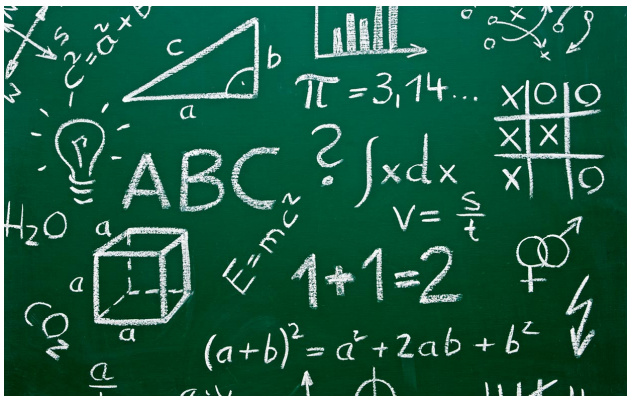
INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

INSA de Rennes

Prévision de dépassement d'un seuil d'ozone

Auteurs :
Romain Colombat
Aymen Ragoubi

Enseignant :
Jean François DUPUY



22 janvier 2016

Table des matières

1 Introduction

L'ozone (O3) est un gaz nocif incolore à l'odeur très forte dont la molécule est formée de trois atomes d'oxygène (O).

La présence de l'ozone dans l'atmosphère est liée à celle de l'ultraviolet produit à partir des oxydes d'azotes (NOx) produit par les voitures, les systèmes de chauffages ou bien à partir d'autres composés chimiques comme certains liquides.

Des niveaux d'alertes et de pics de pollution ont été fixés pour l'ozone :

- supérieur à180 microgrammes d'ozone par m3 d'air : niveau d'information et de recommandation. La population est informée par les préfectures, et des recommandations sanitaires et comportementales sont diffusées.
- supérieur à 240 microgrammes/m3 en moyenne horaire dépassé pendant 3 heures consécutives : niveau d'alerte 1
- supérieur à 300 microgrammes/m3 en moyenne horaire dépassé pendant 3 heures consécutives : niveau d'alerte 2
- supérieur à 360 microgrammes/m3 en moyenne horaire : niveau d'alerte 3 Le niveau d'alerte implique des mesures de restrictions d'activités émettrices d'ozone, industrielles, automobiles et également domestiques.

une exposition à l'ozone occasionne les affections immédiates suivantes :

- respiration courte et précipitée ;
- douleur thoracique reliée à une inhalation profonde ;
- respiration sifflante et toux ;
- prédisposition accrue aux infections respiratoires ;
- inflammation des poumons et des voies respiratoires ;
- risques accrus d'une crise d'asthme ;
- besoin accru d'un traitement médical et d'une hospitalisation chez les personnes atteintes d'une maladie pulmonaire, telle que l'asthme ou la bronchopneumopathie obstructive chronique.

Dans le cadre du projet d'apprentissag et aide à la décision proposé en 4ème année informatique à l'INSA de Rennes, il nous a été demandé de construire et comparer des modèles statistiques de prévision du dépassement d'un seuil de concentration d'ozone dans l'air (150 µg/ m3), en utilisant le logiciel R.

2 Présentation de l'étude

2.1 Le jeu de données

Le jeu de données contient 1024 observations et 10 variables explicatives :

- JOUR : le type de jour (ferié : 1 ou pas : 0)
- O3obs : La concentration d'ozone observée le jour considéré
- MOCAGE : La prévision de la concentration obtenue par un modèle de mécanique des fluides
- TEMPE : La température prévue par Météo France pour le jour donné.
- SMH2O : Le rapport de l'humidité.
- LNO2 : La concentration en dioxyde d'azote.
- LNO : La concentration en monoxyde d'azote.
- STATION : Le lieu des observations ; On dispose ici de cinq stations différentes.
- VentMOD : La force du vent.
- VentANG : Orientation du vent.

2.2 Problématique

L'objectif de cette étude est de modéliser le dépassement du seuil de la concentration d'ozone dans l'air en utlisant les prédicteurs du jeu de données ci-dessus.

