, ces derniers variant trop, nous avons préféré nous rabattre sur la moyenne journalière qui est plus régulière. De plus les effets sur la santé s'accumulent au long de la journée donc la moyenne journalière est même plus pertinente d'un point de vue sanitaire.

2 Les données

2.1 Présentation des données

Pour notre projet, nous nous servons de données de la RATP mesurant la concentration de plusieurs molécules dans l'air, ainsi que la température et l'humidité dans l'air. Ces données sont mesurées dans trois stations de métro de Paris : Auber, Châtelet et Franklin D. Roosevelt. La présence de seulement 3 stations dans ce jeu de données ne nous permet pas d'établir un lien entre les différentes stations du réseau ferroviaire. Notre étude se basera surtout sur l'étude des séries temporelles, nous tâcherons de prédire les quantités d'oxydes d'azote sur un mois à partir des observations passées. Nous nous intéresserons plus particulièrement aux données issues de la station Auber car cette base de données présente moins de données manquantes que les autres. Ces données datent de 2013 à aujourd'hui (on peut télécharger le jeu de données actualisé sur le site de la RATP) avec une mesure effectuée toutes les heures.

2.2 Variables présentes

On retrouve les variables suivantes dans la base de données :
-DATE_HEURE : date et heure de la mesure
-NO : concentration de monoxyde d'azote (NO) en µg/m^3
-NO2 : concentration de dioxyde d'azote (NO2) en µg/m^3
-PM10 : concentration de particules fines dont le diamètre est inférieur à 10 µm en µg/m^3
-PM2.5 : concentration de particules fines dont le diamètre est inférieur à 2.5 µm en µg/m^3
-CO2 : concentration de dioxyde de carbone (CO2) en µg/m^3
-TEMP : température en °C
-HUMI : humidité relative en pourcentage

Une première chose que nous pouvons faire est de regarder la corrélation entre ces différentes variables ainsi qu'entre les jeux de données des différentes stations. On peut visualiser la matrice de corrélation (cf figure 1) :

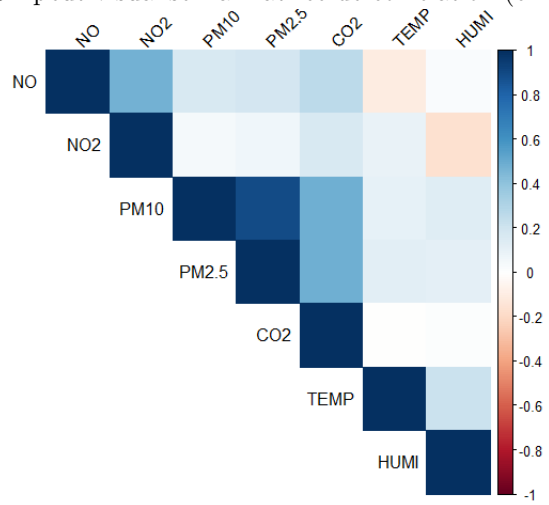

Figure 1 : Matrice de corrélation entre les variables mesurées à la station Auber

Nous n'observons pas de corrélation marquante entre les différentes grandeurs présentes dans nos données excepté pour les taux de PM10 et de PM2.5 mais cela était attendu étant donné que ces taux correspondent aux mêmes particules mais de tailles différentes donc il ne nous a pas paru pertinent d'utiliser l'une pour prédire l'autre.
De plus, lorsque nous comparons le taux de NO2 de la station Auber avec celui de la station Châtelet, nous ne trouvons pas de corrélation notable donc nous ne pouvons pas non plus utiliser ces données comme étant des variables explicatives.

2.3 Traitement des données

Les données que nous avons récupérées sont brutes et beaucoup de variables manquent à l'appel. En effet, à partir du 27/07/2018, les données sont quasi inexistantes, nous devons donc exclure cette partie des données. Avant cette date, le nombre de données manquantes varie de 1523 à 5287 (sur 48787). On doit procéder à l'imputation des données pour pouvoir nous en servir par la suite. Nous avons réalisé l'imputation à l'aide du package mice. Ce dernier offre une multitude de méthodes pour imputer des données. Nous avons opté pour une imputation à l'aide de forêts aléatoires qui semble être le plus adéquat pour nos données.
Cette méthode d'imputation consiste à initialiser chaque valeur manquante à la moyenne des autres valeurs puis à itérativement:
- construire une forêt aléatoire à partir de l'ensemble des données (celles d'origines et celles initialement manquantes)
- remplacer chaque valeur initialement manquante par la moyenne pondérée par les poids de proximité de la forêt
On obtient les résultats suivants qui sont assez satisfaisant, on peut même observer l'apparition d'un pic de pollution vers août 2016 (cf figure 2).

