Méthodes de prévision d'occurrence d'Orage par traitement statistique: Cas d'Alger

BENSABEUR Abdelmadjid [1]*

Abstract
Notre travail porte sur les aspects fondamentaux des méthodes statistiques utilisées dans la prévision météorologique. Il a été élaboré dans le but d'améliorer la prévision d'occurrence d'orage sur l'Algérois.
L'étude met en relation des occurrences d'orages, diagnostiquées à l'aide des valeurs d'indices d'instabilité déduits du modèle ALADIN-Algérie, complétés par des paramètres mesurés au niveau de la station météorologique de Dar El beida (Pression, température, humidité). Le CAPE est l'indice le plus corrélé à l'occurrence d'orage sur l'Algérois.
Les expériences étaient axées sur la méthode Perfect Prog, basée sur les fonctions pronostiques telles-que la régression logistique et l'Arbre de décision (CART). L'utilisation de la CAPE, pour le diagnostic et la prévision d'occurrence d'orage a permis une amélioration substantielle par rapport à une prévision climatologique et, dans une moindre mesure, par rapport à une prévision par persistance.

Keywords
Prévisions météorologiques; indice d'instabilité, prévision d'occurrence d'orage, CAPE.

[1] *Office national de la météorologie, Direction régionale ouest*
*__Correspondant__: a_bensabeur@yahoo.fr

Contents

Les inondations en Algérie font partie des dix risques majeurs auxquels est confronté notre pays. La cause la plus fréquente de ces inondations est un orage qui se déplace lentement et peut déverser d'énormes quantités d'eau sur une zone. La ville d'Alger reste une des zones les plus exposées au risque orageux. Elle représente le lieu de rencontre de deux masses d'air de températures et degrés hygrométriques différents, une maritime et froide, l'autre continentale et sèche. A cet égard, on peut citer quelques cas d'inondations catastrophiques sur les plans humain, matériel et économique, liées à de fortes précipitations ayant généré des crues dévastatrices sur Alger (Publication du conseil national économique et social, mai 2003) :

- Mars 1974 sur les wilayas d'Alger et de Tizi-Ouzou et qui ont favorisé la formation de crues importantes. Ces pluies et crues ont causé des dégâts énormes : 52 morts, 4 570 habitations détruites, 18 000 sinistrés et la destruction de 13 ponts et de nombreuses routes.
- Le 10 novembre 2001, de dévastatrices inondations alimentées par un violent orage ont dévalées, en flots puissants et continus, des hauteurs du quartier de Bab El Oued, emportant sur leurs passagers des centaines de véhicules et se soldant par la mort de plus d'un millier de personnes et de dizaines de disparus.

De tels événements, sans parler des dégâts matériels, montrent la nécessité et l'importance de la prévision des orages. En effet l'anticipation de l'arrivé d'un orage important reste encore assez imprécise dans le temps et dans l'espace. Pour ce faire, les prévisionnistes de l'office national de météorologie utilisent différents produits issus des modèles numériques de prévision du temps, des images satellitaires, des stations d'observations au sol, et des radios sondages. Ceci reste insuffisant sans le traitement statistique qui nous permet d'établir la liaison entre les sorties brutes d'un modèle numérique et les paramètres météorologiques nécessaires aux prévisions d'exploitation. Il s'agit en fait d'une «interprétation» de la sortie du modèle, qui donne des renseignements sur la signification de la prévision numérique en ce qui concerne les paramètres météorologiques observés.

2. Méthodologie

Pour procéder à cette étude nous avons divisé notre travail en deux parties. La première partie est basée sur la littérature. Nous avons donné un résumé des principaux points de la méthodologie statistique utilisée dans le développement d'équations pronostiques. La deuxième partie de ce travail sera consacrée aux principaux résultats obtenus et leurs analyses.

2.1. Méthodes de formulation

En fonction d'un type de données utilisé dans le développement et la mise en œuvre d'équations pronostiques, trois méthodes de formulation sont reconnues:

Méthode classique :

Le développement et l'application des équations pronostiques n'utilisent pas la sortie du modèle de prévision numérique

du temps. Les valeurs des prédicteurs dans un échantillon d'apprentissage sont dérivées des données d'observation disponibles au moment où la prévision doit être publiée.

