MATH 60603 - Apprentissage Statistique

Impact des différentes politiques de confinement sur le taux de NO2 à Paris

Présenté à

Pre. Aurélie Labbe

Par

Christelle George - 11288106
David Lemieux - 11118064
Quentin Tabourin - 11290690

Le 18 décembre 2020

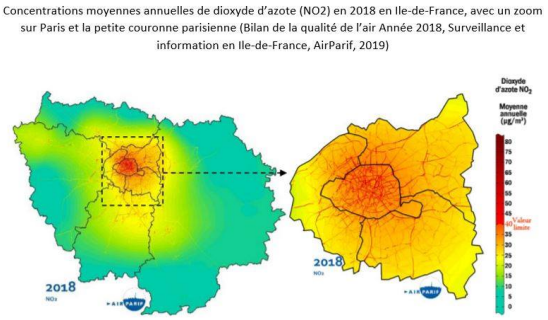
À Montréal

1 Introduction

Les différentes politiques de confinement recommandées face à l'épidémie de la Covid19 ont profondément transformé les habitudes du quotidien, et ce à l'échelle mondiale. Désormais, la grande majorité des interactions humaines (sociales et professionnelles) se fait virtuellement. L'une des conséquences majeures de ce nouveau monde virtuel se reflète dans les habitudes de transports. Or, les voitures à moteur à combustion interne (Diesel) produisent du dioxyde d'azote NO2, un polluant majeur de l'atmosphère terrestre.

Les images satellites et les stations de mesure ont montré la diminution de la concentration de NO2 lors du confinement du printemps 2020(Forster, Forster, and Evans 2020). La variation a notamment été visible dans les zones industrialisées et les grandes villes. En général, la mesure de pollution la plus utilisée est celle de CO2. Cependant, sa variation dans le temps ne peut pas être facilement observée de façon ponctuelle. Pour cette raison, la mesure de pollution choisie est le NO2. À gauche il est possible de voir la Figure 1.1

Figure 1.1: Carte Paris



Notre étude porte sur la ville de Paris : il s'agit d'une des villes les plus denses au monde. La principale source d'émission de NO2 dans cette ville est liée à son trafic routier. Il existe bien entendu de nombreuses règlementations en terme de taux de particules maximales contenues dans l'air. Selon un décret de 2010, la valeur limite annuelle de NO2, pour la protection de la santé humaine, est fixée à 40 $\mu g/m^3$, ce qui correspond à une valeur de 21 particules par million (ppm) approximativement. Ceci dit, si la moyenne nationale de la France respecte ce taux, nous ne pouvons pas en dire de même sur la ville de Paris. Ainsi, à proximité des axes les plus chargés, les niveaux de NO2 sont, en moyenne, deux fois supérieurs à la valeur limite annuelle.

Ce polluant de NO2 est fort complexe puisqu'il provient des émissions directes (transports etc.) mais aussi des équilibres chimiques avec le polluant de l'ozone O3. À titre de simplification, ce projet traite uniquement du NO2 et ne vise pas à quantifier l'effet mutuel des polluants.

1.1 Objectif d'analyse

Le but de ce projet est donc de mieux comprendre l'impact des différentes politiques de confinement à Paris sur le taux d'émission de NO2. Pour y parvenir, les données adéquates sont collectées et transformées afin d'expliciter les caractéristiques recherchées. Différents modèles sont étudiés et entrainés; celui possédant les meilleures prédictions sur l'ensemble de validation est ensuite choisi. Ainsi, deux objectifs peuvent être définis :

- Premièrement, étudier l'effet du confinement à Paris sur le taux de NO2 émis.
- Deuxièmement, trouver le meilleur modèle de prédiction parmi plusieurs modèles différents.

1.2 Revue de la littérature

De nombreuses études ont été faites sur des sujets similaires ou connexes avec la méthode CART et ce, à une échelle mondiale (Bulgarie, Pologne, Canada, Californie etc). Dans la majorité des cas où l'on cherche à prédire une variable cible de pollution (PM10, NOx, CO,O3), ce sont des variables météorologiques qui sont utilisées en guise de variables explicatives : la température minimale et maximale de la journée, l'humidité relative, la vitesse et la direction du vent, la pression, la couverture nuageuse et bien d'autres ((Stoimenova et al. 2017), (Burrows et al. 1995))
Il est intéressant de noter que, quelle que soit l'étude réalisée, le traitement des données est la partie la plus importante et la plus longue. Plusieurs points méritent d'être soulevés ici. Les données temporelles sont en général non stationnaires

en moyenne et en variance, et traduisent des cycles saisonniers multiples (quotidien, hebdomadaire ou annuel). Ainsi, le traitement de ces variables est requis puisqu'il permet de simplifier cette saisonnalité en éliminant la non stationnarité et permet également de filtrer les cycles ((Dudek 2015), (Choi et al. 2013)). À titre d'exemple, l'ajout de prédicteurs particuliers (lag) permet de tenir compte des données temporelles ou encore une transformée WDI de la direction du vent permet de tenir compte de sa périodicité ((Stoimenova et al. 2017)). Aussi, dans le cas d'une variable cible non normale (PM10 par exemple), la transformée de Yeo Johnson y remédie en lui accordant une allure normale. En comparant plusieurs modèles différents (PM10 vs PM10 transformé), la meilleure performance est atteinte avec la variable transformée ((Stoimenova et al. 2017)). Le nombre de prédicteurs utilisé doit être choisi judicieusement : trop peu de variables explicatives engendre de mauvaises prédictions des évènements rares (une forte concentration – anormale – de O3 ne sera pas convenablement prédite par exemple). Finalement, l'horizon temporel des données est important : idéalement, les données sur de nombreuses années sont requises afin de cerner correctement la variabilité liée à certaines variables ((Burrows et al. 1995)). D'autres part, les modèles sont évalués selon plusieurs critères. Si certaines études se basent sur une comparaison des coefficients de détermination R2 des différents modèles ((Stoimenova et al. 2017)), d'autres privilégient la stabilisation de l'indicateur de l'erreur OOB ((Dudek 2015)). Plusieurs études indiquent également l'importance des variables selon le modèle choisi.

La méthode CART est une méthode très populaire quant à la prédiction dans le domaine du climat (météo / pollution) mais s'applique également très bien dans d'autres domaines variés comme la finance (I. Bou-Hamad, 2020), la biostatistique (Kane et al. 2014) et l'environnement (Gocheva-Ilieva et al. 2019). Des auteurs comme (Dudek 2015) et (Mei et al. 2014) ont utilisé les forêts aléatoires afin de prédire la demande d'électricité à court-terme. Comme le mentionne (Dudek 2015), les RF sont des modèles relativement simples avec un nombre restreint de paramètres à calibrer et à évaluer.
In fine, d'autres méthodes sont employées dans un contexte de prédiction. La méthode de Takagi-Sugeno, par exemple, est adaptée pour traiter des systèmes non linéaires (Elayan et al. 2006). Des modèles hybrides CNN LSTM également ont été employés : la LSTM est une version spéciale des RNNs, telle que chaque neurone dans sa structure est une cellule mémoire qui permet le transfert de données. Ce modèle est très adapté pour les séries temporelles (Kaya and Öğüdücü 2020). Nous ne nous y attarderons pas puisque ceci n'est pas le but de notre projet.
Somme toute, la nouveauté de notre projet relève de l'actualité qui y est attachée : si beaucoup d'études ont déjà traité de la relation entre les données météorologique et une certaine variable de pollution, très peu ont considéré leur relation avec les politiques de confinement instaurées pour faire face à la pandémie de la Covid19.

1.3 Présentation des données

Quatre bases de données sont considérées dans le cadre de ce projet, soit une sur des données de mobilité, une sur des données météorologiques, une autre sur des données de confinement et la dernière sur des données de taux de NO2.
La première base de données est liée à la mobilité : elle est fournie par Google et relève des tendances de mobilité mondiale à partir des données Google Maps. Ces données sont classées en fonction du type de déplacement (travail, résidentiel, parc, courses…) et débutent le 15 février 2020. Elles sont exprimées en terme de pourcentage par rapport à la référence, soit la date du 15 février 2020.

La deuxième base de données concerne les variables météorologiques. Elle est tirée des messages internationaux d'observation en surface (SYNOP) circulant sur le système mondial de télécommunication (SMT) de l'Organisation Météorologique Mondiale (OMM). Cette base de données comportant beaucoup de prédicteurs, nous nous limitons à la selecion de 7 variables jugées importantes: la direction moyenne du vent, la vitesse moyenne du vent, les précipitations moyennes, l'humidité moyenne, la pression moyenne, les températures minimales, et les températures maximales. Ces variables ont été mesurées dans un lieu localisé au sud de l'aéroport d'Orly (environ 13km au sud de Paris centre). Pour simplifier notre étude, il est assumé que ces valeurs sont similaires à celles observées à Paris.

La troisième base de données comporte les observations du niveau de confinement imposé en France. Encore une fois, nous assumons que ces mêmes niveaux de confinement s'appliquent à la ville de Paris. Une première base de données sur le niveau de confinement en France était disponible pour les dates du 1er janvier 2020 au 11 juin 2020. Nous avons

donc complété manuellement cette base de données pour les dates manquantes en se basant sur les annonces de confinement par le gouvernement français.

Finalement, la dernière base de données concerne notre variable cible, soit le taux de NO2. Elle provient de l'organisation à but non lucratif «OpenAQ» et contient la variable réponse du taux de NO2, en ppm, mesurée au cours des 5 dernières années à une échelle mondiale. Les données sont classées par localisation (pays ou région). Dans le cadre de ce projet, nous avons sélectionné les observations sur la ville de Paris à partir de 14 points d'observations différents:

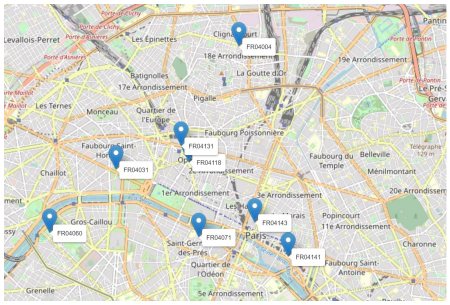
2 Méthodologie

2.1 Méthodologie sur les données

Un regard critique est porté sur la base de données "aq" contenant notre variable cible de NO2. Il s'agit de sélectionner un ou plusieurs points d'observations pertinents parmi les 14 à disposition.