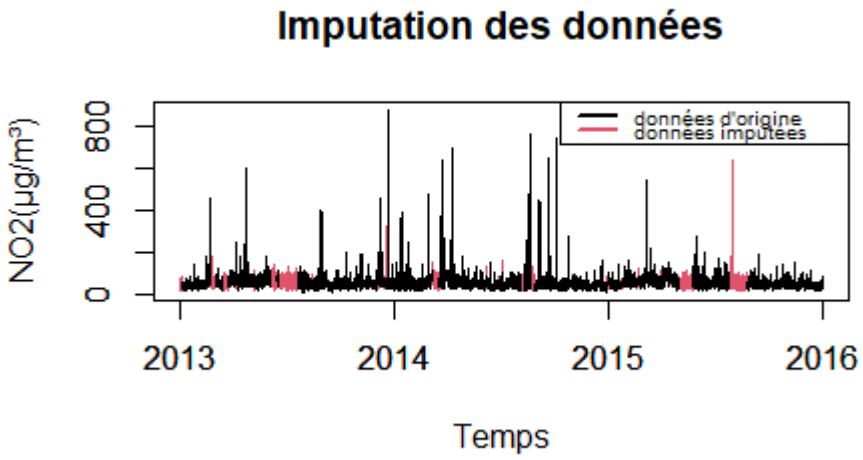

Figure 2 : Graphique représentant la concentration de NO2 en fonction du temps

Une autre méthode pour pallier ce nombre important de données manquantes consiste à baser nos prévisions sur les moyennes journalières. Cette méthode peut être considérée comme satisfaisante car elle nous évite de rajouter des données "factices" et contribue à régulariser nos données. Nous utiliserons le jeu de données composé des dates, des moyennes et des maxima journaliers par la suite. Ici le minimum journalier est inutile car nous voulons anticiper les journées où la qualité de l'air est la plus mauvaise.

2.4 Variables explicatives

Nous avons beaucoup cherché à obtenir des données explicatives pour nos données sans résultat. En effet, les variables qui auraient pu expliquer nos données sont l'affluence dans les gares ainsi que le trafic ferroviaire. Malheureusement, la RATP ne peut pas fournir au public des données précises sur ces aspects, celles qui sont disponibles sont inutilisables dans le cadre de ce rapport.
Une autre possibilité était d'utiliser une base de données fournie par Airparif mesurant la qualité de l'air avec les mêmes variables mais à l'extérieur de la station de métro. Deux problèmes sont survenus. Le premier est qu'il n'est pas crédible de se servir du même type de données que celles que l'on veut prédire en tant que variable explicative. Le second est simplement que la corrélation entre les données n'est pas très importante.
Notre deuxième piste a été d'étudier des données de trafic routier à proximité des stations où ont lieux nos mesures. Une base de données fournie par la ville aurait pu nous suffire, mais une nouvelle fois les corrélations avec nos données étaient mauvaises. Cette étude prévisionnelle sera donc principalement basée sur les méthodes de séries temporelles.

2.5 Evaluation des prédictions

Pour évaluer nos prédictions, nous allons essentiellement utiliser deux critères, le MAPE (Mean Absolute Percentage Error) et le RMSE (Root Mean Square Error). Si on a n points à évaluer et qu'on note x_i les valeurs observées et y_i les valeurs prédites, on a les formules suivantes :

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{x_i - y_i}{x_i} \right| \tag{1}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

L'intérêt du MAPE est que la différence entre la valeur prédite et celle observée est rapportée à leur échelle. Cette normalisaton permet d'avoir une quantité qui s'interprète de la même façon pour des jeux de données différents. Comme son nom l'indique on l'exprime sous forme de pourcentage en multipliant par 100. Une valeur de 0% correspond à une parfaite adéquation, mais le MAPE peut dépasser 100% si la valeur prédite est beaucoup trop grande. Afin d'éviter des divisions par 0, le MAPE est généralement calculé sur des données éloignées de 0 et donc de même signe, ce sera toujours le cas dans le cadre de notre étude. Il y a alors une dissymétrie entre les prédictions trop hautes et trop basses. Les prédictions basses ne donneront jamais un MAPE de plus de 100% et auront tendance à être moins pénalisées que les prédictions hautes.
Le RMSE quant à lui présente l'avantage d'être symétrique entre valeur prédite et observée, ce qui induit aussi une symétrie entre prédictions hautes et basses. L'utilisation du moment d'ordre 2 permet de pénaliser davantage les grosses

erreurs de prédictions que les petites. Intuitivement, une prédiction à faible MAPE mais fort RMSE a sans doute quelques erreurs importantes tandis qu'une prédiction à faible RMSE mais fort MAPE devrait avoir beaucoup de petites erreurs. Cependant le RMSE n'est pas normalisé par rapport à l'échelle des données, donc sa valeur intrinsèque est difficilement interprétable, mais elle peut être utilisée pour comparer plusieurs prédictions sur un même jeu de données.

3 Prédiction du jour au lendemain

Dans toute cette partie, nous allons prédire les moyennes journalières de concentration de NO2 sur une année jour après jour. Pour ce faire, nous allons utiliser la base de données moyennées de 2013, 2014 et 2015 comme échantillon pour prédire les moyennes de 2016. Ainsi, nous fournissons les données de chaque jour au fur et à mesure, afin que le modèle s'adapte.

3.1 Le modèle SARIMA

Dans cette partie, nous allons utiliser le modèle pour séries temporelles SARIMA pour prédire les concentrations moyennes de NO2. Pour se donner une idée de la qualité des résultats, prédire l'année suivante grâce à la moyenne amène à un MAPE d'environ 19%. Le modèle SARIMA va permettre de repérer les tendances et les saisonalités de notre série temporelle pour prédire l'évolution de la série.