Méthode Perfect Prog [1]

Dans le développement des équations pronostiques, la prédiction est liée aux valeurs des prédicteurs observées dans l'intervalle de prévision cible. En appliquant les équations Perfect Prog, les prédicteurs sont dérivés de la sortie du modèle de prévision numérique du temps.

Méthode des statistiques de modèles de production (MOS) (Glahn et Lowry, 1972):

Dans les deux cas, les prédicteurs sont dérivés de la sortie du modèle de prévision numérique du temps.

2.2. Type de la fonction pronostique

La fonction pronostique donne une estimation d'une probabilité conditionnelle d'occurrence d'un orage. La fonction pronostique a été déterminée à partir de l'échantillon de données d'apprentissage par 2 méthodes, dans notre cas :

La régression logistique :

La régression logistique (Ricco R., 2011) est une technique prédictive. Elle vise à construire un modèle permettant de prédire / expliquer les valeurs prises par une variable cible qualitative (le plus souvent binaire, on parle alors de régression logistique binaire ; si elle possède plus de 2 modalités, on parle de régression logistique polytomique) à partir d'un ensemble de variables explicatives quantitatives ou qualitatives.

Arbre de décision :

Est un outil d'aide à la décision représentant un ensemble de choix sous la forme graphique d'un arbre. Les différentes décisions possibles sont situées aux extrémités des branches (les « feuilles » de l'arbre), et sont atteints en fonction de décisions prises à chaque étape. L'arbre de décision est un outil utilisé dans beaucoup de domaines variés. Un avantage majeur des arbres de décision est qu'ils peuvent être calculés automatiquement à partir de bases de données par des algorithmes d'apprentissage supervisé. Ces algorithmes sélectionnent automatiquement les variables discriminantes à partir de données non-structurées et potentiellement volumineuses. Ils peuvent ainsi permettre d'extraire des règles logiques de cause à effet (des déterminismes) qui n'apparaissaient pas initialement dans les données brutes. Les algorithmes conçus pour créer des arbres de décision optimisés sont la CART, ASSISTANT, CLS et ID3/ C4.5 (Ricco R., 2011).

2.3. Conditions et données d'étude

Notre étude propose une méthode statistique pour prévoir l'occurrence d'orage sur l'aéroport Houari Boumedienne et ces alentours. Cette étude permettra de déduire les indices d'instabilité, ainsi que les paramètres mesurés les plus corrélés à l'occurrence d'orage. La période d'étude porte sur deux ans, de janvier 2007 au décembre 2008. L'étude met en relation des occurrences d'orages, diagnostiquées à l'aide des valeurs d'indices d'instabilité déduits du modèle ALADIN-Algérie pour la période 2007 et 2008 avec un pas du temps de trois heures, complétés par des paramètres mesurées au niveau de la station météorologique de Dar El Beida au même heure. Nous avons utilisé pour cette étude la méthode de formulation « Méthode Perfect Prog» (Allan J.B., 1999), et la fonction pronostiques : Régression logistique et l'arbre de décision C. RT (CHART) (Der Mégréditchian G., 1979). En ce qui concerne le logiciel de traitement Nous avons utilisé TANAGRA (Ricco R. 2005).

a. Optimisation du prédictand

Le prédictand a été définie comme une occurrence d'orage au moins dans une des deux stations (Dar-El-Beida ou Alger port) (source : centre climatologique Météo algérie). La valeur prédite est considérée comme une variable binaire, représentant l'occurrence d'orage ($y = 1$) ou la non-occurrence ($y = 0$).

b. Optimisation des prédicteurs

L'ensemble des prédicteurs comprend quelques indices d'instabilités dérivés de l'analyse du modèle ALADIN-Algérie, complétés par des observations des paramètres pression, température et l'humidité mesurés au niveau de la station météorologique de Dar El Beida Alger. Les prédicteurs dérivés du modèle ALADIN-Algérie concernent les indices d'instabilités suivants :

La CAPE (Convective Avaible Potentiel Energy) qui est l'énergie potentielle convective disponible, représente l'énergie potentielle convective susceptible d'être transformée en énergie cinétique dans les mouvements ascendants.

