MARTIN Arnaud & GEANTET Etienne

# Projet d'Apprentissage et Aide à la Décision

## Plan



## 1. Présentation du sujet

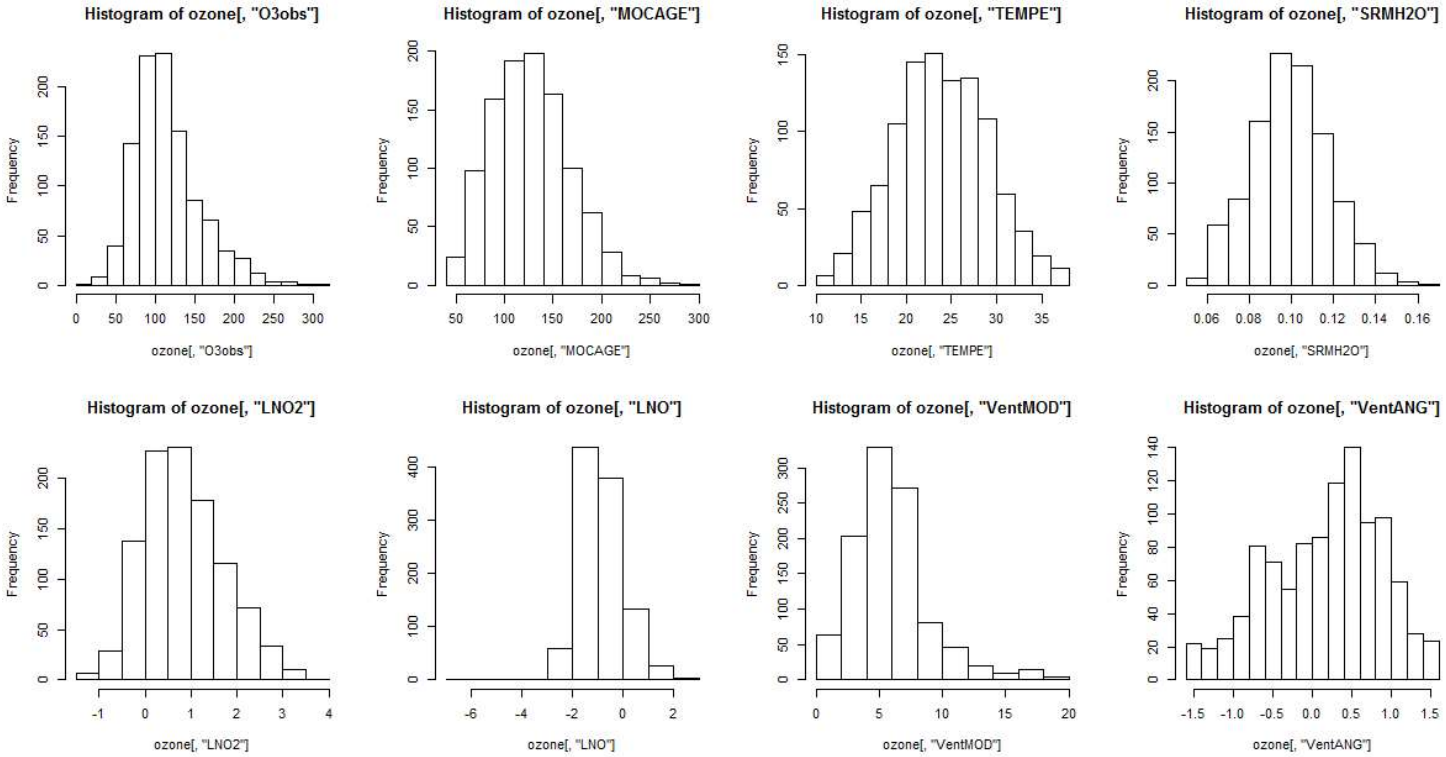
Ce projet nous propose d'étudier un échantillon de relevés météorologiques afin de déterminer un modèle capable de prédire le dépassement d'un seuil de 150 µg/m3 d'O3 dans l'air. Le jeu de données à disposition contient les variables suivantes :

- **JOUR** : le type de jour (ferié : 1 ou pas : 0)
- **O3obs** : concentration d'ozone effectivement observée le jour considéré (relevée à 17h et correspondant souvent au maximum observé de concentration)
- **MOCAGE** : prevision de la concentration obtenue par un modèle de mécanique des fluides
- **TEMPE** : température prévue par Météo France pour le jour donné
- **RMH2O** : rapport d'humidité
- **NO2** : concentration en dioxyde d'azote
- **NO** : concentration en monoxyde d'azote
- **STATION** : lieu d'observation (5 stations différentes)
- **VentMOD** : force du vent
- **VentANG** : orientation du vent

La variable réponse dépendra donc de **O3obs.**
Avant d'analyser les données avec différentes méthodes, avons tranformé certaines variables :

**JOUR** est mis sous forme de variable qualitative (`as.factor` dans R).
On prend la racine de **RMH2O.** Nouvelle variable : **SRMH2O.**
On prend les log de **NO2** et **NO.** Nouvelles variables : **LNO2** et **LNO.**
De plus, on ajoute une variables réponse de type qualitative **DepSeuil** qui a pour valeur TRUE si O3obs est au dessus de 150 µg/m3 et FALSE sinon. C'est la variable réponse que nous choisiront pour les méthodes de régression logistique, LDA et QDA.