3.1.1 Prédiction grâce à la méthode de Box-Jenkins

Premièrement, nous allons estimer les paramètres de notre modèle grâce à la méthode de Box-Jenkins, en étudiant les autocorrélogramme et autocorrélogramme partiel de notre série temporelle. Une fois les paramètres estimés, nous générons notre prévision grâce à la fonction Arima de R. Voici le résultat obtenu (cf figure

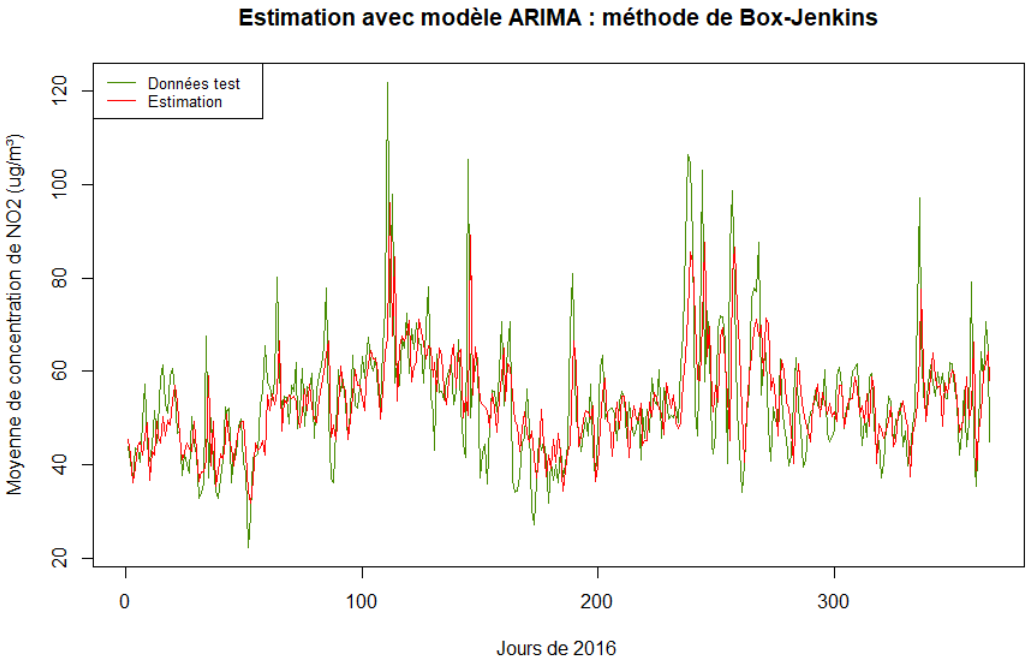

Figure 3 : Graphique représentant la moyenne de concentration de NO2 réelle (issue des données) et celle estimée avec le modèle ARIMA, en fonction du temps

Le résultat est plutôt satisfaisant. En effet, le modèle suit bien les tendances, mais prévoit mal les valeurs extrêmes, ce qui n'est pas inattendu. On obtient un MAPE de 13.6% et un RMSE de 10.62. Le résultat n'est pas parfait mais encourageant.

3.1.2 Prédiction grâce à auto.arima et au lissage exponentiel

Nous avons essayé d'améliorer nos résultats en utilisant d'autres méthodes pour déterminer le modèle. On a utilisé la fonction auto.arima (qui détermine le meilleur modèle SARIMA) associée aux critères AIC et BIC mais les résultats obtenus avaient un MAPE supérieur. Nous avons ensuite testé les méthodes de lissage exponentiel mais une fois de plus, les résultats étaient moins bons que notre première estimation. Voici les prévisions obtenues (cf figure 4):

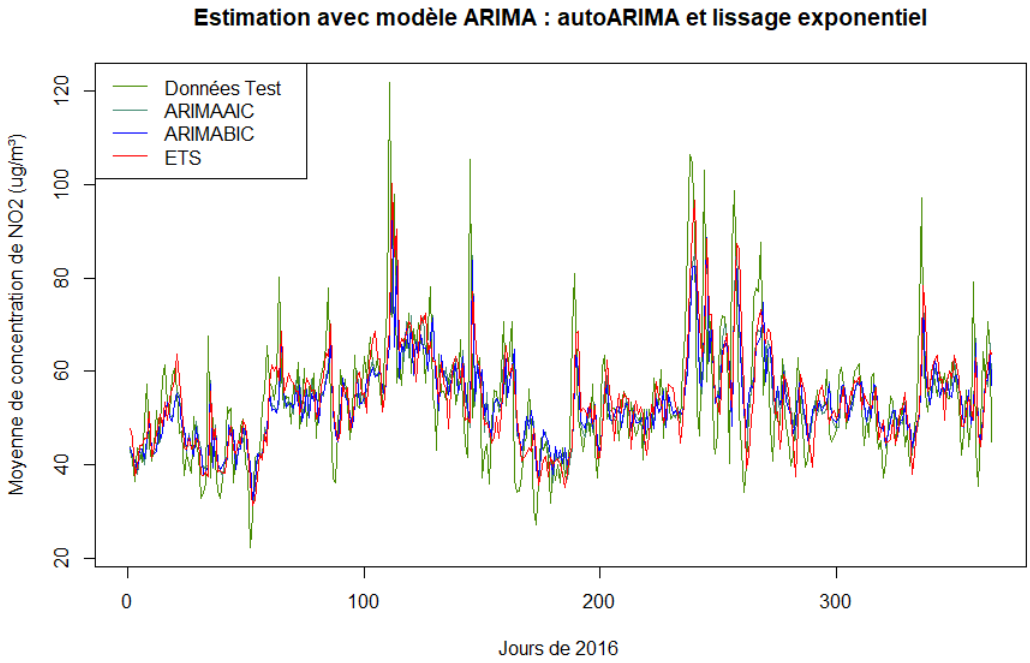

Figure 4 : Graphique représentant la moyenne de concentration de NO2 réelle (issue des données) et celle estimée avec le modèle SARIMA, en fonction du temps

3.2 Les forêts aléatoires

Dans cette partie, nous allons utiliser les modèles de forêts aléatoires afin de prédire les concentrations de NO2 jour par jour sur l'année 2016. Nous utiliserons pour cela le package ranger.

