

UNIVERSITÉ DE BORDEAUX

MASTER 1 — 2024/2025

UE Projet d'expertise

---

## Rapport de TER

**Le Boosting des Prédicteurs Statistiques et son Application à la  
Prédiction de l'Ozone**

---

**Réalisé par :**

Marwa Dades

Amélie de Maillard

Mathis Rita

**Encadrant :**

Vincent Coualier

## Table des matières

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Introduction aux Forêts Aléatoires . . . . .	2
1.2	Définition du Boosting . . . . .	2
1.3	Selection de variables et approche itérative : cadre théorique . . . . .	4
1.4	Contexte et Problématique . . . . .	4
1.5	Objectifs du travail . . . . .	5
<b>2</b>	<b>Méthodes de Régression et Théorie du Boosting</b>	<b>6</b>
2.1	Rappels sur les Modèles de Régression . . . . .	6
2.2	Méthodes standards : Linéaire simple et Lasso . . . . .	7
2.2.1	Méthode de régression linéaire multiple . . . . .	7
2.2.2	Méthode de régression pénalisée Lasso . . . . .	8
2.3	Algorithmes classiques de Boosting : Gradient Boosting et AdaBoost . . . . .	9
2.3.1	Gradient Boosting . . . . .	9
2.3.2	AdaBoost . . . . .	9
<b>3</b>	<b>Présentation des Données et Prétraitement</b>	<b>11</b>
3.1	Description du Jeu de Données . . . . .	11
3.2	Préparation, Nettoyage et Prétraitement des Données . . . . .	11
<b>4</b>	<b>Implémentation des Méthodes et Comparaison</b>	<b>12</b>
4.1	Présentation des algorithmes . . . . .	12
4.1.1	Régression linéaire multiple . . . . .	12
4.1.2	Régression pénalisée Lasso . . . . .	13
4.1.3	Méthode type Gradient Boosting . . . . .	18
4.2	Analyse des résultats et comparaisons des modèles . . . . .	20
4.2.1	Méthode de régression multiple . . . . .	20
4.2.2	Méthode de régression pénalisée . . . . .	29
4.2.3	Méthode type boosting . . . . .	38
4.2.4	Comparaison des résultats avec jeu de données complet . . . . .	47
4.2.5	Comparaison des résultats avec jeu de données réduit . . . . .	49
<b>5</b>	<b>Avantages et limites</b>	<b>52</b>
<b>6</b>	<b>Conclusion</b>	<b>54</b>
<b>7</b>	<b>Bibliographie et Références</b>	<b>56</b>

# 1. Introduction

## 1.1. Introduction aux Forêts Aléatoires

Parmi les méthodes d'agrégation de modèles prédictifs, les forêts aléatoires (*Random Forests*) occupent une place essentielle, notamment en tant qu'exemple emblématique de la technique de bagging. Dans l'article écrit par Leo Breiman en 2001 [3], cette méthode repose sur la construction d'un ensemble de classificateurs, chacun construit à partir d'un échantillon des données d'apprentissage, et injectant de l'aléa supplémentaire par la sélection aléatoire de sous-ensembles de variables explicatives à chaque division de nœud.

Plus formellement, selon la définition donnée par Breiman :

*"A random forest is a classifier consisting of a collection of tree-structured classifiers  $h(x, k)$ ,  $k = 1, \dots$ , where the  $k$  are independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input  $x$ ."* [3]

Chaque arbre  $h(x, k)$  agit comme un classificateur entraîné sur un sous-échantillon différent, avec une structure influencée par un vecteur aléatoire  $k$ , assurant la diversité des modèles. En classification, les prédictions sont agrégées par vote majoritaire (ou moyenne en régression), ce qui permet de stabiliser le modèle global, de réduire la variance et d'améliorer la robustesse face au sur-apprentissage.

Contrairement à un seul arbre de décision, qui peut être très instable et sensible aux variations des données, une forêt aléatoire bénéficie d'un ensemble d'arbres diversifiés, tout en conservant l'interprétabilité des structures arborescentes.

Bien que les forêts aléatoires aient démontré leur efficacité dans de nombreux contextes en réduisant la variance des modèles instables comme les arbres de décision, elles ne s'attaquent pas directement au biais introduit par les *weak learners* (apprenants faibles). C'est dans ce contexte qu'intervient une autre méthode d'agrégation : **le boosting**.

## 1.2. Définition du Boosting

Avant d'introduire le boosting, il est essentiel de comprendre une autre méthode d'agrégation de modèles : le bagging.

**Le bagging** (*Bootstrap Aggregating*), introduit par Breiman [1], consiste à entraîner plusieurs modèles de manière indépendante sur des échantillons aléatoires du jeu de données. Chaque modèle, appelé *weak learner* (apprenant faible), a une performance limitée individuellement. Cependant, en combinant leurs prédictions selon une méthode de votation (moyenne pour les régressions, majorité pour les classifications), on obtient un modèle global plus robuste, appelé *strong learner* (apprenant fort).

Cette technique repose sur l'idée que la diversité des modèles entraîne une réduction de la variance et améliore la généralisation. Un exemple typique d'algorithme utilisant le bagging est la forêt aléatoire (*Random Forest*) [3], qui agrège plusieurs arbres de décision.

Contrairement au bagging, où les modèles sont entraînés indépendamment les uns des autres, **le boosting** adopte une approche itérative où chaque nouveau modèle est construit pour corriger les erreurs des modèles précédents [2].

Dans le bagging, la diversité des modèles est obtenue par resampling, c'est-à-dire en créant des sous-ensembles aléatoires de données pour chaque modèle. Cependant, dans le boosting, cette diversité est introduite de manière contrôlée, en attribuant plus de poids aux observations mal classées par les modèles précédents. Ainsi, chaque modèle se concentre sur des erreurs spécifiques, ce qui permet d'augmenter la performance globale en corrigeant progressivement les faiblesses des modèles précédents [1]. L'idée est d'affecter plus d'importance aux observations mal prédites afin d'améliorer progressivement la performance globale.

En reprenant les similarités de la forêt, si dans le bagging nous avons une forêt constituée de plusieurs arbres indépendants, dans le boosting, chaque arbre est soigneusement planté en fonction des faiblesses des arbres précédents, rendant ainsi l'ensemble plus solide et performant. Par exemple, dans un scénario où des erreurs spécifiques persistent dans les prédictions d'un arbre de décision, le modèle suivant va être "guidé" pour mieux traiter ces erreurs, plutôt que de travailler sur un ensemble aléatoire d'observations.

Une idée fondamentale du boosting est l'utilisation de *weak learners* (apprenants faibles), qui, bien qu'ils aient une performance inférieure à celle d'un classificateur optimal, peuvent, lorsqu'ils sont combinés, créer un modèle global puissant. Schapire [5] a démontré que, même un apprenant qui fait à peine mieux qu'un tirage au sort peut contribuer de manière significative à la réduction de l'erreur globale.

Le boosting repose sur les étapes suivantes :

1. Un modèle faible est entraîné sur l'ensemble des données.
2. Les erreurs de ce modèle sont identifiées.
3. Un second modèle est entraîné en accordant plus d'importance aux erreurs du premier modèle.
4. Ce processus est répété plusieurs fois jusqu'à ce que l'erreur globale diminue significativement.
5. Enfin, les prédictions de tous les modèles entraînés sont combinées selon un schéma de pondération spécifique.

L'un des algorithmes de boosting les plus populaires est AdaBoost (*Adaptive Boosting*), introduit par Freund et Schapire [2]. Cet algorithme ajuste dynamiquement les poids des observations à chaque itération. En effet, dans le cadre du boosting, l'algorithme ajuste dynamiquement les poids des observations en fonction des erreurs commises par les modèles précédents. En particulier, les observations mal classées voient leur poids augmenter, ce qui oblige les modèles suivants à se concentrer davantage sur ces erreurs et si les observations sont bien prédites, leur poids est diminué. Ce mécanisme différencie le boosting des autres méthodes comme le bagging, où la diversité est obtenue via des sous-échantillons des données, mais sans une pondération spécifique sur les erreurs. Dans le boosting, les erreurs sont utilisées pour guider l'entraînement des modèles suivants, en mettant l'accent sur les points les plus difficiles à prédire.

AdaBoost permet une amélioration exponentielle de la performance sous certaines conditions, notamment lorsque chaque modèle faible est légèrement meilleur qu'un choix aléatoire [5].

Le boosting est une méthode clé en apprentissage automatique, offrant une approche efficace pour améliorer la précision des prédictions en combinant plusieurs modèles faibles.

Contrairement au bagging, qui réduit la variance, le boosting vise à réduire le biais des modèles faibles en leur permettant d'apprendre de leurs erreurs successives.

Dans le cadre du boosting, à chaque itération, l'algorithme se concentre davantage sur les observations mal classées, ce qui permet de réduire progressivement l'erreur globale. Ce mécanisme de pondération dynamique joue un rôle crucial dans l'amélioration de la performance des modèles, car il permet de corriger les faiblesses des itérations précédentes.

Dans la suite de ce travail, nous approfondirons les différentes variantes du boosting, telles que AdaBoost et le Gradient Boosting [4] ainsi que leur application en régression.

### 1.3. Sélection de variables et approche itérative : cadre théorique

La sélection de variables constitue une étape cruciale dans la construction de modèles prédictifs, notamment lorsqu'il s'agit de modèles linéaires. Elle vise à identifier un sous-ensemble pertinent de variables explicatives qui contribuent significativement à la qualité de la prédiction, tout en améliorant l'interprétabilité du modèle et en évitant le surapprentissage dû à la présence de variables inutiles ou redondantes [7].

Les méthodes de sélection de variables peuvent être divisées en plusieurs catégories : la sélection manuelle basée sur des critères a priori, les méthodes automatiques classiques telles que la sélection forward, backward ou stepwise [10], ainsi que les méthodes basées sur la pénalisation, comme le Lasso (régularisation L1), qui introduisent un terme de contrainte favorisant la parcimonie du modèle [8].

Au-delà de ces approches classiques, l'utilisation d'une sélection itérative des variables s'est révélée particulièrement efficace pour améliorer la performance des modèles linéaires. Cette approche consiste à ajuster progressivement le modèle en sélectionnant ou en pondérant les variables au fil des itérations, souvent dans un cadre où le modèle est amélioré par étapes successives. Par exemple, dans le boosting, chaque étape itérative corrige les erreurs du modèle précédent, en modifiant implicitement l'importance des variables [9].

L'intérêt d'une sélection itérative repose sur sa capacité à intégrer de manière dynamique l'information apportée par chaque variable, en tenant compte des interactions complexes et des corrélations qui peuvent exister entre les prédicteurs. Cela permet notamment de réduire le biais du modèle, d'affiner la sélection des variables pertinentes, et d'améliorer la robustesse face au surapprentissage.

Dans le cadre des modèles linéaires, la sélection itérative peut ainsi permettre d'obtenir un compromis optimal entre complexité et performance, en adaptant continuellement la composition du modèle à la structure des données. Ce mécanisme est très utilisé en machine learning, telles que le boosting linéaire, où la sélection progressive des variables participe à la construction d'un modèle global plus performant et plus interprétable.

### 1.4. Contexte et Problématique

Dans un contexte environnemental, la qualité de l'air constitue un enjeu majeur pour la santé publique et l'écosystème. Parmi les polluants atmosphériques, **l'ozone ( $O_3$ )** joue un rôle crucial, car il peut avoir des effets néfastes sur la santé humaine, notamment sur les voies respiratoires, ainsi que sur la végétation et les matériaux. La prévision de la concentration en ozone permet ainsi d'anticiper les épisodes de pollution et de mettre en place des mesures de prévention adaptées.

Cependant, prédire précisément la concentration en ozone est une tâche complexe. De nombreux facteurs influencent cette variable, notamment la concentration d'autres polluants ( $PM_{2.5}$ ,  $NO_2$ , NO), les conditions météorologiques (température, humidité, pression atmosphérique) et des phénomènes naturels (vent, précipitations). Trouver une méthode de régression optimal pour effectuer cette prédition est donc un défi, en raison de la multiplicité des variables explicatives et de leurs interactions possibles.

Les méthodes de régression classiques, telles que la régression linéaire simple, permettent de modéliser les relations entre les variables, mais elles sont souvent limitées lorsqu'il s'agit de capturer des relations non linéaires et des interactions complexes. Pour pallier ces limites, des méthodes avancées ont été développées, dont le boosting, une approche du machine learning qui repose sur la combinaison de plusieurs modèles simples, appelés prédicteurs faibles, pour construire un modèle plus performant.

Ainsi, face aux limites des méthodes de régression classiques et aux enjeux liés à la prévision de la concentration en ozone, il est essentiel d'évaluer l'apport des approches avancées comme le Boosting.

**Comment la sélection de variables, appliquée de manière itérative, contribue-t-elle à l'amélioration d'un modèle linéaire ?**

## 1.5. Objectifs du travail

L'objectif principal de cette étude est d'évaluer l'efficacité du boosting dans un cadre de régression, appliqué à la prévision de la concentration d'ozone. Cette approche sera mise en perspective avec d'autres méthodes de régression, afin de comparer leurs performances et leur capacité à fournir une meilleure prédition.

Plus précisément, notre travail vise à :

1. **Étudier le boosting** dans un cadre simple afin de comprendre son fonctionnement et ses avantages par rapport aux modèles de régression classiques.
2. **Comparer les performances du boosting** avec celles de deux autres méthodes :
  - La **régression linéaire multiple**
  - La **régression pénalisée L1 (Lasso)**, qui intègre un mécanisme de sélection de variables et de régularisation pour améliorer la prédition.
3. **Comparer les variables** de chacune des méthodes.
4. **Appliquer des modèles linéaires sur un jeu de données réelles**, constitué de mesures atmosphériques collectées à Périgueux. L'objectif est d'utiliser 80 % des données disponibles pour entraîner les modèles et d'évaluer leur capacité de généralisation en prédisant les 20 % restants, avec un focus particulier sur la prévision du maximum de concentration d'ozone.

Ce travail vise ainsi à fournir une vision comparative et critique des méthodes étudiées, en mettant en lumière la pertinence du boosting pour la modélisation de phénomènes environnementaux. Les conclusions permettront de mieux orienter les choix méthodologiques pour la prévision de la qualité de l'air dans des contextes similaires.

## 2. Méthodes de Régression et Théorie du Boosting

### 2.1. Rappels sur les Modèles de Régression

Un modèle de régression linéaire suppose que la fonction de régression  $E(X|Y)$  est linéaire par rapport aux variables explicatives  $X_1, \dots, X_p$ . Ils sont simples et offrent souvent une description adéquate et interprétable de la manière dont les variables explicatives influencent la variable de sortie. Pour la prédiction, ils peuvent parfois surpasser des modèles non linéaires plus sophistiqués, en particulier dans des situations où le nombre d'échantillons d'entraînement est faible. Enfin, les méthodes linéaires peuvent être appliquées à des transformations des variables explicatives, ce qui élargit considérablement leur champ d'application.

La régression linéaire est un outil fondamental pour l'analyse des résultats numériques. Elle permet d'expliquer et de quantifier le lien entre une variable d'intérêt et des variables explicatives. En tant que première approche de la statistique inférentielle, elle constitue la source de nombreuses modélisations statistiques. Un de ses principaux atouts est sa capacité à représenter un grand nombre d'expériences, offrant ainsi une large applicabilité dans divers domaines. De plus, sa simplicité permet une compréhension approfondie de ses propriétés, facilitant son interprétation et son utilisation. Cependant, son usage repose sur des hypothèses mathématiques parfois contraignantes, ce qui peut limiter son applicabilité dans certains contextes.

Un modèle de régression linéaire est défini par une équation de la forme :

$$Y = X\beta + \epsilon$$

où :

- $Y$  est un vecteur aléatoire de dimension  $n$  ;
- $X$  est une matrice connue de taille  $n \times p$ , appelée matrice design ;
- $\beta$  est le vecteur de dimension  $p$  des paramètres inconnus du modèle ;
- $\epsilon$  est le vecteur aléatoire de dimension  $n$ , appelé erreur du modèle.

La notion de linéarité fait référence au fait que ces modèles sont linéaires en leurs paramètres.

L'équation du modèle peut s'écrire aussi de la manière suivante :

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \quad i = 1, \dots, n$$

$$Y_i = \sum_{j=1}^p \beta_j x_{ij} + \epsilon_i.$$

Généralement, la constante fait partie du modèle tel que :

- $x_{i1} = 1$  pour tout  $i = 1, \dots, n$
- $\beta_1$  représente la constante (intercept dans les logiciels).

Ainsi les hypothèses du modèle sont les suivantes :

$$\begin{cases} H_1 : \text{rg}(X) = p \\ H_4 : \epsilon \sim \mathcal{N}(0, \sigma^2 I_n) \end{cases}$$

## 2.2. Méthodes standards : Linéaire simple et Lasso

### 2.2.1 Méthode de régression linéaire multiple

La régression linéaire multiple est une extension naturelle de la régression linéaire simple. Elle permet de modéliser la relation entre une variable cible  $Y$  et plusieurs variables explicatives  $x_1, x_2, \dots, x_p$  simultanément. La méthode est exprimé par :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i, i = 1, \dots, n$$

où :

- $y_i$  est la valeur de la variable cible pour l'observation  $i$ ,
- $x_{ij}$  représente la  $j$ -ème variable explicative pour l'observation  $i$ ,
- $\beta_0, \beta_1, \dots, \beta_p$  sont les coefficients du modèle à estimer,
- $\epsilon_i$  est le terme d'erreur aléatoire supposé suivre une loi normale centrée et de variance constante.

Le but est d'estimer les coefficients  $\beta_j$  afin de modéliser au mieux la relation entre  $y$  et l'ensemble des variables explicatives  $(x_1, \dots, x_p)$ .

En régression multiple, il est important de vérifier que les variables explicatives sont pertinentes collectivement et individuellement pour expliquer la variabilité de  $y$ . L'analyse simultanée permet de prendre en compte les effets des variables, et d'ajuster les estimations pour éviter les biais dus aux corrélations entre les variables explicatives.

Pour évaluer et comparer différents modèles de régression multiple, on utilise notamment le critère d'information d'*Akaike Information Criterion* (AIC). L'AIC permet de mesurer la qualité d'un modèle en prenant en compte à la fois son ajustement aux données et sa complexité. Il est défini par :

$$AIC = -2\ln(\hat{L}) + 2|m|$$

où :

- $m$  est le nombre de paramètres estimés,
- $2|m|$  le terme de pénalisation,
- $\hat{L}$  est la vraisemblance maximale du modèle,
- $-2\ln(\hat{L})$  est le terme d'attache aux données.

Un AIC plus faible est préféré, car il indique un meilleur compromis entre précision et simplicité des données.

Il est aussi utilisé des procédures automatiques de sélection de variables :

- **Forward** : à partir d'un modèle vide (sans variables), on ajoute successivement la variable qui améliore le plus l'ajustement du modèle selon un critère (par exemple, l'AIC), jusqu'à ce qu'aucune amélioration significative ne soit obtenue.
- **Backward** : à partir du modèle complet (incluant toutes les variables), on retire progressivement la variable la moins significative (souvent avec la plus haute p-valeur ou le plus faible impact sur l'AIC), jusqu'à stabilisation.

- **Both** : à partir du modèle complet (incluant toutes les variables) on retire la variable la moins significative et/ou on ajoute une variable précédemment retirée qui permet finalement un meilleur ajustement du modèle.

Ces méthodes permettent d'identifier un sous-ensemble de variables explicatives pertinentes tout en maintenant un équilibre entre complexité du modèle et qualité de la prédiction. Ce processus est important, en particulier lorsque le nombre de variables est élevé par rapport au nombre d'observations.

La combinaison entre la régression linéaire multiple, l'analyse des corrélations et la sélection pas à pas (forward/backward) constitue une approche classique mais robuste pour construire progressivement un modèle interprétable, tout en évitant le sur-apprentissage.

### 2.2.2 Méthode de régression pénalisée Lasso

La régression Lasso (*Least Absolute Shrinkage and Selection Operator*) est une méthode de régression pénalisée qui ajoute une contrainte de régularisation  $L_1$  à la somme des moindres carrés. L'objectif est d'imposer une pénalité sur la somme des valeurs absolues des coefficients, ce qui peut conduire à l'annulation de certains d'entre eux.

L'estimateur du Lasso est donné par [6] :

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \sum_{i=1}^N \left( y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 \quad \text{sous la contrainte} \quad \sum_{j=1}^p |\beta_j| \leq t.$$

Et l'expression équivalente en forme de Lagrangien est :

$$\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left( \frac{1}{2} \sum_{i=1}^N \left( y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right).$$

où :

- $y_i$  est la variable cible,
- $x_{ij}$  représente les variables explicatives,
- $\beta_j$  sont les coefficients du modèle,
- $\beta_0$  représente la valeur de la variable cible  $y$  lorsque toutes les variables explicatives sont nulles.
- $\lambda$  est un paramètre de régularisation qui contrôle la pénalité appliquée aux coefficients.

Lorsque  $\lambda$  augmente, plus de coefficients sont réduits à zéro, ce qui entraîne une sélection automatique de variables.