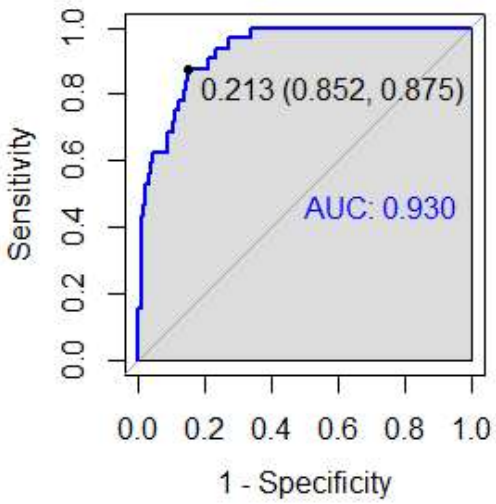
Apperçu des variables après transformations :




## 2. Regression Logistique

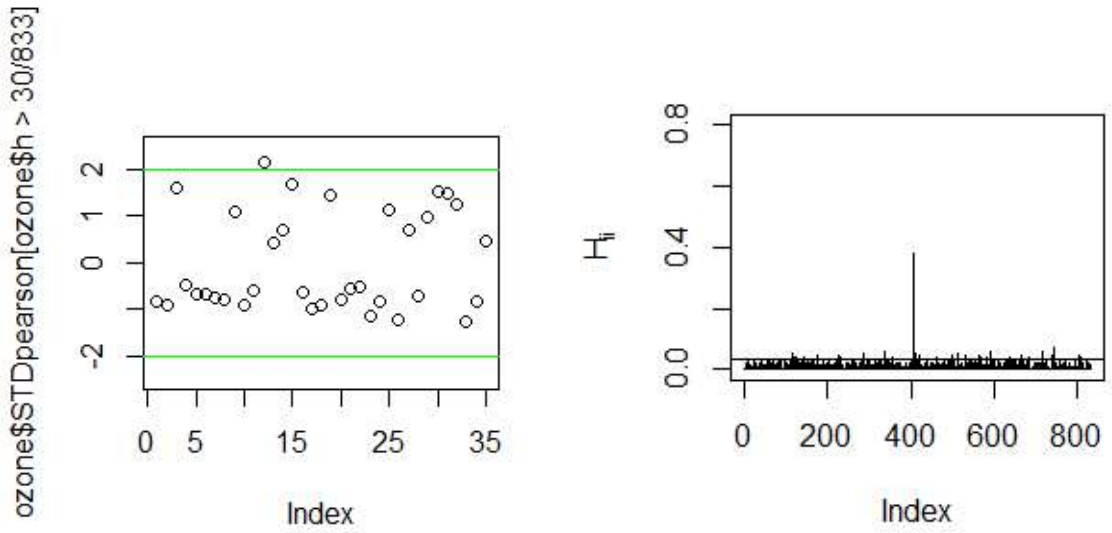
Nous avons tout d'abord crée un modèle logistique utilisant toutes les variables. Via des tests à 5%, nous avons déterminé que les variables **JOUR**, **MOCAGE** et **VenANG** n'etaient pas impliquées. Nous avons donc crée un nouveau modèle en enlevant ces variables, ce modèle s'appelera m.logit2 par la suite.

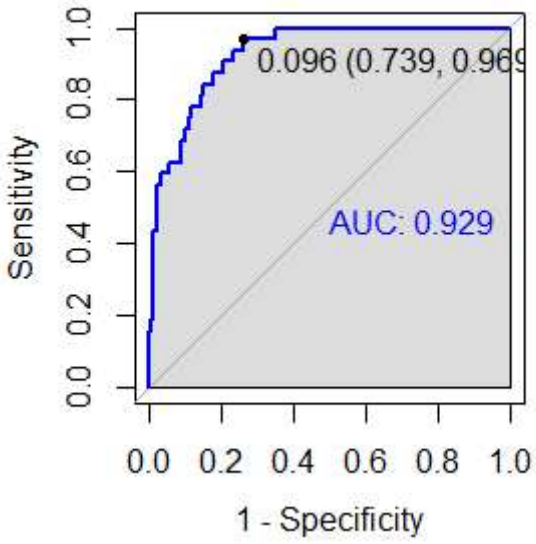
*Ci dessous: courbes ROC du modèle m.logit2*




Après une analyse des valeurs dans notre échantillons qui pourraient être abérrantes et leviée, nous en avons trouvé une. Après son retrait on obtient le modèle m.logit4.