Nous souhaitons prédire une variable qualitative à deux modalités :

$$\text{DepSeuil} = \begin{cases} = 1 \; si \; la \; concentration \; d'ozone \; dpasse \; le \; seuil \; 150 g/m^3 \\ = 0 \; sinon \end{cases}$$

3 Modélisation statistique

Pour modéliser la variable qualitative binaire nous avons utilisé deux méthodes appropriées à notre étude : La régression logistique et l'analyse discriminante.
La régression logistique est une méthode utilisée pour expliquer et prédire une variable binaire. Dans notre cas on s'intéresse à la variable DepSeuil.
Quant à l'analyse discriminante, c'est une technique statistique qui vise à décrire, expliquer et prédire l'appartenance à des groupes prédéfinis d'un ensemble d'observations.

3.1 Régression logistique

3.1.1 Construction d'un échantillon d'apprentissage et d'un échantillon test

Nous partageons nos données en deux échantillons :

- Echantillon de test : sur lequel nous allons réaliser l'analyse de la qualité prédictive des modèles construits en comparant les valeurs prédites au données de l'échantillon. Il constitut 20% de nos données totales sélectionnées aléatoirement.

- Echantillon d'apprentissage : sur lequel nous allons construire les modèles de prévisions. Il constitut 80% des données totales.

3.1.2 Construction d'un premier modèle de prévision d'un dépassement de seuil d'ozone

A partir de notre échantillon d'apprentissage, nous alons utiliser la technique "backward-selection" pour comparer plusieurs modèles. Cette technique consiste à commencer par le modèle saturé qui contient toutes les variables explicatives et tester l'élimination de chacune variables en se basant sur le critère d'AIC, on itére jusqu'à ce qu'il n'y ai plus d'améliorations possibes. Les deux tables obtenues sous R en appliquant cette méthode (voir Table1 et Table2).

On a une déviance de 395.01 pour le modèle saturé et de 744.70 pour le modèle null (sans aucune variable). L'AIC du modèle saturé est de 421.01.

Le modèle retenu à la fin de la procédure backward comporte les prédicteurs JOUR, TEMPE, STATION, VentMOD, SRMH20, LNO et LNO2. L'AIC du modèle obtenu est de 417.77 ce qui est inférieur au modèle saturé de base.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -19.124216   1.798695 -10.632  < 2e-16 ***
JOUR1                     0.453032   0.281281   1.611 0.107267
MOCAGE                    0.005109   0.006309   0.810 0.418081
TEMPE                     0.441995   0.047469   9.311  < 2e-16 ***
as.factor(STATION)Als    -0.313960   0.527128  -0.596 0.551439
as.factor(STATION)Cad     0.403837   0.414823   0.974 0.330297
as.factor(STATION)Pla     2.149429   0.602453   3.568 0.000360 ***
as.factor(STATION)Ram    -1.183426   0.549942  -2.152 0.031405 *
VentMOD                  -0.187415   0.078040  -2.402 0.016326 *
VentANG                   0.059860   0.208214   0.287 0.773734
SRMH2O                   28.281625   7.950313   3.557 0.000375 ***
LNO                      -1.313534   0.531698  -2.470 0.013494 *
LNO2                      1.564919   0.625265   2.503 0.012321 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 744.70  on 832  degrees of freedom
Residual deviance: 395.01  on 820  degrees of freedom
AIC: 421.01

Number of Fisher Scoring iterations: 7
```

TABLE 1 – Modèle logistique saturé

```
Step:  AIC=417.77
DepSeuil ~ JOUR + TEMPE + as.factor(STATION) + VentMOD + SRMH2O +
    LNO + LNO2

                    Df Deviance    AIC
<none>                   395.77 417.77
- JOUR               1   398.34 418.34
- VentMOD            1   404.31 424.31
- SRMH2O             1   409.35 429.35
- LNO                1   409.89 429.89
- LNO2               1   416.03 436.03
- as.factor(STATION) 4   439.18 453.18
- TEMPE              1   541.88 561.88
```

TABLE 2 – Selection automatique du modèle avec la méthode backward

D'après les résultats fourni par R (voir Table1), on remarque que les variables explicatives STATION (pour Als et Cad), VentMOD et JOUR ne sont a priori pas significatives car les p-value associées à leurs coefficients sont inférieurs à 5%. En prenant en compte cette constatation, nous décidons de confronter le modèle a un nouveau modèle obtenu sans les variables STATION, VentMOD et JOUR.

Pourt faire le choix entre ces deux modèles nous allons faire un test de déviance sur des modèles emboîtés. Les hypothèses du test sont :

H0 : { Le modèle M1 avec le moins de paramètre est approprié}
H1 : { Le modèle M2 avec le plus de paramètre est approprié}

Voici les résultats obtenus sous R :