La régression Lasso est une méthode de réduction qui impose une contrainte sur la somme des valeurs absolues des coefficients. Contrairement à la régression Ridge, qui applique une pénalité sur la somme des carrés des coefficients, le lasso peut réduire certains coefficients à zéro, effectuant ainsi une sélection de variables. Cette méthode est particulièrement utile pour choisir un sous-ensemble de variables pertinentes, ce qui la rend efficace pour la sélection de modèles. Le choix du paramètre de pénalité est essentiel et doit être fait de manière adaptative pour minimiser l'erreur de prédiction.

## 2.3. Algorithmes classiques de Boosting : Gradient Boosting et AdaBoost

Les premiers algorithmes de boosting développés par Schapire et Freund étaient avant tout des constructions théoriques visant à démontrer le concept de boosting plutôt que des algorithmes réellement adaptés à une utilisation pratique. Cependant, ces travaux ont ouvert la voie à des algorithmes de boosting concrets. En effet, aujourd’hui nous utilisons de nombreux algorithmes de boosting les algorithmes Gradient Boosting, AdaBoost, et XGBoost sont les trois algorithmes les plus utilisés.

### 2.3.1 Gradient Boosting

Le **Gradient Boosting** est une approche du boosting qui a été introduite par Friedman en 2001 [4]. C'est une méthode d'apprentissage supervisé visant à construire un modèle prédictif fort à partir d'une somme de modèles faibles. Contrairement à d'autres approches de Boosting comme AdaBoost, qui ajustent les poids des observations, le Gradient Boosting considère l'apprentissage comme un problème d'optimisation de fonction de perte, et applique une descente de gradient fonctionnelle pour réduire l'erreur à chaque itération.

Soit  $F(x)$  le modèle prédictif que l'on cherche à construire. Plutôt que d'apprendre ce modèle en une seule fois, on le construit comme une somme de modèles faibles  $h_m(x)$ , chacun venant corriger les erreurs du précédent :

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x)$$

où  $\nu \in (0, 1]$  est un taux d'apprentissage.

À chaque étape, le nouvel apprenant  $h_m$  est entraîné pour minimiser une fonction de perte  $L(y, F(x))$  à l'aide du gradient de cette fonction par rapport aux prédictions précédentes :

$$r_{im} = - \left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F=F_{m-1}}$$

Le modèle  $h_m$  est ensuite ajusté pour approximer ces résidus  $r_{im}$ .

Dans le cas de la régression linéaire, si la fonction de perte est la perte quadratique  $L(y, F(x)) = \frac{1}{2}(y - F(x))^2$ , alors le gradient correspond simplement à l'erreur de prédiction :

$$r_{im} = y_i - F_{m-1}(x_i)$$

Chaque itération consiste donc à corriger les erreurs précédentes de manière additive.

Le terme **gradient** fait référence à l'idée d'utiliser la direction du plus grand changement de la fonction de perte (le gradient) pour guider l'apprentissage, généralisant ainsi la descente de gradient classique.

### 2.3.2 AdaBoost

AdaBoost (*Adaptive Boost*), a été le premier algorithme de boosting concret à être introduit. Il s'agit d'une technique d'apprentissage supervisé qui combine plusieurs apprenants faibles (*weak learners*) afin de former un classificateur robuste et performant.

Contrairement aux premières méthodes de boosting, qui étaient principalement théoriques, AdaBoost a introduit un mécanisme adaptatif qui ajuste dynamiquement les poids des données en fonction des erreurs commises lors des itérations successives.

AdaBoost est un algorithme itératif dont le principe repose sur l'entraînement de plusieurs apprenants faibles généralement des arbres de décision simples appelés *stumps* (arbres avec seulement une racine et deux feuilles) sur un ensemble d'apprentissage, puis sur leur combinaison afin de former un classificateur plus robuste. L'algorithme fonctionne en modifiant la distribution des données en fonction de la correction des erreurs de classification et de la précision globale du modèle. Chaque instance d'apprentissage reçoit un poids, et à chaque itération, l'algorithme ajuste les poids des observations mal classées en fonction de la capacité du classificateur à bien la prédire afin que les prochains classificateurs se concentrent davantage sur ces erreurs. Enfin, les classificateurs entraînés sont fusionnés pour former la décision finale.

Le processus d'entraînement suit plusieurs étapes clés :

1. **Initialisation des poids** : Toutes les observations du jeu de données reçoivent un poids initial égal.
2. **Apprentissage du premier classificateur faible** : Un premier modèle (*weak learner*) est entraîné sur les données pondérées.
3. **Évaluation et pondération du classificateur** : Une importance ( $\alpha$ ) est attribuée à ce classificateur en fonction de son taux d'erreur. Plus il est performant, plus il aura d'influence dans la classification finale.
4. **Mise à jour des poids des observations** : Les poids des observations mal classées sont augmentés afin de forcer le prochain classificateur à mieux les apprendre.
5. **Répétition du processus** : De nouveaux classificateurs sont entraînés en prenant en compte les nouveaux poids des données, et le processus continue jusqu'à obtenir un ensemble de classificateurs combinés en un modèle final.
6. **Classification finale** : La prédiction est obtenue en agrégant les décisions des classificateurs faibles en fonction de leur poids respectif.

La figure 1 illustre le fonctionnement de l'algorithme AdaBoost. Ce dernier augmente progressivement le poids des points de données mal classés afin de corriger les erreurs et, à la fin, combine les résultats des différents modèles de classification pour créer un classificateur global plus précis.

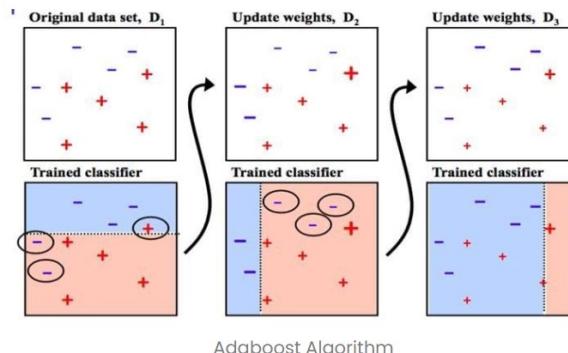


FIGURE 1 – Fonctionnement de l'algorithme Adaboost

### 3. Présentation des Données et Prétraitement

#### 3.1. Description du Jeu de Données

Dans cette étude, nous nous intéressons à la prévision des niveaux maximaux d'ozone ( $O_3$ ) à partir de données atmosphériques collectées dans la ville de Périgueux. L'objectif est d'évaluer la performance de trois approches de régression : une régression linéaire multiple, une régression pénalisée  $L_1$ , et un modèle de boosting.

Le jeu de données utilisé comprend 8 697 observations, correspondant à des mesures horaires réalisées entre le 8 novembre 2023 à minuit et le 7 novembre 2024 à minuit. Chaque observation représente un enregistrement horaire de diverses variables, totalisant 27 variables explicatives. Ces variables incluent :

- des concentrations de polluants atmosphériques ( $PM_{2.5}$ ,  $NO_2$ ,  $NO$ , etc.) ;
- des paramètres météorologiques (température, humidité relative, vitesse et direction du vent, pression atmosphérique) ;
- des informations temporelles (date, heure).

Une visualisation du début du jeu de données est proposée pour faciliter la compréhension de sa structure et de ses caractéristiques principales.

	date_debut	heure	NO2	O3	NO	PM10	temperature	pression	pression_variation_3h	humidite	point_de_rosse	vent_moyen	vent_rafales	vent_rafales_10min	
1	2023-11-08 00:00:00	0	3	34	3.6	11	7.1	1018.6	0.0	94	6.1	0.0	0.0	0.0	
2	2023-11-08 01:00:00	1	3	32	3.9	11	7.5	1018.1	0.0	95	6.7	0.0	0.0	0.0	
3	2023-11-08 02:00:00	2	3	34	3.4	10	8.1	1018.5	0.0	95	7.2	0.0	0.0	0.0	
4	2023-11-08 03:00:00	3	6	31	5.9	10	8.4	1018.3	-0.3	95	7.8	0.0	0.0	0.0	
5	2023-11-08 04:00:00	4	4	26	3.7	12	8.6	1018.6	0.5	95	7.8	0.0	0.0	0.0	
6	2023-11-08 05:00:00	5	6	22	7.1	12	8.9	1018.6	0.1	95	8.3	0.0	0.0	0.0	
7	2023-11-08 06:00:00	6	6	25	9.7	12	8.7	1018.6	0.3	95	7.8	0.0	0.0	0.0	
8	2023-11-08 07:00:00	7	9	47	11.7	9	8.9	1018.9	0.3	94	8.3	0.0	0.0	0.0	
9	2023-11-08 08:00:00	8	13	48	20.3	10	9.6	1019.1	0.5	94	8.9	1.6	0.0	8.0	
10	2023-11-08 09:00:00	9	7	54	11.5	10	10.7	1018.9	0.3	91	9.4	3.2	0.0	9.7	
11	2023-11-08 10:00:00	10	5	58	8.6	9	11.2	1018.8	-0.1	89	9.4	3.2	0.0	11.3	
12	2023-11-08 11:00:00	11	3	64	6.1	8	12.2	1018.1	-1.0	87	10.0	3.2	0.0	9.7	
13	2023-11-08 12:00:00	12	4	67	9.2	9	13.5	1017.3	-1.6	80	10.0	4.8	0.0	14.5	
14	2023-11-08 13:00:00	13	3	73	4.4	7	14.8	1016.2	-2.6	77	10.6	3.2	0.0	11.3	
15	2023-11-08 14:00:00	14	3	74	5.1	5	14.9	1015.4	-2.7	72	10.0	3.2	0.0	9.7	
16	2023-11-08 15:00:00	15	3	73	5.4	5	14.9	1014.5	-2.8	68	8.9	3.2	0.0	17.7	
17	2023-11-08 16:00:00	16	5	68	7.2	6	13.8	1013.8	-2.4	75	9.4	3.2	0.0	9.7	
18	2023-11-08 17:00:00	17	7	62	9.7	7	12.4	1013.1	-2.3	81	9.4	3.2	0.0	16.1	
19	2023-11-08 18:00:00	18	7	59	8.7	8	11.7	1012.9	-1.6	84	8.9	1.6	0.0	9.7	
20	2023-11-08 19:00:00	19	6	61	7.1	8	11.5	1013.0	-0.8	84	8.9	1.6	0.0	4.8	
21	2023-11-08 20:00:00	20	4	62	5.0	8	11.6	1012.2	-0.9	83	8.9	3.2	0.0	9.7	
22	2023-11-08 21:00:00	21	3	65	3.4	7	11.7	1011.6	-1.3	82	8.9	1.6	0.0	8.0	
23	2023-11-08 22:00:00	22	3	63	3.2	7	10.9	1011.3	-1.7	88	8.9	0.0	0.0	0.0	

#### 3.2. Préparation, Nettoyage et Prétraitement des Données

##### Préparation des données

Avant d'appliquer les différentes méthodes, une phase essentielle de nettoyage et de préparation des données a été effectuée afin de garantir la qualité, la fiabilité et la pertinence de l'analyse.

Formellement, l'ensemble de données peut être représenté par :

$$D = \{(x_i, y_i)\}_{i=1}^n$$

où :

- $x_i \in \mathbb{R}^p$  est un vecteur des variables explicatives pour l'observation  $i$ ,
- $y_i \in \mathbb{R}$  représente la concentration maximale d'ozone observée.

Sous forme matricielle, cela s'écrit :

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n.$$

Autrement dit,  $X$  représente la matrice contenant les vecteurs explicatifs  $x_i$  en lignes, et  $y$  est le vecteur colonne des réponses observées.

Les principales étapes de préparation ont été les suivantes :

- **Suppression de variables non pertinentes** : certaines colonnes, telles que *jour* et *PM2.5*, ont été écartées car elles n'apportaient pas d'information utile et pour *PM2.5* il y avait de nombreuses valeurs manquantes.
- **Gestion des valeurs manquantes** : toutes les observations comportant des valeurs manquantes ont été supprimées afin d'éviter les biais liés à l'imputation et d'assurer une base d'entraînement homogène.
- **Sélection des variables numériques pertinentes** : seules les variables quantitatives directement exploitables pour la régression ont été conservées.
- **Division du jeu de données** : l'ensemble des données a été aléatoirement séparé en deux sous-ensembles :
  - un jeu d'entraînement ( $X_{\text{train}}, y_{\text{train}}$ ) représentant 80% des données,
  - un jeu de test ( $X_{\text{test}}, y_{\text{test}}$ ) représentant les 20% restants.

Cette séparation aléatoire vise à éviter un biais saisonnier ou temporel qui pourrait fausser l'évaluation des modèles.

En complément, une démarche similaire a été appliquée sur un sous-échantillon du jeu de données initial. Plus précisément, 500 observations ont été extraites aléatoirement pour constituer un jeu d'entraînement, et 125 pour le test, sans chevauchement entre les deux ensembles.

Cette réduction volontaire vise à explorer l'impact du ratio *nombre de variables / nombre d'observations* sur les performances des méthodes de sélection. Elle permet notamment de mieux observer les effets de surapprentissage ainsi que la stabilité des modèles dans des conditions de données plus restreintes, un cas courant en pratique.

La phase de préparation des données est cruciale car elle conditionne la qualité et la robustesse des modèles prédictifs appliqués par la suite.

## 4. Implémentation des Méthodes et Comparaison

### 4.1. Présentation des algorithmes

#### 4.1.1 Régression linéaire multiple

**Approche 1 : Régression linéaire multiple avec sélection par corrélation**

**1<sup>er</sup> cas : Application sur le jeu de données complet**

Dans un premier temps, nous avons entraîné un modèle de régression linéaire multiple en sélectionnant uniquement les 12 variables les plus corrélées avec la variable  $O_3$ . Cette

approche repose sur l'idée que les variables présentant une forte corrélation linéaire avec  $O_3$  sont potentiellement les plus informatives pour la prédiction. Pour cela, nous avons calculé la matrice de corrélation sur l'ensemble des données d'apprentissage (80 % des observations), en excluant la variable cible  $O_3$ . Nous avons ensuite extrait les variables explicatives dont la corrélation absolue avec  $O_3$  dépassait un seuil arbitraire de 0,4. Ce seuil permet de filtrer les variables peu informatives tout en conservant un nombre raisonnable de prédicteurs.

Une fois les variables sélectionnées, nous avons construit un modèle de régression linéaire multiple que nous avons utilisé pour faire des prédictions sur les données d'entraînement et sur les données de test. Pour chaque ensemble, nous avons ensuite calculé l'erreur quadratique moyenne (RMSE) afin d'évaluer la qualité des prédictions.

### **2<sup>ème</sup> cas : Application sur un sous-échantillon réduit**

Dans un second temps, nous avons appliqué la même approche sur un sous-échantillon du jeu de données initial, composé de 500 observations pour l'entraînement et 125 pour le test. L'objectif était d'analyser le modèle dans un contexte de données limitées, où le ratio entre le nombre de variables et le nombre d'observations devient plus défavorable.

Comme précédemment, les variables les plus corrélées à  $O_3$  (avec un seuil de 0,4) ont été sélectionnées, puis intégrées dans un modèle de régression linéaire multiple. Ce modèle a ensuite été utilisé pour générer des prédictions sur les deux ensembles. Les performances ont été évaluées en calculant le RMSE sur les données d'apprentissage et de test.

## **Approche 2 : Régression linéaire multiple avec sélection par AIC**

### **1<sup>er</sup> cas : Application sur le jeu de données complet**

Dans ce premier cas, nous avons utilisé une sélection automatique de variables selon le critère AIC. Une fois les 12 premières variables sélectionnées, nous avons entraîné un modèle de régression linéaire multiple, puis effectué des prédictions sur les données d'entraînement et de test. Nous avons ensuite tracé un nuage de points comparant les valeurs réelles et prédites, et calculé le RMSE pour évaluer la qualité des prédictions.

### **2<sup>ème</sup> cas : Application sur un sous-échantillon réduit**

Nous avons repris la même méthode de sélection par AIC sur un sous-échantillon du jeu de données, composé de 500 observations pour l'entraînement et 125 pour le test. Après avoir sélectionné les 12 premières variables, nous avons entraîné un modèle de régression linéaire multiple, réalisé les prédictions sur les deux ensembles, tracé les nuages de points, puis calculé les RMSE pour évaluer les performances.

#### **4.1.2 Régression pénalisée Lasso**

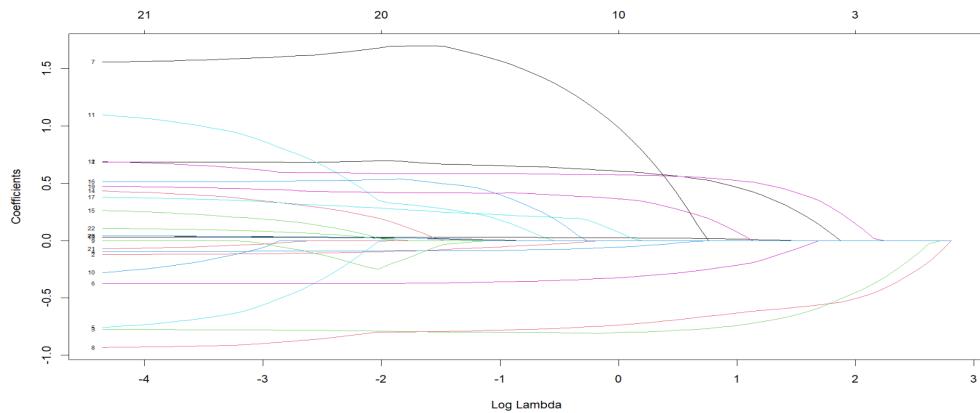
**Entraînement de la méthode :** Afin de sélectionner les variables explicatives les plus pertinentes pour prédire la concentration maximale d'ozone, nous avons appliqué une régression pénalisée de type Lasso ( $L_1$ ) en utilisant la bibliothèque `glmnet`.

La méthode Lasso introduit une pénalisation  $L_1$  sur les coefficients de la régression, favorisant la mise à zéro des coefficients associés aux variables moins importantes. Cette propriété permet de réaliser à la fois une sélection de variables et une régularisation du modèle.

L'entraînement du modèle s'est déroulé comme suit :

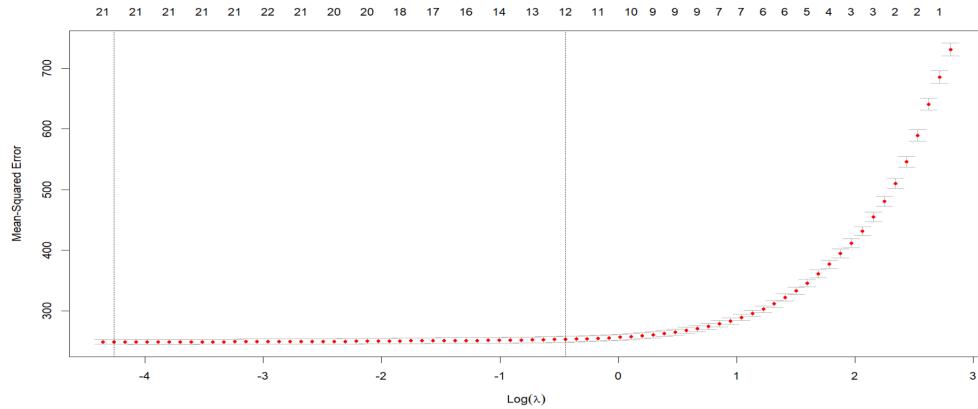
- Une graine aléatoire (`set.seed(456)`) a été fixée pour garantir la reproductibilité des résultats.
- La méthode Lasso a été ajusté sur les données d'entraînement via `glmnet`, avec  $\alpha = 1$  pour spécifier une régularisation purement  $L_1$ .

La fonction `glmnet` ajuste ainsi un ensemble de modèles correspondant à différentes intensités de pénalisation. L'évolution des coefficients en fonction de  $\lambda$  a été visualisée grâce au graphique des chemins de régularisation :



Le graphique obtenu permet de mieux comprendre le rôle du paramètre de régularisation  $\lambda$  dans les méthodes de type Lasso. Lorsque  $\lambda$  est faible, les coefficients des variables explicatives sont peu pénalisés, ce qui conduit à un modèle plus complexe, incluant la majorité des prédicteurs. À mesure que  $\lambda$  augmente, la pénalisation devient plus forte : certains coefficients sont progressivement contraints à zéro, ce qui simplifie le modèle en éliminant les variables jugées peu pertinentes. Enfin, les variables dont les coefficients restent non nuls même pour des valeurs élevées de  $\lambda$  apparaissent comme les plus importantes pour la prédiction, car elles conservent leur pouvoir explicatif malgré une forte régularisation.

**Sélection du paramètre de régularisation  $\lambda$  :** Afin de déterminer une valeur optimale pour le paramètre de régularisation  $\lambda$ , une validation croisée a été réalisée à l'aide de la fonction `cv.glmnet`. Cette procédure consiste à estimer l'erreur quadratique moyenne pour différentes valeurs de  $\lambda$ , en utilisant des sous-échantillons du jeu d'entraînement. Elle permet ensuite d'identifier deux valeurs clés :  $\lambda_{\min}$ , qui correspond à la valeur minimisant l'erreur de validation croisée, et  $\lambda_{1se}$ , qui représente le modèle le plus simple dont l'erreur reste inférieure à la valeur minimale augmentée d'un écart-type. Cette dernière option est souvent privilégiée pour sa capacité à limiter la complexité du modèle tout en conservant de bonnes performances.



