Approche Statistique de la modélisation des phénomènes lithométéores

Hakim ABANE [1]*,Papa Ngor NDIAYE [1], Mouhamadou Moustaphaa KAMARA [1]

Résumé
Les lithométéores représentent des types de temps caractéristiques des régions arides et semi-arides. Malgré l'existence d'une typologie variée, l'identification de ces phénomènes reste toujours liée à l'appréciation de l'observateur. Les aérosols, en aggravant les conditions météorologiques, peuvent intervenir dans divers aspects socio-économiques, notamment en nuisant au trafic aérien ou à la télétransmission. Comme la modélisation statistique est un outil contemporain de la prévision ; l'étude menée a permis d'élaborer un modèle statistique de régression logistique binaire dédié à la prévision de l'occurrence ou non des phénomènes lithométéores résultant des émissions d'aérosols désertiques. Les résultats ainsi obtenus expliquent la performance de la méthode utilisée.

Mots Clés:
Phénomènes lithométéores — Modélisation statistique — Régression logistique binaire

[1] *Institut Hydrométéorologique de Formation et de Recherches, Oran, Algérie*
***Correspondant**: hakim73_a@yahoo.fr

1. Introduction

Le développement des outils d'aide à la prévision est devenu une nécessité pour remédier aux problèmes dus parfois à la divergence des modèles numériques ou à la complexité du phénomène à prévoir. Le présent article est dédié à l'élaboration d'un modèle statistique qui servira à la prévision de l'occurrence d'un phénomène lithométéore sur la région de Bechar.

2. Données et méthodes

Les données

Ce sont des données quotidiennes de la station d'observation de Bechar, provenant de la banque climatique de l'office national de la météorologie(ONM).Elles couvrent la période 1990-2000, soit 22 mois (juillet et aout) de 11 saisons estivale ou ont été extraites, soit un total de 682 relevés journaliers.

Une gestion de données a été menée afin d'aboutir à un seul fichier de forme enregistrement tabulaire. Cette forme est propice pour mettre en œuvre les algorithmes d'apprentissage. Elle permet d'associer la valeur à prédire avec les variables prédictives.

Les paramètres météorologiques utilisés comme prédicteurs sont :

- La température T de l'air à 2m du sol
- La force et la direction du vent à 10m du sol
- La nébulosité N
- La pression à la station

Le prédictant est l'occurrence de phénomène lithométéore.

A partir de la base de données élaborée, trois échantillons ont été construits

- Echantillon d'apprentissage constitué de 342 jours d'observations, qui va servir par la suite à la sélection des prédicteurs.
- Echantillon test constitué de 278 jours d'observations, qui vont servir à la validation des prédicteurs choisis.
- Echantillon validation constitué de 62 jours pour valider le modèle élaboré.

Description de la Méthodologie

L'approche statistique utilisée consiste à :

- Faire une sélection de variables à la liste de prédicteurs déjà établie pour en choisir les meilleurs
- Utiliser la régression logistique binaire qui consiste à établir une expression de régression entre une ou plusieurs variables indépendantes appelées « prédicteurs » et une variable dépendante dichotomique appelée « prédictant »
- Chercher à valider ce modèle en soumettant les résultats obtenus aux observations faites à des stations spécifiques

La sélection des variables avec le logiciel R : la stratégie WRAPPER

La sélection de variables est un aspect essentiel de l'apprentissage supervisé. Nous devons déterminer les variables pertinentes pour la prédiction des valeurs de la variable à prédire, pour différentes raisons : un modèle plus simple sera plus facile à comprendre et à interpréter ; le déploiement sera facilité, nous aurons besoin de moins d'informations à recueillir pour la prédiction ; enfin, un modèle simple se révèle souvent plus robuste en généralisation c.-à-d. lorsqu'il est appliqué sur la population.

L'approche WRAPPER utilisée sur R cherche à optimiser un critère de performance pour la recherche du sous-ensemble de prédicteurs pertinents en présentant à la méthode d'apprentissage des scénarios de solutions. Le plus souvent, il s'agit du taux d'erreur. Mais en réalité, tout critère peut convenir par exemple l'introduction d'une matrice de coûts de mauvais classements ; le calcul de l'aire sous la courbe ROC Didacticiel Tanagra et R (2009).