```
Analysis of Deviance Table

Model 1: DepSeuil ~ TEMPE + SRMH2O + LNO + LNO2
Model 2: DepSeuil ~ JOUR + TEMPE + as.factor(STATION) + VentMOD + SRMH2O +
    LNO + LNO2
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       828     443.02
2       822     395.77  6   47.257 1.662e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TABLE 3 – Test de déviance entre des modèles emboîtés

D'après le test du χ^2 à 6 degrés de libertés on peut rejeter l'hypothèse H0 avec un niveau de confiance de 5%. Donc le modèle choisi est :

$$\textit{logit[P(DepSeuil=1|G)]= } \beta 0 + \beta 1 * \textit{TEMPE} + \beta 2 * \textit{SRMH2O} + \beta 3 * \textit{LNO} + \beta 4 * \textit{LNO2} + \beta 5j * \textit{STATION} + \beta 6 * \textit{VentMod} + \beta 7 * \textit{JOUR}$$

avec G l'ensemble des variables explicatives TEMPE, SRMH2O, LNO, LNO2, STATION, VentMOD et JOUR.

Les valeurs $\beta 0$,$\beta 1$,$\beta 2$,$\beta 3$, $\beta 4$, $\beta 5$, $\beta 6$ et $\beta 7$ sont estimés par les valeurs suivantes :

$\widehat{\beta}0$ = -19.22 , $\widehat{\beta}1$ = 0.45 , $\widehat{\beta}2$ = 28.92 , $\widehat{\beta}3$ = -1.59 , $\widehat{\beta}4$ = 1.95 , $\widehat{\beta}5$Als=-0.39 , $\widehat{\beta}5$Cad = 0.46 , $\widehat{\beta}5$Pla = 2.37 , $\widehat{\beta}5$Ram = -1.25 , $\widehat{\beta}6$ = -0.20 , $\widehat{\beta}7$ = 0.44.

On retrouve ces résultats en affichant les caractéristiques de ce modèle sur R.

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)             -19.22873    1.80053 -10.679  < 2e-16 ***
JOUR1                     0.44488    0.27648   1.609 0.107595
TEMPE                     0.45048    0.04682   9.621  < 2e-16 ***
as.factor(STATION)Als    -0.39128    0.51372  -0.762 0.446258
as.factor(STATION)Cad     0.46224    0.40896   1.130 0.258363
as.factor(STATION)Pla     2.37174    0.54451   4.356 1.33e-05 ***
as.factor(STATION)Ram    -1.25957    0.53514  -2.354 0.018587 *
VentMOD                  -0.20503    0.07515  -2.728 0.006366 **
SRMH2O                   28.92781    7.88132   3.670 0.000242 ***
LNO                      -1.59976    0.47079  -3.398 0.000679 ***
LNO2                      1.95627    0.47234   4.142 3.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 744.70  on 832  degrees of freedom
Residual deviance: 395.77  on 822  degrees of freedom
AIC: 417.77