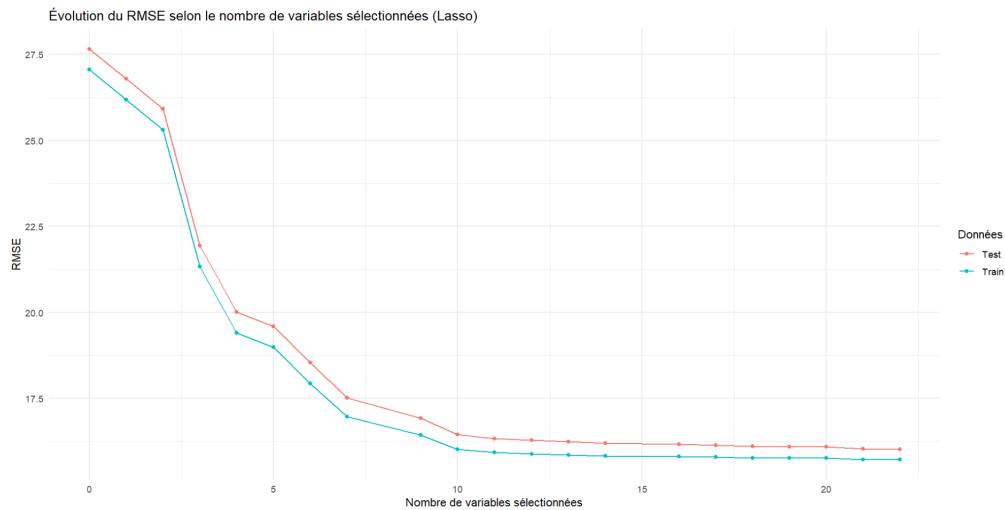
Le graphique obtenu représente l'évolution de l'erreur quadratique moyenne en fonction de  $\log(\lambda)$ . Cette étape de validation croisée et de visualisation des résultats est essentielle pour choisir le meilleur compromis entre la complexité du modèle et la performance sur des données non vues. Le graphique nous aide à visualiser comment la pénalisation affecte l'erreur du modèle et comment choisir le meilleur  $\lambda$  pour éviter à la fois le surapprentissage et le sous-apprentissage.

**Construction du modèle final :** La méthode Lasso final a été réentraîné sur l'ensemble des données d'entraînement en fixant  $\lambda$  à la valeur optimale sélectionnée.

Les coefficients du modèle final permettent d'identifier les variables ayant un effet significatif sur la concentration d'ozone  $O_3$ , c'est-à-dire celles dont les coefficients ne sont pas nuls. Ces coefficients renseignent à la fois sur l'importance et sur le sens de l'effet de chaque variable explicative.

**Analyse de l'évolution du modèle selon le nombre de variables sélectionnées :** Au-delà de la simple sélection de la meilleure valeur de  $\lambda$  via la validation croisée, nous avons souhaité étudier plus en détail l'effet de la régularisation sur la structure du modèle. En particulier, nous nous sommes intéressés à deux aspects : le nombre de variables explicatives conservées en fonction du niveau de pénalisation  $\lambda$ , et l'évolution de l'erreur de prédiction (RMSE) sur les jeux d'entraînement et de test.

Pour cela, nous avons procédé comme suit. Pour chaque valeur de  $\lambda$  générée automatiquement par la fonction `glmnet`, nous avons compté le nombre de variables dont les coefficients restaient non nuls (hors constante). Nous avons ensuite calculé le RMSE correspondant sur les données d'entraînement et de test. Enfin, nous avons visualisé l'évolution conjointe du RMSE et du nombre de variables sélectionnées, afin de mieux comprendre le compromis entre la complexité du modèle et la performance prédictive.

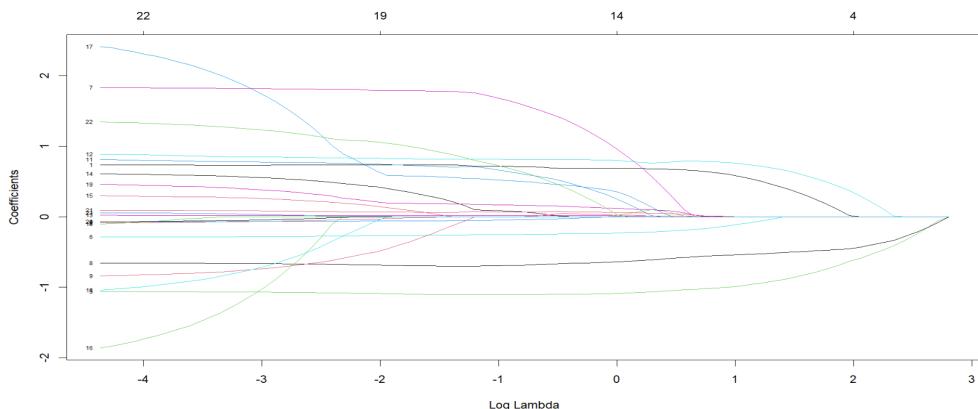


Le graphique obtenu met en évidence une dynamique classique du compromis biais-variance. Lorsque peu de variables explicatives sont sélectionnées, le modèle reste très simple mais l'erreur de prédiction reste relativement élevée, en raison d'un manque d'information. À mesure que l'on augmente le nombre de variables retenues, l'erreur décroît rapidement, avant de se stabiliser sur un plateau. Ce comportement indique qu'au-delà d'un certain seuil, l'ajout de variables supplémentaires n'apporte plus de gain significatif en termes de précision prédictive. Cette analyse permet de déterminer un compromis pertinent entre complexité du modèle et performance prédictive.

**Implémentation sur un sous-échantillon réduit** Dans un second temps, nous avons souhaité appliquer la régression Lasso à un jeu de données réduit, afin d'évaluer la robustesse de la méthode dans un contexte de données limitées.

**Entraînement du modèle sur sous-échantillon** Le même protocole que précédemment a été appliqué pour l'ajustement du modèle Lasso. Les modèles ont d'abord été entraînés à l'aide de la fonction `glmnet` sur le sous-échantillon d'apprentissage. La valeur optimale de  $\lambda$  a ensuite été déterminée par validation croisée, à l'aide de la fonction `cv.glmnet`, afin d'assurer un bon compromis entre complexité et performance. Enfin, les chemins de régularisation ont été visualisés afin d'analyser l'évolution des coefficients en fonction de la pénalisation appliquée.

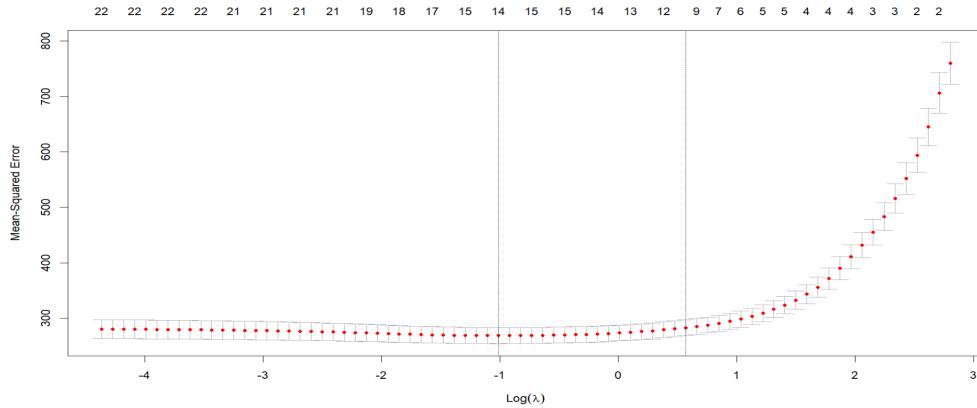
### Chemins de régularisation sur le sous-échantillon :



Le graphique issu de l'ajustement sur le sous-échantillon présente une évolution similaire à celle observée sur l'échantillon complet. On constate que les coefficients diminuent progressivement à mesure que la pénalisation  $\lambda$  augmente. Certaines variables sont éliminées très tôt dans le processus, tandis que seules les plus informatives conservent un effet significatif pour des valeurs de  $\lambda$  élevées. Cette dynamique illustre clairement le rôle de la régularisation dans la simplification du modèle.

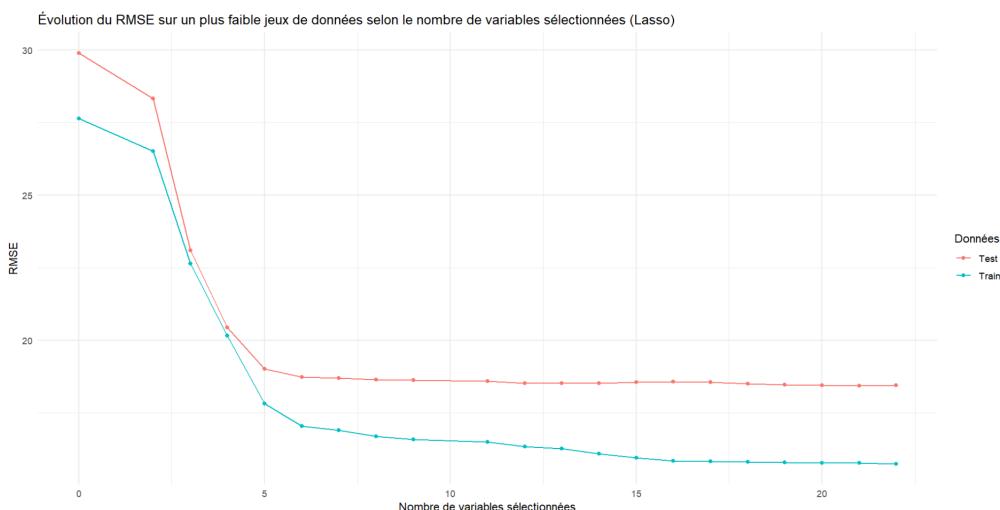
**Validation croisée** Le graphique de validation croisée permet d'identifier les deux valeurs de  $\lambda$  clés :

- $\lambda_{\min}$  : minimisant l'erreur moyenne quadratique ;
- $\lambda_{1se} = 1.77$  : garantissant un modèle plus simple mais avec une erreur comparable.



Comme dans le cas précédent, nous avons choisi de retenir  $\lambda_{1se}$  afin de privilégier la parcimonie et la robustesse. Cette décision est encore plus pertinente ici, car l'échantillon réduit rend le modèle plus sensible au bruit.

**Évolution du RMSE en fonction du nombre de variables sélectionnées** Enfin, nous avons mesuré les performances du modèle pour chaque  $\lambda$ , en suivant l'évolution du RMSE et du nombre de variables sélectionnées :



Le graphique confirme plusieurs éléments importants. Un modèle très simple, ne retenant que deux ou trois variables, conduit à des erreurs de prédiction élevées. L'ajout

progressif de variables explicatives permet d'améliorer les performances du modèle, jusqu'à atteindre un certain plateau au-delà duquel les gains deviennent marginaux. Enfin, une pénalisation trop forte, correspondant à des valeurs de  $\lambda$  trop élevées, entraîne une détérioration des performances, illustrant les effets négatifs de la surpénalisation.

L'objectif de cette expérimentation est d'analyser l'effet d'une réduction de taille d'échantillon sur le comportement d'un modèle Lasso concernant la différence entre les erreurs d'entraînement et de test ainsi que la stabilité du modèle avec un nombre de données réduites. Cette analyse permet de visualiser s'il y a un risque de surapprentissage lorsque le nombre d'observations diminue. Elle permet aussi d'évaluer la robustesse de la sélection de variables dans un contexte réaliste de données limitées.

#### 4.1.3 Méthode type Gradient Boosting

**Entraînement du modèle :** Pour identifier les variables explicatives les plus pertinentes dans la prédiction de la concentration maximale d'ozone ( $O_3$ ), nous avons implémenté une méthode inspirée du boosting, qui consiste à construire un modèle de régression comme une somme itérative de prédicteurs faibles, ici des modèles linéaires simples à une seule variable. Le but est de simplifier au maximum la sélection de variable.

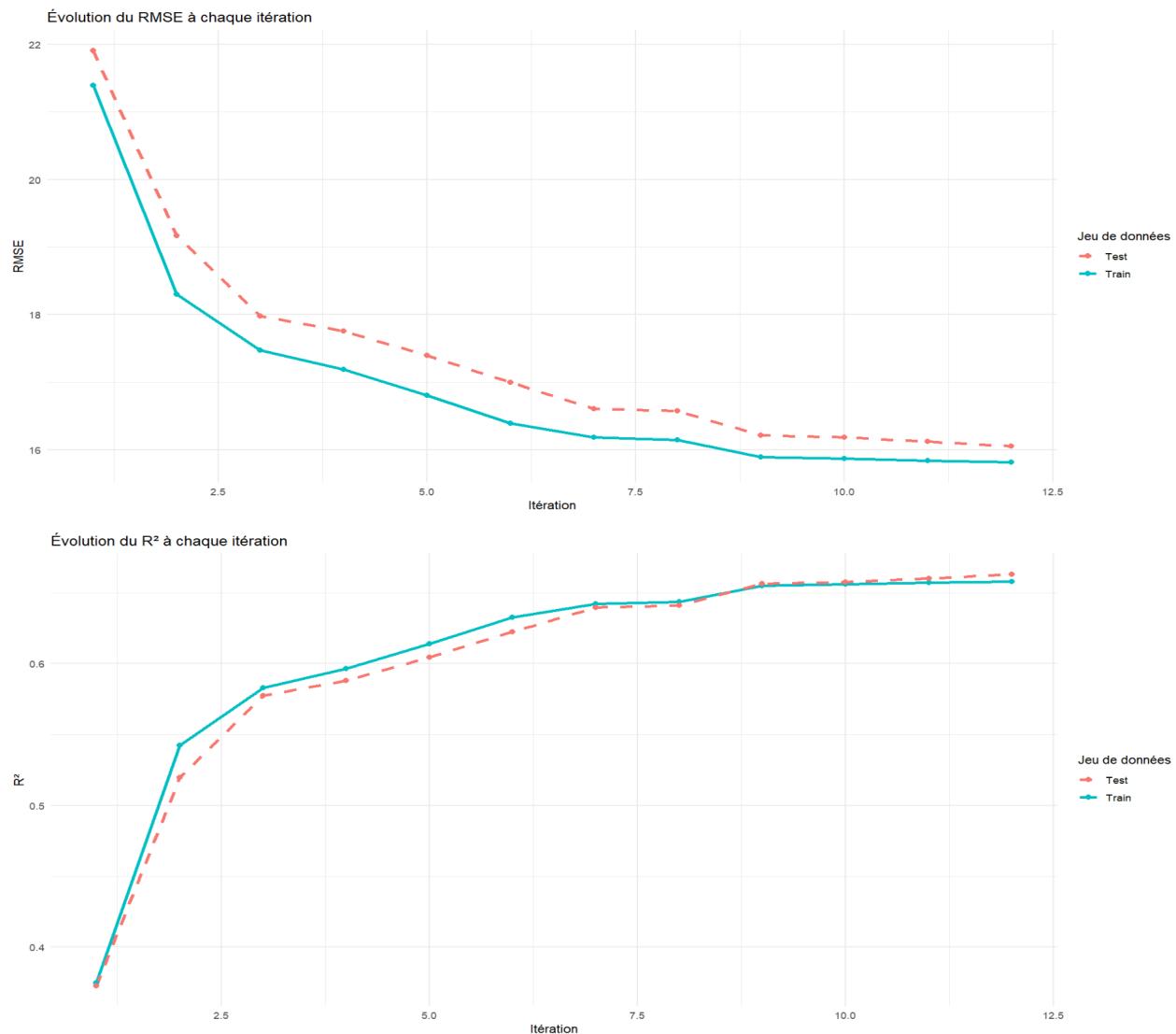
Le boosting est une approche d'apprentissage séquentielle dans laquelle, à chaque itération, un nouveau modèle est ajusté pour corriger les erreurs résiduelles du modèle précédent. Contrairement à des méthodes classiques de sélection ou de régularisation directe (comme le Lasso), le boosting s'appuie sur un processus additif qui affine progressivement la prédiction, tout en permettant une sélection implicite des variables au fil des itérations.

Dans notre implémentation, nous avons adopté une version simplifiée du gradient boosting appliquée à une perte quadratique (erreur quadratique moyenne). Chaque étape sélectionne la variable explicative la plus corrélée au résidu courant, puis ajuste un modèle linéaire simple sur cette variable, avant de mettre à jour la prédiction globale.

Le protocole mis en œuvre est le suivant :

- Fixation d'une graine aléatoire pour la reproductibilité (`set.seed(456)`).
- Séparation des données en un jeu d'apprentissage et un jeu de test (6281 observations pour l'apprentissage, 1571 pour le test), dans un deuxième temps nous avons aussi fait un sous-échantillon réduit (500 observations pour l'apprentissage, 125 pour le test).
- À chaque itération :
  - sélection de la variable explicative la plus corrélée aux résidus,
  - ajustement d'une régression linéaire simple sur cette variable,
  - mise à jour du modèle comme somme pondérée des modèles précédents.
  - calcul des résidus (différence entre la cible et la prédiction courante),
- Ce processus est répété pendant un nombre fixé d'itérations, sans réutiliser une même variable.

**Visualisation de l'apprentissage itératif :** Le graphe suivant illustre l'évolution du RMSE et du  $R^2$  au fil des itérations, à la fois sur les jeux d'apprentissage et de test.



On observe une amélioration progressive des performances, moins importante au fil du temps, traduisant une saturation de la capacité explicative du modèle. Ce comportement caractéristique du boosting illustre bien le compromis entre complexité et précision : chaque ajout de variable permet un gain marginal, mais trop d'itérations pourraient aussi conduire au surapprentissage.

**Structure du modèle final :** Le modèle final sélectionne successivement les 11 variables suivantes : humidite, NO, pluie\_cumul\_0h, pression\_variation\_3h, pression, NO2, vent\_direction, pluie\_intensite\_max\_1h, vent\_rafales\_10min, pluie\_24h, temperature.

Chaque variable est associée à un modèle linéaire simple (degré 1), dont les coefficients sont combinés de manière additive. Contrairement à la régression linéaire classique, ou au Lasso où tous les coefficients sont ajustés simultanément via une pénalisation, le boosting agit ici par étapes, chaque variable étant sélectionnée en fonction de l'information résiduelle qu'elle apporte.

Ce processus permet une interprétation directe du rôle de chaque variable dans la dynamique prédictive, tout en limitant les effets de multicolinéarité : les variables sont

intégrées une par une en fonction de leur capacité à corriger les erreurs résiduelles précédentes.

Une attention particulière doit être portée au nombre d’itérations : un trop grand nombre de termes peut conduire à un surapprentissage. Ici, es jeux de données étant suffisamment important, nous avons pu choisir le nombre de variable que nous voulions pour correspondre aux autres méthodes. Dans certains cas, il faut limiter volontairement le nombre d’itérations pour rester dans une zone de généralisation optimale.

**Évaluation des performances du modèle** À l’issue des 12 itérations de boosting, les performances du modèle ont été évaluées à l’aide de deux indicateurs classiques : la racine de l’erreur quadratique moyenne (RMSE) et le coefficient de détermination  $R^2$ . Ces mesures ont été calculées à la fois sur le jeu d’apprentissage (6281 observations) et sur le jeu de test (1571 observations), mais également sur un ensemble de données de 500 observations d’entraînement et 125 de test permettant ainsi de diagnostiquer la capacité de généralisation du modèle. Nous parlerons ultérieurement des résultats obtenus.

**Correction des prédictions négatives** Certaines prédictions issues du modèle peuvent, du fait de la nature linéaire des prédicteurs faibles, prendre des valeurs négatives. Or, une concentration d’ozone négative n’a pas de sens physique. Pour remédier à cela, nous avons corrigé les prédictions en remplaçant toutes les valeurs négatives par 0, ce qui a permis d’améliorer légèrement les performances du modèle. Cette étape simple permet d’aligner le modèle sur des contraintes de réalisme tout en réduisant l’erreur globale.

## 4.2. Analyse des résultats et comparaisons des modèles

Dans cette partie, nous présentons et comparons les résultats obtenus à l’aide des trois méthodes de modélisation utilisées. Pour chacune d’elles, nous détaillons les variables sélectionnées, visualisons les performances à l’aide de nuages de points représentant les prédictions sur les ensembles d’entraînement et de test, et calculons l’erreur de prédiction à l’aide de la métrique RMSE.