Stratégie de recherche

La stratégie de recherche des solutions joue un rôle très important dans la stratégie WRAPPER. Si l'on s'en tient au taux d'erreur, on pourrait subdiviser les données en

échantillons d'apprentissage et de test : construire le modèle sur les premières données, évaluer les performances sur les secondes, en vue de sélectionner les variables pertinentes. Séduisante a priori, cette approche n'est pas exempte de reproches. L'échantillon test, censé constituer un juge impartial pour l'évaluation des performances, devient partie prenante dans l'apprentissage. Il est explicitement exploité pour choisir la meilleure solution. Il est de ce fait inutilisable pour estimer objectivement l'erreur en généralisation.

Il semble qu'une approche viable serait toujours de subdiviser les données en apprentissage et test, mais de se baser sur une méthode de ré échantillonnage (la validation croisée la plupart du temps) sur la première partie des données pour évaluer les différentes solutions et sélectionner celle qui paraît la plus pertinente. L'ensemble test ne servira que pour mesurer les performances du modèle finalement sélectionné. Ainsi, il joue le rôle qui lui est normalement dévolu : évaluer objectivement la performance sans prendre part ni à la construction, ni à la sélection des modèles.

La méthode « NAIVE BAYES » (modèle d'indépendance conditionnelle) est utilisée dans le processus de sélection WRAPPER. Elle recherche des causalités entre les variables prédictives et la variable à prédire, plus le nombre de variables retenu est faible, meilleure sera la lisibilité du modèle prédictif. Elle est bien adaptée aux variables prédictives catégorielles, c'est le cas de nos données ; elle n'intègre pas un processus interne de sélection ; et elle est sensible aux variables non pertinentes. Ainsi, l'influence de la sélection de variables sur la performance sera particulièrement visible Didacticiel Tanagra et R (2009).

La Régression Logistique binaire

La Régression Logistique est une technique de modélisation statistique qui, dans sa version la plus répandue, vise à prédire et expliquer les valeurs d'une variable catégorielle binaire Y (variable à prédire, variable expliquée, variable dépendante, attribut classe, variable endogène) à partir d'une collection de variables X continues ou binaires (variables prédictives, variables explicatives, variables indépendantes, descripteurs, variables exogènes).

Elle fait partie des méthodes d'apprentissage supervisé ; elle peut s'inscrire dans le cadre de la régression linéaire généralisée ; elle peut être vue comme une variante de la régression linéaire multiple, bien connue en économétrie **?**.

Problématique En apprentissage supervisé l'objectif est de prédire les valeurs prises par la variable aléatoire Y définie dans $\{y_1, y_2, ..., y_k\}$. Pour la régression logistique binaire Y , prend uniquement deux modalités $\{+,-\}$ ou $\{1,0\}$.

Nous disposons d'un échantillon Ω de taille n . La valeur prise par Y pour un individu ω est notée $Y(\omega)$. Le fichier comporte j descripteurs $\{X_1, X_2, ..., X_j\}$. Le vecteur de valeurs pour un individu ω s'écrit $(X_1(\omega), ..., X_j(\omega))$.

Dans le cadre binaire, pour un individu donné supposé être positif, sa probabilité s'écrit :

$$P[Y(\omega) = +] = p(\omega) \tag{1}$$

Lorsque l'échantillon est issu d'un tirage aléatoire dans la population, sans distinction des classes d'appartenance, si n_+ est le nombre d'observations positives dans Ω , p peut être estimée par $\frac{n_+}{n}$. La probabilité a posteriori d'un individu ω d'être positif c.-à-d. sachant les valeurs prises par les descripteurs est notée .

$$P[Y(\omega) = +/X(\omega)] = p(\omega) \tag{2}$$

Ce dernier terme est très important. En effet, c'est la probabilité que l'on cherche à modéliser en apprentissage supervisé.

Le LOGIT d'un individu ω s'écrit

$$\ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = a_0 + a_1 X_1(\omega) + ... + a_j X_j(\omega) \tag{3}$$

$a_0, a_1, ..., a_j$ sont les paramètres que l'on souhaite estimer à partir des données.