Indice K :

Cet indice a été développé par J.J. George (1960), mesure le risque orageux à partir d'une série d'informations sur la température verticale, sur l'apport d'humidité près du sol et l'apport d'air sec en altitude.

$$K = T(850) - T(500) + Td(850) - T(700) + Td(700) \quad (1)$$

avec : T = température du point d'état, Td = température du point de rosée et les niveaux considérés 850 hPa, 700 hPa et 500 hPa.

Indice Totals Total : Cet indice est en fait la somme de deux indices distincts, soit le total vertical (VT) et le total transversal (CT). Le total vertical représente le profil de la température verticale, tandis que le total transversal donne une image du profil d'humidité. L'équation de l'indice Total-Total est la suivante :

$$TT = Td(850) + T(850) - 2T(500) \quad (2)$$

Avec : T = température du point d'état, Td = température du point de rosée et les niveaux considérés 850 hPa et 500 hPa.

Indice Lifted Index (LI) :

LI, est généralement défini par la différence de température à 500 hPa, le milieu de la troposphère, entre la température de l'environnement (T 500 hpa) et celle d'un parcelle d'air

soulevée adiabatiquement depuis la surface (T particule) en degrés Celsius (^oC) :

$$LI = T(500hPa) - T(particule) \quad (3)$$

2.4. Méthodes statistiques de sélection des meilleurs prédicteurs

La sélection de variables est une étape clé de la modélisation. Dans notre cas, nous sommes confrontés à sept descripteurs des variables explicatives potentielles. Certaines d'entre elles sont redondantes, d'autres n'ont aucun rapport avec l'orage. Nous avons choisi pour la sélection la méthode STEPDISC (Stepwise Discriminant Analysis). Cette méthode repose sur le critère du LAMBDA de WILKS. Pour évaluer le rôle significatif d'une variable, nous utilisons :

- La statistique F qui, a priori, suit une loi de Fisher, nous choisissons comme règle d'arrêt de comparer la valeur de F calculée avec le seuil 3.842.
- La p-value calculée pour la variable à évaluer et la comparer avec le niveau de signification de seuil (5%).
- Le lambda de Wilks (L), est l'indicateur privilégié pour l'évaluation statistique du modèle. Il indique dans quelle mesure les centres de classes sont distincts les uns des autres dans l'espace de représentation. Il varie entre 0 et 1 : vers 0, le modèle sera bon parce que les nuages sont bien distincts ; vers 1, les nuages sont confondus, il est difficile de discerner les individus appartenant à des classes différentes. Il s'agit en réalité d'un test d'analyse de variance multivariée.

a- Stepdisc Approche Foward

Pour le Stepdisc Forward (Tableau 1), nous observons que la variable CAPE a été introduite à la première étape, le F calculé est de 1769,50. A la seconde étape, la variable T avec un F de 98,44.

CAPE	T	KI	Tot	LI
L : 0,3995	0,3686	0,3566	0,3551	0,3512
F : 1769,50	98,44	39,68	4,87	12,87
p : 0,0000	0,0000	0,0000	0,0275	0,0003

Tableau 1 : Classement des prédicteurs

N^o	prédicteurs
1	CAPE
2	T
3	KI
4	Tot
5	LI

b- Stepdisc Approche Backward

Pour le StepdiscBakward (Fig.2), nous observons que la variable CAPE a été introduite à la première étape, le F calculé est de 1192,95 à la seconde étape, la variable LI avec un F de 55,86

CAPE	LI	KI	T
L : 0,7088	L : 0,3683	L : 0,3573	L : 0,3640
F : 1192,95	F : 55,86	F : 19,10	F : 41,64
p : 0,0000	p : 0,0000	p : 0,0000	p : 0,0000