Afin de comparer objectivement les méthodes, nous avons retenu l’erreur quadratique moyenne (RMSE) comme critère principal d’évaluation. Cette métrique permet de mesurer l’écart moyen entre les valeurs prédites par le modèle et les valeurs réelles. Elle est définie par la formule suivante :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (1)$$

où :

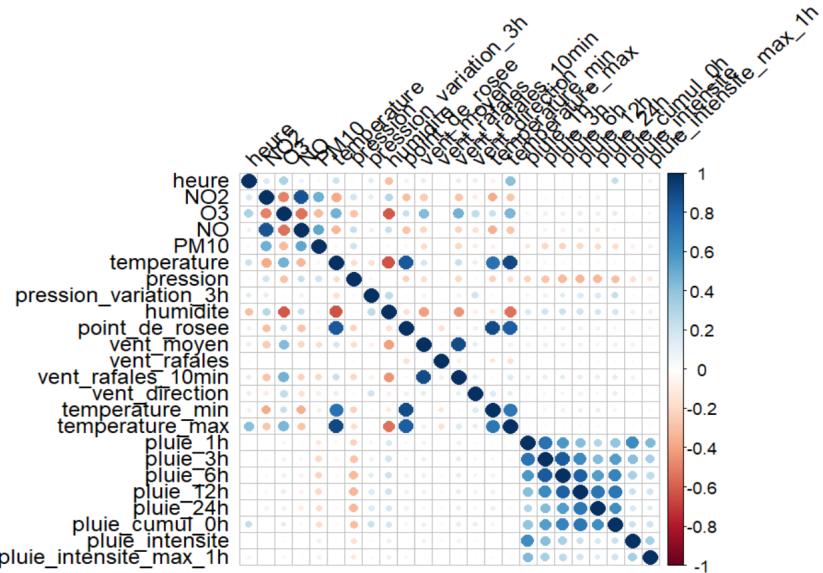
- $y_i$  désigne la valeur réelle de l’observation  $i$ ,
- $\hat{y}_i$  la valeur prédite par le modèle pour cette même observation,
- $n$  le nombre total d’observations considérées.

### 4.2.1 Méthode de régression multiple

#### A) Résultats de l’approche par corrélation

##### a) Application sur le jeu de données complet

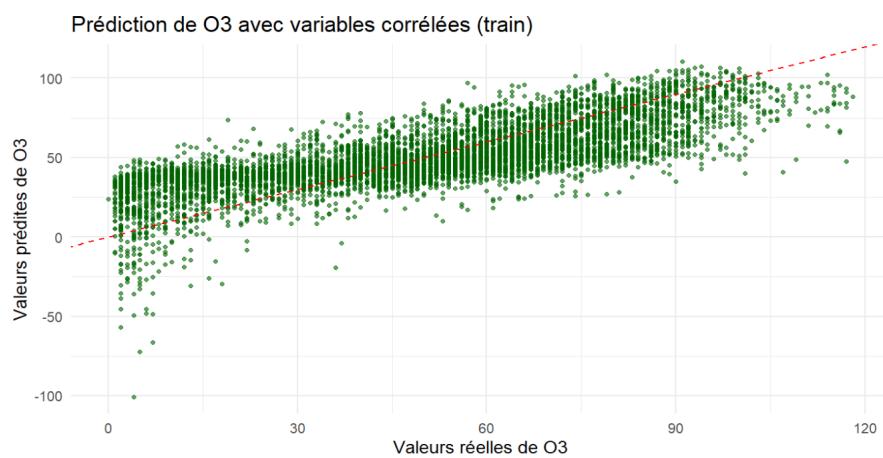
**Strucure du modèle :** Nous avons calculé la matrice de corrélation entre toutes les variables explicatives afin d'identifier celles qui sont le plus fortement associées à la concentration d'ozone ( $O_3$ ). L'objectif est de retenir uniquement les variables ayant une relation linéaire significative avec la variable cible.

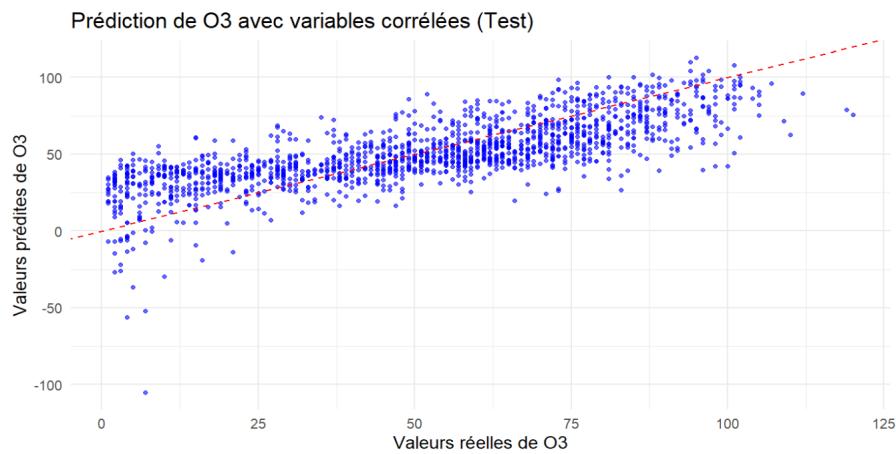


À partir de cette analyse, nous avons sélectionné les variables dont la corrélation absolue avec l'ozone dépasse un seuil de 0,4. Ce critère permet de filtrer les variables peu informatives tout en conservant un nombre raisonnable de prédicteurs pertinents.

Les 12 variables retenues sont : heure, NO<sub>2</sub>, NO, PM10, température, pression, humidité, point\_de\_rosee, vent\_moyen, vent\_rafales\_10min, vent\_direction, température\_min, température\_max.

**Visualisation des prédictions :** Nous montrons dans cette section les deux nuages de points selon les données d'entraînement et les données test.





Le nuage de points en vert montre les prédictions de la méthode sur les données d'entraînement. On voit que la plupart des points sont bien alignés le long de la diagonale, ce qui signifie que le modèle parvient à bien apprendre la relation entre les variables explicatives et la concentration d'ozone.

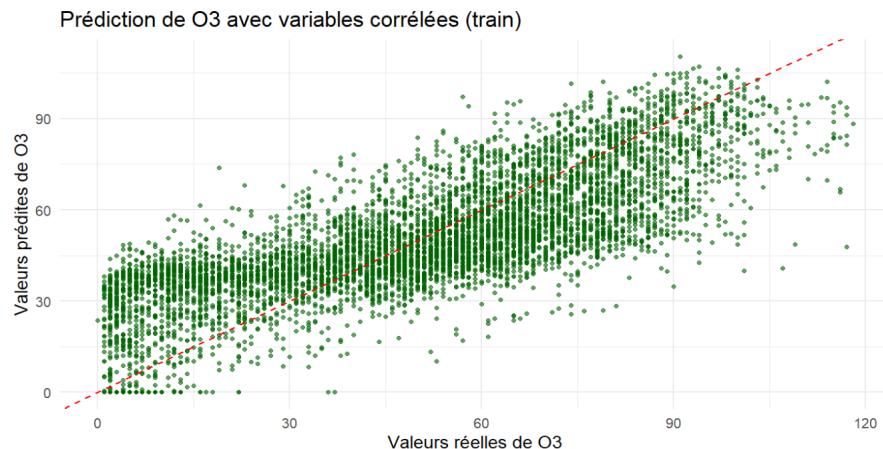
Ensuite, le nuage de points en bleu compare les valeurs prédites par le modèle aux vraies concentrations d'ozone sur les données test. On remarque que, dans l'ensemble, la méthode suit la tendance, mais il y a une dispersion notable, surtout pour les faibles valeurs de concentration. Certains points sont très éloignés de la diagonale, traduisant des erreurs de prédiction importantes.

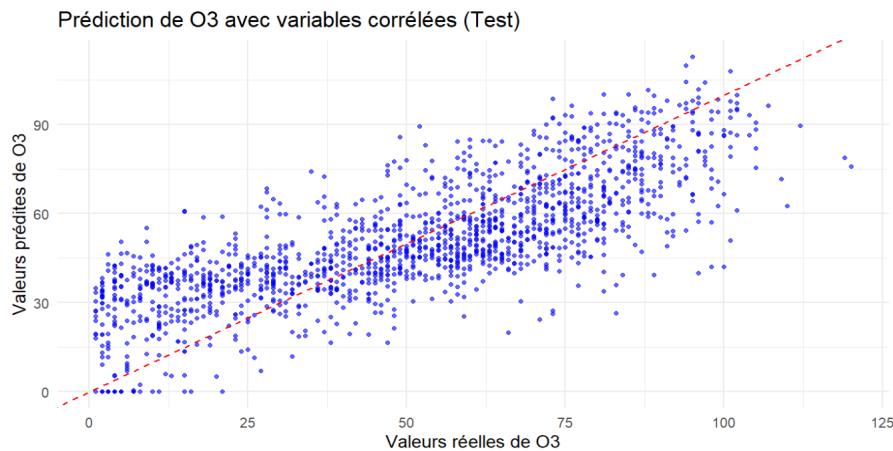
### Performances de la méthode :

- RMSE sur les données d'entraînement : 17.23
- RMSE sur les données de test : 18.00

Ces résultats reflètent une erreur moyenne modérée, avec des performances légèrement meilleures sur les données d'entraînement.

**Visualisation des prédictions (après correction des valeurs négatives) :** Dans cette section, nous avons voulu retirer les prédictions négatives car en réalité il n'est pas possible d'interpréter une concentration en ozone négativement. Alors nous avons décidé de mettre les prédictions négatives à zéro.





Le premier nuage de points montre les prédictions corrigées sur les données d'entraînement, où toutes les valeurs négatives ont été remplacées par zéro. La majorité des points suit bien la diagonale, indiquant que le modèle capture globalement la structure des données. Une certaine dispersion persiste pour les faibles concentrations.

Le second graphique illustre les prédictions corrigées du modèle sur les données de test. On observe une meilleure cohérence avec les vraies valeurs, notamment l'élimination des valeurs physiquement incohérentes (inférieures à zéro). La qualité globale des prédictions s'améliore visuellement.

#### **Effet de la correction :**

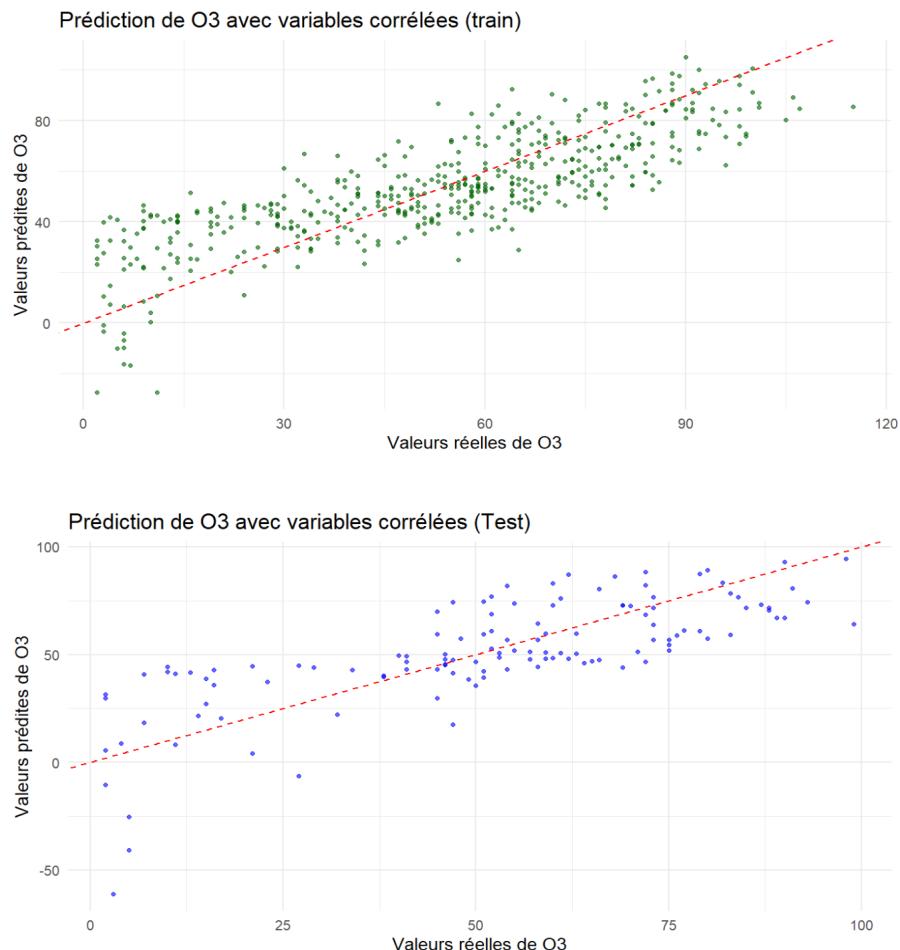
- RMSE corrigé sur les données d'entraînement : 17.03
- RMSE corrigé sur les données de test : 17.65

La correction a permis une légère amélioration des performances sur l'ensemble des données, en réduisant l'erreur moyenne et en rendant les prédictions plus réalistes.

#### **b) Application sur un jeu de sous-échantillons**

**Strucutre du modèle :** Comme dans le cas précédent, nous avons utilisé la matrice de corrélation pour identifier les variables les plus associées à la concentration d'ozone ( $O_3$ ) dans chaque sous-échantillon. Les variables sélectionnées sont celles dont la corrélation absolue avec  $O_3$  dépasse un seuil de 0,4. Ce critère permet de conserver uniquement les prédicteurs les plus pertinents dans chaque sous-jeu.

#### **Visualisation des prédictions**



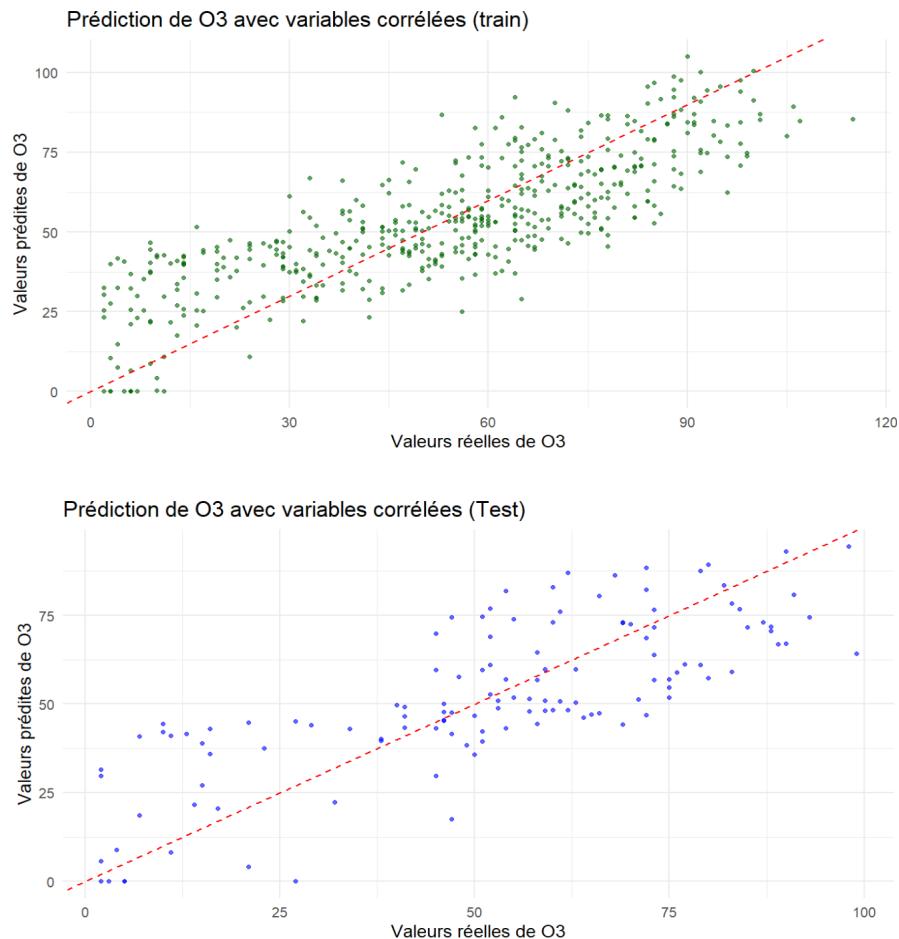
Le nuage de points en vert montre les prédictions de la méthode de régression linéaire sur les données d'apprentissage issues des sous-échantillons. On observe un alignement général des points autour de la diagonale, mais avec une dispersion plus marquée que sur le jeu complet, en particulier pour les faibles concentrations.

Le second nuage de points compare les prédictions de la méthode aux vraies valeurs d'ozone sur les données de test du sous-échantillonnage. La tendance générale est conservée, mais l'écart entre les valeurs prédites et observées est plus variable. On observe également la présence de quelques valeurs négatives.

### Performances de la méthode :

- RMSE sur les données d'entraînement : 15.45
- RMSE sur les données de test : 17.70

### Visualisation des prédictions (après correction des valeurs négatives)



Les prédictions négatives ont été remplacées par zéro. Cela améliore visuellement la cohérence des résultats. Les points se rapprochent davantage de la diagonale sur les données d'entraînement, en particulier pour les plus faibles concentrations.

La correction des valeurs négatives sur les données de test conduit à des prédictions plus réaliste. On constate une meilleure conformité avec les concentrations réelles d'ozone, et la suppression des valeurs négatives contribue à réduire les erreurs extrêmes.

#### Effet de la correction :

- RMSE corrigé sur les données d'entraînement : 15.20
- RMSE corrigé sur les données de test : 15.35

La correction permet une amélioration modeste mais significative de la performance du modèle, en supprimant les prédictions physiquement incohérentes.

## B) Résultats de l'approche par l'AIC

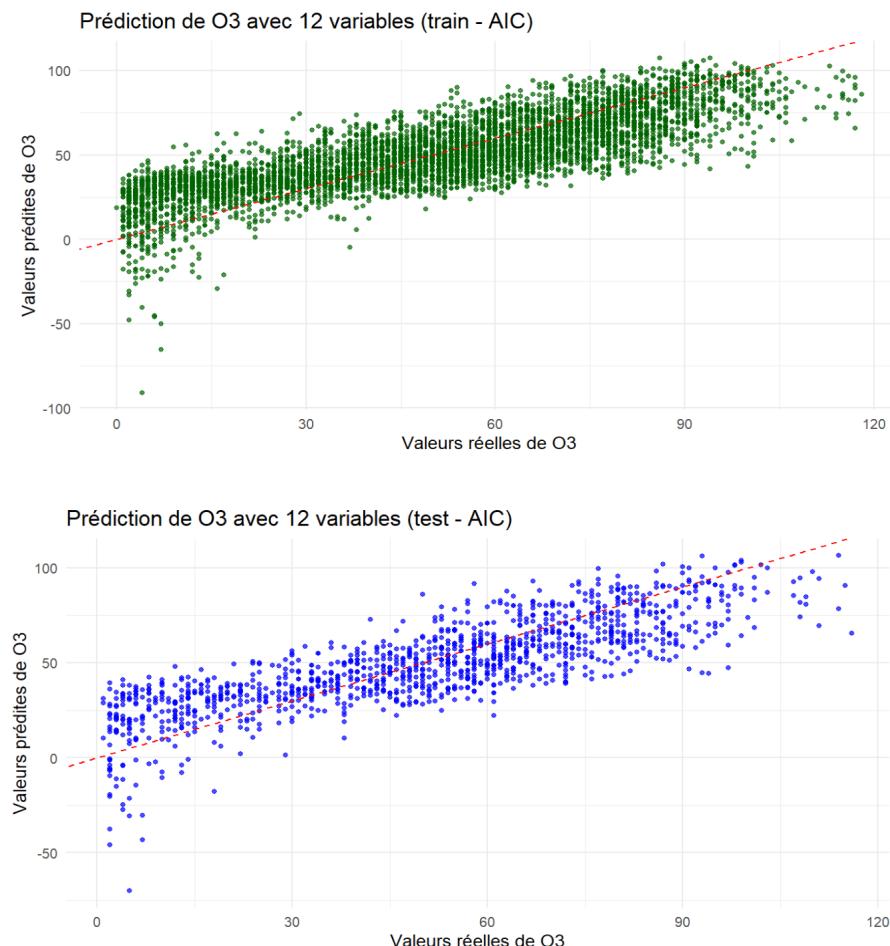
### a) Application sur le jeu de données complet

**Structure du modèle :** Dans cette approche, les variables explicatives ont été sélectionnées automatiquement à l'aide du critère d'information d'Akaike (AIC), via une procédure pas à pas (stepwise) dans les deux directions. Pour rester cohérent avec l'approche par corrélation, nous avons conservé les 12 premières variables sélectionnées qui sont les suivantes :

humidite, NO, heure, vent\_rafales\_10min, vent\_direction, pression, pression\_variation\_3h, PM10, pluie\_cumul\_0h, pluie\_1h, pluie\_12h, pluie\_24h.

Ces variables ont ensuite servi à entraîner un modèle de régression linéaire multiple sur le jeu de données complet.

### Visualisation des prédictions :



Le premier nuage de points montre les prédictions de la méthode AIC sur les données d'entraînement. La majorité des points est bien alignée autour de la diagonale, ce qui signifie que le modèle parvient à bien apprendre la relation entre les variables explicatives et la concentration d'ozone. Quelques dispersions apparaissent, notamment pour les faibles valeurs.