Number of Fisher Scoring iterations: 7
```

TABLE 4 – Caractéristiques du modèle retenu

La fonction logit est une fonction de lien permettant de passer en mode linéaire.

$$\boxed{logit(x) = \frac{x}{(1+x)}}$$

Les variables dont les coefficients sont négatifs, contrairement au variables dont les coefficients sont positives, ont un comportement inverse sur le depassement du seuil d'ozone. Par exemple quand la valeur de la variable LNO est grande, le risque de dépassement du seuil est réduit.

3.2 Analyse discriminante

Afin de compléter notre analyse du jeu de donnée par régression logistique, nous utilisons une seconde méthode basée sur la classification et qui utilise l'analyse discriminante. Bien que notre variable réponse soit binaire (DepSeuil à TRUE ou FALSE), cette technique possède des avantages sur la régression logistique. Elle permet d'établir un modèle dans le cas d'un jeu de donnée ou les groupes sont bien séparé, ce n'est à priori pas le cas ici. Mais aussi d'augmenter la stabilité de la décision pour un échantillon de petite taille et avec des prédicteurs distribués selon une loi normale.

3.2.1 Sélections des variables

L'analyse de la variance (ANOVA) permet de tester l'effet de la variable DepSeuil sur les différentes variables continues. Il est donc nécessaire de traîter les variables discrètes JOUR et STATION a part. Nous allons tester l'indépendances de ces dernières par rapport à la variabe DepSeuil grâce à des test du χ^2.

Nos hypothèses sont les suivantes :

H0 : {La variable étudiée est indépendante avec la variable DepSeuil.}
H1 : {La variable étudiée est n'est pas indépendante avec DepSeuil.}

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  t
X-squared = 2.5985, df = 1, p-value = 0.107
```

Table 5 – Test d'indépendance des variables JOUR et DepSeuil

On peut conclure avec un risque de 5% que JOUR et DepSeuil sont indépendants. On conserve H0.

```
        Pearson's Chi-squared test

data:  t
X-squared = 19.662, df = 4, p-value = 0.0005824
```

Table 6 – Test d'indépendance des variables STATION et DepSeuil

On peut conclure avec un risque de 5% que STATION et DepSeuil ne sont pas indépendants. On rejette H0.

Avant de faire une ANOVA sur nos variables continues vérifions qu'elles sont bien distribuée selon une loi normale.

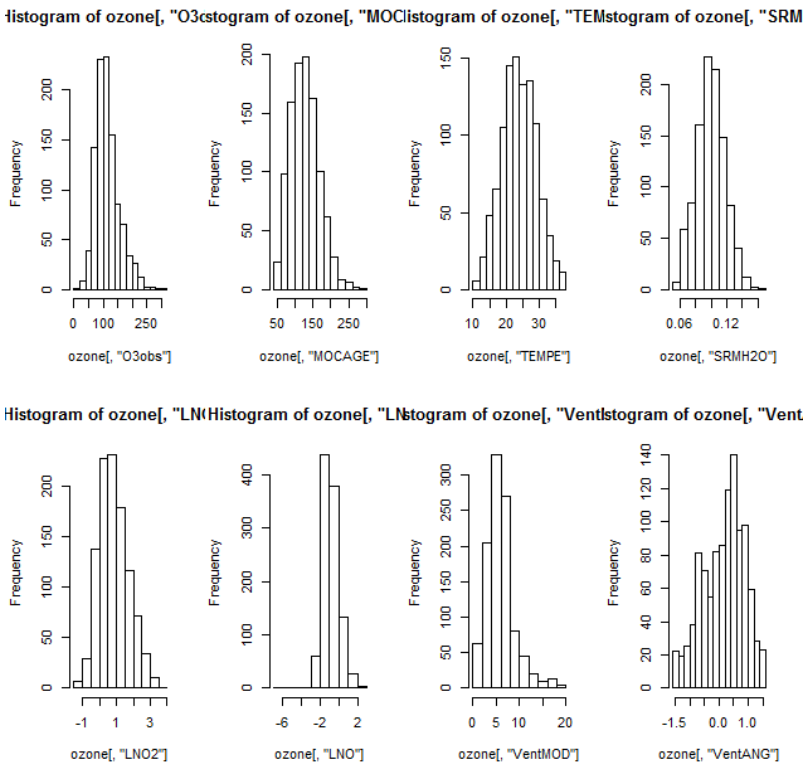


FIGURE 1 – Histogramme des variables continues, les valeurs en fonctions de leur fréquence d'apparition

Les variables son distribuées, avec plus ou moins de justesse, selon une loi normale.