2.1.1 Traitement de la base de données "AQ"

Figure 2.1: Carte Paris



Comme il est possible de le voir dans la carte ci-contre (Figure 2.1), les points d'observations sont répartis sur tout le territoire de paris, ce qui peut entraîner une grande variabilité dans les données observées. Les 3 points d'observations les plus au centre de Paris sont choisis, soit "FR04143","FR04071" et "FR04141". Il est important de noter que ces points de prélèvement se situent à proximité de grands axes routiers, dans une zone ou la circulation est souvent congestionnées afin de garantir des prédictions les plus réalistes possibles.

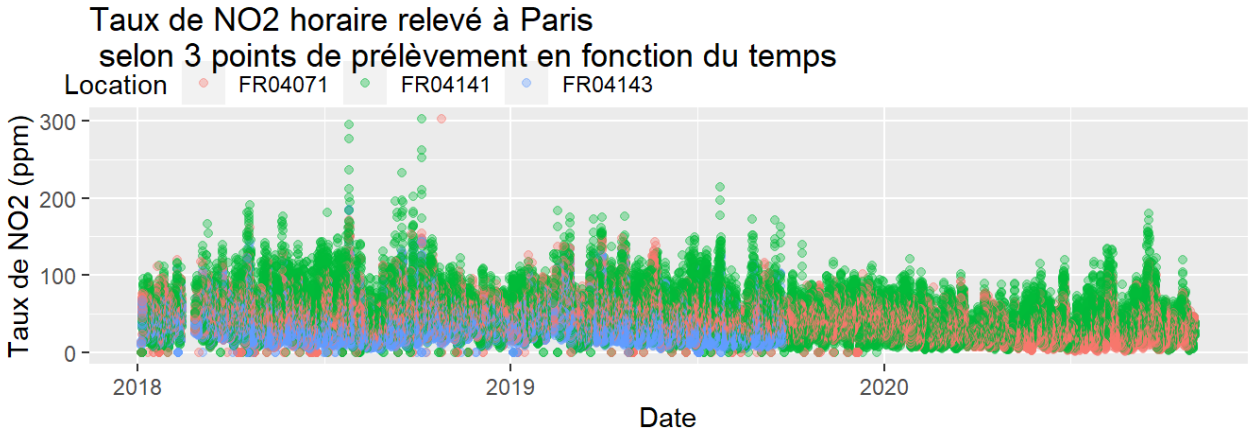


Figure 2.2: Taux horaire

Il est possible de voir que la station FR04141 est souvent beaucoup plus extrême que les deux autres stations. Cette raison nous pousse a croire que les lectures sont possiblement biaisées ou que cette station est très sensible aux évènements sporadiques et imprévisibles. Il est aussi possible de voir que la station FR04143 n'a pas toutes les données disponibles pour la période à l'étude. Cette station sera donc enlevée. Au final, la station FR04071, située tout près de l'ile de la cité, est retenue. À titre indicatif, il n'y a pas une seule bonne station dans ce contexte, tant que la station de prélèvement se situe sur un axe bien fréquenté et représentatif de la réalité (FR04131 par exemple).

Pour aller encore plus loin, deux observations peuvent être émises à partir du graphique ci-dessus (Figure 2.2) pour la station FR04071. Premièrement, plusieurs observations horaires affichent 0, ce qui est impossible, même par exemple la nuit lorsque la circulation est presque nulle. Deuxièmement, il est possible de voir une observation extrême avoisinant les

300ppm, très éloignée de l'observation la plus proche. Ces données seront donc éliminées. Comme il s'agit d'observations mesurées à chaque heure, éliminer ces observations extrêmes n'est pas problématique: cela ne fera que corriger la moyenne journalière et ainsi donnera plus de fiabilité à nos données.

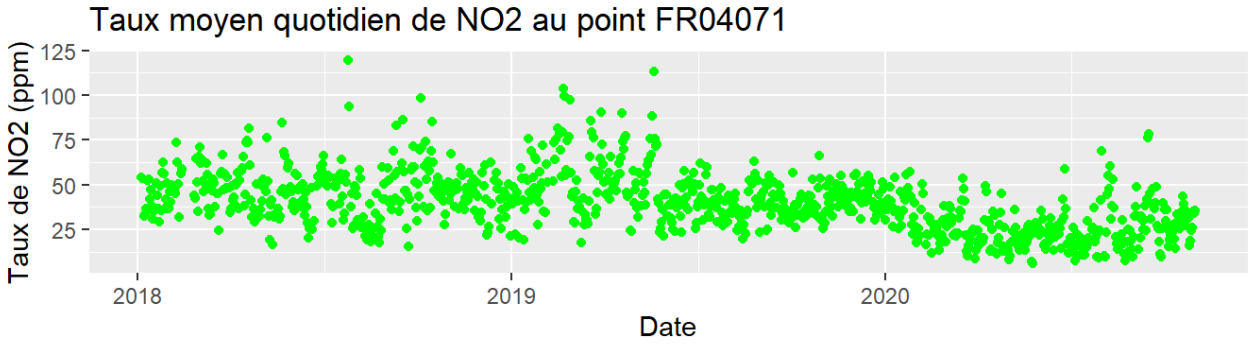


Figure 2.3: Outliers

Il est possible de remarquer sur le graphique ci-dessus (Figure 2.3) que Paris a toujours dépassé la limite de régulation autorisée du taux de NO2 de 21 ppm, on voit cependant que l'allure du graphique semble décroissante. Entre 2018 et 2020, le graphique indique une légère baisse de ce taux, en terme d'allure générale.

L'année 2020 marque une diminution plus marquée du taux moyen quotidien de NO2 à Paris. Ceci est accentué dès lors que le niveau 3 le plus sévère est atteint, en mars 2020. Cette décroissance nette est marquée par l'absence de valeurs extrêmes. Le relâchement du confinement au niveau 1 (le moins sévère) début juin 2020 s'accompagne d'une augmentation, à nouveau, du taux de NO2.

Somme toute, ceci indique déjà que le confinement semble avoir eu un impact sur le taux d'émission de NO2. Ceci dit, l'horizon temporel étant limité aux données à cette date, il faudrait idéalement accumuler davantage de données de confinement afin de garantir plus de robustesse des résultats.

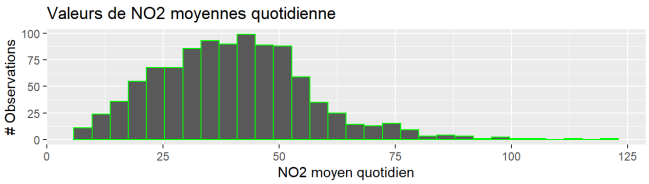
2.1.2 Catégories de variables

Les variables explicatives peuvent être divisées en 4 catégories: Variables de calendrier, Variables de décalage, Variable de climats et Variables sociales

2.1.2.1 Variables de calendrier

Inspiré de la revue de littérature, 5 variables sont ajoutées à notre base de données - en plus de la variable principale "Date". Elles permettront de contextualiser les observations et ainsi de mieux interpréter les résultats. La variable "day" indique le jour du mois dont il est question. Les variables "weekday" et "month" indiquent respectivement le jour de la semaine ainsi que le mois propre à l'observation. La variable "julian" représente le "id" de l'observation. Finalement, la variable "jour_ferie" est une variable binaire qui vaut 0 les jours ordinaires et 1 les jours fériés.

Figure 2.4: Histogramme moyenne



Il est posible de voir sur le graphique ci-contre (Figure 2.4) que le taux moyen de NO2 quotidien le plus fréquent au point de prélèvement FR04071 est de 37.5 ppm. Ces données confirment ce qui a été dit dans l'introduction, à savoir que la ville de Paris ne respecte pas les réglementation de 21 ppm. Le taux émis est quasiment le double!

2.1.2.2 Variables de décalage

Des variables représentant la variable dépendante à plusieurs moments dans le temps "lag" sont utilisées. Cette technique permet de considérer la dépendance temporelle dans les observations de la série. Jusqu'à présent, la variable cible NO2 n'a pas été filtrée en prenant en compte sa saisonnalité. En d'autres termes, une hausse de ce taux à un

certain moment de l'année pourrait ne pas être significative s'il s'agit d'une "tendance" quotidienne, hebdomadaire ou mensuelle. Afin de garantir des résultats fiables et interprétables, les variables décalées de 1, 7 et 28 jours sont utilisées.

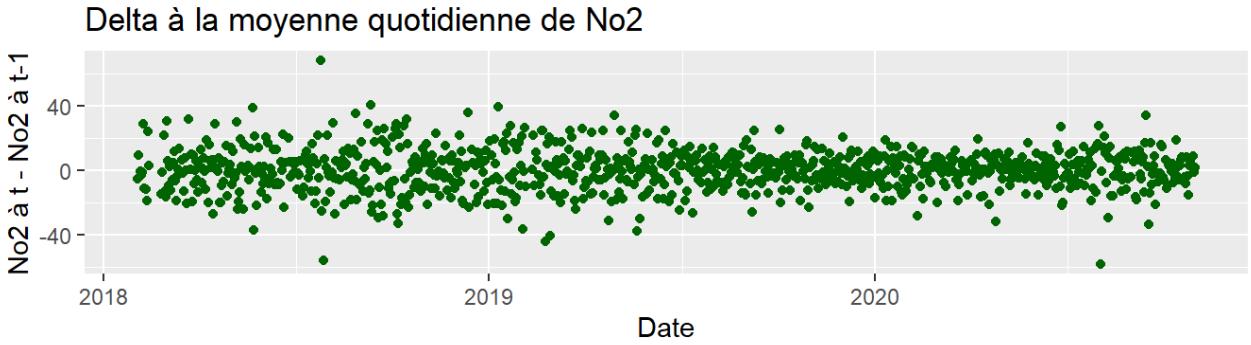


Figure 2.5: Delta moyenne

On remarque sur le graphique précédent (Figure 2.5) que la différence entre la moyenne de NO2 et la moyenne de No2 de la veille ne varie pas avec le temps. Cela illustre le fait que l'ajout de la variable de décalage stationnarise le problème.