Lorsque nous adoptons une écriture matricielle, nous écrirons

$$\ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = X(\omega) * a \tag{4}$$

Avec $X(\omega) = (1, X_1(\omega), X_2(\omega), ..., X_j(\omega))$, la première constante $(X_0(\omega) = 1, \forall\omega)$ symbolise la constante ; $a = (a_0, a_1, ..., a_j)$ est le vecteur des paramètres.

Enfin, toujours pour alléger l'écriture, nous omettrons le terme ω lorsque cela est possible.

Nous obtenons une série d'indicateurs lorsque nous le traitons avec le logiciel R. Certaines permettent d'évaluer la qualité globale de la régression, d'autres permettent de juger la contribution individuelle de chaque variable. Expliciter les principes qui régissent la méthode et décrire les formules associées pour que nous sachions lire en connaissance de cause les résultats constituant les objectifs de ce support.

Le modèle LOGIT

La régression logistique peut être décrite d'une autre manière. Pour un individu , on appelle transformation LOGIT de $\pi(\omega)$ l'expression

$$LOGIT = \ln\left[\frac{\pi(\omega)}{1-\pi(\omega)}\right] = a_0 + a_1 X_1(\omega) + ... + a_j X_j(\omega) \tag{5}$$

Posons : $C(X) = a_0 + a_1 X_1(\omega) + ... + a_j X_j(\omega)$

Nous pouvons revenir sur π avec la fonction logistique.

$$\pi = \frac{\exp(C(X))}{1+\exp(C(X))} = \frac{1}{1+\exp(-C(X))} \tag{6}$$

Quelques commentaires et remarques

A propos de la fonction de transformation,

- Le $LOGIT = C(X)$ est théoriquement défini entre $-\infty$ et $+\infty$
- En revanche, $0 \leq \pi \leq 1$ issue de la transformation de $C(X)$ représente une probabilité.
- $C(X)$ et π permettent tous deux de "scorer" les individus, et ainsi de les classer selon leur propension à être "positif".
- π représente une probabilité, avec les propriétés inhérentes à une probabilité, entres autres

$$P(Y = +/X) + P(Y = -X/X) = 1 \tag{7}$$

— la fonction de transfert logistique est non linéaire, c'est en ce sens que l'on qualifie la régression logistique de régression non-linéaire dans la littérature.

La règle d'affectation peut être basée sur π de différentes manières

— Si $\frac{\pi}{1-\pi} > 1$ alors $Y = +$
— Si $\pi > 0.5$ alors $Y = +$

Elle peut être aussi basée simplement sur $C(X)$ avec :

— Si $C(X) > 0$ alors $Y = +$

Evaluation de la régression

Maintenant que nous avons construit un modèle de prédiction, il faut en évaluer l'efficacité. Nous pouvons le faire de différentes manières :

— Confronter les valeurs observées de la variable dépendante avec les prédictions .
— Comparer les vraies valeurs π avec celles prédites par le modèle . En effet, n'oublions pas que la régression logistique sait fournir une bonne approximation de cette quantité. Elle peut se révéler très utile lorsque nous souhaitons classer les individus selon leurs degrés de positivité ou introduire d'autres calculs ultérieurementRakotomalala (2011).

La matrice de confusion La matrice de confusion confronte toujours les valeurs observées de la variable dépendante avec celles qui sont prédites, puis comptabilise les bonnes et les mauvaises prédictions. Son intérêt est qu'elle permet à la fois d'appréhender la quantité de l'erreur (le taux d'erreur) et de rendre compte de la structure de l'erreur (la manière de se tromper du modèle).

Table 1. Forme générique de la matrice de confusion

Y \ $\widehat{Y}$	$\widehat{+}$	$\widehat{-}$	Total
$+$	a	b	$a+b$
$-$	c	d	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

Dans un problème à 2 classes $(+vs.-)$, à partir de la forme générique de la matrice de confusion (Tableau1), plusieurs indicateurs peuvent être déduits pour rendre compte de la concordance entre les valeurs observées et les valeurs prédites. Nous nous concentrons sur les ratios suivants :

a sont les vrais positifs c.-à-d. les observations qui ont été classées positives et qui le sont réellement. b sont les faux positifs c.-à-d. les individus classés positifs et qui sont en réalité des négatifs. De la même manière, c sont les faux négatifs et d sont les vrais négatifs. Mais ces termes sont peu utilisés en pratique car les positifs et les négatifs n'ont pas le même statut dans la majorité des études (ex. les positifs sont les fraudeurs que l'on cherche à isoler ; les positifs sont les personnes atteintes d'une maladie que l'on chercher à détecter ; etc.). Le taux d'erreur est égal au nombre de mauvais classement rapporté à l'effectif total c.-à-d.