Tableau 2 : Classement des prédicteurs

N^o	prédicteurs
1	CAPE
2	LI
3	Tot
4	T
5	KI

c- Comparaison FORWARD – BACKWARD

En comparant les résultats fournis par les deux approches, nous observons que le sous-ensemble final diffère légèrement. Nous recensons les variables dans le tableau suivant :

N^o	foward	backward
1	CAPE	CAPE
2	T	LI
3	KI	Tot
4	Tot	T
5	LI	KI

Seule la CAPE est sélectionné dans les deux cas en premier rang. Les résultats montrent que la valeur du F calculée pour la CAPE dépasse largement le second sélectionné. On peut conclure que l'indice CAPE est le prédicteur le mieux corrélé à l'occurrence d'orage.

2.5. Fonctions pronostiques

2.5.1 Régression logistique

Nous avons isolé les observations dédiées à l'apprentissage, 50 % d'individus sont maintenant sélectionnés pour l'apprentissage. La première information fournie est la matrice de confusion (Tableau 3). Nous disposons du taux d'erreur global en resubstitution (erreur = 0.0051), puis de la sensibilité (Rappel- RECALL) et de (1- Précision) pour chaque modalité de la variable à prédire. Si les YES sont les positifs que l'on cherche à détecter en priorité, nous constatons que notre modèle est plus précis (Précision = 143/143 =1.0000) que sensible (143/146 = 0.9795).

La section Statistiques du modèle (Tableau 4), confronte le modèle étudié (MODEL) avec le modèle trivial composé uniquement de la constante (INTERCEPT). L'idée est d'évaluer la contribution des variables prédictives dans l'explication des valeurs de la variable occurrence d'orage, il faut que les valeurs pour MODEL soient plus faibles que celles de INTERCEPT.

Les indicateurs les plus intéressants sont AIC (Akaike) et SC (BIC de Schwartz) car ils tiennent compte de la complexité du modèle. La déviance (-2LL) du modèle étudié est mécaniquement plus petit que celui du modèle trivial.

Error rate			0.0051			
Values prediction			Confusion matrix			
Value	Recall	1- précision				
				Y	N	Sum
			Y	143	3	146
Y	0.9795	0.0000	N	0	443	443
N	1.0000	0.0067	Sum	143	446	589

Tableau 3. Matrice de confusion.

Predicted attribute	ORAGE	
Positive value	Y	
Number of examples	589	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	661.670	30. 954
SC	666.049	39.710
-2LL	659.670	26.954
Model Chi test (LR)		
Chi-2	632,7165	
d.f.	1	
P(>Chi-2)	0,0000	
R-like		
McFadden's R	Cox and Snell's	
0,9591	0,6584	

Tableau 4. Statistiques du modèle

Concernant notre fichier, si l'on s'en tient au critère SC, nous constatons que les prédictives contribuent effectivement. En effet $AIC(MODEL) = 30.954 < AIC(INTERCEPT) = 661.670$ $SC(MODEL) = 39.710 < SC(INTERCEPT) = 666.049$ La section $MODELCHI^2$ TEST (LR) implémente le test du rapport de vraisemblance pour la significativité globale de la régression. La statistique $CHI{-}2 = LR = -2LL[INTERCEPT] - (-2LL[MODEL]) = 632.7165$ Le degré de liberté est égal au nombre de variables explicatives (1).

Nous obtenons une p-value de 0.0000 avec la loi du KHI-2 à 1 degrés de liberté. Le modèle est donc globalement significatif au risque 5R²-LIKE fournit les pseudo-R². Ils confrontent d'une manière ou d'une autre la vraisemblance du modèle étudié et du modèle trivial. Nous disposons de 3 indicateurs différents (Mc Fadden, Cox and Snell, Nagelkerke). Ils sont proches de 1, à l'exception de Cox and Snell's, la régression est significative.