2.1.2.3 Variables de climats

Le traitement des variables de climat se base sur les différentes études de modélisation d'un polluant (Voir Revue de littérature). La transformation de la variable de direction du vent est faite suivant les travaux de (Stoimenova et al. 2017) qui utilise la transformation suivante

$$WDI = 1 + \sin(Wind\_direction + \frac{\pi}{4})$$

D'autre part, concernant la variable de température, plusieurs auteurs comme (Hor, Watson, and Majithia 2005) ont utilisé des transformations de cette variable afin de prendre en considération la sévérité et la durée des cycles de température. Cette logique est utilisée ici afin de linéariser davantage cette variable. Ceci est fort pertinent puisqu'à titre d'exemple, le modèle de régression linéaire assume la linéarité entre les variables.

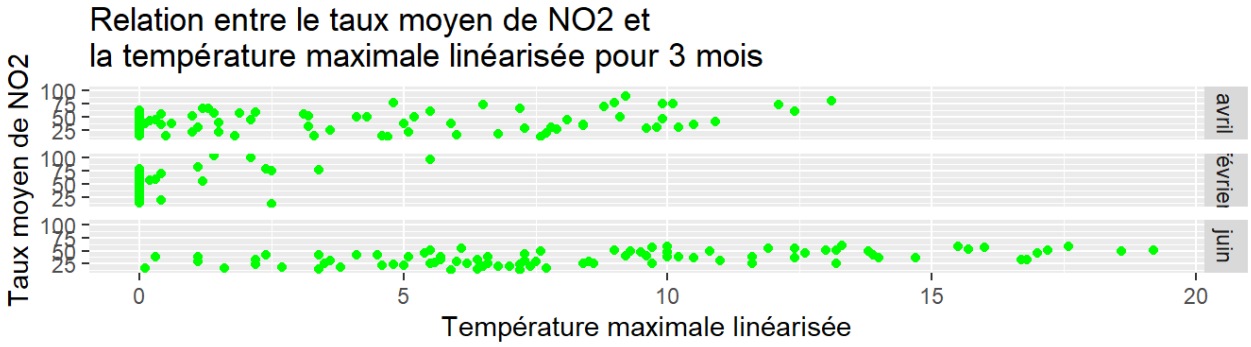


Figure 2.6: Temperature

Somme toute, on voit à la Figure 2.6 que l'augmentation de la température (maximale et minimale) induit une augmentation du taux moyen de NO2. De plus, les valeurs de températures nulles sont celles qui n'auront pas de pouvoir prédictif, mais nous ne nous attarderons pas sur ce point ici.

2.1.2.4 Variables sociales

Parmi les variables sociales, la variable principale à l'étude est le niveau de confinement associé à Paris : nous nous intéressons à sa significativité. Les variables de transport selon plusieurs catégories de déplacements seront aussi utilisées.

2.2 Méthodologie sur les méthodes considérées

Plusieurs modèles sont utilisés dans le cadre de notre projet, le but étant de trouver celui qui offre les meilleurs résultats.

Dans un premier temps, un modèle linéaire est considéré. Il s'agit du modèle de référence, qui s'inscrit dans l'optique d'un apprentissage supervisé. Il suppose une relation linéaire entre chacun des prédicteurs et la variable d'intérêt Y, soit le taux de NO2. Les termes d'erreurs suivent une loi normale centrée de moyenne 0 et sont indépendants. Ce modèle est entrainé deux fois: avec et sans la variable de confinement, puisque ce projet tente de souligner l'effet de cette variable en particulier.

Ensuite, une deuxième méthode d'apprentissage supervisé est considérée : la méthode du support vector classifier. Les kernels linéaires sont choisis puisqu'une multitude de **features** est à disposition. Aussi, le noyau linéaire permet de mieux généraliser comparativement à d'autres noyaux (polynomial, par exemple). Dans cette méthode, la validation croisée est employée afin de rechercher le paramètre de tuning C. Plus C augmente, plus la tolérance aux observations mal classifiées augmente.

Finalement, l'approche des forêts aléatoires est étudiée. En fait, comme la revue de littérature a pu le souligner, il s'agit d'une approche qui est très commune et sans doute une des plus puissantes pour la classification et l'étude des relations, et ce quel que soit le domaine étudié (environnemental, financier, médical etc.). L'idée globale est simple : une moyenne des valeurs prédites par différents modèles est souvent plus précise et plus stable que la valeur prédite à l'aide d'un seul modèle. De plus, cette méthode permet de prédire la valeur d'une variable cible en utilisant un nombre de "prédicteurs candidats", quelle que soit leur relation (linéaire ou pas) avec la variable cible. Somme toute, l'approche des forêts aléatoires permet de modéliser des liens compliqués entre une variable cible et des prédicteurs tout en assurant une interprétation simple!

Lors de l'implémentation de cette méthode, il s'agit de calibrer 3 paramètres essentiels au moyen de la validation croisée (Mei et al. 2014) : le nombre de prédicteurs candidats (mtry), le nombre d'arbres dans la forêt (ntree) et la taille minimale des noeuds terminaux (node_size). Dans la revue de littérature, ceux utilisés sont 500 arbres, 3 variables par split, et un minimum de 5 observations par noeud terminal. Dans notre cas, nous souhaitons élargir la recherche de paramètres afin d'inclure d'autres facteurs observés dans des travaux connexes. Ainsi, nous fixons le nombre d'arbres de la forêt à 300. Le nombre de prédicteurs candidats sera varié parmi une grille de valeurs entre 3 et 16, et la taille minimale des noeuds terminaux sera variée entre les valeurs de 3 et 7. Le modèle est créé avec le package RandomForest de R (Liaw and Wiener 2002) .

Puisque plusieurs modèles différents sont étudiés, le jeu de données est divisé en un ensemble d'apprentissage et un ensemble de test. Ceci permet, ultimement, de comparer les modèles entre eux avec un indicateur non biaisé de performance.

En résumé, trois familles de modèles différentes sont considérées dans notre étude. Dans la première famille de modèles, la régression *linéaire simple* est employée. Avec très peu de **fine tuning**, il s'agit de notre modèle de référence. Deuxièmement, la famille des *SVM* est testée, plus précisément celle avec un noyau linéaire, avec une validation croisée sur la paramètre de coût. Troisièmement, les modèles de *forêts aléatoire* sont utilisés.

L'ensemble de ces modèles sont entraînés et évalués au moyen de l'ensemble des prédicteurs à disposition hormis les 5 variables de mobilité: leur utilisation restreint la qualité de notre analyse, car elles ne sont disponibles qu'à partir du 15 février 2020. Leurs effets coïncident avec celui de la variable de confinement, dont les données sont disponibles dans les premiers mois de 2020 également. Puisque cette étude souhaite souligner l'effet du confinement principalement, les variables de mobilité seront dans un premier temps écartées. La méthode des forêt aléatoire, quant à elle, sera également testé avec le jeu de données complet (5 variables de mobilité inclues). Il serait intéressant de comparer les résultats obtenus avec et sans ces variables.

La variable cible étant une variable continue, l'erreur quadratique moyenne (RMSE) est la mesure considérée pour quantifier la performance des modèles et les comparer entre eux. Il s'agit du critère utilisé par (Gocheva-Ilieva et al. 2019), basé sur la vraisemblance du modèle.

2.2.1 Création de la partition

La partition est séparée en ensemble d'entrainement et de test, contenant respectivement 75% et 25% des observations. Trois partitions sont mise à disposition: la première enlève les 5 variables de mobilité, la deuxième enlève les 5 variables de mobilité ainsi que la variable de confinement et la troisième contient la totalité des variables (elle subit un traitement de données afin d'enlever les valeurs manquantes).

3 Résultats

Pour chaque modèle, trois graphiques présentent en ordre aux figures (3.1, 3.2), la comparaison des taux moyens observés à ceux prédits à partir du modèle entraîné, les résidus du modèle en fonction du temps, et l'importance des variables du modèle (lorsque cette option est possible). Le graphique de densité des résidus est également présenté pour les deux modèles de forêts aléatoires. Le sommaire de chaque modèle est disponible en annexe.