$$\in = \frac{b+c}{n} = 1 - \frac{a+d}{n} \tag{8}$$

Il estime la probabilité de mauvais classement du modèle. Le taux de succès correspond à la probabilité de bon classement du modèle, c'est le complémentaire à 1 du taux d'erreur

$$\theta = \frac{a+d}{n} = 1 - \in \tag{9}$$

La sensibilité (ou le rappel, ou encore le taux de vrais positifs $[TVP]$) indique la capacité du modèle à retrouver les positifs

$$Se = Sensibilite = TVP = rappel = \frac{a}{a+b} \tag{10}$$

La précision indique la proportion de vrais positifs parmi les individus qui ont été classés positifs .

$$precision = \frac{a}{a+c} \tag{11}$$

Elle estime la probabilité d'un individu d'être réellement positif lorsque le modèle le classe comme tel. Dans certains domaines, on parle de valeur prédictive positive ($[VPP]$) .

La spécificité, à l'inverse de la sensibilité, indique la proportion de négatifs détectés

$$Sp = Specificite = \frac{d}{c+d} \tag{12}$$

Parfois, on utilise le taux de faux positifs ($[TFP]$) , il correspond à la proportion de négatifs qui ont été classés positifs c.-à-d.

$$TFP = Specificite = \frac{c}{c+d} = 1 - Sp \tag{13}$$

Quelques remarques sur le comportement de ces indicateurs : Un "bon" modèle doit présenter des valeurs faibles de taux d'erreur et de taux de faux positifs (proche de 0) ; des valeurs élevées de sensibilité, précision et spécificité (proche de 1).

Le taux d'erreur est un indicateur symétrique, il donne la même importance aux faux positifs (c) et aux faux négatifs (b).

La sensibilité et la précision sont asymétriques, ils accordent un rôle particulier aux positifs.

Enfin, en règle générale, lorsqu'on oriente l'apprentissage de manière à améliorer la sensibilité, on dégrade souvent la précision et la spécificité. Un modèle qui serait meilleur que les autres sur ces deux groupes de critères antinomiques est celui qu'il faut absolument retenir **?**.

Test de Hosmer-Lemeshow Le test de Hosmer-Lemeshow relève à peu près de la même logique que le diagramme de fiabilité. A la différence qu'au lieu de se baser simplement sur une impression visuelle, on extrait du tableau de calcul un indicateur statistique qui permet de quantifier la qualité des estimations .

La courbe ROC : La courbe ROC est un outil très riche. Son champ d'application dépasse largement le cadre de l'apprentissage supervisé. Elle est par exemple très utilisée en épidémiologie. Pour nous, elle présente surtout des caractéristiques très intéressantes pour l'évaluation et la comparaison des performances des classificateurs :

1. Elle propose un outil graphique qui permet d'évaluer et de comparer globalement le comportement des classifieurs.

2. Elle est indépendante des coûts de mauvaise affectation. Elle permet par exemple de déterminer si un classifieur surpasse un autre, quelle que soit la combinaison de coûts utilisée.
3. Elle est opérationnelle même dans le cas des distributions très déséquilibrées. Mieux, même si les proportions des classes ne sont pas représentatives des probabilités a priori dans le fichier - c'est le cas lorsque l'on procède à un tirage rétrospectif c.-à-d. on fixe le nombre de positifs et négatifs à obtenir, et on tire au hasard dans chaque sous-population - la courbe ROC reste valable.
4. Enfin, on peut lui associer un indicateur synthétique, le critère (aire sous la courbe, en anglais area undercurve), que l'on sait interpréter.