Attribute	Coef.	Std-dev	Wald	Signif
constant	-29,624142	8,0029	13,7025	0,0002
CAPE	0,224711	0,0648	12,0296	0,0005

Nous disposons ensuite du tableau des coefficients (Tableau 5). 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. Nous constatons que toutes les variables sont significatives à 5.

Modèle de prédiction

$$C(cape) = -29.624142 + (0.224711 CAPE)$$

$$ORAGE = exp(C(cape))/(1 + exp(C(cape)))$$

$0.5 < \text{ORAGE} \leq 1$ Présence d'orage.
$0 < \text{ORAGE} \leq 0.5$ Pas d'orage.

Évaluation du modèle

Diagramme de fiabilité

Le diagramme de fiabilité (Fig.1), (https://eric.univ-lyon2.fr/ricco/cours/cours/pratique_regression_logistique.pdf) cherche également à confronter les probabilités prédites par le modèle (les scores, en abscisse) et les probabilités observées (la proportion de positifs, en ordonnée) dans des groupes d'individus. Si le modèle est bien calibré, les points doivent former une droite. Dans ce cas notre modèle est moins bon, présente une surévaluation (du deuxième au cinquième point).

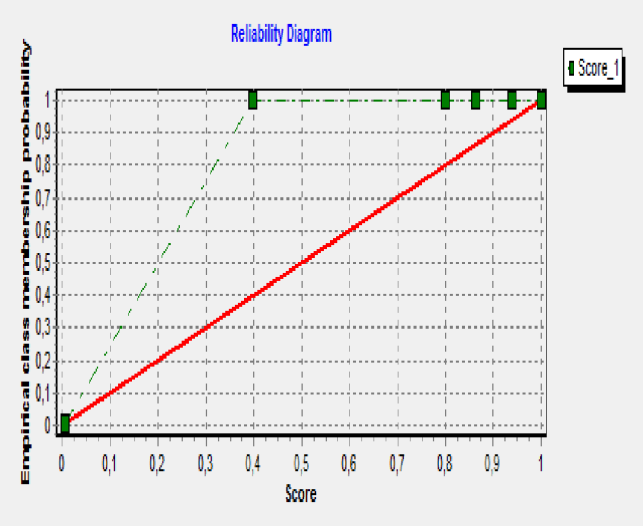

Figure 1. diagramme de fiabilité

Courbe ROC

La courbe ROC (Fig.2), (http://tutoriels-data-mining.blogspot.fr/), évalue la capacité du modèle à placer les positifs devant les négatifs à partir des scores. Nous obtenons la courbe ROC. L'intéressant dans notre contexte est le critère AUC. Nous avons $AUC = 0.998$. La discrimination fournie par le modèle est « acceptable » .

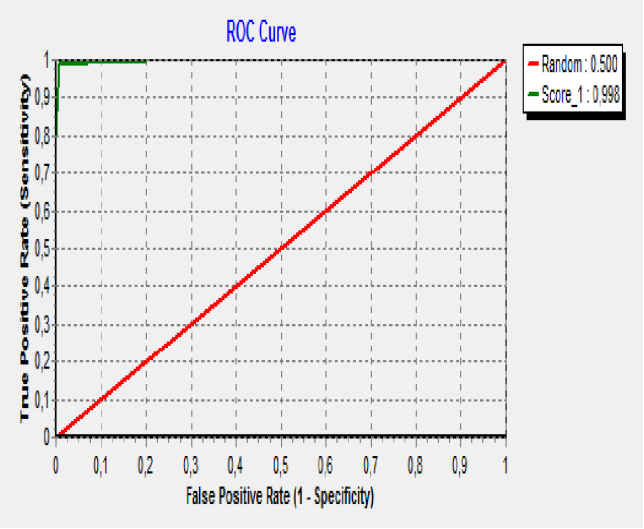

Figure 2. Courbe ROC

Évaluation sur l'échantillon test

Le modèle sur l'échantillon test (fig.8),est précis mais moins sensible ; avec un taux d'erreur acceptable de 0.0119.