*Ci dessous: Analyse des points leviers, détections des valeurs abérantes parmis les points leviers, courbes ROC du modèle m.logit4*




On peut voir à leur courbes ROC que ces deux modèles on leur avantages et leur inconvéniants, nous détaillerons cela dans la conclusion.

## 3. LDA

Le premier modèle créé contient toutes les variables. On obtient la matrice de confusion :

| obs\pred | FALSE | TRUE |
|---|---|---|
| FALSE | 168 | 8 |
| TRUE | 14 | 18 |

On effectue ensuite une analyse de variance (grâce à la fonction `anova`) pour déterminer les variables les plus discriminantes du modèle. C'est à dire les variables qui ont le plus d'importance à l'explication du modèle. On en déduit que les variables **SRMH2O**,**VentANG**, **LNO**,**LNO2** n'ont pas une grande influence et qu'on peut les enlever du modèle. De même, la variable **VentMOD** semble assez peu influente.

Ce choix se fait en observant d'abord la valeur du test de l'égalité à 0 du coefficient de la variable, puis en observant l'estimateur en question : si la valeur est trop petite, on retire la variable du modèle car son influence est faible. Le but étant de trouver un modèle suffisamment précis, expliqué par le moins de variables possibles. C'est le principe de parcimonie.

La variables **JOUR** et **STATION** ne peuvent pas êtres testées par cette méthode car elles ne sont pas continues. On crée donc une série de modèle, qu'on compare afin de trouver celui donnant les meilleurs résultats au test. On fini par garder le modèle expliqué par les variables **JOUR**, **MOCAGE**, **TEMPE**, **STATION**, qui donne la matrice de confusion :

| obs\pred | FALSE | TRUE |
|---|---|---|
| FALSE | 160 | 16 |
| TRUE | 9 | 23 |

## 4. QDA

Nous avons d'abord fait un modèle à partir de toutes les variables, il donnais pour matrice de confusion:

| obs\pred | FALSE | TRUE |
|---|---|---|
| FALSE | 160 | 16 |
| TRUE | 9 | 23 |

Ensuite avec une anova, nous avons déterminé quels étaient dans les variables continues celles qui influaient le moins sur O3. Les variables **VentANG**,**SRMH2O**,**LNO2** et **LNO** ont ainsi été excluent.

Nous avons ensuite comparé les modèles avec et sans les variables discrètes (Jour et Station) il en ressort que ces variables sont nécessaires au modèle.

Le modèle final as une matrice de confusion comme suis:

| obs\pred | FALSE | TRUE |
|---|---|---|
| FALSE | 160 | 16 |
| TRUE | 9 | 23 |

## 5. Autres méthodes

### Régression linéaire

On pourrait imaginer que la concentration d'O3 dans l'air suivrait une relation linéaire avec les variables explicatives. En réalisant une première régression linéaire avec toutes les variables (fonction `lm`), on trouve un $R^2$ de 0.5405. Cela signifie de le modèle est assez mauvais. Un modèle linéraire est généralement bon pour un $R^2$ supérieur à 0.8. On crée un dexième modèle en retirant les variables dont le test d'égalité à zéro est vérifié.
Le modèle prenant les variable **MOCAGE, TEMPE, STATION, VentMOD, VentANG, LNO2** et **LNO** nous donne un ajusted $R^2$ de 0.5325.
Un modèle de régression linéaire ne semble pas adapté à la représentation de ce jeu de données.

## 6. Conclusion

Nous avons donc ressorti 4 modèles de nos analyses, deux logistiques, un qda et un lda.
Leur matrices de confusions sont:

| QDA | LDA | m.logit2 | m.logit4 |
|---|---|---|---|
| JOUR+MOCAGE+TEMPE+STATION+VentMOD | JOUR+MOCAGE+TEMPE+STATION | TEMPE+STATION+VentMOD+SRMH2O+LNO2+LNO | TEMPE+STATION+VentMOD+SRMH2O+LNO2+LNO |
| FALSE TRUE<br>FALSE 160 16<br>TRUE 9 23 | FALSE TRUE<br>FALSE 169 7<br>TRUE 16 16 | FALSE TRUE<br>FALSE 150 26<br>TRUE 4 28 | FALSE TRUE<br>FALSE 130 46<br>TRUE 1 31 |
| sensibilité :0.71875 | sensibilité :0.5 | sensibilité :0.875 | sensibilité :0.96875 |
| spécificité :0.9090909 | spécificité :0.9602273 | spécificité :0.8522727 | spécificité :0.7386364 |

La LDA n'est donc pas adapté du tout, ensuite cela dépend de ce que l'on veut.
Si l'on souhaite prédire dans la majorité des cas correctement si l'O3 va dépasser le seuil et que cela ne dérange pas de faire de fausse alerte, alors on prend le modèle m.logit4
Si au contraire on veut faire le moins de fausse alerte, alors on préfèrera le modèle QDA.
Mais si l'on veut optimiser les deux, nous prendrions le modèle m.logit2