La courbe ROC met en relation le taux de vrais positifs (la sensibilité, le rappel) et le taux de faux positifs dans un graphique nuage de points. Habituellement, nous comparons à un seuil pour effectuer une prédiction . Nous pouvons ainsi construire la matrice de confusion et en extraire les 2 indicateurs précités. La courbe ROC généralise cette idée en faisant varier sur tout le continuum des valeurs possibles entre 0 et 1. Pour chaque configuration, nous construisons la matrice de confusion et nous calculons. C'est l'idée directrice. Elle est un peu lourde à mettre en place. Dans la pratique, il n'est pas nécessaire de construire explicitement la matrice de confusion, nous procédons de la manière suivante :

1. Calculer le score $\widehat{\pi}(\omega)$ de chaque individu à l'aide du modèle de prédiction.
2. Trier le fichier selon un score décroissant.
3. Considérons qu'il n'y a pas d'ex-æquo. Chaque valeur du score peut être potentiellement un seuil s . Pour toutes les observations dont le score est supérieur ou égal à s , les individus dans la partie haute du tableau, nous pouvons comptabiliser le nombre de positifs $n_+(s)$ et le nombre de négatifs n_-s . Nous en déduisons $TVP = \frac{n_+(s)}{n_+}$ et $TFP = \frac{n_-(s)}{n_-}$.
4. La courbe ROC correspond au graphique nuage de points qui relie les couples (TVP,TFP) . Le premier point est forcément $(0,0)$, le dernier est $(1,1)$.

Deux situations extrêmes peuvent survenir. Soit la discrimination est parfaite et tous les positifs sont situés devant les négatifs, la courbe ROC est collée aux extrémités Ouest et Nord du repère. Soit les scores sont totalement inopérants, le classifieur attribue des valeurs au hasard, dans ce cas les positifs et les négatifs sont mélangés. La courbe ROC se confond avec la première bissectrice Wikipedia.

Le critère AUC : Il est possible de caractériser numériquement la courbe ROC en calculant la surface située sous la courbe. C'est le critère AUC . Elle exprime la probabilité de placer un individu positif devant un négatif. Ainsi, dans le cas d'une discrimination parfaite, les positifs sont sûrs d'être placés devant les négatifs, nous avons $AUC = 1$. A contrario, si le classifieur attribue des scores au hasard, il y a autant de chances de placer un positif devant un négatif que l'inverse, la courbe ROC se confond avec la première bissectrice, nous avons $AUC = 0,5$. C'est la situation de référence, notre classifieur doit faire mieux. On propose généralement différents paliers pour donner un ordre d'idées sur la qualité de la discrimination (Tableau 2).

Table 2. Valeur de l'AUC et qualité de la discrimination

Valeur de l'AUC	Commentaire
$AUC = 0,5$	Pas de discrimination
$0.7 \leq AUC \leq 0.8$	Discrimination acceptable
$0.8 \leq AUC \leq 0.9$	Discrimination excellente
$AUC \geq 0,9$	Discrimination exceptionelle

Pour calculer l'AUC , nous pouvons utiliser une bête intégration numérique, la méthode des trapèzes par exemple. Au final, il apparaît que le critère AUC est un résumé très commode. Il permet, entre autres, les comparaisons rapides entre les classifieurs. Mais il est évident que si l'on souhaite analyser finement leur comportement, rien ne vaut la courbe ROC.

Tests de significativité des coefficients : L'objectif des tests de significativité est d'éprouver le rôle d'une variable explicative. Formellement, l'hypothèse nulle peut se décliner comme suit : évaluer la contribution individuelle d'une variable

$$H_0 : aj = 0 \tag{14}$$

Ce test de significativité est systématiquement donné par les logiciels (Tanagra,R). Nous verrons plus loin que seule une de ses formes (test de Wald) est en réalité proposée. L'autre (test du rapport de vraisemblance) est passée sous silence. Or ces approches ne se comportent pas de la même manière. Il faut le savoir pour interpréter les résultats en connaissance de cause.

Tests de Wald :

Matrice de variance-covariance des coefficients : Matrice Hessienne : Lors de la description de l'algorithme d'optimisation de Newton-Raphson, il a été défini une matrice des dérivées partielles secondes, dite matrice hessienne. Nous en reprenons l'expression matricielle.

$$H = X'VX \tag{15}$$

Où X est la matrice des données, la première colonne correspondant à la constante. Elle est de dimension $n * (j+1)$.

La matrice V est une matrice diagonale de taille $n*n$, composée des valeurs de $\pi(\omega)x(1-\pi(\omega))$ les probabilités $\pi(\omega)$ étant obtenues après estimation des paramètres.