On fait une ANOVA sur les variables continues suivantes : MOCAGE, TEMPE, VentMOD, VentANG, SRMH2O ,LNO2, LNO.

```
Response: ozodis[, k]
           Df  Sum Sq Mean Sq F value    Pr(>F)
DepSeuil    1  198715  198715  154.02 < 2.2e-16 ***
Residuals 831 1072162    1290
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

TABLE 7 – Analyse de la table de variance de la variable MOCAGE

A la suite de cette analyse on conserve les 3 variables les plus discriminantes, c'est à dires les 3 variables comportant les F-value les plus élevés, TEMPE, MOCAGE et SRMH2O avec des valeurs respectivement à 271.83, 154.02, 71.835.

A la suite de ces analyses ont décide de conserver les 4 variables MOCAGE, TEMP, SRMH2O et STATION pour une analyse graphique.

3.2.2 Analyse graphique

Avec les 4 variables choisies nous pouvons représenter graphiquement leurs contributions à la séparabilité des deux classes étudiées. Sur la figure nous avons choisi de les réprésenter 2 à 2. Le groupe rouge correspond a un non dépassement du seuil, et le groupe vert à un dépassement du seuil.

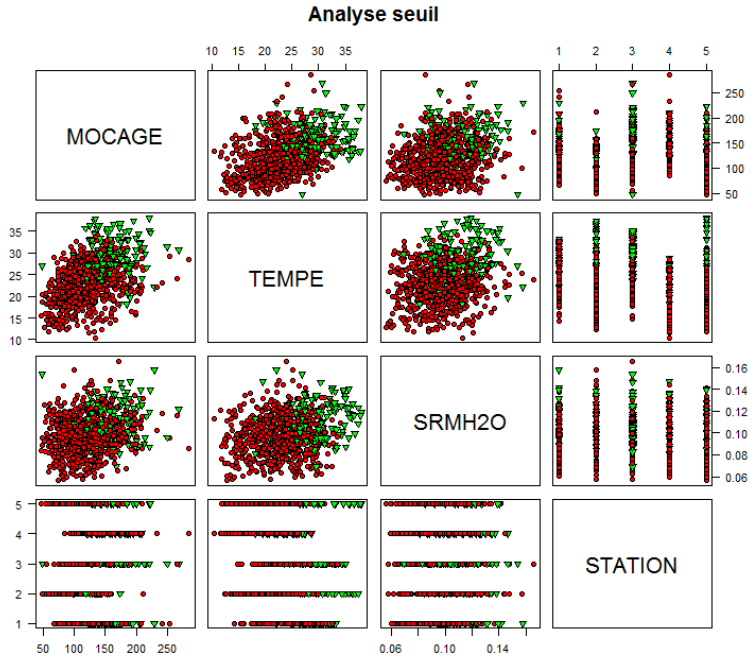


FIGURE 2 – Répresentation des deux classes selont 4 variables croisées

Suite à cette première répresentation graphique on décide d'éliminer la variable discrète STATION, elle apporte peu à la classification.

Maintenant représentons les 3 variables restantes MOCAGE, TEMP et SRMH2O en dimension 3.