Les résultats montrent que le modèle est généralement acceptable, si l'on considère la majorité des indicateurs d'évaluation utilisés jusqu'à présent (matrice de confusion, courbe ROC, évaluation sur test), moins bon sur le diagramme de fiabilité.

2.5.2. Arbre de décision – La méthode CART

La première information fournie est la matrice de confusion (Fig.9). Nous disposons du taux d'erreur global (erreur = 0.0068), puis de la sensibilité (Rappel – RECALL) et de (1-Précision) pour chaque modalité de la variable à prédire. Si les YES sont les positifs que l'on cherche à détecter en priorité, nous constatons que notre modèle est sensible (288/295 = 0.9763) et précis (Précision = 288/289 =0.9989).

Résultat :

CAPE < 125.5000 then ORAGE = *N*, Pas d'orage
CAPE ≥ 125.5000 then ORAGE = *Y* , Présence d'orage.

Évaluation sur l'échantillon test:

Le modèle sur l'échantillon test (Tableau 10),est précis mais moins sensible ; avec un taux d'erreur acceptable de 0.0102.

3. Conclusions

Nous avons donc étudié objectivement, sur un échantillon de taille moyenne, les distributions d'un nombre d'indices d'instabilité calculés à partir des sorties du modèle ALADIN-Algérie, complétés par certains paramètres météorologiques observés à la station météorologique de Dar-El-Beida Alger. Au vu des résultats obtenus, on peut conclure que l'utilisation d'un indice d'instabilité - en particulier celui de la CAPE, est une aide efficace pour la prévision des orages. Ces résultats permettent d'envisager une utilisation pour la prévision opérationnelle. Les seuils proposés peuvent constituer un outil pratique pour les prévisionnistes (Voir chapitre 2.5.1 " modèle de prédiction», ou le chapitre 2.5.2 " Résultat "), c'est une amélioration substantielle par rapport à une prévision climatologique et, dans une moindre mesure, par rapport à la prévision par persistance.

Cette étude présente plusieurs limitations. Parmi ces limitations le faite qu'elle est basée uniquement sur la méthode statistique. En outre certains choix retenus dans cette étude restent arbitraires, comme celui de la taille du domaine spatial autour de la station et celui des plages horaires retenues.

[2] [3] [4] [5]

References

[1] William H Klein, Billy M Lewis, and Isadore Enger. Objective prediction of five-day mean temperatures during winter. *Journal of Meteorology*, 16(6):672–682, 1959.

[2] Ricco Rakotomalala. Tanagra: un logiciel gratuit pour l'enseignement et la recherche. In *EGC*, pages 697–702, 2005.

[3] Allan J Bussey, Michael F Remeika, Brian Newton, Daniel DeBenedictis, and Donald C Norquist. A thunderstorm prediction technique based on a perfect-prog approach. Technical report, AIR FORCE RESEARCH LAB KIRTLAND AFB NMSPACE VEHICLES DIRECTORATE, 1999.

Error rate			0.0119			
Values prediction			Confusion matrix			
Value	Recall	1- précision				
				Y	N	Sum
			Y	142	7	149
Y	0.9530	0.0000	N	0	441	441
N	1.0000	0.0156	Sum	142	448	590

Tableau 8. Matrice de confusion (Echantillon test)

Error rate			0.0068			
Values prediction			Confusion matrix			
Value	Recall	1- précision				
				Y	N	Sum
			Y	288	7	295
Y	0.9763	0.0035	N	1	883	884
N	0.9989	0.0079	Sum	289	890	1179

Tableau 9. Matrice de confusion

[4] Stéphane Sénési and Rose-May Thepenier. Indices d'instabilité et occurrence d'orage: le cas de l'île-de-france. *La Météorologie*, 1997.

[5] Joseph J George. *Weather forecasting for aeronautics*. Academic press, 2014.

Error rate			0.0068			
Values prediction			Confusion matrix			
Value	Recall	1- précision				
				Y	N	Sum
			Y	143	6	149
Y	0.9597	0.0000	N	0	441	441
N	1.0000	0.0134	Sum	143	447	590

Tableau 10. Matrice de confusion (Echantillon test)