3.2.1 Sélection de nos paramètres

Pour utiliser la dépendance chronologique de nos données, nous séparons notre variable Date en trois variables : Day, Month et Year. Ces variables seront utilisées dans nos modèles de forêts aléatoires. Nous allons aussi utiliser les moyennes journalières des jours précédents pour prédire les jours futurs.
Afin d'obtenir un modèle de prédiction optimal, nous devons d'abord déterminer plusieurs paramètres. Nous séparons notre base de données échantillon en deux. Ainsi les données de 2013 et 2014 serviront de données d'entraînement tandis que les données de 2015 serviront de données de validation.
Le premier paramètre à estimer est le nombre de jours à prendre en compte, précédant la date que l'on veut prédire. Pour cela, on va étudier le MAPE en faisant seulement varier le nombre de jours de 1 à 7. Suite au résultat obtenu (cf

figure 5), prendre les 7 jours qui précèdent le jour désiré semble être la meilleure alternative. Ce qui est cohérent car les données semblent avoir une corrélation hebdomadaire.

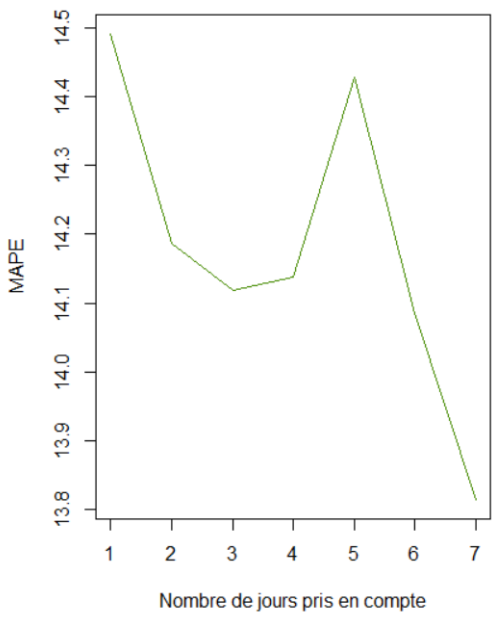

Figure 5 : Graphique représentant l'erreur de validation MAPE selon le nombre de jours pris en compte dans le modèle

Le paramètre suivant à déterminer est mtry qui représente le nombre de variables pouvant être aléatoirement échantillonnées pour construire nos arbres. Nous faisons varier ce paramètre entre 1 et 11. Suite à la courbe obtenue (cf figure 6), mtry = 11 semble optimal.

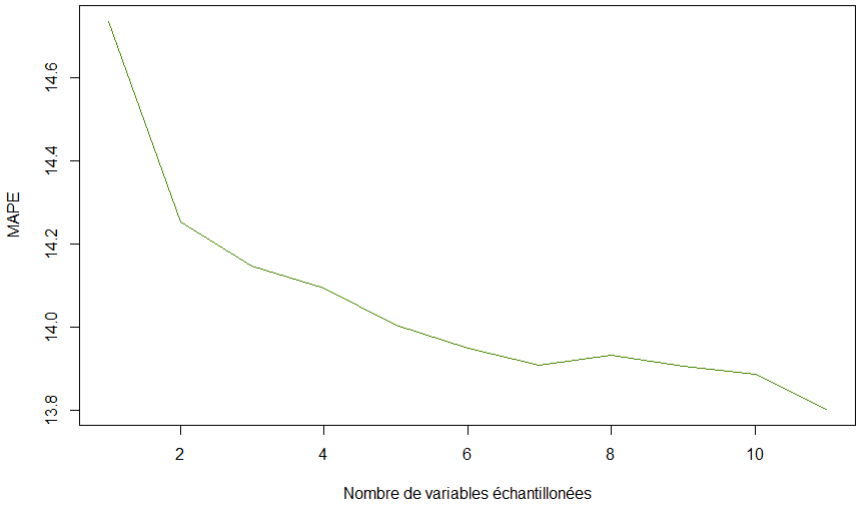

Figure 6 : Graphique représentant l'erreur de validation MAPE selon le nombre de variables échantillonées dans le modèle

Enfin, il faut sélectionner le nombre d'arbres ntree dans notre forêt. Nous faisons donc varier ce paramètre entre 1 et 1000. Au vu de la figure obtenue (cf figure 7), ntree=400 semble convenable.

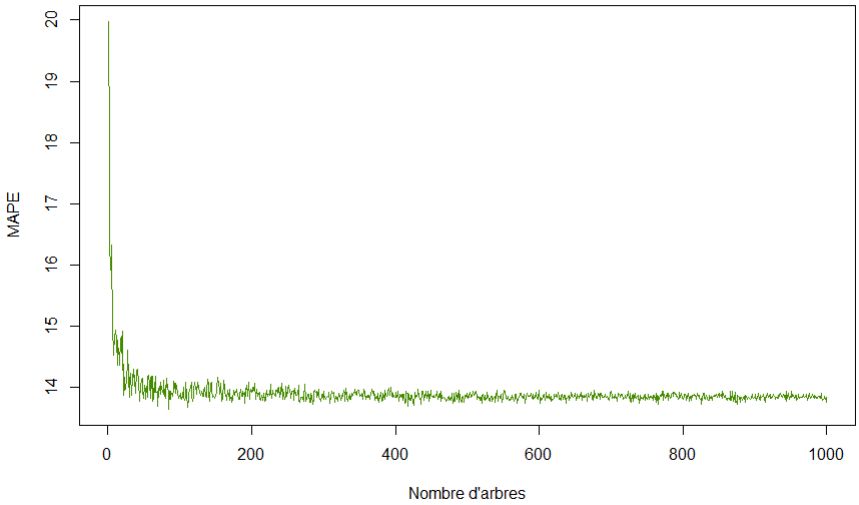

Figure 7 : Graphique représentant l'erreur de validation MAPE en fonction du nombre d'arbres