3.0.1 Modèle Linéaire - Avec variable de confinement

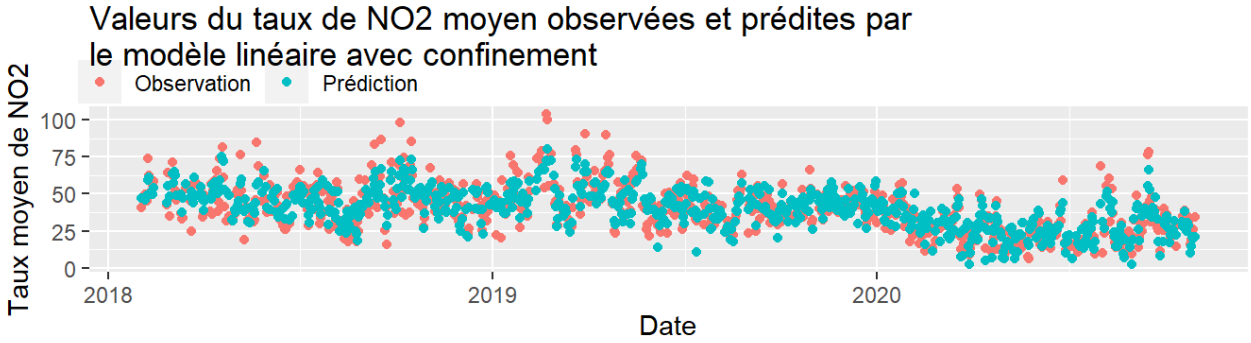


Figure 3.1: Modèle linéaire1

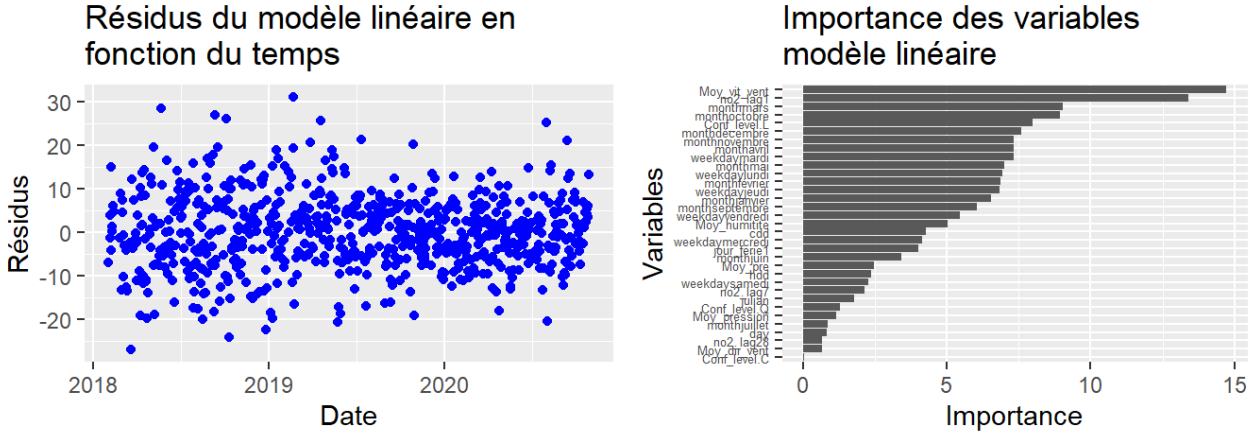


Figure 3.2: Modèle linéaire1

Les résultats sommaires du modèle linéaire sont disponible à l'annexe 6.0.0.1 Il est possible de voir le coefficient de détermination R^2 du modèle sur l'ensemble d'entrainement est de 72%. Il s'agit d'une valeur satisfaisante. Deux variables se démarquent en terme d'importance : la vitesse moyenne du vent (importance = 14.3) et la variable de lag sur 1 jour (importance = 14.1). En d'autres termes, c'est la valeur de la veille du taux de NO2 ainsi que la vitesse moyenne du vent qui influenceraient le plus sur la prédiction de la valeur actuelle - d'aujourd'hui - du taux de NO2.

D'autre part, le graphique indiquant les valeurs du taux moyen de NO2 prédites (Figure 3.1) versus mesurées est satisfaisant. Ces valeurs forment un ensemble homogène. On note cependant que le modèle linéaire n'est pas en mesure de bien prédire les valeurs se situant au-dessus de 80 ppm, soit les valeurs "exceptionnelles": le modèle linéaire ne permet de généraliser les valeurs extrêmes. Ceci dit, il est intéressant de noter que le modèle a bien prédit la baisse du taux de NO2 en 2020.

Par ailleurs, le graphique des résidus (Figure 3.2) obtenu confirme la validité des hypothèses du modèle linéaire. Les résidus semblent être répartis aléatoirement, mais de façon symétrique par rapport à l'axe 0 : les postulats de linéarité et d'homoscédasticité sont bien respectés. De plus, la majorité des résidus se situent entre + et - 20 ppm : ceci est convenable.

Le traitement des données a été crucial: l'introduction de variables de saisonnalité a permis de bien considérer la dépendance temportelle et les différentes étapes de linéarisation des données permettent d'obtenir un modèle linéaire satisfaisant.

Finalement, le RMSE pour ce modèle est de 8.51. À première vue, cette valeur semble convenable, bien que hors contexte. Elle sera interprétée plus tard, lorsque les modèles seront comparés entre eux, à partir de l'échantillon test.

3.0.2 Modèle Linéaire - Sans variable de confinement

Les graphiques résultats sont disponible aux figures 3.3 et 3.4 suivantes.

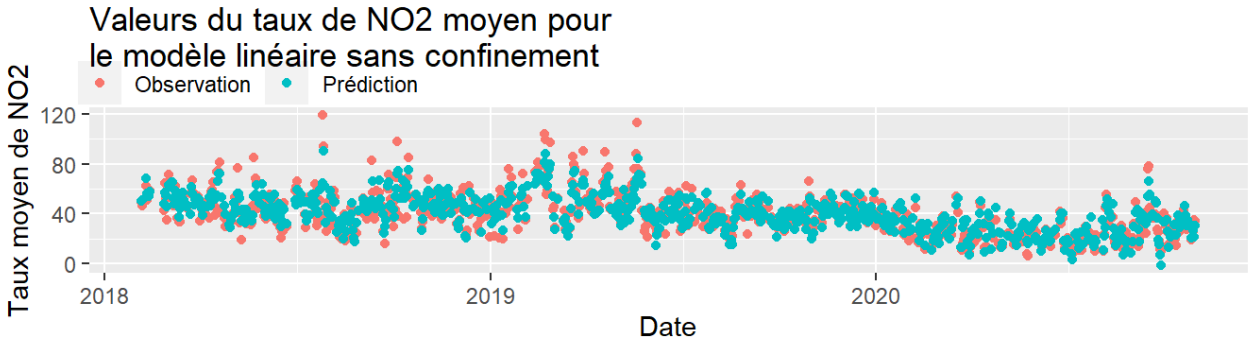


Figure 3.3: Modèle linéaire1

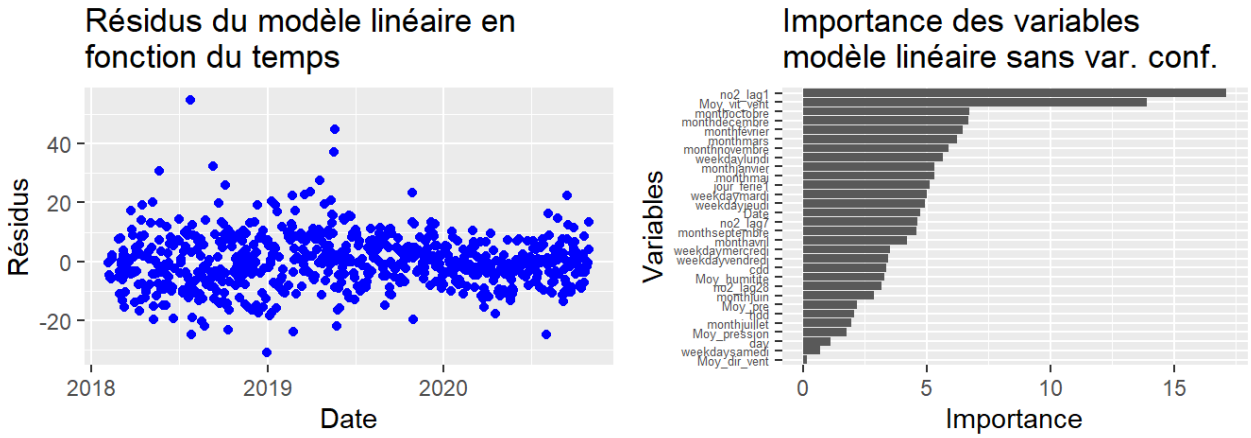


Figure 3.4: Importance

Il est possible de voir le sommaire des résultats pour ce modèle en annexe 6.0.0.2 Le $R^2 ajusté$ est de 71%. Les deux variables les plus importantes restent inchangées et se démarquent encore plus des autres : il s'agit de la vitesse moyenne du vent (importance = 11.8) et de la valeur de la veille du taux de NO2 (importance = 15.6). La même observation se fait sur le graphique des valeurs prédites versus mesurées de NO2 : le modèle linéaire ne permet de bien généraliser les valeurs extrêmes et s'adapte bien à la diminution du taux en 2020. Le graphique des résidus est saisfaisant.

Finalement, le RMSE pour ce modèle est de 9,16: la variable de confinement semble impacter très marginalement la qualité de la prédiction. Ceci sera approfondi par la suite, lors de la comparaison de modèles sur l'échantillon test. À présent, la variable de confinement est utilisée dans tous les modèles suivants.

3.0.3 SVM

Comme énoncé dans la méthodologie, la validation croisée effectuée afin de déterminer le paramètre C optimal donne une valeur de 300. Il s'agit d'une valeur assez grande : la frontière se situe loin des points, ce qui devrait engendrer un petit taux d'erreur. À titre indicatif, la validation croisée a été effectuée sur une grille de valeurs allant jusqu'à 1000.

Les graphiques résultats sont disponible dans les graphiques de la figure 3.5 ci-dessous

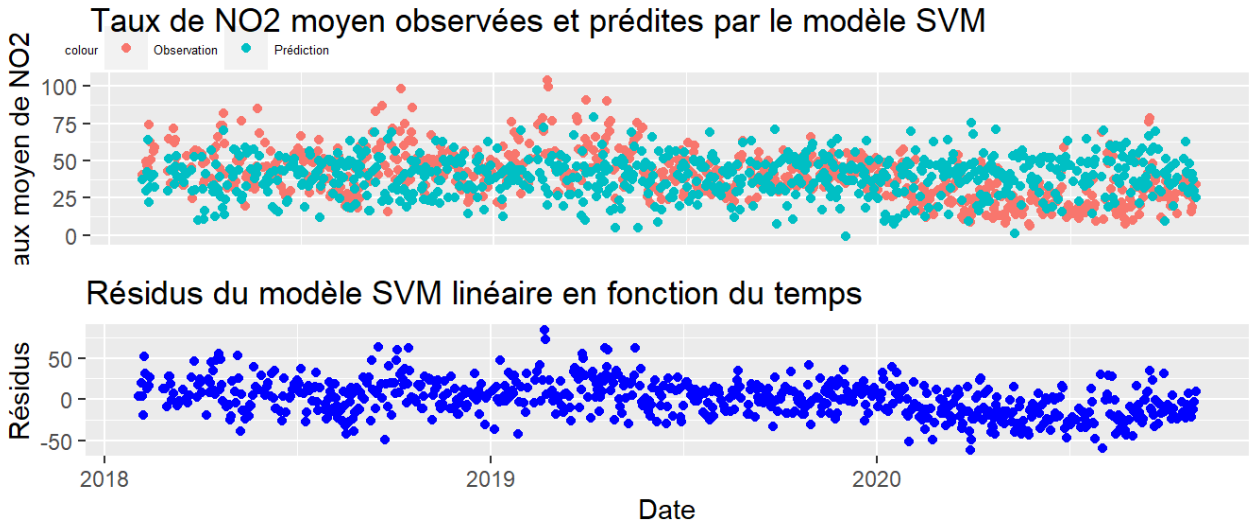


Figure 3.5: SVM

Plusieurs observations peuvent être faites à ce stade :