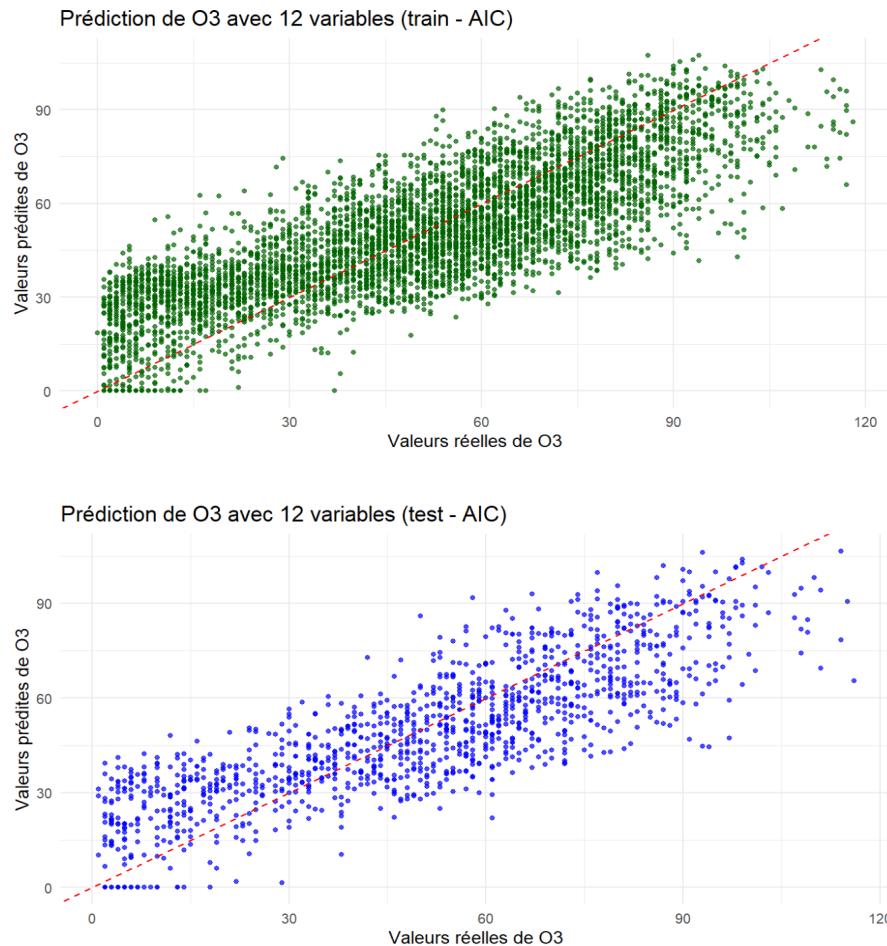
Le second graphique présente les prédictions de la méthode AIC sur les données de test. L'alignement global des points sur la diagonale indique que la tendance est bien capturée, mais certaines prédictions s'écartent sensiblement, notamment dans les faibles concentrations, avec quelques valeurs négatives.

### Performances de la méthode :

- RMSE sur les données d'entraînement : 15.77
- RMSE sur les données de test : 15.84

Ces résultats montrent une erreur moyenne modérée, avec des performances relativement équilibrées entre l'entraînement et le test.

**Visualisation des prédictions (après correction des valeurs négatives) :** Dans cette section, les prédictions négatives ont été remplacées par zéro, car une concentration d'ozone ne peut pas être inférieure à zéro dans la réalité.



Le nuage de points corrigé pour les données d'entraînement (en vert) montre un bon alignement avec la diagonale. Le remplacement des valeurs négatives par zéro améliore la cohérence physique de la méthode, avec un impact visuel positif sur la qualité des prédictions.

Le graphique en bleu illustre les prédictions corrigées de la méthode AIC sur les données de test. La majorité des points suit bien la diagonale, et les prédictions sont toutes non négatives, assurant ainsi une meilleure plausibilité physique. La dispersion pour les faibles concentrations reste visible mais réduite.

#### Effet de la correction :

- RMSE corrigé sur les données d'entraînement : 15,45
- RMSE corrigé sur les données de test : 15,48

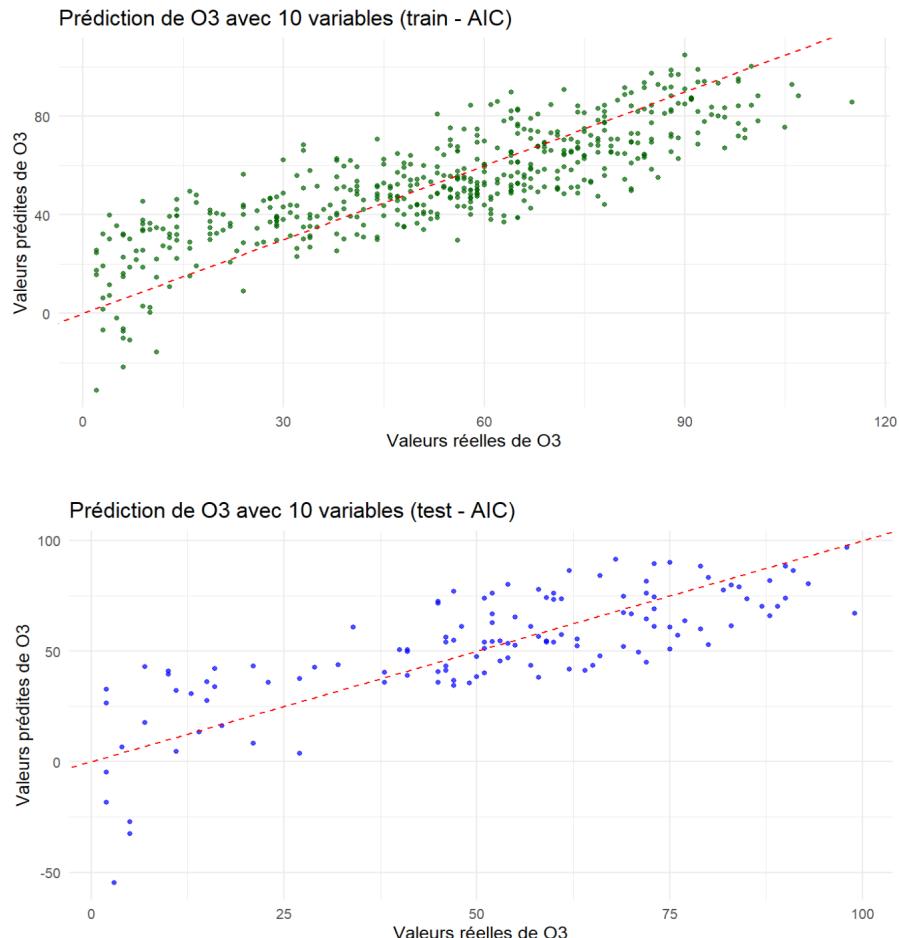
La correction a permis une légère amélioration sur les données d'entraînement mais aussi sur les données test.

#### b) Application sur un jeu de sous-échantillons

**Structure du modèle :** La méthode AIC a également été appliquée à un sous-échantillon de données (500 observations pour l'apprentissage et 125 pour le test). Le processus de sélection pas à pas a identifié un ensemble de variables légèrement différent de celui utilisé pour le jeu complet.

Les variables sélectionnées par le modèle sont : humidité, NO, heure, vent\_rafales\_10min, pression, pression\_variation\_3h, NO2, pluie\_12h, vent\_direction, pluie\_intensite\_max\_1h.

### Visualisation des prédictions :



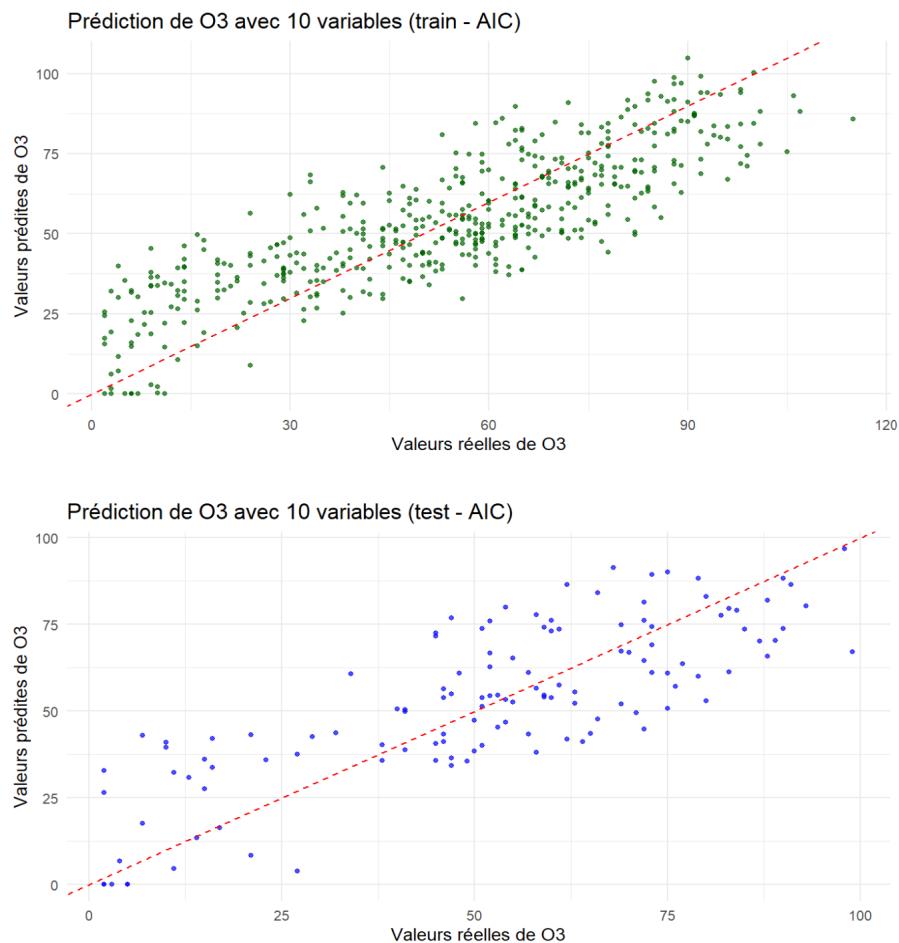
Le modèle AIC appliqué aux données d'entraînement (en vert) restitue bien la tendance des vraies valeurs, avec un bon alignement des points sur la diagonale. Quelques écarts persistent, notamment pour les faibles valeurs de concentration d'ozone mais aussi pour les grandes valeurs.

Sur les données de test, le modèle montre une cohérence générale avec la tendance des observations. Toutefois, des écarts sont présents pour les faibles concentrations, avec quelques prédictions négatives.

### Performances de la méthode :

- RMSE sur les données d'entraînement : 14.43
- RMSE sur les données de test : 15.84

## Visualisation des prédictions (après correction des valeurs négatives) :



Après correction, les prédictions sur les données d'entraînement sont toutes non négatives, avec un alignement amélioré sur la diagonale. Cela montre que le modèle reste fidèle à la tendance globale tout en supprimant les valeurs physiquement incohérentes.

Sur les données de test, la correction améliore la plausibilité des résultats. L'ensemble des points est désormais dans la zone  $O_3 \geq 0$ , avec une bonne restitution de la tendance globale.

### Effet de la correction :

- RMSE corrigé sur les données d'entraînement : 14.23
- RMSE corrigé sur les données de test : 14.31

La correction a permis une amélioration significative sur les deux jeux, en réduisant les erreurs extrêmes et en assurant la cohérence physique des prédictions.

### 4.2.2 Méthode de régression pénalisée

Dans le but de construire un modèle explicatif simple et robuste, nous avons utilisé la régression Lasso sur l'ensemble des échantillons. Nous présentons ci-dessous les résultats obtenus à l'aide de ce type de méthode.

### Modèle sélectionné : Lasso avec $\lambda_{1se}$

**Pourquoi choisir  $\lambda_{1se}$  ?** Dans notre étude, nous avons testé plusieurs valeurs de pénalisation pour la régression Lasso. Finalement, nous avons retenu le modèle correspondant à  $\lambda_{1se} = 0.64$ , car il présente le meilleur compromis entre performance et parcimonie. Ce choix repose sur le principe du "one standard error rule" : il s'agit du plus grand  $\lambda$  dont l'erreur de validation reste dans une borne d'un écart-type autour de l'erreur minimale. Cela permet d'éviter le surapprentissage, en préférant un modèle plus simple mais presque aussi performant. Ce choix favorise un modèle plus parcimonieux. L'analyse portera sur :

- La simplification du modèle par rapport à  $\lambda_{\min}$ ,
- L'impact sur la performance prédictive,
- L'interprétabilité du modèle.

### Interprétation du modèle avec $\lambda_{1se}$

**Principe de  $\lambda_{1se}$  :** Le paramètre  $\lambda_{1se} = 0.64$  est choisi car il maximise la régularisation tout en garantissant une erreur de validation croisée inférieure à la somme de l'erreur minimale et de son écart-type. Cela signifie que ce modèle, bien que légèrement moins précis, est plus simple et donc plus robuste (moins sensible au bruit).

**Structure du modèle :** Sur les 24 variables explicatives, seules 12 ont des coefficients non nuls dans le modèle final.

Voici les variables retenues et leur effet :

Variable	Coefficient estimé
(Intercept)	455.76
heure	0.63
NO2	-0.03
NO	-0.80
PM10	-0.07
pression	-0.35
pression_variation_3h	1.32
humidite	-0.76
vent_rafales_10min	0.58
vent_direction	0.03
pluie_1h	0.13
pluie_3h	0.19
pluie_12h	0.39

**Interprétation des effets des variables sélectionnées :** Le modèle Lasso final a retenu 12 variables explicatives ayant un effet non nul sur la concentration maximale d'ozone. L'interprétation des coefficients repose sur leur signe et leur amplitude, en considérant que toutes les autres variables sont maintenues constantes. Une valeur positive du coefficient indique qu'une augmentation de la variable correspondante est associée à une augmentation attendue de l'ozone, tandis qu'un coefficient négatif traduit une relation inverse. Voici les interprétations principales :

- **heure (0,63)** : L'ozone augmente au fil de la journée, ce qui est cohérent avec son pic observé généralement en début d'après-midi en raison d'une activité photochimique plus intense.
- **NO2 (-0,03) et NO (-0,80)** : Les oxydes d'azote ont un effet négatif sur la concentration d'ozone. Le monoxyde d'azote (NO) réagit directement avec l'ozone, ce qui explique son impact fortement négatif. Le dioxyde d'azote ( $\text{NO}_2$ ) joue également un rôle opposé plus modéré.
- **PM10 (-0,07)** : Les particules en suspension semblent corrélées à une légère baisse de l'ozone, possiblement en raison d'interactions chimiques ou d'effets de dispersion atmosphérique.
- **pression (-0,35)** : Une pression atmosphérique plus élevée est associée à une baisse modérée de l'ozone, ce qui pourrait refléter des conditions stables limitant la dispersion verticale des polluants.
- **pression\_variation\_3h (1,32)** : Une variation rapide de la pression sur 3 heures est liée à une hausse de l'ozone. Cela peut traduire des changements de masse d'air ou des situations météorologiques favorables à la formation d'ozone.
- **humidite (-0,76)** : L'humidité agit négativement sur l'ozone. Une atmosphère humide réduit l'intensité du rayonnement solaire nécessaire à la formation de l'ozone.
- **vent\_rafales\_10min (0,58)** : Les rafales de vent à court terme peuvent peut-être amener de l'air contenant beaucoup d'ozone ou disperser localement les substances qui en produisent.
- **vent\_direction (0,03)** : L'effet de la direction du vent, bien que faible, pourrait refléter des apports spécifiques d'air pollué selon l'orientation dominante (ex. : origine urbaine ou industrielle).
- **pluie\_1h (0,13), pluie\_3h (0,19) et pluie\_12h (0,39)** : Contrairement aux attentes, ces variables présentent un effet légèrement positif. Bien que la pluie tende à éliminer les polluants, cette corrélation peut refléter des effets indirects liés à des changements météorologiques ou à une structure de données particulière. Il peut également s'agir d'un artefact statistique, ces variables pouvant être corrélées à d'autres facteurs non pris en compte dans le modèle.

Le coefficient d'interception (455,76) représente la concentration d'ozone prédite lorsque toutes les variables explicatives sont nulles. Cette valeur n'a pas d'interprétation physique directe, mais elle est nécessaire pour ajuster le modèle linéaire.

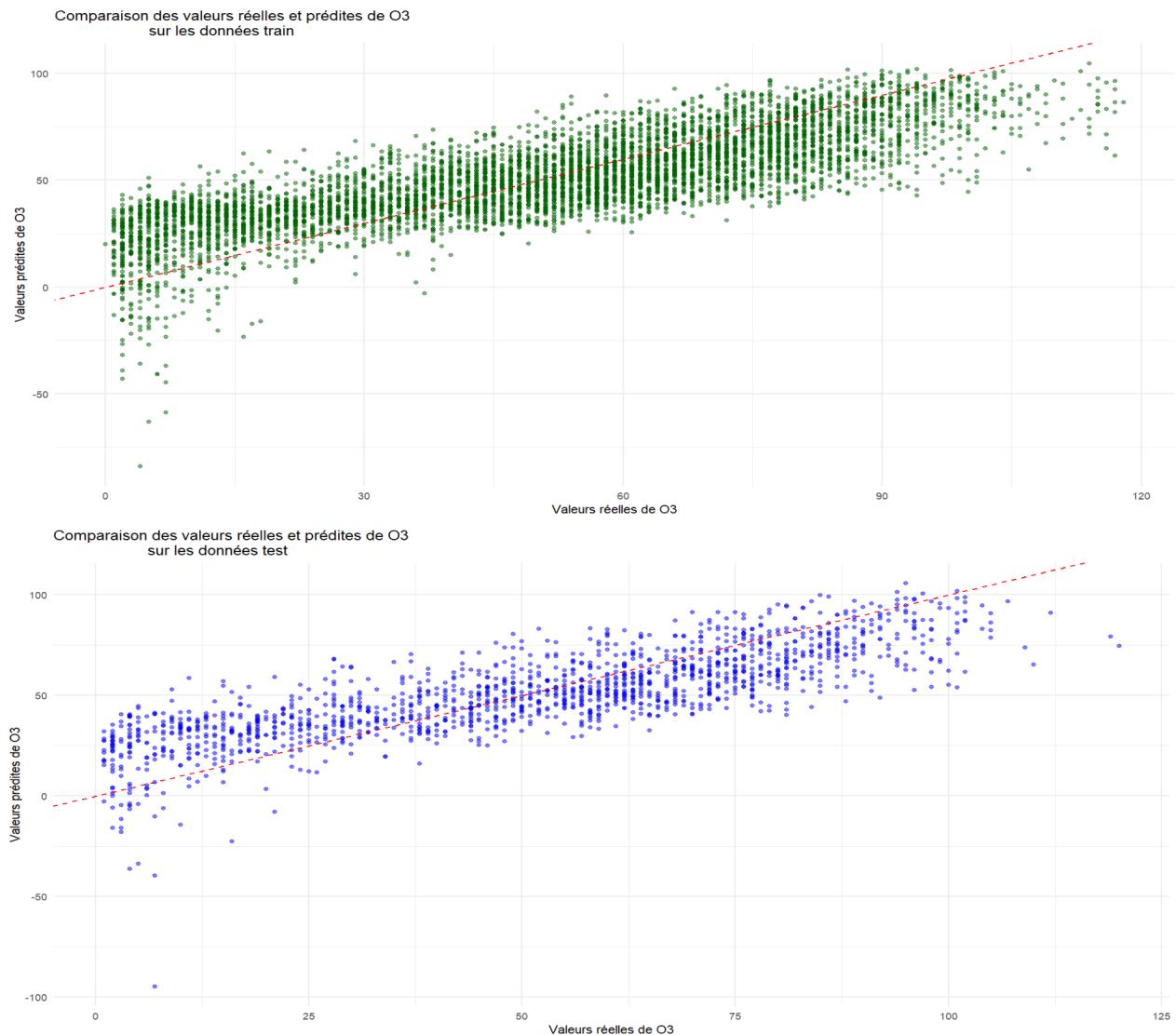
Ainsi, même si le modèle est purement linéaire, les signes et l'intensité des coefficients sont en cohérence avec certaines dynamiques atmosphériques connues.

Toutes les autres variables (comme la température, le point de rosée, les vents moyens, les températures min/max, etc.) ont été éliminées par la pénalisation, c'est-à-dire qu'elles ont des coefficients nuls.

Nous avons moins de variables, donc le modèle est plus facile à interpréter. En supprimant les variables les moins pertinentes, on réduit le bruit. Mais certaines variables potentiellement informatives ont été écartées (comme la température). Le modèle reste linéaire, ce qui limite sa capacité à capturer des relations complexes, par exemple, si la température et l'humidité sont corrélées, le Lasso peut en choisir une et pas l'autre sans que ce soit forcément "mieux".

Il est important de souligner que l'interprétation des coefficients dans un modèle Lasso doit être faite avec prudence, en raison des corrélations éventuelles entre les variables explicatives et de la nature pénalisée de l'estimation, qui induit un biais dans les coefficients sélectionnés. Le modèle met en évidence les variables les plus utiles pour la prédition, mais pas nécessairement les relations causales réelles.