3.2.2 La prédiction

Maintenant que tous les paramètres sont estimés sur les données de validation, nous pouvons réaliser la prédiction sur les données test. Comme le modèle précédent, le résultat n'est pas parfait mais reste satisfaisant (cf figure 8). Le RMSE est de 10.67 et le MAPE est de 13.96%.

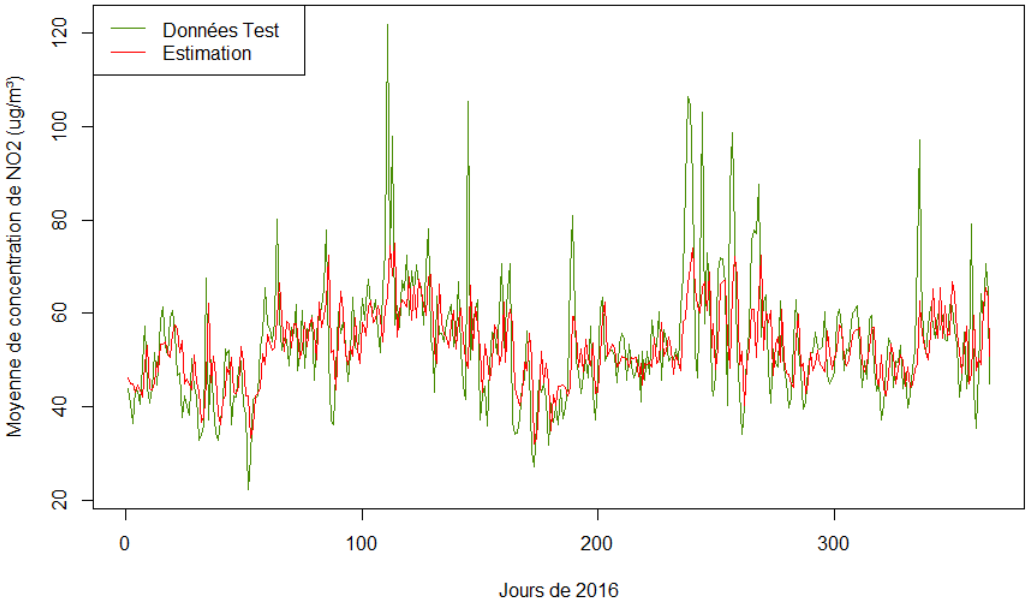

Figure 8 : Graphique représentant la moyenne de concentration en NO2 des données réelles et celles estimées par forêts aléatoires en fonction du temps

3.3 Le modèle GAM

Dans cette partie, nous utilisons les modèles GAM (Generalized Additive Model) pour prédire les concentrations moyennes de NO2 de 2016. Nous utilisons la fonction gam pour générer le modèle, issue du package mgcv. L'idée est de représenter la concentration moyenne en NO2 d'une journée comme fonction de celles des jours précédents. Ces fonctions sont approximées sur une base de fonctions lisses dont la dimension est un paramètre à choisir.

3.3.1 Sélection des paramètres

Tout comme pour les forêts aléatoires, nous devons donc d'abord choisir nos paramètres pour avoir un modèle optimal. Nous séparons le jeu de données échantillon en deux jeux d'entraînement et de validation de données de la même façon que précédemment.

Nous devons déterminer les différents facteurs k, qui sont les dimensions des bases pour représenter nos termes lisses. Ici, nos termes lisses seront les mesures moyennes des concentrations en NO2 remontant de un à sept jours avant la prévision. Nos variables Day, Month et Year seront laissées en variables linéaires. Voici les différents MAPE que nous avons obtenu (cf figure 9):

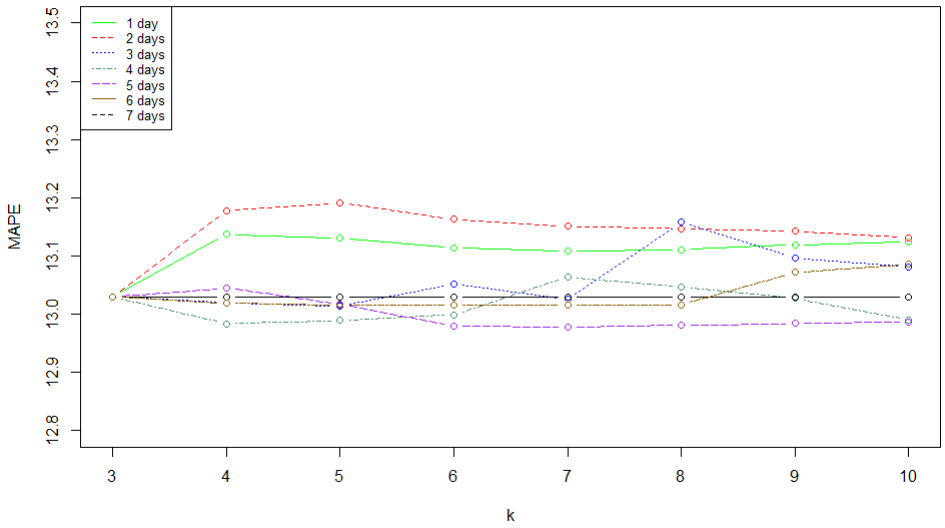

Figure 9 : Graphique représentant l'erreur de validation MAPE en fonction de la dimension de la base pour chaque jours précédents considérés