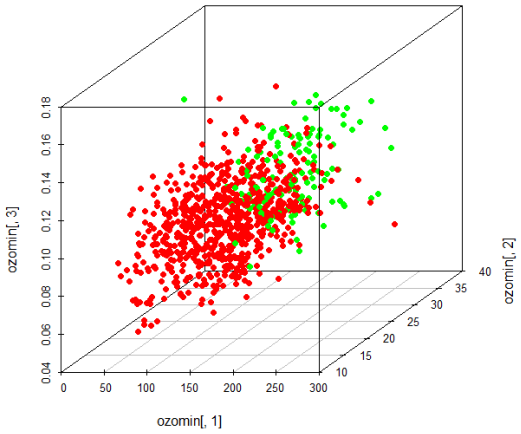


FIGURE 3 – Répresentation des 2 groupes selons les variables MOCAGE, TEMP et SRMH2O

On peut aussi analyser les distributions marginales des variables. Cela confirme le caractère discriminant des variables selectionnées par ANOVA.

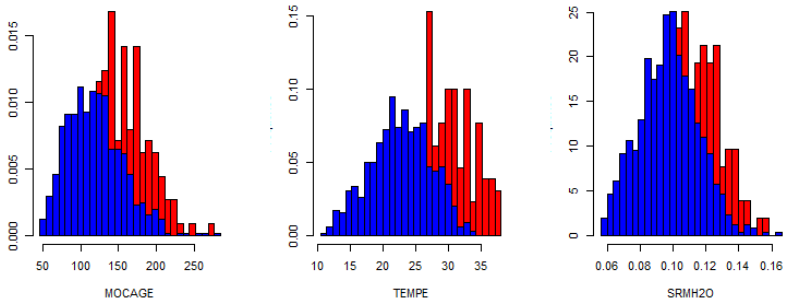


FIGURE 4 – Histogramme de la distribution des variables par classe

3.2.3 Classification sous hypothèse de normalité

On procède maintenant à une analyse discriminante en utilisant les variables MOCAGE, TEMPE et SRMH2O. On commence par une LDA, c'est à dire une analyse discriminante linéaire, on en déduit une première matrice de confusion pour l'échantillon d'apprentissage. On a un taux de mauvais classement de 11.3%.

obs \ pred	False	True	Total
False	669	27	696
True	67	70	137

TABLE 8 – Matrice de confusion sur l'échantillon d'apprentissage avec le modèle à 3 variables

Si on utilisais deux variable on aurait un taux de mauvais classement sur l'échantillon de test similaire.

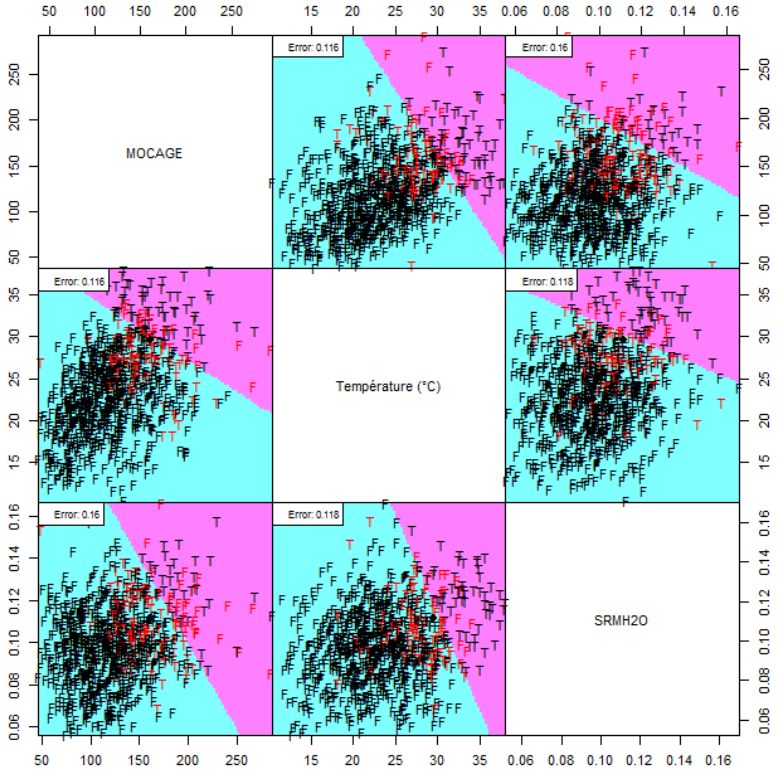


FIGURE 5 – Analyse discriminante linéaire en fonctions des 3 variables

On fait un test d'équalité des matrices de varance-covariance des variables MOCAGE,TEMPE et SRMH2O pour chacun des groupes. Les hypothèses sont les suivantes :
H0 : {Les matrices de variances-covariances sont égales.}
H1 : {Les matrices de variances-covariances ne sont pas égales.}