- Le RMSE obtenu est de 8,61.
- Le graphique des valeurs prédites versus observées du taux de NO2 moyen indique une moins bonne adéquation des données, particulièrement à partir de 2020. Le modèle SVM prédit mal les valeurs après 2020, ceci se reflète par la diminution de la tendance dans le graphique des résidus.
- Le modèle ne permet pas de bien généraliser les valeurs extrêmes, puisqu'au delà des valeurs mesurées de 75 ppm, les prédictions du modèle ne dépassent pas cette valeur (approximativement). Somme toute, il pourrait s'agit de valeurs extrêmes que notre traitement de données n'a pas pu éliminer.
- Contrairement aux modèles de régression ou de random forests, la méthode de validation croisée pour déterminer l'importance de chaque variable (fonction varImp) ne peut pas être réalisée sur un modèle SVC.
- Finalement, le postulat de linéarité n'est pas respecté car les résidus sont en moyenne inférieurs à 0 à partir de l'année 2020.

De plus, le sommaire du modèle est présenté à l'annexe 6.0.0.3

3.0.4 Forêts aléatoires - Modèle 1

La validation croisée est utilisée sur 2 paramètres : le nombres de prédicteurs utilisés et le nombre de noeuds terminaux.

Les graphiques résultats sont présent à aux figures 3.6, 3.7 et 3.8 ci-dessous:

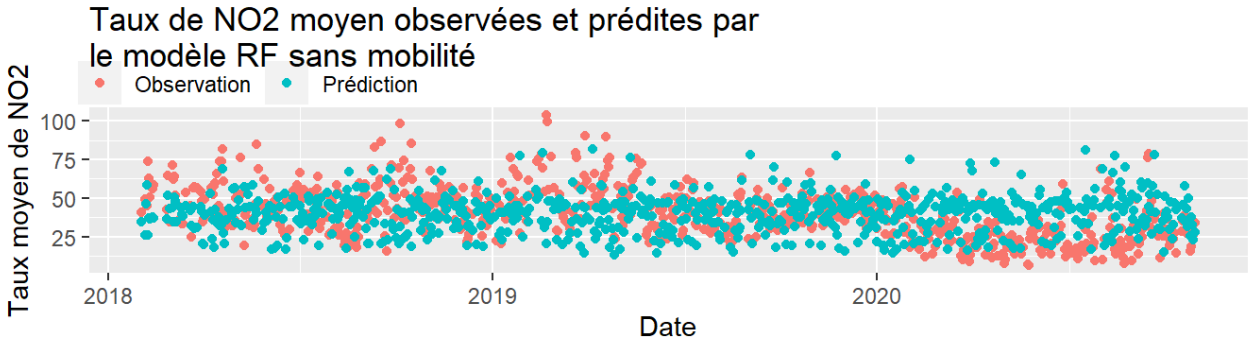


Figure 3.6: RF1

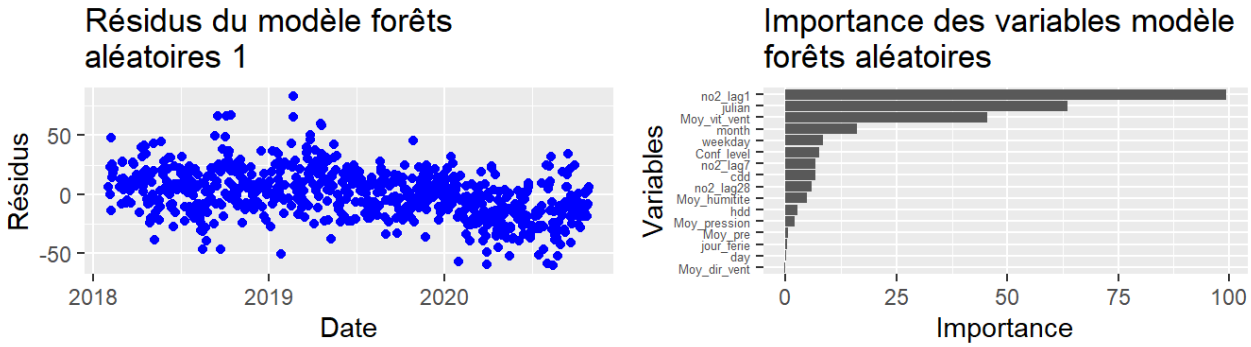


Figure 3.7: Importance RF1

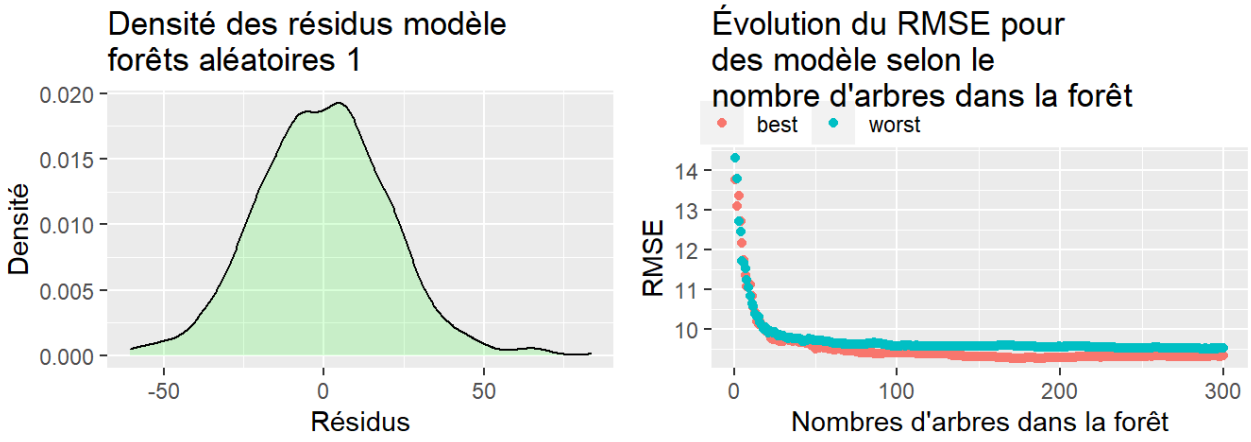


Figure 3.8: residus et RMSE RF1

Le sommaire du meilleur modèle est présenté à l'annexe 6.0.0.4. il est possible d'y voir que le meilleur modèle obtenu est celui pour lequel 11 prédicteurs sont utilisés pour chaque split. Il renvoie une valeur de RMSE de 9,25.

Le graphique obtenu des valeurs prédites versus observées est intéressant : on voit que les valeurs prédites sont plus "centrées" que les valeurs observées. Les valeurs extrêmes observées,soit les taux moyens inférieures à 30ppm avant 2020 et supérieures à 75ppm, ne sont pas généralisées par notre modèle de prédiction.

La courbe de densité des résidus tracée traduit la normalité des résidus. Par contre, similairement au modèle SVC, le postulat de linéarité ne semble pas respecté car les résidus sont en moyenne inférieurs à 0 à partir de 2020.

Finalement, un graphique illustrant l'évolution de la RMSE entre le pire et le meilleur modèle obtenu par validation croisée en fonction du nombre d'arbres est tracé (figure 3.8. Plus le nombre d'arbres augmente, plus la valeur de la RMSE diminue, avant d'atteindre un plateau. Le nombre choisi de 300 arbres est donc plus que suffisant, puisque la RMSE se stabilise autour de 100 arbres. Il s'agit de trouver un bon équilibre entre la robustesse de la méthode et le risque de surapprentissage.

Les variables les plus importantes sont "no2_lag1" (figure 3.7, soit la valeur du taux de NO2 moyen de la veille (importance = 99) et "julian" (importance = 63). La variable de confinement est 6ème en terme d'importance: le confinement a un effet notable sur la prédiction du taux de NO2.

3.0.5 Forêts aléatoires - Modèle 2

Dans ce dernier modèle considéré, les 5 variables de mobilités sont incluses. C'est le seul modèle qui comporte ces variables. Les graphiques résultats sont disponible aux figures 3.9 et 3.10 ci-dessous:

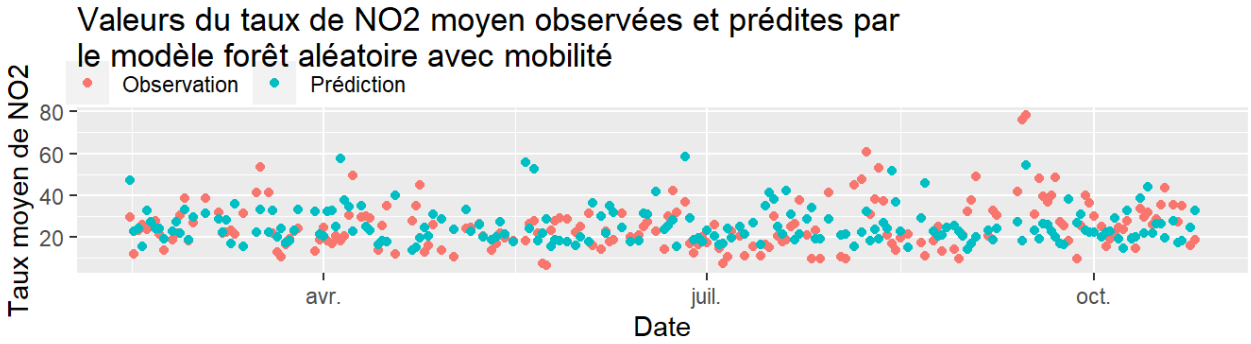


Figure 3.9: RF2

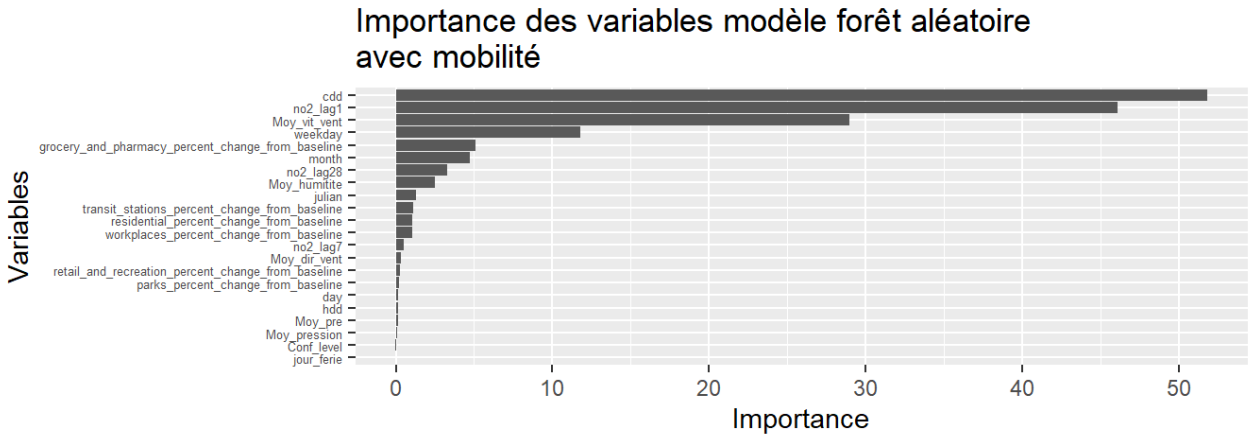


Figure 3.10: RF2_2

Le sommaire du meilleur modèle est disponible à l'annexe 6.0.0.5 Pour ce modèle, la validation croisée indique que le nombre optimal de prédicteur utilisé à chaque split est de 13, et renvoie une valeur de RMSE de 7,24. Il s'agit de la meilleure valeur obtenue parmi l'ensemble des modèles considérés! Les variables les plus importantes sont la température minimale (importance = 51,8) et la valeur du taux de NO2 de la veille (importance = 46).

Le graphique des prédictions versus observations indique que le modèle semble mieux performer puisqu'il ne semble pas démontrer de lacune pour la période critique de 2020 comme les autres modèles précédents. Le modèle parvient donc à correctement généraliser les observations, avec encore une légère différence pour les valeurs les plus extrêmes. Finalement, l'ensemble des graphiques des résidus (annexes) confirment que les postulats de linéarité et d'homoscédasticité sont respectés. Il est intéressant de noter que l'allure des résidus varie en fonction du jour considéré : si la normalité semble parfaite pour les jours du mercredi et du jeudi, elle semble légèrement moins normale pour les jours du vendredi et du dimanche.

Les variables de mobilité sont donc très importantes quant à la prédiction du taux de NO2 : le taux de NO2 émis par les véhicules est directement lié à la circulation des voitures. En fait, il s'agit quasiment de la même variable, exprimée différemment! De par la colinéarité de ces variables avec la variable cible ainsi que leurs effets qui coïncident avec celui du confinement.

3.0.6 Comparaison des différents modèles

	lineaire_avec_confinement	**lineaire_sans_confinement**	**SVM**	**RF_sans_mobilite**	**RF_avec_mobilite**
RMSE_train	8.51	9.16	8.61	9.25	7.24
RMSE_test	10.14	9.39	10.06	10.56	7.03
variables_importantes	vitesse moyenne du vent et valeur de la veille du taux de NO2	vitesse moyenne du vent et valeur de la veille du taux de NO2	•	valeur de la veille du taux de NO2	température minimale et valeur de la veille du taux de NO2

voir annexe 6.0.0.6 pour explication du message d'erreur
L'ensemble des modèles ont un RMSE de test supérieur au RMSE d'entraînement: la performance est surévaluée sur l'échantillon d'entraînement. Il est possible de noter que le modèle de forêts aléatoires avec variables de mobilité a une erreur test plus faible que lors de la période d'entrainement. Cette différence est non significative et peut être causée par une complexité moindre pour cette période.

D'après le critère de sélection du RMSE de test, le modèle des forêts aléatoires contenant les six variables de mobilité et la variable d'indice de confinement est le plus performant avec une RMSE de 7,03. De plus, ce modèle a aussi le RMSE d'entraînement le plus faible. Par rapport au modèle de référence (régression linéaire avec variable de confinement), le RMSE test est réduit de 30%. Le fait d'ajouter les variables de mobilité dans le modèle des forêts aléatoires améliore la performance du modèle de 33% (passe de 10,56 à 7.03): ces variables de mobilité ont un effet notable sur le RMSE test.

Comme le modèle est sujet à être actualisé avec de nouvelles données, un point faible des forêts aléatoires est sa difficulté à interpréter des données entrantes qui n'ont jamais été "vue" par le modèle dans la phase d'entraînement. Ainsi, si la France invoque un niveau de confinement 4 (jusqu'ici un maximum de 3), le modèle performera très mal.

4 Analyse des variables

L'analyse de l'importance des variables du modèle de forêts aléatoires incluant les variables de mobilité indique les effets qui influencent le plus sur la concentration de NO2 à Paris. Sur les 22 variables utilisées par le modèle, les 5 variables principales associées à leur augmentation en terme de MSE, sont les suivantes:

- la température minimale (52%)
- la valeur du taux de NO2 de la veille (46%)
- la vitesse moyenne du vent (29%)
- la journée de la semaine (12%)
- la variation des déplacements vers la pharmacie et l'alimentation (5%)

Des graphiques sont présenté en annexe Figures (6.1,6.2,6.3,6.4) illustrant les relations spécifiques entre ces variables et la variable cible.

Ces variables forment deux catégories. La première porte sur les variables hors du contrôle humain. Elle regroupe la température minimale et la vitesse du vent. La deuxième catégorie, à l'inverse, regroupe les variables sous contrôle humain, dont la journée de la semaine et la variation des déplacements vers la pharmacie et l'alimentation. Finalement, la valeur du taux de NO2 de la veille ne peut pas réellement être classée dans une des catégories.

La variable qui influence le plus la concentration de NO2 est la température minimale. Lorsque les températures sont plus froides, il est probable que plus de personnes prennent la voiture ou le bus pour se déplacer, ce qui contribue à augmenter la concentration de NO2. Le froid a également pour conséquence la hausse de la demande de chauffage, qui entraîne une hausse de l'utilisation des chaudières au gaz ou au fioul.

L'augmentation de la concentration de NO2 en hiver liée à l'augmentation de l'utilisation des véhicules à combustion interne n'est cependant pas la seule explication. En effet, l'action du rayonnement solaire enclenche des réactions chimiques, avec l'O2, qui transforment le NO2 en monoxyde d'azote (NO) et en ozone (O3). Le *Guide d'estimation de la concentration de dioxyde d'azote (NO2) dans l'air ambiant lors de l'application des modèles de dispersion atmosphérique* indique par ailleurs que "ces réactions sont influencées par les facteurs météorologiques tels que l'intensité du rayonnement solaire, la température et la vitesse du vent"(Couture, Y 2008). Une étude québécoise sur les *Effets du dioxyde d'azote et de l'ozone sur les maladies respiratoires à Montréal*, a notamment montré que "les maximums de NO2 sont observés en février et mars, correspondant plutôt à des températures froides" et que "les concentrations moyennes journalières de l'O3 les plus élevées sont observées en été (juin, juillet et août)"(Caouette 2010).

La concentration de NO2 de la veille est la deuxième variable en importance qui influence la concentration de NO2 de la journée. Cela montre que le NO2 a la caractéristique de s'accumuler dans l'air d'une journée à l'autre sans conditions météorologiques particulières (pluie ou vent). Ainsi, appliquer à Paris des mesures de courte durée de réduction de l'utilisation de la voiture ne permettrait pas de réduire la concentration de NO2 dans l'air.

La dernière variable météorologique est la vitesse du vent. Une journée avec un vent fort a en moyenne une concentration de NO2 plus faible qu'une journée avec un vent faible (annexe). Cette observation est logique car puisque la variable cible est le taux de NO2 dans l'air, le vent fort empêche l'accumulation de NO2.

Les deux dernières variables ne sont pas causées par des phénomènes météorologiques. Elles reflètent donc l'impact de la variation des habitudes d'utilisation de la voiture sur le taux de pollution. Cependant, ces variables une importance beaucoup plus faible que les trois autres variables. Ainsi, la concentration de NO2 est plus faible la fin de semaine que dans les jours de semaine. En effet, la concentration atteint respectivement en moyenne 37.5ppm et 32.8 ppm le samedi et le dimanche, alors qu'elle dépasse les 39ppm tous les jours de la semaine, atteignant près de 43 ppm en moyenne les mardi et vendredi. Ceci est du aux déplacements pour aller au travail la semaine. Ceci dit, malgré le confinement complet du printemps 2020, la concentration de NO2 ne semble pas diminuer avec la diminution des déplacements en voiture vers le lieu de travail. La variation des déplacements vers la pharmacie et l'alimentation exerce quant à elle une certaine influence sur le taux de NO2. Il semble que moins les personnes se déplacent pour aller à la pharmacie ou à l'épicerie, plus le taux de NO2 est faible (annexe).

L'analyse de l'importance des variables montre que la variation de l'utilisation de la voiture a un effet négligeable sur la concentration de NO2 comparativement aux variables météorologiques de température et de vent. De ce fait, les mesures de confinement qui limitent l'utilisation de la voiture auraient un faible impact sur la concentration de NO2.