**Visualisation des prédictions :** Les nuages de points suivants (le premier concernant les données train en vert et le deuxième concernant les données test en bleu) comparent les valeurs réelles et les valeurs prédictives de la concentration maximale d'ozone. Une ligne de référence  $y = x$  a été ajoutée pour faciliter l'interprétation. Plus les points sont proches de cette ligne, meilleure est la performance prédictive du modèle. :



On observe une tendance générale des points à se regrouper autour de la ligne diagonale (rouge pointillée), ce qui indique que le modèle parvient à reproduire les tendances globales des valeurs de concentration d'ozone.

Ensuite, on remarque que pour de faibles concentrations ( $O_3$  proche de 0), le modèle a tendance à surestimer les valeurs et pour les fortes concentrations (au-delà de  $80 \mu\text{g}/\text{m}^3$ ), le modèle sous-estime légèrement. Cela peut refléter un manque de représentativité de ces

cas extrêmes dans les données d'entraînement, ou une difficulté du modèle à capturer des effets non linéaires.

De plus, certains points sont très éloignés de la diagonale, traduisant des erreurs de prédiction marquées. Ces écarts pourraient être liés à des conditions météorologiques ou des pics de pollution atypiques que le modèle ne parvient pas à bien modéliser. Alors qu'entre 30 et 70  $\mu\text{g}/\text{m}^3$ , les prédictions semblent plus fiables, avec une faible dispersion autour de la diagonale. Cela est cohérent avec le fait que cette plage couvre la majorité des observations (valeurs les plus fréquentes).

**Performance de la méthode :** Les erreurs quadratiques moyennes (RMSE) obtenues sur les jeux d'entraînement et de test sont relativement proches : 15,87 pour l'apprentissage et 16,27 pour le test. Cette faible différence suggère que le modèle Lasso présente une bonne capacité de généralisation. Autrement dit, le modèle ne souffre ni de surapprentissage (overfitting), ni de sous-apprentissage (underfitting) marqués.

Cette stabilité des performances entre les deux ensembles est un effet direct de la régularisation  $L_1$ , qui limite la complexité du modèle en éliminant les variables peu informatives. En contrignant certains coefficients à zéro, le modèle évite de s'ajuster de manière excessive aux fluctuations spécifiques des données d'apprentissage, ce qui améliore sa robustesse.

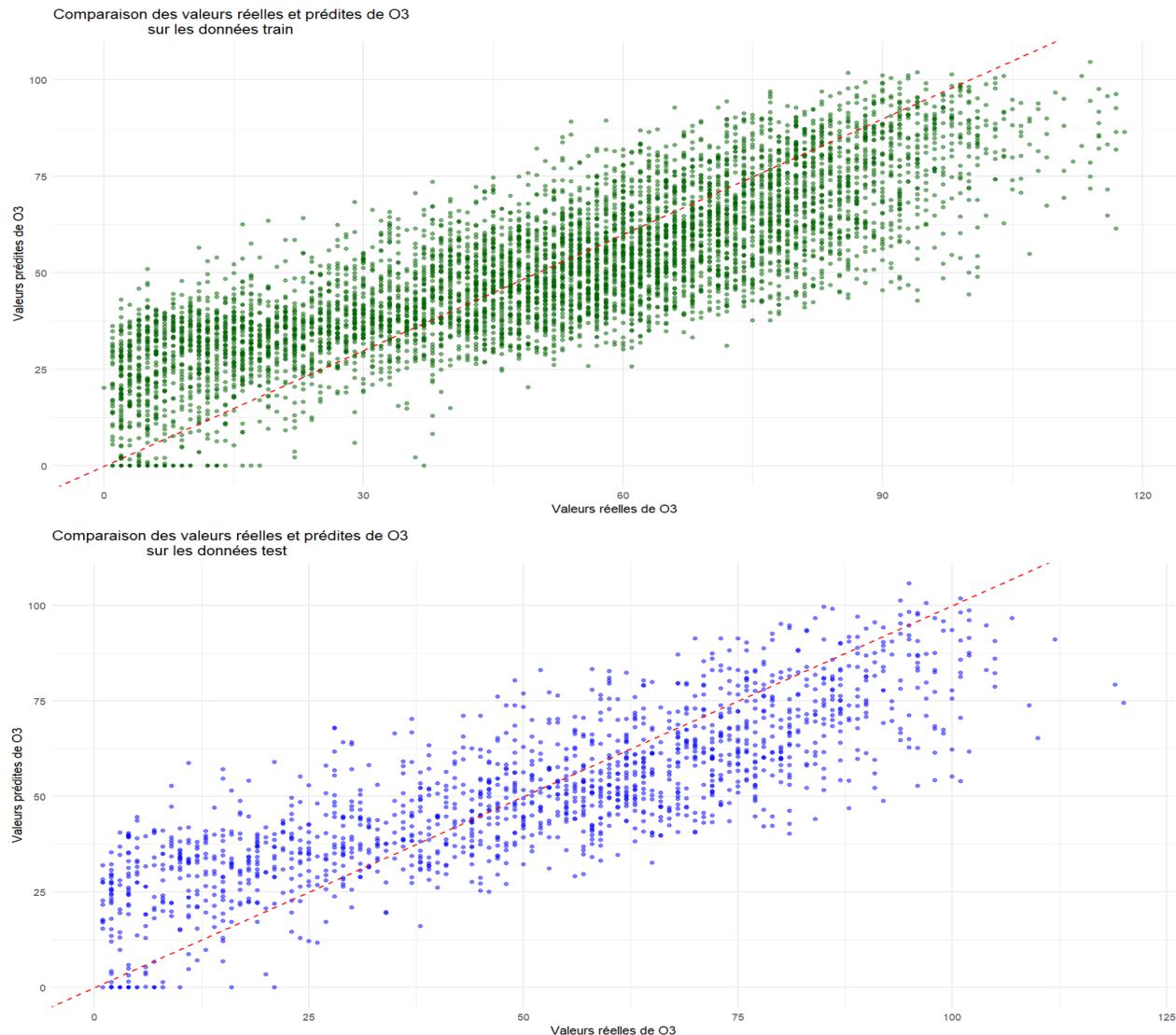
En pratique, un RMSE d'environ 16 peut être considéré comme acceptable dans le cadre de la prédiction de la concentration d'ozone, compte tenu de la variabilité naturelle des mesures environnementales. Toutefois, cette valeur reste à mettre en perspective avec les plages typiques de concentration observées, afin d'évaluer la précision relative du modèle.

Enfin, notons que l'approche utilisée ici repose sur une validation empirique, et que l'interprétation des résultats doit également tenir compte d'un éventuel effet d'instabilité dans la sélection des variables, notamment si des corrélations fortes existent entre les prédicteurs.

**Pourquoi certaines valeurs prédites sont négatives ?** Dans les graphiques, on remarque que certaines prédictions sont inférieures à zéro. Cela n'a pas de sens physiquement, car une concentration en ozone est nécessairement positive. En effet, le modèle de régression Lasso est un modèle linéaire, sans contrainte sur la positivité des sorties. Si la combinaison linéaire des variables explicatives donne un résultat négatif, le modèle le renverra tel quel. Cela se produit pour des observations avec des valeurs extrêmes (hors distribution principale) et cela peut être dû à un manque d'information explicative (exemple : météo perturbée, erreurs de mesure).

## Correction des prédictions négatives

Pour corriger ce problème, nous avons décidé que toute valeur prédite strictement inférieure à zéro est ramenée à 0



Visuellement, la distribution des points ne change pas significativement, mais les prédictions situées en dessous de zéro sont désormais projetées sur l'axe horizontal. Cette opération améliore l'interprétabilité des résultats et la crédibilité des prévisions.

#### Effets de cette correction :

- Les performances globales du modèle s'améliorent légèrement :
  - RMSE (train) passe de 15.87 à 15.61,
  - RMSE (test) passe de 16.27 à 15.88.

Cette opération garantit la cohérence physique des résultats sans modifier l'architecture du modèle. Elle est particulièrement utile dans les plages de faibles concentrations, où les erreurs de prédiction sont les plus fréquentes.

#### Évaluation du modèle Lasso sur un sous-échantillon

Afin d'étudier l'impact de la taille de l'échantillon sur la stabilité et la robustesse des méthodes de sélection de variables, nous avons extrait un sous-échantillon du jeu de données initial :

- $n_{\text{train}} = 500$  observations pour l'entraînement,
- $n_{\text{test}} = 125$  observations pour le test,
- $p = 24$  variables explicatives.

Le ratio *nombre de variables / taille de l'échantillon* devient ici plus défavorable, ce qui accentue les risques de surapprentissage, et rend la régularisation d'autant plus cruciale.

Nous avons à nouveau ajusté un modèle Lasso avec validation croisée, en retenant la régularisation correspondant à  $\lambda_{\text{lse}}$ .

#### Paramètre sélectionné :

- $\lambda_{\text{lse}} = 1.77$

Ce  $\lambda$  plus élevé que précédemment reflète le besoin de pénalisation plus forte pour un ensemble de données plus petit. En effet, moins d'observations signifie plus de variance dans l'estimation des coefficients, ce que le Lasso compense en forçant davantage de coefficients à zéro.

**Structure du modèle sélectionné :** Seules 11 variables ont été conservées sur les 24 initiales :

Variable	Coefficient estimé
(Intercept)	286.50
heure	0.66
NO	-1.04
pression	-0.19
pression_variation_3h	0.10
humidite	-0.58
vent_moyen	0.01
vent_rafales_10min	0.79
vent_direction	0.01
temperature_max	0.01
pluie_12h	0.05
pluie_cumul_0h	0.03

Comparé au modèle précédent entraîné sur l'ensemble des données, on observe quelques différences notables. Le modèle issu du sous-échantillon est légèrement plus restreint, avec une variable explicative en moins. De nouvelles variables à effet faible, comme `vent_moyen` ou `pluie_cumul_0h`, apparaissent, ce qui peut s'expliquer par des corrélations particulières à l'échantillon utilisé. À l'inverse, certaines variables jugées importantes dans le modèle global, telles que NO2 ou PM10, ne sont plus sélectionnées, ce qui illustre bien la sensibilité du Lasso aux variations d'échantillonnage.

#### Visualisation des prédictions :



Les nuages de points présentés ci-dessus comparent les valeurs réelles aux valeurs prédictes, à la fois sur les données d'apprentissage (en vert) et sur les données test (en bleu). On observe une plus forte dispersion autour de la diagonale, en particulier pour les données de test, ce qui indique une baisse de précision sur les observations non vues par le modèle. Certaines prédictions test apparaissent même en dehors de la plage plausible, avec des valeurs inférieures à 0 ou supérieures à 100, signalant une instabilité du modèle face aux cas atypiques. Néanmoins, l'alignement global autour de la diagonale reste raisonnable pour les concentrations situées dans l'intervalle central ( $30\text{--}70 \mu\text{g}/\text{m}^3$ ), ce qui montre une certaine capacité de généralisation dans les cas les plus fréquents.

### Performances de la méthode :

- RMSE sur les données d'entraînement : 16.50
- RMSE sur les données test : 18.58

On observe ici un écart plus marqué entre entraînement et test, signalant une légère perte de généralisation. Cela est attendu, car la taille réduite du jeu d'entraînement limite l'apprentissage des relations structurelles dans les données. La performance reste toutefois acceptable, grâce à la régularisation qui limite le surapprentissage.

**Conclusion sur le sous-échantillon :** Cette expérience confirme l'importance d'un bon rapport  $n/p$  pour les modèles de régression. Sur un petit échantillon, les performances se dégradent et les coefficients deviennent plus sensibles aux fluctuations de l'échantillonnage. Toutefois, la régularisation (via  $\lambda_{1se}$ ) permet de maintenir une certaine robustesse et interprétabilité, même en contexte défavorable.

**Correction des valeurs négatives :** Comme observé précédemment, le modèle Lasso étant une régression linéaire, il peut produire des prédictions négatives, ce qui est incohérent sur le plan physique puisque la concentration en ozone est nécessairement positive.

Ce phénomène est encore plus visible avec un sous-échantillon réduit : la variabilité accrue et la moindre représentativité des cas extrêmes peuvent accentuer les extrapolations erronées.

Pour remédier à cela, nous avons appliqué une correction simple : toute prédition négative est ramenée à 0.



Ces deux nouveaux nuages de points conservent la même structure que précédemment, mais avec les prédictions négatives projetées sur la ligne horizontale  $y = 0$ . Cette correc-

tion est visuellement discrète mais contribue à renforcer la robustesse globale du modèle, notamment pour les cas à faibles concentrations d'ozone.

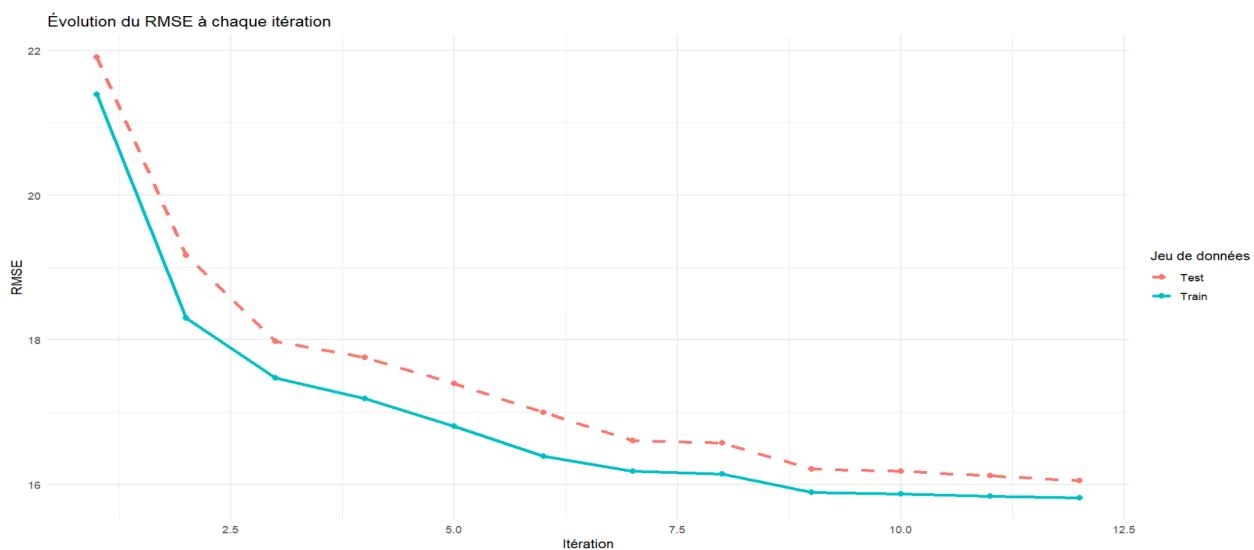
**Effets de cette correction :** L'application d'une contrainte de non-négativité sur les prédictions permet de corriger les valeurs incohérentes, notamment les concentrations négatives, tout en préservant les prédictions déjà physiquement plausibles. Cette correction a eu un effet bénéfique sur les performances du modèle : le RMSE sur l'ensemble d'apprentissage passe de 16.50 à 16.24, et sur l'ensemble test de 18.58 à 16.23. Ainsi, au-delà de l'amélioration quantitative, cette approche garantit une cohérence physique des prédictions sur l'ensemble du domaine de variation.

#### 4.2.3 Méthode type boosting

La méthode de boosting linéaire utilisée repose sur un enchaînement de régressions linéaires successives, chacune visant à corriger les résidus de la prédiction précédente. À chaque itération, une unique variable est sélectionnée afin de modéliser les erreurs restantes, ce qui permet de construire progressivement un modèle additif. Bien que chaque étape soit simple (régression univariée), l'ensemble du modèle peut capturer une structure plus complexe.

**Visualisation des prédictions** Comme pour les autres méthodes, nous avons représenté graphiquement les valeurs réelles de concentration d'ozone ( $O_3$ ) en fonction des valeurs prédites, séparément pour les jeux d'entraînement et de test. Ces graphiques permettent de juger de la qualité des prédictions et de détecter des biais.

**Performances de la méthode** Le modèle a été entraîné sur un ensemble de 6281 observations, et évalué sur 1571 données test. Nous avons réalisé 12 itérations afin d'obtenir le même nombre de variable que les autres modèles. Comme nous pouvons choisir le nombre d'itération, cela permettra par la suite de comparer les différentes méthodes. Il est alors possible de connaître le RMSE à chaque étape.



Nous pouvons voir clairement une baisse du RMSE en ajoutant des variables, ce qui est parfaitement cohérent. Nous constatons aussi que le RMSE des données de test ne s'éloignent pas trop du RMSE d'entraînement.

Nous allons nous intéresser tout particulièrement aux valeurs et variables obtenues après 12 opérations. Voici les performances obtenues avant et après correction des prédictions négatives :

— **Avant correction :**

- RMSE (train) : 15,82
- RMSE (test) : 16,06

— **Après correction (prédictions négatives ramenées à zéro) :**

- RMSE corrigé (train) : 15,49
- RMSE corrigé (test) : 15,56

**Variables sélectionnées et interprétation des coefficients** Au total, 12 variables ont été sélectionnées, au fil des itérations successives. À chaque étape, un modèle de régression univariée est ajusté sur les résidus précédents, en sélectionnant la variable qui explique le mieux les erreurs restantes. Voici la liste des variables sélectionnées dans l'ordre, ainsi que le coefficient estimé à chaque itération :

Itération	Variable	Intercept	Coefficient
1	humidite	141,23	-1,11
2	NO	9,20	-0,92
3	pluie_cumul_0h	-2,30	1,37
4	pression_variation_3h	-0,08	2,80
5	pression	340,54	-0,34
6	heure	-4,13	0,36
7	vent_direction	-4,60	0,02
8	temperature	3,71	-0,28
9	vent_rafales_10min	-1,76	0,31
10	pluie_24h	0,38	-0,15
11	temperature_min	-0,95	0,10
12	vent_rafales	-0,05	0,53

La première variable sélectionnée est humidité, avec un coefficient négatif, ce qui confirme le lien négatif entre humidité et concentration d'ozone. L'ozone se forme plus facilement dans des conditions sèches et ensoleillées, ce qui explique pourquoi une augmentation de l'humidité tend à freiner ce processus.

Les deux variables sélectionnées ensuite sont NO (-0,92) et pluie\_cumul\_0h (1,37). Le monoxyde d'azote est un précurseur de l'ozone, mais il peut aussi le détruire par réaction directe. Son effet négatif dans le modèle suggère que, dans les conditions observées, son rôle destructeur prédomine. L'effet positif de pluie\_cumul\_0h est plus difficile à interpréter directement, mais pourrait refléter des conditions météorologiques particulières qui n'empêchent pas la formation d'ozone dans les heures précédentes.

La variable pression\_variation\_3h est sélectionnée avec un fort effet positif, cela peut refléter des conditions atmosphériques propices à l'accumulation des polluants. À l'inverse, la pression a un effet négatif. Les coefficients, étant adaptées aux erreurs résiduelles, sont

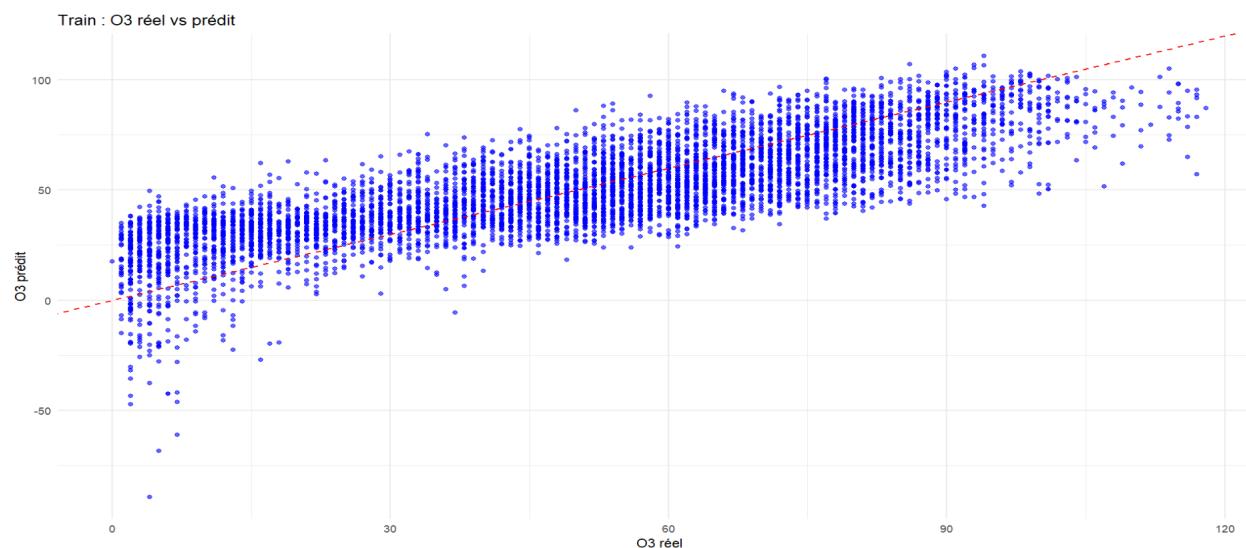
aussi très largement impactés par ces dernières et peuvent donc être très différents de ceux auxquels on s'attendrait.