```
        Box's M-test for Homogeneity of Covariance Matrices

data:  Y
Chi-Sq (approx.) = 16.407, df = 6, p-value = 0.01173
```

TABLE 9 – Test d'égalité des variances pour les 3 variables par LDA

On rejette l'hypothèse H0 au risque de 5%. Les matrices ne semblent pas égales. On va effectuer une analyse quadratique.
A la suite de l'analyse quadratique on obtient un taux de mauvais classement de 11.8%, qui est très similaire à l'analyse linéaire mais légèrement moins bon.

obs \ pred	False	True	Total
False	665	31	696
True	67	70	137

TABLE 10 – Matrice de confusion sur l'échantillon d'apprentissage avec le modèle à 3 variables par QDA

On retient donc le modèle LDA, qui en plus d'être plus adapté est plus simple à mettre en place. Les coefficients permettant de tracer le plan de séparation des classes sont résumés ci-dessous.

```
Call:
lda(DepSeuil ~ MOCAGE + TEMPE + SRMH2O, data = ozomin)

Prior probabilities of groups:
    FALSE      TRUE
0.8355342 0.1644658

Group means:
        MOCAGE    TEMPE     SRMH2O
FALSE 118.9283 22.59684 0.09731782
TRUE  160.5934 29.58467 0.11128399