Ainsi, nous pouvons former la matrice hessienne H de taille $(j+1)*(j+1)$, et l'inverse de la matrice hessienne correspond à la matrice de variance covariance des coefficients estimés. En particulier, nous obtenons les variances des coefficients sur la diagonale principale.

Ainsi nous disposons de $\widehat{a}$, vecteur des estimations des paramètres de la régression logistique ; nous savons qu'il suit une loi normale multidimensionnelle ; nous disposons de la matrice de variance covariance associée. Tout est en place pour que nous puissions réaliser les tests de significativité Didacticiel Tanagra et R (2009) Didacticiel Tanagra et R (2009).

Tester la nullité d'un des coefficients : Très facile à mettre en oeuvre puisque l'on dispose directement de la variance des coefficients, le test s'appuie sur la statistique de Wald W_j qui, sous H_0 , suit une loi du χ^2 à un degré de liberté.

$$W_j = \frac{\widehat{a}_{aj}^2}{\widehat{\sigma}_{aj}^2} \tag{16}$$

Où $\widehat{\sigma}_{aj}^2$ est la variance du coefficient a_j , lue sur la diagonale principale de la matrice de variance covariance de coefficients.

Le logiciel R, lui, propose la statistique Z_j à la place de W_j , avec :

$$Z_j = \frac{\widehat{a}_j^2}{\widehat{\sigma}_j^2} = signe\left(\widehat{a}_j^2\right) * \sqrt{W_j} \sim N(1,0) \tag{17}$$

Z_j peut prendre des valeurs négatives. Le test étant bilatéral, nous retrouvons exactement les mêmes probabilités critiques $p-value$ qu'avec la statistique de Wald W_j.

Intervalle de confiance de Wald pour un coefficient : a_j suit asymptotiquement une loi normale que l'on soit ou non au voisinage de $a_j = 0$. De fait, nous pouvons construire l'intervalle de confiance au niveau de confiance $a-\alpha$ pour tout coefficient pris individuellement. Les bornes sont obtenues de la manière suivante

$$\widehat{a}_j \pm u_{1-\alpha/2} * \widehat{\sigma}_{aj} \tag{18}$$

$u_{1-\alpha/2}$ est le fractile d'ordre $1-\alpha/2$ de la loi normale centrée réduite.

3. Résultats et discussions

Sélection des variables

La sélection des variables a été effectuée par le package RWeka sur R Didacticiel Tanagra et R (2009) dans lequel est inclus la méthode NAIVE BAYES qui effectue la phase d'apprentissage et le test d'évaluation. Ce dernier révèle un taux d'erreur de 28.4173% .

Une autre phase d'apprentissage est effectuée et les prédicteurs FF, T , et N ont été choisis. Ainsi un nouveau test d'évaluation est fait avec les prédicteurs choisis et le taux d'erreur trouvé est de 28.8561% .

Table 3. Matrice de confusion obtenue sur la base test avec l'ensemble des prédicteurs

Occurrence	absence	présence	total
absence	176	25	201
présence	58	19	77
total	234	44	278

Table 4. Matrice de confusion obtenue sur la base test avec les prédicteurs choisis

Occurrence	absence	présence	total
absence	181	20	201
présence	59	18	77
total	240	38	278

Après cette sélection des variables pertinentes nous allons entreprendre la régression logistique pour établir l'équation et essayer de la valider.

Coefficients de la régression

Table 5. Coefficients issus de modélisation logistique

Coefficient	écart	type	Stat. de Wald	p-value
Intercept	1.80265	1.57535	1.144	0.25250
T	-0.14771	0.04671	-3.162	0.00157 **
N	0.33745	0.05748	5.870	4.35 e-9 ***
FF	0.27128	0.05564	5.876	1.08 e-6 ***

Le tableau ci-dessus résume les coefficients obtenus pour chaque descripteur, y compris la constante, nous avons l'estimation de la valeur du coefficient, son écart type, la statistique de Wald destinée à en évaluer sa significativité et la (p-value) s'y rapportant. Ainsi nous constatons que les variables T, N, FF sont respectivement significatives aux niveaux 0.01 ,0.001 et 0.001, mais en revanche la constante Intercept n'est pas significative même à 10%.