Cette observation va dans le même sens des conclusions d'une étude française de 1994 sur l'influence du trafic et des conditions météorologiques à Paris. Cette recherche conclue que "les mesures de restriction de circulation éventuelles auraient un effet plutôt réduit sur les teneurs en NO2". L'étude conclue de plus que "la pente [de la régression linéaire] est nettement plus forte entre le trafic et NO que celle obtenue pour NO2" ce qui fait du "monoxyde d'azote NO est un très bon indicateur de la pollution d'origine automobile"(Alary, Donati, and Viellard 1994).

5 Conclusion

Ce projet cherchait à étudier l'effet du confinement à Paris, à travers la réduction de l'utilisation de la voiture, sur le taux de NO2 émis. Pour ce faire, l'étude a porté sur la période du 2 février 2018 au 30 octobre 2020. Les mesures de NO2 ont été captées quotidiennement par des stations localisées dans Paris, tandis que l'indice de confinement a été évalué sur cette période à partir des annonces du gouvernement. Les variables de mobilité ne portaient que sur la période réduite du 15 février 2020 au 30 octrobre 2020. Les effets de temps et de la météo ont également été captés par l'ajout de variables supplémentaires. Une fois les données traitées, plusieurs modèles de prédiction du taux quotidien de NO2 ont été entraînés et comparés. Par rapport au modèle linéaire de base, le modèle de prédiction le plus performant est le modèle de forêts aléatoires contenant les variables de mobilité. Il a été 30% plus performant avec un RMSE sur l'échantillon test de 7.03.

Finalement, grâce à l'analyse des variables qui influencent le plus le taux de NO2, il est apparu que les variables qui indiquent une variation de l'utilisation de la voiture ont un effet négligeable sur la variable cible comparativement aux variables météorologiques. Par conséquent, les mesures de confinement qui ont restreints la circulation à Paris ont eu un effet négligeable sur le taux de NO2.

Ce projet a permis de mettre en évidence que le monoxyde d'azote NO est un meilleur indicateur de la pollution provenant des véhicules que ne l'est le NO2. Il serait intéressant dans le futur de tester la relation entre les mesures de confinement et le taux de NO.

6 Annexes

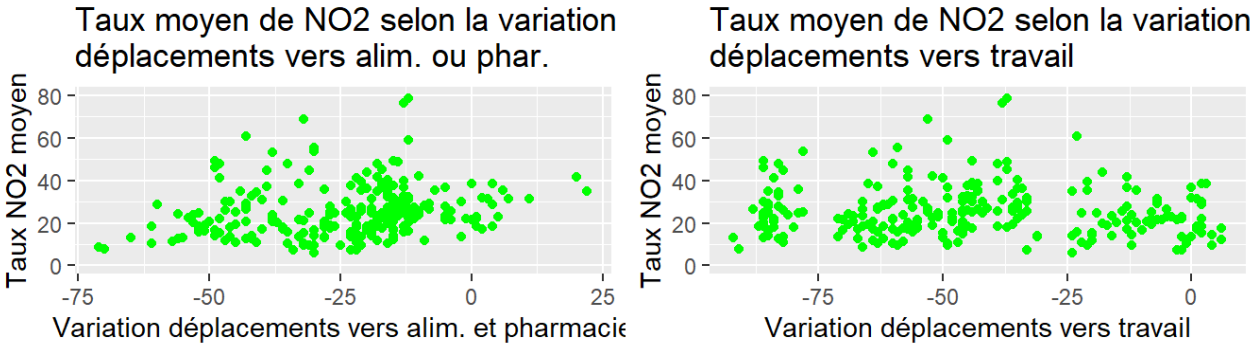


Figure 6.1: Alimentation et travail

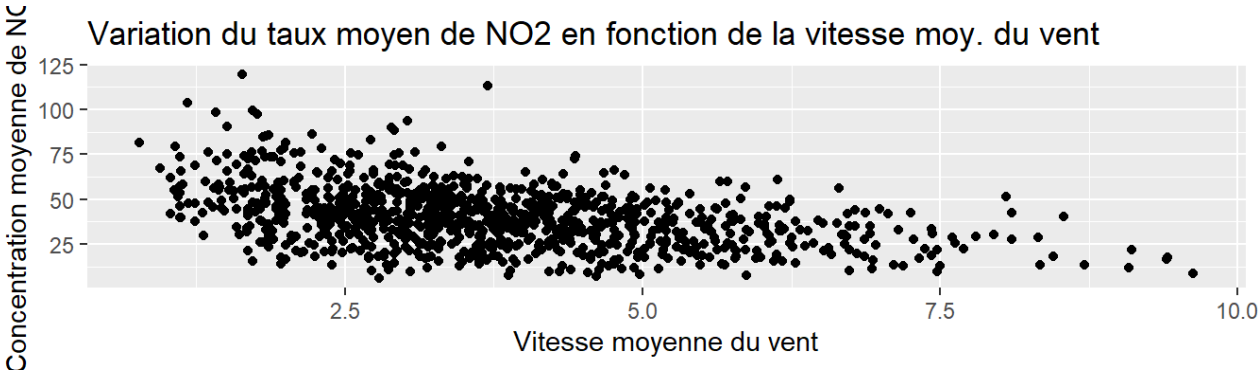


Figure 6.2: No2 et vent

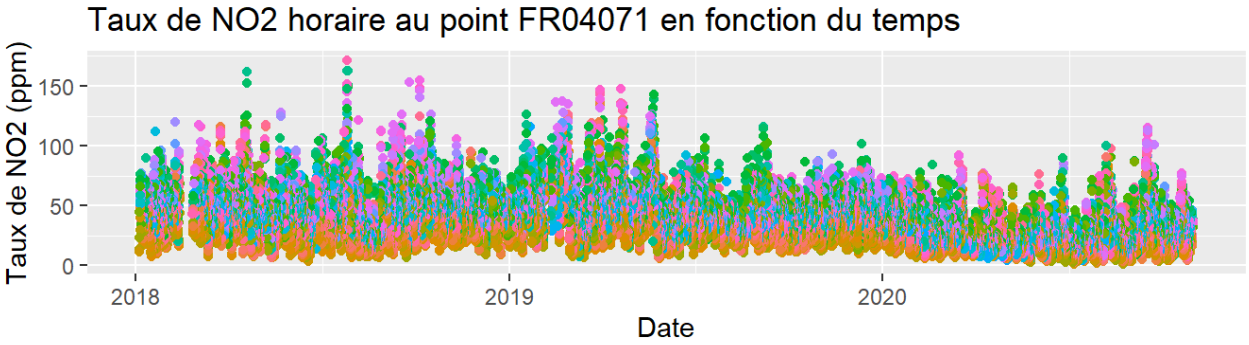


Figure 6.3: FR04071 et temps

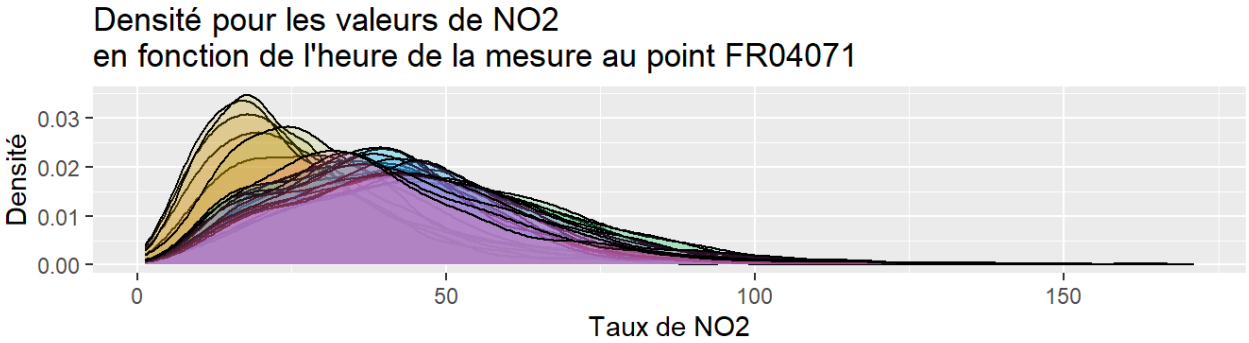


Figure 6.4: Densité et heure

6.0.0.1 Sommaire du modèle linéaire

```
##
## Call:
## lm(formula = mean_no2 ~ ., data = train_df_1 %>% select(-Date))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.0744  -5.2099   0.0819   5.1622  31.1031
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.894e+01  4.592e+01   1.719 0.086015 .
## Moy_dir_vent    -2.992e-01  4.540e-01  -0.659 0.510119
## Moy_pre          1.561e+00  6.328e-01   2.467 0.013851 *
## Moy_vit_vent    -3.683e+00  2.500e-01 -14.731  < 2e-16 ***
## Moy_humitite    -2.114e-01  4.212e-02  -5.019 6.60e-07 ***
## Moy_pression    -5.140e-04  4.466e-04  -1.151 0.250145
## Conf_level.L    -1.172e+01  1.469e+00  -7.975 6.32e-15 ***
## Conf_level.Q     1.712e+00  1.332e+00   1.285 0.199104
## Conf_level.C     5.395e-02  1.378e+00   0.039 0.968769
## day              3.075e-02  3.758e-02   0.818 0.413481
## weekdayjeudi     8.392e+00  1.230e+00   6.824 1.94e-11 ***
## weekdaylundi     8.225e+00  1.186e+00   6.936 9.28e-12 ***
## weekdaymardi     8.882e+00  1.214e+00   7.313 7.24e-13 ***
## weekdaymercredi  5.226e+00  1.266e+00   4.128 4.11e-05 ***
## weekdaysamedi    2.804e+00  1.231e+00   2.277 0.023085 *
## weekdayvendredi  6.857e+00  1.260e+00   5.442 7.32e-08 ***
## monthavril       1.399e+01  1.912e+00   7.318 7.04e-13 ***
## monthdécembre    1.678e+01  2.214e+00   7.579 1.13e-13 ***
## monthfévrier     1.578e+01  2.293e+00   6.881 1.33e-11 ***
## monthjanvier     1.516e+01  2.320e+00   6.534 1.24e-10 ***
## monthjuillet     1.284e+00  1.487e+00   0.863 0.388215
## monthjuin        5.170e+00  1.521e+00   3.400 0.000712 ***
## monthmai         1.243e+01  1.776e+00   7.001 6.03e-12 ***
## monthmars        1.866e+01  2.064e+00   9.043  < 2e-16 ***
## monthnovembre    1.608e+01  2.190e+00   7.340 6.01e-13 ***
## monthoctobre     1.644e+01  1.837e+00   8.950  < 2e-16 ***
## monthseptembre   9.439e+00  1.558e+00   6.057 2.28e-09 ***
## julian          -3.540e-03  2.006e-03  -1.765 0.078071 .
## jour_ferie1     -7.982e+00  1.996e+00  -3.999 7.04e-05 ***
## no2_lag1         3.680e-01  2.744e-02  13.411  < 2e-16 ***
## no2_lag7         5.085e-02  2.390e-02   2.128 0.033720 *
## no2_lag28       -1.646e-02  2.459e-02  -0.669 0.503458
## hdd              3.416e-01  1.435e-01   2.380 0.017587 *
## cdd              5.521e-01  1.288e-01   4.285 2.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.514 on 690 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.7207
## F-statistic: 57.52 on 33 and 690 DF,  p-value: < 2.2e-16
```