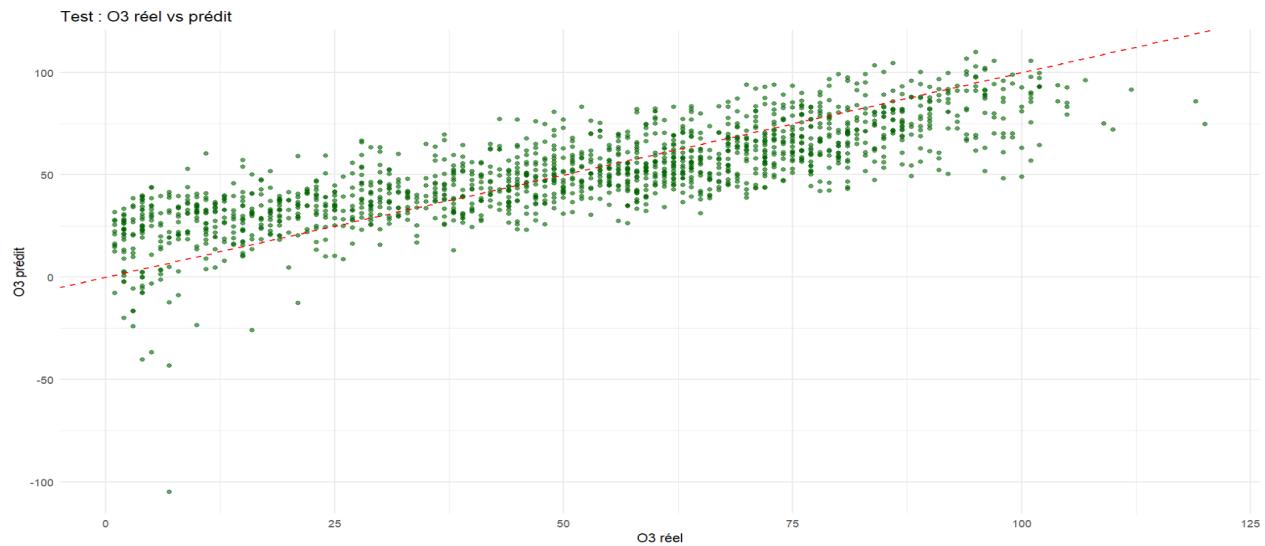
L'heure de la journée heure est également sélectionnée, traduisant les fortes variations journalières typiques de l'ozone, avec des concentrations plus élevées l'après-midi. Son effet dépend du codage de la variable, mais son inclusion confirme l'importance du rythme journalier dans la dynamique de l'ozone.

Plusieurs variables météorologiques liées au vent et à la température apparaissent ensuite. `vent_direction` et `vent_rafales_10min` ont des effets positifs modestes, indiquant que certaines orientations ou intensités du vent peuvent favoriser ou empêcher la dispersion locale. La température, bien que souvent associée à une hausse de l'ozone en journée, présente ici un effet négatif. Ce paradoxe apparent pourrait être dû à des effets de compensation ou d'interactions complexes avec d'autres variables.

Enfin, des variables liées aux précipitations ou aux extrêmes sont intégrées en fin de modèle : `pluie_24h`, `temperature_min` et `vent_rafales`. Ces variables ont des effets plus faibles, mais contribuent à affiner les résidus du modèle. Leur sélection tardive reflète leur utilité marginale dans l'explication directe de l'ozone, tout en soulignant la capacité du boosting à ajuster des aspects plus fins du phénomène.

Dans l'ensemble, la dynamique des coefficients suit une logique attendue : les premières variables ont des coefficients plus marqués, et les suivantes affinent le modèle de manière incrémentale. Cette hiérarchisation naturelle des effets est typique de la méthode de boosting, qui sélectionne progressivement les variables les plus explicatives, puis améliore les prédictions en ajoutant des termes correcteurs plus subtils.

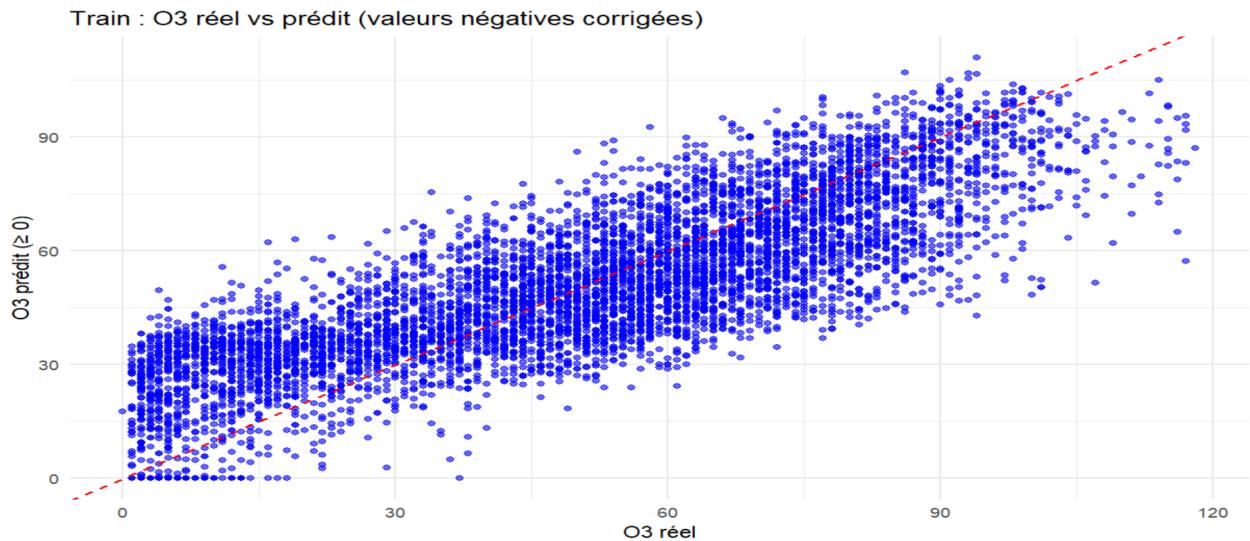


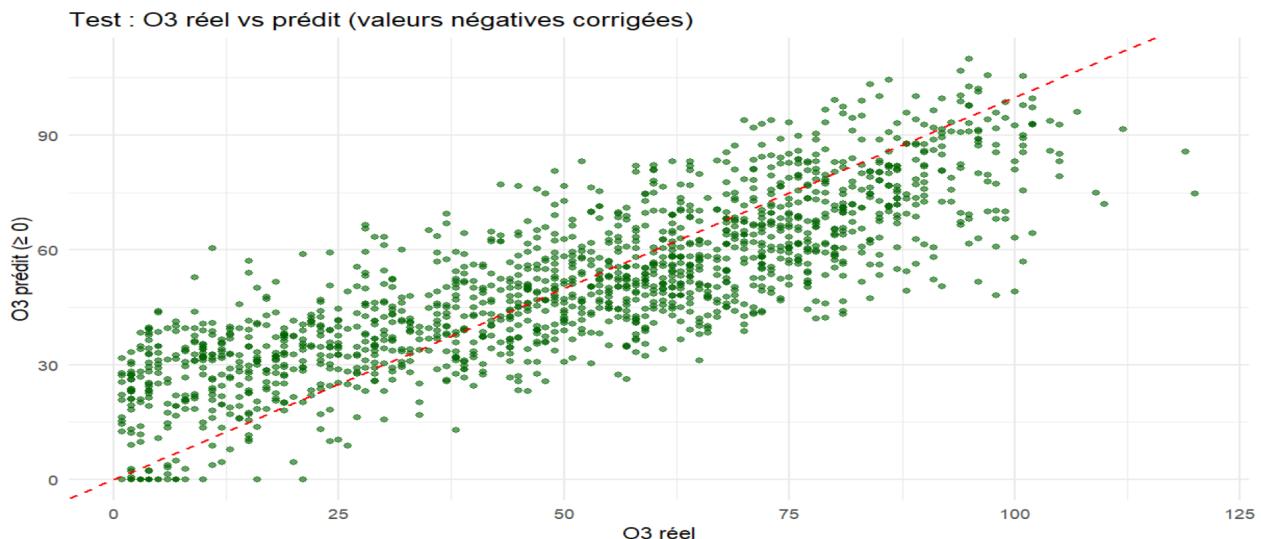


**Correction des prédictions négatives** Certaines prédictions produites par ce modèle sont négatives, ce qui n'est pas physiquement interprétable. Une concentration d'ozone ne peut être inférieure à zéro. Nous avons donc appliqué une simple transformation :

*Toute prédiction < 0 est ramenée à 0.*

Cela garantit la cohérence physique des résultats, sans modifier la structure du modèle.





Visuellement, la correction ne modifie que peu la forme globale des prédictions, mais permet de ramener certaines valeurs à un niveau physiquement valide. Cela améliore légèrement le RMSE, en particulier sur les données test, et renforce la crédibilité du modèle. Nous n'utiliserons cependant pas ce résultat modifié manuellement car il n'est pas directement dû aux méthodes utilisées. De plus, cette modification n'est possible qu'en tenant compte du contexte. Afin de faire une comparaison plus générale, nous préférerons donc conserver les résultats initiaux.

En revanche, cette amélioration des résultats après remise à 0 des résultats négatifs nous montrent qu'il est important de tenir compte du contexte afin de prédire au mieux les valeurs des variables souhaitées.

**Conclusion sur la méthode de boosting** La méthode de type boosting linéaire offre un compromis intéressant entre simplicité et performance. Elle reste interprétable, chaque itération n'ajoutant qu'un seul terme linéaire. De plus, le fait d'itérer successivement permet de mieux capter la structure des résidus et de choisir plus simplement la complexité du modèle.

Comparée à d'autres méthodes, elle est :

- Plus flexible qu'un modèle linéaire classique, car elle peut capturer des interactions faibles de manière additive.
- Moins sensible au surapprentissage que des modèles trop complexes, tout en gardant une bonne performance.
- Interprétable, car chaque ajout de variable est identifiable et les effets sont mesurables.

Dans notre cas, cette méthode atteint un RMSE comparable à celui des autres méthodes, tout en maintenant une structure simple et explicative. C'est une alternative rigoureuse et efficace pour modéliser les concentrations d'ozone. Nous comparerons les méthodes de manière plus approfondie par la suite.

### Évaluation du modèle Boosting Linéaire sur un sous-échantillon

Afin d'évaluer la robustesse de la méthode de boosting linéaire dans un contexte de données réduites, nous avons appliqué la même procédure sur un sous-échantillon du jeu initial :

- $n_{\text{train}} = 500$  observations pour l'entraînement,
- $n_{\text{test}} = 125$  observations pour le test,
- $p = 24$  variables explicatives.

Le ratio défavorable  $n/p$  rend la sélection de variables et l'évitement du surapprentissage encore plus critiques. Le boosting linéaire, en construisant progressivement un modèle via l'agrégation de régressions univariées sur les résidus, offre une approche adaptative bien adaptée à ce cadre.

**Structure du modèle sélectionné :** Le modèle final est obtenu après 11 itérations. À chaque étape, une nouvelle variable est ajoutée, sélectionnée comme la plus corrélée aux résidus courants. Les coefficients estimés à chaque itération révèlent la contribution spécifique de chaque variable au processus de prédiction. Le tableau ci-dessous résume les variables sélectionnées et les coefficients associés :

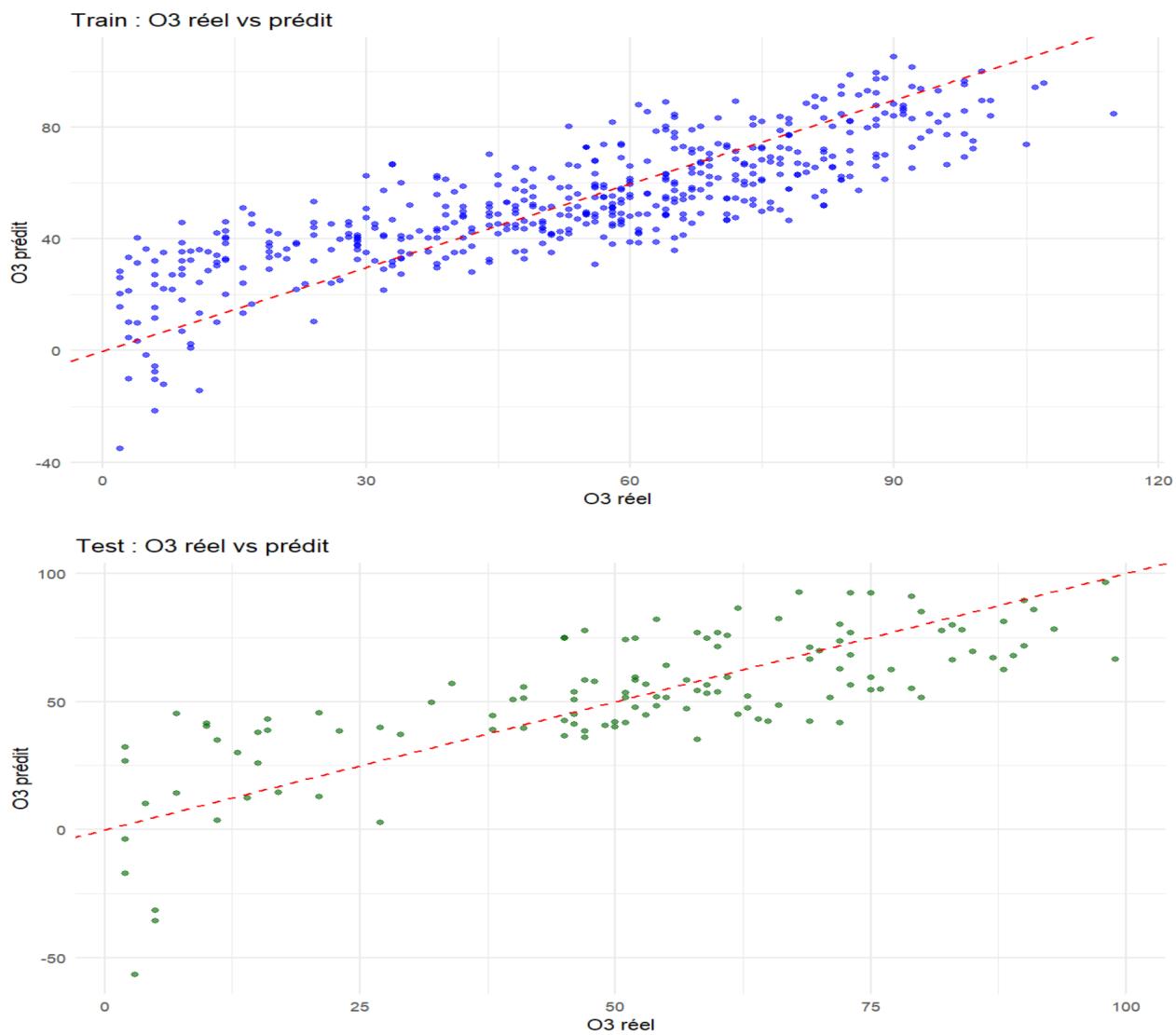
Itération	Variable	Intercept	Coefficient
1	humidite	141.79	-1.09
2	NO	10.88	-1.26
3	pluie_cumul_0h	-1.90	1.07
4	pression_variation_3h	-0.17	2.91
5	pression	342.69	-0.34
6	NO2	-2.03	0.31
7	vent_direction	-5.01	0.02
8	pluie_intensite_max_1h	-0.21	0.41
9	vent_rafales_10min	-1.48	0.26
10	pluie_24h	0.66	-0.25
11	temperature	2.29	-0.17

Sur le jeu de données réduit, le modèle sélectionne également humidité en premier, avec un effet négatif marqué (-1,09), cohérent avec l'idée que l'ozone est favorisé par des conditions sèches. Suivent les polluants NO (-1,26) et NO2 (0,31), reflétant leurs rôles opposés dans la chimie de l'ozone : effet destructeur pour le premier, contribution à la formation pour le second.

Plusieurs variables météorologiques sont ensuite introduites : pression\_variation\_3h a un effet positif important (2,91), tandis que pression et température ont des effets plus modérés et négatifs. Ces coefficients peuvent traduire à la fois des relations physiques plausibles et des ajustements liés aux résidus du modèle.

Les variables liées aux précipitations et au vent (pluie\_cumul\_0h, pluie\_24h, vent\_rafales\_10min, etc.) ont des effets plus modestes, mais contribuent à affiner les prédictions. Certaines associations peuvent paraître contre-intuitives (comme l'effet positif de la pluie instantanée), mais s'interprètent dans le contexte de la dynamique résiduelle du boosting.

On retrouve enfin une hiérarchisation similaire à celle observée sur le jeu complet : les premières variables apportent l'essentiel de l'explication, tandis que les suivantes viennent corriger les erreurs résiduelles. Cela confirme la capacité du boosting à construire un modèle explicatif de manière progressive et structurée, même sur un sous-échantillon.



**Performances de la méthode :** Les performances initiales du modèle (sans contrainte sur les prédictions) sont les suivantes :

- RMSE (train) : 14,70
- RMSE (test) : 17,59

Comme précédemment, nous avons ramené à 0 toute prédition négative pour assurer la cohérence physique du modèle (la concentration d'ozone ne pouvant être négative).



Après correction, les performances sont légèrement améliorées :

- RMSE corrigé (train) : 14,45
- RMSE corrigé (test) : 15,96

Cette correction simple améliore non seulement la cohérence du modèle mais aussi ses performances globales sur les deux ensembles.

**Conclusion sur le sous-échantillon :** Le boosting linéaire démontre ici une bonne capacité de généralisation, même avec un échantillon réduit. L'approche itérative permet de régulariser implicitement l'apprentissage, et la sélection progressive des variables favorise une interprétabilité naturelle du modèle. La correction des prédictions négatives reste une étape essentielle pour garantir la cohérence physique du modèle final.

En somme, le boosting linéaire apparaît comme une méthode robuste et performante dans des contextes de données limitées, et se distingue favorablement du Lasso dans cette configuration.

**Comparaison entre l'échantillon complet et le sous-échantillon :** L'analyse comparative entre les deux tailles d'échantillons permet de mieux comprendre les effets de la

quantité de données sur le comportement du boosting linéaire. Une comparaison des coefficients obtenus sur l'échantillon complet et sur le sous-échantillon permet d'évaluer la stabilité du processus de sélection de variables et des effets estimés dans le modèle de boosting linéaire.

Les deux modèles présentent une grande cohérence quant aux variables sélectionnées : 10 des 11 variables choisies dans le sous-échantillon figurent également parmi les 12 variables sélectionnées dans le modèle ajusté sur l'échantillon complet. Cette convergence atteste de la robustesse du processus de sélection séquentielle, malgré la réduction drastique de la taille de l'échantillon.

Les signes des coefficients sont également largement cohérents entre les deux jeux de données. Par exemple, les variables humidité, NO, pression ou encore température sont associées à des coefficients négatifs dans les deux cas, traduisant un effet défavorable sur la concentration d'ozone. De même, pluie\_cumul\_0h ou pression\_variation\_3h sont systématiquement associées à un effet positif.

Cependant, on note que les valeurs précises des coefficients peuvent varier d'un jeu à l'autre, en particulier pour les variables sélectionnées dans les dernières itérations, comme vent\_rafales\_10min ou pluie\_24h. Cela reflète une certaine sensibilité aux fluctuations de l'échantillon.

En résumé, cette comparaison montre que le modèle de boosting linéaire reste relativement stable en termes de sélection de variables et de sens des effets, même lorsqu'il est entraîné sur un sous-échantillon réduit. Cela constitue un argument en faveur de la robustesse de cette méthode dans un contexte de données limitées.

D'un autre côté, nous pouvons également comparer les RMSE obtenus :

- Sur l'échantillon complet ( $n_{\text{train}} = 6281$ ,  $n_{\text{test}} = 1571$ ), le modèle a un RMSE de 15,82 sur l'entraînement et 16,06 sur le test. Les performances sont équilibrées, avec un bon pouvoir de généralisation.
- En revanche, sur le sous-échantillon ( $n_{\text{train}} = 500$ ,  $n_{\text{test}} = 125$ ), bien que le RMSE d'entraînement soit légèrement meilleur (14,70), celui obtenu sur le test est sensiblement plus élevé (17,59), signalant une généralisation moins efficace.

Ces différences s'expliquent principalement par la variabilité accrue sur les petits échantillons. Le modèle y est plus sensible aux fluctuations aléatoires et aux corrélations particulières du sous-échantillon. Cela se reflète également dans la structure des modèles : bien que plusieurs variables soient communes (humidité, NO, pression, vent\_direction, pluie\_cumul\_0h, température), d'autres divergent (ex. NO2 sélectionnée uniquement dans le sous-échantillon, heure et température\_min uniquement dans le modèle global). Les jeux de données ayant moins d'observations sont aussi bien plus sensibles au surapprentissage pour un même nombre de variable, expliquant la plus grande différence entre les données d'entraînement et celles de test.

En résumé, si le boosting linéaire conserve une bonne efficacité sur des jeux de données réduits, la stabilité des variables sélectionnées et la performance en test s'améliorent clairement avec un plus grand volume de données. Cela souligne l'intérêt de disposer d'échantillons suffisamment riches pour assurer une sélection robuste et une bonne capacité de généralisation.