En notant k_i le k gardé pour le i-ème jour, nous obtenons donc :
$k_1 = 3,\ k_2 = 3,\ k_3 = 5,\ k_4 = 4,\ k_5 = 7,\ k_6 = 6,$ et $k_7 = 10.$

3.3.2 La prédiction

Nos paramètres estimés, nous pouvons réaliser notre prédiction sur les données test. Voici le résultat obtenu (cf figure 10):

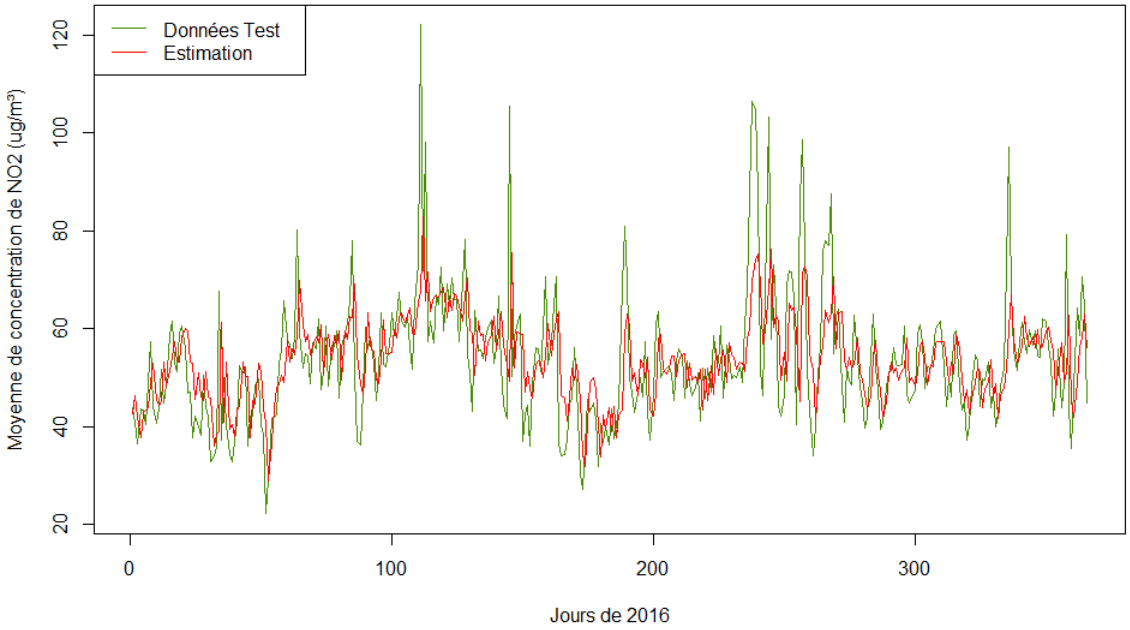

Figure 10 : Graphique représentant la moyenne de concentration en NO2 des données réelles et celles estimées par GAM en fonction du temps

Nous obtenons un MAPE de 14.01% et un RMSE de 10.72. Ce modèle est à peu près aussi bon que celui des forêts aléatoires tout en étant différent.

3.4 Résultat final et conclusions

Finalement, nous agrégeons nos trois modèles afin d'obtenir un meilleur résultat qui prend en compte les points forts de chaque modèle. Nous utilisons la fonction mixture pour effectuer cette agrégation. Nous obtenons la prédiction suivante (cf figure 11):