6.0.0.2 Sommaire du modèle linéaire sans variable confinement

```
##
## Call:
## lm(formula = mean_no2 ~ ., data = train_df_11)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -30.875  -5.337  -0.200   5.052  54.747
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.403e+02  5.636e+01   4.264 2.28e-05 ***
## Date            -7.714e-03  1.634e-03  -4.721 2.84e-06 ***
## Moy_dir_vent     7.599e-02  4.951e-01   0.154 0.878048
## Moy_pre          1.433e+00  6.582e-01   2.177 0.029817 *
## Moy_vit_vent    -3.658e+00  2.635e-01 -13.885  < 2e-16 ***
## Moy_humitite    -1.433e-01  4.340e-02  -3.302 0.001009 **
## Moy_pression    -8.005e-04  4.539e-04  -1.764 0.078211 .
## day              4.479e-02  3.992e-02   1.122 0.262272
## weekdayjeudi     6.557e+00  1.327e+00   4.940 9.80e-07 ***
## weekdaylundi     7.694e+00  1.362e+00   5.650 2.35e-08 ***
## weekdaymardi     6.472e+00  1.290e+00   5.018 6.64e-07 ***
## weekdaymercredi  4.599e+00  1.314e+00   3.499 0.000497 ***
## weekdaysamedi    9.134e-01  1.305e+00   0.700 0.484049
## weekdayvendredi  4.589e+00  1.330e+00   3.450 0.000594 ***
## monthavril       8.044e+00  1.907e+00   4.217 2.80e-05 ***
## monthdécembre    1.523e+01  2.279e+00   6.680 4.92e-11 ***
## monthfévrier     1.540e+01  2.383e+00   6.461 1.95e-10 ***
## monthjanvier     1.282e+01  2.415e+00   5.310 1.48e-07 ***
## monthjuillet     3.143e+00  1.608e+00   1.954 0.051078 .
## monthjuin        4.610e+00  1.619e+00   2.848 0.004529 **
## monthmai         9.188e+00  1.731e+00   5.306 1.51e-07 ***
## monthmars        1.296e+01  2.077e+00   6.239 7.65e-10 ***
## monthnovembre    1.329e+01  2.258e+00   5.883 6.27e-09 ***
## monthoctobre     1.270e+01  1.892e+00   6.714 3.94e-11 ***
## monthseptembre   7.699e+00  1.685e+00   4.569 5.81e-06 ***
## julian                 NA         NA      NA       NA
## jour_ferie1     -1.066e+01  2.091e+00  -5.097 4.45e-07 ***
## no2_lag1         4.612e-01  2.694e-02  17.119  < 2e-16 ***
## no2_lag7         1.170e-01  2.534e-02   4.616 4.66e-06 ***
## no2_lag28        7.852e-02  2.490e-02   3.154 0.001682 **
## hdd              3.209e-01  1.566e-01   2.049 0.040822 *
## cdd              4.430e-01  1.319e-01   3.358 0.000828 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.169 on 693 degrees of freedom
## Multiple R-squared:  0.721,  Adjusted R-squared:  0.7089
## F-statistic: 59.69 on 30 and 693 DF,  p-value: < 2.2e-16
```

6.0.0.3 Sommaire du modèle SVM

```
##
## Call:
## best.svm(x = mean_no2 ~ ., data = train_df_1 %>% select(-Date), cost = c(10^(-4:2),
##     seq(150, 300, 50)), kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  eps-regression
##  SVM-Kernel:  linear
##        cost:  300
##       gamma:  0.02941176
##     epsilon:  0.1
##
##
## Number of Support Vectors:  595
```

6.0.0.4 Sommaire du modèle 1 de forêt aléatoire

```
##
## Call:
##  randomForest(formula = mean_no2 ~ ., data = train_df_1 %>% select(-Date),      ntree = 300, mtry =
tune_grid$mtry[i], node_size = tune_grid$node_size[i],      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 300
## No. of variables tried at each split: 11
##
##           Mean of squared residuals: 86.78139
##                     % Var explained: 66.59
```

6.0.0.5 Sommaire du modèle 2 de forêt aléatoire avec variables mobilité

```
##
## Call:
##  randomForest(formula = mean_no2 ~ ., data = train_df_2 %>% select(-Date),      ntree = 300, mtry =
tune_grid2$mtry[i], node_size = tune_grid2$node_size[i],      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 300
## No. of variables tried at each split: 13
##
##           Mean of squared residuals: 52.37937
##                     % Var explained: 66.31
```

6.0.0.6 Message d'erreur

Un message d'avertissement apparait lors de la prédiction pour les deux modèles de régression linéaire ("prediction from a rank-deficient fit may be misleading") : la variable "julian" est entièrement corrélée avec une ou plusieurs autres variables présentes dans le modèle. Il serait pertinent de la retirer des modèles dans un travail futur.

References

Alary, R., J. Donati, and H. Viellard. 1994. “La Pollution Automobile à Paris. Influence Du Trafic et Des Conditions Météorologiques.” *Pollution Atmosphérique*, no. 141: 55–65.

Burrows, William R, Mario Benjamin, Stephen Beauchamp, Edward R Lord, Douglas McCollor, and Bruce Thomson. 1995. “CART Decision-Tree Statistical Analysis and Prediction of Summer Season Maximum Surface Ozone for the Vancouver, Montreal, and Atlantic Regions of Canada.” *Journal of Applied Meteorology* 34 (8): 1848–62.

Caouette, C. 2010. “Effets du dioxyde d’azote, de l’ozone et de la météorologie sur les maladies respiratoires à Montréal.” Master’s thesis, Institut national de la recherche scientifique, Centre Eau Terre Environnement.

Choi, W, SE Paulson, J Casmassi, and AM Winer. 2013. “Evaluating Meteorological Comparability in Air Quality Studies: Classification and Regression Trees for Primary Pollutants in California’s South Coast Air Basin.” *Atmospheric Environment* 64. Elsevier: 150–59.

Couture, Y. 2008. *Guide d’estimation de La Concentration de Dioxyde d’azote (No2) Dans L’air Ambiant Lors de L’application Des Modèles de Dispersion Atmosphérique*. Québec: Ministère du Développement durable, de l’Environnement et des Parcs, Direction du suivi de l’état de l’environnement.

Dudek, Grzegorz. 2015. “Short-Term Load Forecasting Using Random Forests.” In *Intelligent Systems’ 2014*, 821–28. Springer.

Elayan, E, F Giri, E Pigeon, and JF Massieu. 2006. “Ozone Concentration Modeling Using a Fuzzy Model over the Region of Basse-Normandie.” In *IFAC Symposium (Sysid’06)-Newcastle*, 28–31.

Forster, P. M., H. I. Forster, and M. J. Evans. 2020. “Current and Future Global Climate Impacts Resulting from Covid-19.” *Nature Climate Change* 10: 913–19.

Gocheva-Ilieva, Snezhana Georgieva, Desislava Stoyanova Voynikova, Maya Plamenova Stoimenova, Atanas Valev Ivanov, and Iliycho Petkov Iliev. 2019. “Regression Trees Modeling of Time Series for Air Pollution Analysis and Forecasting.” *Neural Computing and Applications* 31 (12). Springer: 9023–39.

Hor, Ching-Lai, Simon J Watson, and Shanti Majithia. 2005. “Analyzing the Impact of Weather Variables on Monthly Electricity Demand.” *IEEE Transactions on Power Systems* 20 (4). IEEE: 2078–85.

Kane, Michael J, Natalie Price, Matthew Scotch, and Peter Rabinowitz. 2014. “Comparison of Arima and Random Forest Time Series Models for Prediction of Avian Influenza H5n1 Outbreaks.” *BMC Bioinformatics* 15 (1). Springer: 276.

Kaya, Kiymet, and Şule Gündüz Öğüdücü. 2020. “Deep Flexible Sequential (Dfs) Model for Air Pollution Forecasting.” *Scientific Reports* 10 (1). Nature Publishing Group: 1–12.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/ (https://CRAN.R-project.org/doc/Rnews/).

Mei, Jie, Dawei He, Ronald Harley, Thomas Habetler, and Guannan Qu. 2014. “A Random Forest Method for Real-Time Price Forecasting in New York Electricity Market.” In *2014 Ieee Pes General Meeting| Conference & Exposition*, 1–5. IEEE.

Pórtoles, Javier, Camino González, and Javier M Moguerza. 2018. “Electricity Price Forecasting with Dynamic Trees: A Benchmark Against the Random Forest Approach.” *Energies* 11 (6). Multidisciplinary Digital Publishing Institute: 1588.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).

Stoimenova, M, D Voynikova, A Ivanov, S Gocheva-Ilieva, and I Iliev. 2017. “Regression Trees Modeling and Forecasting of Pm10 Air Pollution in Urban Areas.” In *AIP Conference Proceedings*, 1895:030005. 1. AIP Publishing LLC.