Les coefficients associés à chaque prédicteur, coefficients permettant d'établir l'équation de la régression.

Une simulation a été effectuée sur la base d'apprentissage puis en confrontant valeurs observées et valeurs prédites dans un tableau de contingence.

Ainsi la matrice de confusion obtenue permettra de calculer quelques indicateurs concernant notre modèle.

Table 6. Matrice de confusion obtenue par régression logistique

Occurrence	absence	présence	total
absence	419	26	445
présence	135	18	175
total	554	66	620

- Le taux d'erreur « probabilité de mauvais classement du modèle »
- Le taux de succès« probabilité de bon classement
- Le test statistique de Hosmer Lemeshow effectué montre que la $(p-value) = 0.6098$ est supérieure au risque usuel de 5%. Donc le modèle est compatible avec les données.

Validation du modèle

Pour la validation, l'équation de la régression logistique binaire établie pour utiliser à trouver l'occurrence ou non des phénomènes lithométéores. Cette validation se fera pour les mois de juillet et août 2000. Ainsi les équations, $LOGIT$ et π, établies sont :

$$\begin{aligned} LOGIT &= \ln\frac{\pi}{1-\pi} \\ &= 1.80265 - 0.14771.T + O.33745.N + 0.27128.FF \end{aligned} \tag{19}$$

$$\pi = \frac{1}{1+\exp\left(-C(X)\right)} \tag{20}$$

Avec :

$$C(X) = 1.80265 - 0.14771.T + O.33745.N + 0.27128.FF \tag{21}$$

Table 7. Matrice de confusion sur la base de validation obtenue par régression logistique

Occurrence	absence	présence	total
absence	46	4	50
présence	8	4	12
total	54	8	62

A partir de la matrice de confusion les indicateurs suivants sont calculés :

- Le taux d'erreur $\epsilon = 0.19$
- Le taux de succès $\theta = 0.81$
- La sensibilité $Se = 0.33$
- La précision $p = 0.5$
- La spécificité $Sp = 0.92$

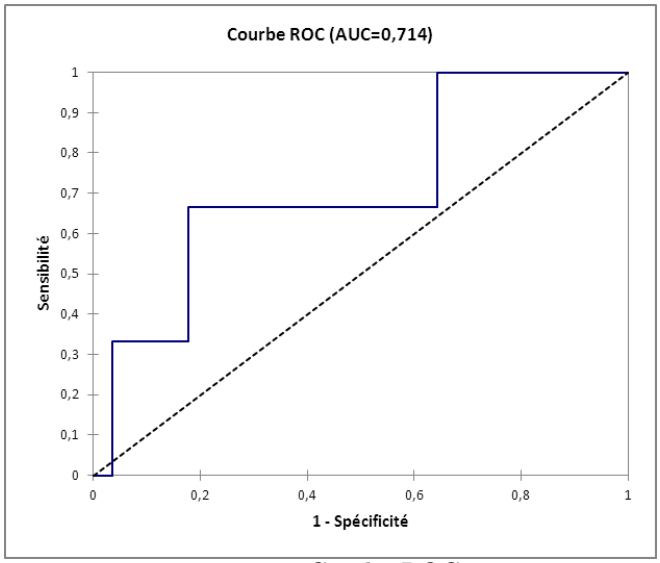

Figure 1. Courbe ROC

La courbe ci-dessus est la courbe ROC, un autre indicateur pour évaluer la régression établie .En se basant sur cette courbe, l'AUC et les indicateurs calculés en haut il peut être affirmé que le modèle statistique est fiable pour la période considérée. Et le test effectué nous donne un taux de succès de 81 .

4. Conclusion

Le taux de succès obtenu nous laisse à penser que des résultats plus optimistes sur des régions où la fréquence d'occurrence du phénomène est plus importante vue la sensibilité de notre méthode statistique sur l'échantillon des deux modalités (présence et absence) peuvent être envisagées.

Références

Didacticiel Tanagra et R (2009). Stratégie "wrapper" pour la sélection de variables.

Didacticiel Tanagra et R (2009). Diagnostic de la régression logistique.

Rakotomalala, R. (2011). Pratique de la regression logistique : Regression logistique binaire et polytomique,. Universite Lumiere Lyon 2.

Wikipedia. Regression lineaire — wikipedia, l'encyclopedie libre.