Coefficients of linear discriminants:
                LD1
MOCAGE  0.01124552
TEMPE   0.15554457
SRMH2O 17.35032287
```

TABLE 11 – Résumé du modèle d'analyse discriminante linéaire

3.3 Validation des modèles

Dans cette partie nous allons comparer les deux modèles obtenus, le premier par régression linéaire et le second par analyse discriminante afin de déterminer

le modèle final. Pour cela nous allons confronter nos deux modèles au jeu de test défini au départ.

3.3.1 Sensibilité et spécificité

Pour le premier modèle obtenu par régression linéaire nous le confrontons au jeu de donnée test afin d'en extraire la matrice de confusion. A la fin on obtient le taux de mauvais classement qui indique le pourcentage des observations qui ont été mal prédites. On choisit le seuil de décision à 0.5 par défaut.

$$sensibilite: \ proportion\ de\ vrais\ positifs = \frac{vrais\ positifs}{vrais\ positifs + \ faux\ negatives}$$

$$specificite: \ proportion\ de\ vrais\ negatives = \frac{vrais\ negatives}{vrais\ negatives + \ faux\ positifs}$$

obs \ pred	False	True	Total
False	156	11	167
True	22	19	41

Table 12 – Matrice de confusion pour un seuil de 50%

On obtient un taux de mauvais classement de 15.8%, une sensibilité de 46.3% et une spécificité de 93.4% avec un seuil de 50%.

On va maintenant valider nos modèles sur l'échantillon de test pour les modèles obtenus par analyse discriminante. Dans notre cas, nous obtenons la même matrice de confusion avec le modèle contenant les prédicteurs MOCAGE, TEMPE et SRMH2O avec la méthode LDA.

obs \ pred	False	True	Total
False	157	10	167
True	21	20	41

Table 13 – Matrice de confusion pour un seuil de 50% par LDA

On obtient un taux de mauvais classement de 14.9%, une sensibilité de 48.8%, et une spécificité de 94%.

Avec des seuils fixés à 50% les méthodes LDA et QDA sont meilleures que par régression linéaire.

On va faire varier le seuil pour maximiser les valeurs de sensibilité et de spécificité.

3.3.2 Courbes ROC

La courbe ROC mesure la performance de notre modèle. AUC c'est l'air en dessous de la courbe.
En faisant varier le seuil de décision (seuil à partir duquel on va prédire qu'il y'a un dépassement du seuil d'ozone) la spécificité et la sensibilité varient.
La courbe ROC permet alors de représenter toutes les possibilités de seuils afin de déterminer un seuil optimal.

Représentons nos courbes ROC pour le seuil de 50% et le seuil optimal.

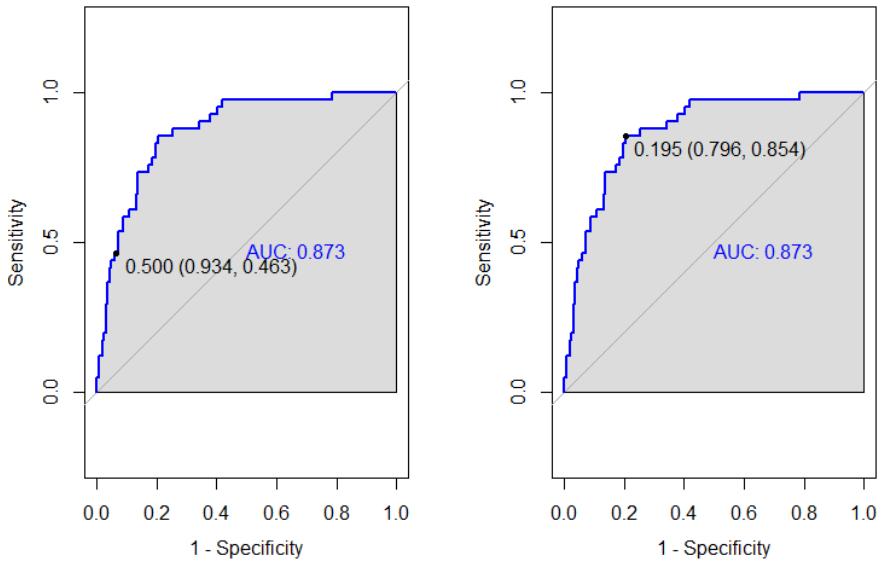


FIGURE 6 – Courbes ROC pour le modèle obtenu par régression logistique

obs \ pred	False	True	Total
False	133	34	167
True	6	35	41

TABLE 14 – Matrice de confusion pour un seuil de 19.5% (modèle logistique)

Le seuil optimal est de 0.195, les valeurs entre parenthèses représentent la sensibilité et la spécificité obtenues en utilisant ce seuil. Pour un test parfait, l'air sous la courbe vaut 1. Dans notre cas, l'AUC vaut 0.873 donc on conclut que notre modèle est assez performant. A partir de la matrice de confusion d'un seuil de 19,5%, on calcule la proportion des vrais positifs et la proportions des vrais négatifs.

Pour notre modèle avec un seuil de 0.194, on trouve un taux de mauvais classement de 19.2%, une sensibilité de 85.4% et une spécificité de 79,6%.

On affiche les courbes ROC pour le modèle basée sur de la LDA.

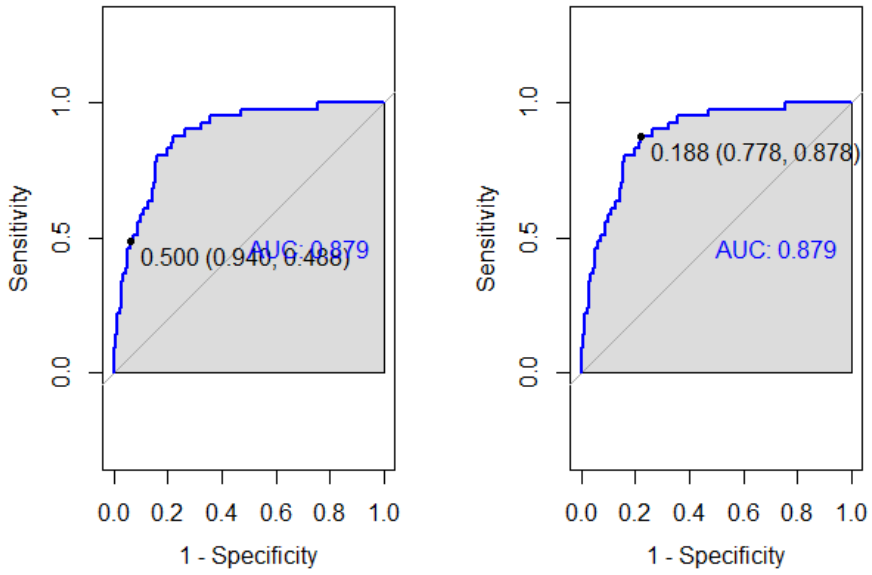


Figure 7 – Courbes ROC pour le modèle LDA

obs \ pred	False	True	Total
False	130	37	167
True	5	36	41

Table 15 – Matrice de confusion pour un seuil de 18.8% (modèle LDA)

Pour un seuil de 18.8%, on a un taux de mauvais classement de 20.1%, une sensibilité de 77.8% et une spécificité de 87.8%.

3.3.3 Modèle retenu

En conclusion, on retient le modèle basée sur l'analyse discriminante linéaire, il nous permet d'atteindre de meilleures valeurs de sensibilité et de spécificité. De plus, nous n'avons besoin que de 3 variables, le MOCAGE, la TEMPE et le SRMH2O.

4 Conclusion

En conclusion, c'est avec la prévision de la concentration obtenue, la température prévue par Météo France et la racine carré du rapport d'humidité que nous pouvons prédire le plus efficacement le dépassement du niveau d'ozone dans l'air. Et ce par une analyse discriminante linéaire.