#### 4.2.4 Comparaison des résultats avec jeu de données complet

##### Variables sélectionnées par méthode

Dans cette section, nous comparons les variables sélectionnées par différentes méthodes de modélisation appliquées à la prédiction des concentrations d'ozone ( $O_3$ ) à Périgueux. Le tableau ci-dessous synthétise les variables retenues par chaque approche. Une croix (✓) indique que la variable a été retenue par la méthode correspondante.

Variable	Corrélation	AIC	Lasso	Boosting
heure	✓	✓	✓	✓
NO	✓	✓	✓	✓
pression	✓	✓	✓	✓
vent_moyen	✓			
temperature_min	✓			✓
NO2	✓		✓	
PM10	✓	✓	✓	
humidite	✓	✓	✓	✓
vent_rafales_10min	✓	✓	✓	✓
temperature_max	✓			
temperature	✓			✓
point_de_rosee	✓			
vent_direction	✓	✓	✓	✓
pluie_cumul_0h		✓		✓
pression_variation_3h		✓	✓	✓
pluie_12h		✓	✓	
pluie_1h		✓	✓	
pluie_24h		✓		✓
pluie_3h			✓	
vent_rafales				✓

TABLE 1 – Variables sélectionnées par méthode de modélisation

##### Analyse des variables sélectionnées

On observe que certaines variables, telles que NO, humidité, pression et heure, sont systématiquement sélectionnées par toutes les méthodes. Cela traduit une forte corrélation avec la variable cible  $O_3$  et une contribution explicative robuste, quel que soit l'algorithme utilisé.

À l'inverse, d'autres variables comme vent\_moyen, point\_de\_rosee ou encore pluie\_3h ne sont retenues que par une seule méthode, suggérant une importance plus marginale ou conditionnelle. Il est également intéressant de noter que les méthodes comme l'AIC ou le Lasso tendent à intégrer davantage de variables météorologiques relatives aux précipitations ou aux variations de pression, souvent moins bien captées par une approche uniquement corrélative.

Ces différences peuvent également s'expliquer par les mécanismes internes des méthodes : par exemple, la régression pénalisée Lasso favorise les variables explicatives non redondantes en introduisant une régularisation, ce qui peut expliquer l'exclusion de variables très corrélées entre elles.

### Comparaison des performances (RMSE)

Nous présentons ci-dessous les erreurs quadratiques moyennes (RMSE) obtenues pour chaque méthode, à la fois sur les données d'entraînement et de test.

Méthode	RMSE Entraînement	RMSE Test
Corrélation	17,23	18,00
AIC	15,77	15,84
Lasso	15,87	16,27
Boosting (type)	15,82	16,06

TABLE 2 – Performances des modèles selon la méthode (RMSE)

Les résultats montrent que les méthodes basées sur l'optimisation du critère AIC et sur le boosting produisent des performances nettement meilleures que la sélection simple par corrélation, tant sur les données d'entraînement que sur les données test. Cela s'explique par le fait que la méthode par corrélation ne tient pas compte des interactions ou de la redondance entre variables.

La régression Lasso donne des performances proches de l'AIC et du boosting, ce qui confirme son efficacité pour la sélection automatique de variables pertinentes.

Il est intéressant de remarquer que l'écart entre les erreurs d'entraînement et de test est faible pour toutes les méthodes, ce qui indique que le sur-apprentissage est limité. Néanmoins, la méthode par corrélation affiche une erreur test plus élevée, montrant une moins bonne généralisation.

### Conclusion

La comparaison des différentes méthodes de sélection de variables et de modélisation met en évidence plusieurs éléments importants à considérer.

Tout d'abord, les méthodes automatiques comme le critère AIC, la régression pénalisée (Lasso) et le boosting offrent des performances supérieures à la sélection par corrélation simple, tant en termes de RMSE que de pertinence des variables sélectionnées. Ces approches tiennent compte des effets combinés entre variables et évitent de conserver des variables redondantes, ce qui améliore la capacité de généralisation du modèle.

La méthode AIC se distingue par une excellente performance et une complexité modérée. Elle permet de construire un modèle relativement compact et explicatif, sans recours à des pénalisations ou à des algorithmes itératifs complexes. Le Lasso, quant à lui, est robuste face à la multicolinéarité, et il est bien adapté lorsqu'on suspecte qu'un sous-ensemble réduit de variables suffit à expliquer la variable cible.

La méthode type boosting, bien qu'elle repose uniquement sur des régressions linéaires simples à chaque étape, s'avère extrêmement compétitive. Elle construit le modèle de manière incrémentale, ce qui permet de corriger progressivement les erreurs de prédiction. Cette approche présente l'avantage d'être très simple à implémenter, tout en étant performante. De plus, le fait qu'elle sélectionne les variables une par une rend le processus de sélection plus interprétable.

En revanche, la sélection par corrélation brute, bien que rapide, se montre moins efficace et ne tient pas compte des interactions complexes entre variables. Elle peut être utile pour une première exploration des données, mais reste insuffisante pour produire un modèle robuste.

En résumé, les méthodes les plus adaptées dans notre contexte sont le critère AIC et la méthode type boosting. Elles présentent un bon compromis entre performance prédictive, stabilité des résultats, complexité de mise en œuvre et interprétabilité. Le Lasso reste une alternative solide, notamment dans les contextes de données plus larges et plus bruitées. La méthode par corrélation, en revanche, devrait être utilisée avec prudence, car elle peut conduire à des modèles trop simples et moins fiables.

#### 4.2.5 Comparaison des résultats avec jeu de données réduit

Afin d'améliorer notre analyse, nous répliquons notre démarche précédente sur une sous-partie du jeu de données initial, comportant 500 observations pour l'entraînement et 125 pour le test, sans chevauchement entre les deux ensembles.

#### Variables sélectionnées selon les méthodes

Le tableau ci-dessous récapitule les variables sélectionnées par chacune des quatre méthodes (corrélation, AIC, Lasso, Boosting).

Variable	Corrélation	AIC	Lasso	Boosting
heure	✓	✓	✓	
NO	✓	✓	✓	✓
pression	✓	✓	✓	✓
vent_moyen	✓		✓	
temperature_min	✓			
NO2	✓	✓		✓
PM10	✓			
humidite	✓	✓	✓	✓
vent_rafales_10min	✓	✓	✓	✓
temperature_max	✓		✓	
temperature	✓			✓
point_de_rosee	✓			
vent_direction	✓	✓	✓	✓
pluie_cumul_0h		✓	✓	✓
pression_variation_3h		✓	✓	✓
pluie_12h		✓	✓	
pluie_1h				
pluie_3h				
pluie_24h				✓
pluie_intensite_max_1h		✓		✓

TABLE 3 – Variables sélectionnées selon les différentes méthodes sur le jeu de données réduit (500/125)

### Analyse des variables sélectionnées

On observe dans le tableau 3 des similitudes intéressantes entre les méthodes, notamment sur certaines variables clés comme NO, pression, humidité, vent\_rafales\_10min et vent\_direction, qui sont sélectionnées par presque toutes les méthodes. Cela suggère qu'elles ont une influence significative sur la concentration d'ozone (O3) et qu'elles sont robustes à la méthode de sélection.

En revanche, d'autres variables comme pluie\_intensite\_max\_1h, temperature, ou pluie\_24h n'apparaissent que dans certaines méthodes (boosting ou AIC), ce qui peut refléter des spécificités méthodologiques. Par exemple, le boosting tend à inclure certaines variables peu corrélées globalement mais fortement explicatives localement dans une régression simple à une étape précise.

Le lasso, en tant que méthode pénalisée, équilibre parcimonie et performance, ce qui explique pourquoi il évite certaines variables fortement corrélées entre elles, comme temperature et temperature\_max. À l'inverse, la méthode par corrélation tend à surreprésenter les variables redondantes en raison de l'absence de mécanisme de régularisation.

Enfin, l'AIC privilégie un compromis entre qualité d'ajustement et complexité, ce qui peut expliquer la sélection d'un nombre modéré de variables (10 ici) tout en incluant à la fois des variables météorologiques (pluie, vent) et des polluants.

## Comparaison des RMSE

Les performances des différentes méthodes sont résumées dans le tableau suivant :

Méthode	RMSE Entraînement	RMSE Test
Corrélation	15,45	17,70
AIC	14,43	15,84
Lasso	16,50	18,58
Boosting (type)	14,70	17,59

TABLE 4 – RMSE d’entraînement et de test sur le jeu de données réduit

La méthode AIC obtient la meilleure performance sur les données de test avec un RMSE de 15,84, meilleur que celui du boosting 17,59 et significativement meilleur que celui obtenu par la méthode de corrélation 17,70 ou le lasso 18,58.

Le lasso, bien qu’élégant théoriquement, semble ici plus sensible à la réduction de la taille d’échantillon, ce qui peut s’expliquer par sa dépendance à un bon réglage du paramètre de régularisation et à la stabilité des corrélations dans les petits jeux de données.

La méthode de corrélation montre des signes de surapprentissage : le RMSE d’entraînement est plutôt bon 15,45, mais la performance se dégrade en test 17,70. Cela montre que la simple corrélation ne garantit pas une généralisation efficace, surtout avec un nombre important de variables dans un jeu réduit.

Enfin, la méthode de type boosting reste très compétitive malgré sa simplicité : en n’utilisant que des régressions simples et sans paramètre à régler, elle obtient un RMSE d’entraînement très bas 14,70 tout en restant raisonnablement performant en test 17,59.

## Conclusion

Dans ce contexte de jeu de données réduit (500 observations pour l’entraînement, 125 pour le test), plusieurs enseignements peuvent être tirés.

Premièrement, on observe que la réduction de la taille de l’échantillon renforce les différences de performances entre les méthodes. Alors que les écarts étaient modérés sur le jeu complet, le lasso et la méthode de corrélation voient leurs erreurs de test s’accroître significativement, suggérant une sensibilité plus marquée au surapprentissage ou à l’instabilité des relations dans les petits échantillons. En revanche, les méthodes AIC et boosting conservent des performances satisfaisantes.

Deuxièmement, la complexité algorithmique joue un rôle important dans l’interprétabilité et la robustesse. La méthode de type boosting, bien qu’extrêmement simple à chaque étape (régression linéaire simple), parvient à rivaliser avec les méthodes classiques. Son principe itératif permet d’ajouter progressivement de l’information, ce qui la rend plus souple et naturellement résistante au surapprentissage, même sans paramètre à ajuster.

Troisièmement, la méthode AIC s’impose ici comme un bon compromis entre complexité, robustesse et performance. Elle sélectionne un nombre modéré de variables et atteint la meilleure généralisation sur les données test, ce qui la rend particulièrement adaptée dans les contextes où la quantité de données est limitée.

En résumé, dans un cadre de données réduites :

- la méthode par AIC semble la plus fiable et performante ;
- le boosting reste très compétitif, simple à mettre en œuvre, et relativement robuste ;
- le lasso nécessite davantage d’observations pour révéler tout son potentiel ;

- la sélection par corrélation est trop naïve dans un petit jeu de données et peut mener à un surapprentissage.

Ce travail illustre l'importance de choisir une méthode adaptée non seulement au problème mais aussi à la taille de l'échantillon et à la nature des variables disponibles.

## 5. Avantages et limites

Avantages et limites des méthodes de régression linéaire classique (corrélation et AIC)

### Avantages :

- **Simplicité et accessibilité** : ces méthodes sont intuitives, faciles à implémenter et interpréter, ce qui les rend utiles pour une première exploration des données ;
- **Rapidité d'exécution** : elles sont peu coûteuses en temps de calcul, même sur de grands jeux de données ;
- **Bonne performance en cas de faible multicolinéarité** : la régression linéaire donne de bons résultats lorsque les variables explicatives sont peu corrélées entre elles et que la relation avec la cible est linéaire ;
- **Modèles parcimonieux** : l'AIC cherche à minimiser l'erreur tout en pénalisant la complexité, ce qui permet de construire des modèles compacts.

### Limites :

- **Sensibilité à la multicolinéarité** : en présence de variables explicatives corrélées, les coefficients peuvent devenir instables, ce qui nuit à l'interprétation et à la généralisation du modèle ;
- **Méthode de sélection parfois naïve** : la sélection via corrélation univariée ignore les effets combinés des variables et peut écarter des variables pertinentes si elles sont corrélées à la cible de manière non linéaire ou en interaction avec d'autres ;
- **Hypothèse forte de linéarité** : la forme du modèle impose une relation linéaire entre chaque variable sélectionnée et la cible, ce qui peut limiter les performances prédictives dans des situations plus complexes ;
- **Risque de surajustement en AIC** : bien que l'AIC pénalise les modèles complexes, il peut parfois sélectionner un trop grand nombre de variables si le critère n'est pas suffisamment sévère, surtout en présence de bruit.

## Avantages et limites de la régression pénalisée L1

### Avantages :

- **Sélection automatique des variables pertinentes** : le Lasso effectue naturellement une sélection de variables en mettant certains coefficients à zéro, simplifiant ainsi l'interprétation du modèle ;
- **Réduction du surapprentissage** : la régularisation introduite par la pénalisation  $L_1$  permet de contrôler la complexité du modèle et d'améliorer sa capacité de généralisation ;
- **Modèle stable et parcimonieux** : en imposant une forme de parcimonie, le Lasso fournit souvent des modèles plus simples, particulièrement utiles dans un contexte avec beaucoup de variables explicatives.

### Limites :

- **Problèmes en présence de variables corrélées** : lorsque plusieurs variables sont fortement corrélées, le Lasso tend à en sélectionner une seule de façon arbitraire, ce qui peut nuire à l'interprétation ;
- **Instabilité en haute dimension** : la méthode peut devenir instable lorsque le nombre de variables dépasse largement le nombre d'observations ;
- **Dépendance au choix du paramètre  $\lambda$**  : la performance du modèle dépend fortement du réglage du paramètre de régularisation, nécessitant une validation croisée rigoureuse pour un choix optimal.

## Avantages et limites de la régression type boosting

### Avantages :

- **Sélection de variables pertinente** : le modèle sélectionne itérativement les variables les plus corrélées à l'erreur, ce qui permet de construire une modélisation explicative concise et efficace ;
- **Facilité d'interprétation** : chaque itération repose sur une régression simple, facilitant la lecture et la compréhension de l'effet de chaque variable sélectionnée ;
- **Réduction des erreurs résiduelles** : l'approche additive permet de corriger progressivement les erreurs, améliorant ainsi la précision des prédictions ;
- **Contrôle de la complexité** : en limitant le nombre d'itérations, on évite de surajuster les données et la méthode est simple à mettre en place.

### Limites :

- **Sensibilité au choix des variables** : la sélection dépend fortement de la corrélation avec les résidus à chaque étape ; certaines variables utiles pourraient être négligées si elles ne sont pas sélectionnées tôt ;
- **Hypothèse de linéarité** : chaque sous-modèle étant une régression linéaire simple, le modèle final suppose des relations linéaires entre les variables sélectionnées et la cible, ce qui peut limiter la capture de relations non linéaires ;
- **Choix des coefficients** : en sélectionnant une variable à chaque étape, les coefficients sont figés et ne sont pas forcément les plus optimisés ;
- **Nécessite un bon prétraitement et une compréhension du contexte** : la qualité du modèle dépend fortement du nettoyage initial des données (suppression des NA, variables non pertinentes, etc.). De plus, le modèle peut produire des prédictions hors domaine (notamment négatives), qu'il convient alors de corriger a posteriori

## 6. Conclusion

Dans ce travail, nous avons exploré différentes approches de sélection de variables appliquées à la prédiction de la concentration d'ozone ( $O_3$ ) à partir de données météorologiques. L'objectif était double : identifier les variables explicatives les plus pertinentes pour le phénomène étudié, et comparer les performances prédictives des modèles issus de ces sélections.

Nous avons tout d'abord étudié deux méthodes classiques, la sélection par corrélation et celle fondée sur le critère d'information AIC. Simples à mettre en œuvre, elles permettent de construire rapidement des modèles linéaires interprétables. Toutefois, leur efficacité est conditionnée à certaines hypothèses, notamment l'absence de multicolinéarité et la linéarité des relations. Nos résultats ont mis en évidence leurs limites : la sélection par corrélation ne prend en compte que les relations univariées avec la cible, ce qui peut écarter des variables utiles dans un cadre multivarié, tandis que l'AIC, bien qu'intégrant une pénalisation de la complexité, peut parfois retenir un trop grand nombre de variables, nuisant à la parcimonie du modèle.

Nous avons ensuite examiné la régression pénalisée Lasso, qui permet une sélection automatique de variables via une pénalisation de norme L1. Cette méthode a montré de bonnes performances, en particulier sur les sous-échantillons de petite taille, grâce à sa capacité à limiter le surajustement. Le Lasso est apparu comme un compromis efficace entre performance et interprétabilité, à condition de bien calibrer le paramètre de régularisation. Néanmoins, sa tendance à ne sélectionner qu'une seule variable parmi des groupes corrélés peut biaiser l'interprétation dans des contextes complexes.

Enfin, nous avons proposé et implémenté une méthode inspirée du gradient boosting linéaire, adaptée spécifiquement à notre problématique. À chaque itération, une variable est sélectionnée en fonction de sa corrélation avec les résidus, et un coefficient est estimé pour corriger l'erreur. Ce processus permet de construire un modèle additif simple et interprétable, tout en améliorant progressivement la précision des prédictions. Sur l'ensemble du jeu de données, cette approche a obtenu de très bons résultats en termes de RMSE, tout en construisant un modèle relativement simple et modulable.

Ce modèle a également mis en évidence certaines variables clés pour la prédiction d'O3, telles que l'humidité, la température, la direction du vent, ou encore la variation de pression. L'analyse des coefficients au fil des itérations a permis d'illustrer l'apport progressif de chaque variable, offrant ainsi une lecture explicative de la dynamique prédictive.

Cependant, cette méthode de boosting linéaire n'est pas sans limites. Elle repose sur l'hypothèse de linéarité à chaque étape, et sa performance peut être affectée par l'ordre de sélection des variables et les coefficients qui se retrouvent figés. De plus, elle nécessite un prétraitement soigné des données pour éviter des effets indésirables comme des prédictions négatives ou aberrantes.

Ainsi, aucune méthode ne peut être considérée comme universellement supérieure : chacune présente des avantages spécifiques et des limites structurelles. Les approches classiques (corrélation, AIC) ont l'avantage de la simplicité et de l'interprétabilité immédiate, mais peuvent manquer de robustesse face à des structures de données plus complexes. Le Lasso apporte une régularisation efficace et une sélection automatique, mais sa sensibilité à la colinéarité et au choix du paramètre de pénalisation en réduit parfois l'efficacité explicative. Le boosting linéaire, quant à lui, s'est démarqué, malgré sa simplicité, par ses performances prédictives globales et sa capacité à construire un modèle progressif, tout en restant transparent, mais il impose une gestion précise de la complexité et du fonctionnement.

Ce travail a mis en évidence l'importance de choisir une méthode de sélection de variables adaptée au contexte d'étude, en tenant compte à la fois des objectifs de prédiction, de la compréhension du phénomène modélisé et de la taille du jeu de données. Il souligne également la nécessité d'un arbitrage entre précision, interprétabilité, robustesse, et complexité.

Afin de prolonger le sujet, ce projet pourrait être prolongé de plusieurs manières. D'une part, l'approche de boosting linéaire pourrait être généralisée à des bases d'apprenants non linéaires (arbres de décision, splines, réseaux de neurones), comme dans le cas du gradient boosting machine (GBM), afin de capturer des relations complexes entre les variables. D'autre part, l'analyse des variables sélectionnées pourrait être enrichie par des méthodes de type "stabilité de sélection" ou de "SHAP values" pour quantifier plus précisément l'importance relative des facteurs prédictifs. Enfin, l'évaluation pourrait être approfondie via des méthodes de validation croisée plus poussées, ou en introduisant des

métriques complémentaires telles que le biais, la variance ou les intervalles de confiance sur les prédictions.

Ce rapport aura ainsi permis d'explorer plusieurs méthodologies modernes de sélection de variables dans un cadre appliqué, illustrant comment des outils statistiques rigoureux peuvent éclairer des problématiques environnementales concrètes telles que la prévision de la pollution à l'ozone.

## 7. Bibliographie et Références

### Références

- [1] Breiman, L. 1996. Bagging predictors. *Machine learning*, 24, pp.123-140.
- [2] Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1), pp.119-139.
- [3] Breiman, L., 2001. Random forests. *Machine learning*, 45, pp.5-32.
- [4] Friedman, J.H., 2001. Greedy function approximation : a gradient boosting machine. *Annals of statistics*, pp.1189-1232.
- [5] Schapire, R.E., 1990. The strength of weak learnability. *Machine learning*, 5, pp.197-227.
- [6] HASTIE, TREVOR, et al. *The elements of statistical learning : data mining, inference, and prediction*. Vol. 2. New York : springer, 2009.
- [7] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar), pp.1157-1182.
- [8] Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), pp.267-288.
- [9] Bühlmann, P. and Hothorn, T., 2007. Boosting algorithms : Regularization, prediction and model fitting.
- [10] Lindsey, C. and Sheather, S., 2010. Variable selection in linear regression. *The Stata Journal*, 10(4), pp.650-669.