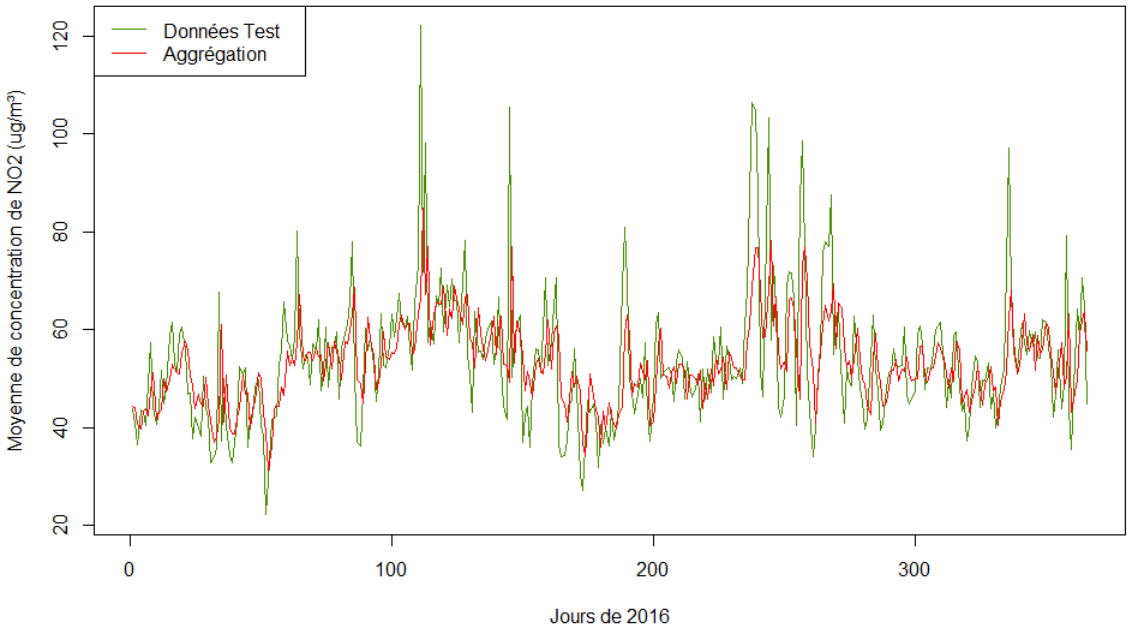

Figure 11 : Graphique représentant la moyenne de concentration en NO2 des données réelles et celles estimées par agrégation des trois modèles en fonction du temps

Cette prédiction admet un MAPE de 13.31% et un RMSE de 10.47. Celui-ci est légèrement améliorée grâce à l'agrégation.
Ainsi, prédire du jour pour le lendemain est faisable avec les modèles de forêts aléatoires, GAM et SARIMA, même si ce dernier est meilleur que les deux autres pour une prévision si proche. En effet, on peut voir que le modèle SARIMA estime relativement bien les valeurs extrêmes contrairement aux deux autres modèles.

4 Prédiction d'une semaine à l'autre

Dans toute cette partie, nous allons prédire les moyennes journalières de concentration de NO2 sur une année une semaine à l'avance. Pour ce faire, nous allons

utiliser la base de données moyennées de 2013, 2014 et 2015 comme échantillon pour prédire les moyennes de 2016. Ainsi, chaque modèle n'aura accès qu'à la quantité mesurée une semaine plus tôt. Les résultats seront nécessairement plus mauvais, voire inutilisables, mais cette analyse est beaucoup plus cohérente d'un point de vue pratique. En effet, comme nous l'affirmions dans l'introduction, ce genre de prédiction sert à anticiper les pics de pollution afin de pouvoir organiser des mesures sanitaires. Une semaine pour prendre des mesures sanitaires est beaucoup plus plausible que seulement une journée.

4.1 Les forêts aléatoires

Nous utilisons d'abord les forêts aléatoires afin de prédire la concentration moyenne de NO2 une semaine en avance. Ainsi, seuls sept variables sont prises en compte par le modèle : Day, Month, Year, Date, et les trois mesures des moyennes de la concentration en NO2 7, 8, et 9 jours avant.
Comme dans la partie précédente, nous sélectionnons nos paramètres de manière optimale en utilisant les années 2013 et 2014 comme données d'échantillon et l'année 2015 comme données de validation. Nous choisissons mtry = 7 et ntree = 402. Voici le résultat obtenu (cf figure 12) :

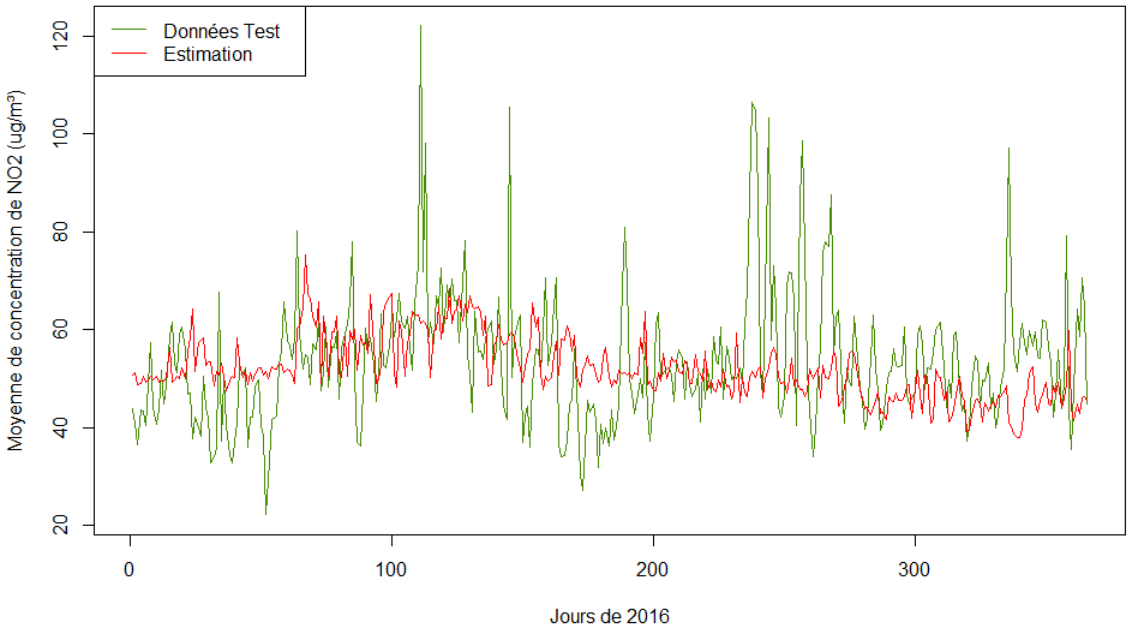

Figure 12 : Graphique représentant la moyenne de concentration en NO2 des données réelles et celles estimées par forêts aléatoires une semaine en avance, en fonction du temps.

Le résultat admet un RMSE de 13.81 et un MAPE de 18.86%, un résultat moins précis que lors de nos prédictions du jour pour le lendemain.

4.2 Le modèle GAM

Nous utilisons maintenant les modèles GAM pour prédire les concentrations en NO2 une semaine à l'avance. Le modèle prend en compte les sept mêmes variables que pour les forêts aléatoires. Comme dans la partie précédente nous sélectionnons les k optimaux pour chacun de nos paramètres à partir des données de validation : on choisit k=10 pour les mesures à 7 et 8 jours, k=4 pour les mesures à 9 jours et k=3 pour les variables jour et mois. Le résultat admet un RMSE de 12.96 et un MAPE de 17.68%. Ce résultat est légèrement meilleur que celui obtenu grâce aux forêts aléatoires. Voici le résultat obtenu (cf figure 13) :

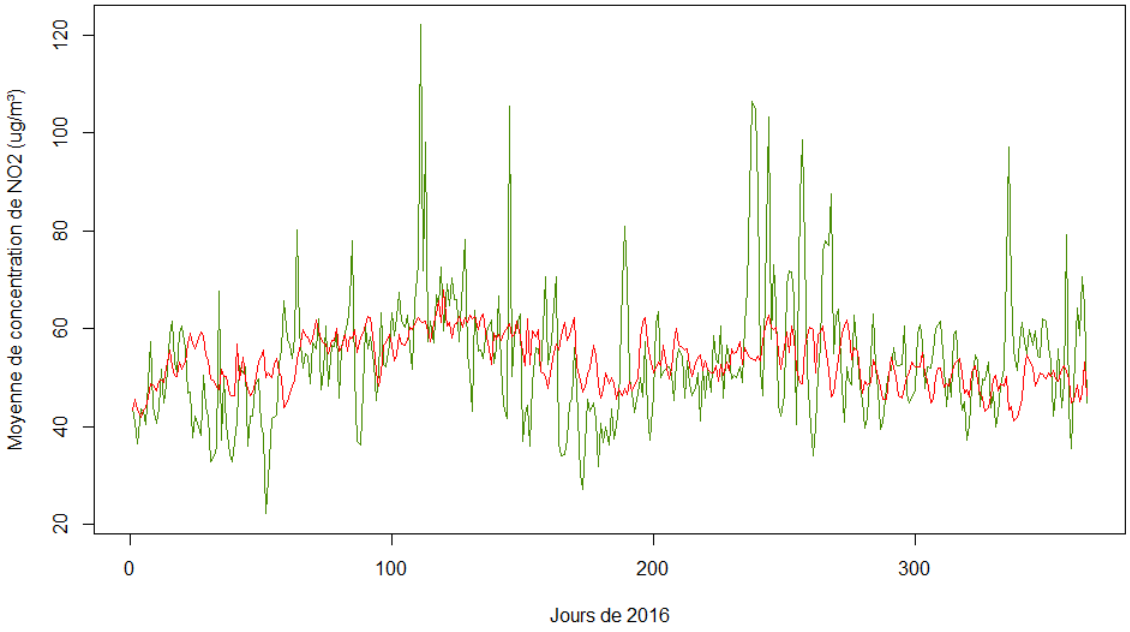

Figure 13: Graphique représentant la moyenne de concentration en NO2 des données réelles et celles estimées par GAM une semaine en avance, en fonction du temps.

4.3 Résultat final et conclusions

Comme dans la partie précédente, nous réalisons une agrégation des deux résultats pour obtenir une meilleure prédiction mais nous obtenons un résultat moins bon qu'avec le modèle GAM. Ce dernier restera l'une des meilleurs prédictions que nous sommes capables de faire.
Malheureusement, les modèles ne suivent pas les tendances comme dans la partie précédente. Les prédictions ne sont pas aberrantes mais elles ne sont pas bonnes pour autant. C'est ici que que notre manque de variable explicative se ressent le plus. En effet, ce genre de prédiction doit s'effectuer en prenant en compte les affluences journalières ainsi que les possibles événements à venir qui sont susceptibles d'augmenter la concentration de polluants dans l'air. Un accès à des données plus précises aurait pu nous permettre de tester d'autres méthodes de prédiction et améliorer nos résultats.

5 Commentaire

La principale critique que nous pouvons faire sur nos prédictions est qu'elles ne prédisent pas très bien les pics de pollution. A part avec le modèle SARIMA, la prédiction des pics est plutôt mauvaise et on a du mal à anticiper leur intensité alors que, dans la réalité, c'est surtout ces pics qui sont intéressants à prévoir. En effet, l'intérêt principal de la prédiction du taux de pollution est de pouvoir prévenir les habitants qu'un pic de pollution aura lieu dans les jours à venir. De plus, quand on essaie de prédire une semaine en avance, on ne voit plus du tout les pics apparaître ce qui est un problème.
Une façon dont on pourrait remédier à ce problème serait d'utiliser la théorie des valeurs extrêmes qui est une branche des statistiques qui cherche à trouver un modèle statistique pour les valeurs extrêmes d'un échantillon. Elle peut être utilisée dans la détection d'anomalies pour prévoir une crue ou, dans notre cas, un pic de pollution. On peut donc imaginer que grâce à cette théorie, on pourrait être capable de faire des prédictions qui seraient moins bonnes de façon globale mais meilleures pour répondre au problème principal de cet exercice.