Prévision de trajectoires de cyclones à l'aide de forêts aléatoires avec arbres de régression

Sterenn Liberge*, Silèye Ba*,
Philippe Lenca*, Ronan Fablet*

*Institut Telecom ; Telecom Bretagne
UMR CNRS 3192 Lab-STICC
Université européenne de Bretagne
nom.prenom@telecom-bretagne.eu,

Résumé. Nous présentons une étude pour la prédiction des trajectoires de cyclones dans l'océan Atlantique Nord à partir de données issues d'images satellites. On y extrait des mesures de vitesses de vent, de vorticité, d'humidité (base JRA-25) et des mesures de latitude, de longitude et de vitesse de vent instantanée des cyclones toutes les 6 heures (base IBTrACS). Les modèles de référence à ce jour ne tiennent pas compte des corrélations entre les données et les prévisions ce qui limite leur intérêt pour certains utilisateurs. Nous proposons ainsi de prédire le déplacement en latitude et le déplacement en longitude au même instant à un horizon de 120 h toutes les 6 h à l'aide de forêts aléatoires avec arbres de régression. Sur le long terme, à partir de 18 h, la méthode proposée donne de meilleurs résultats que les méthodes existantes.

1 Introduction

Les cyclones sont des évènements dynamiques rares et complexes caractérisés par des vents violents tourbillonnant autour d'une région de basse pression. Leur appellation varie selon la région du monde. Ainsi dans l'Atlantique Nord, ils sont appelés ouragan alors que dans le Pacifique Ouest leur dénomination est typhon. Depuis 40 ans, nous observons en moyenne 80 phénomènes de ce type par an, dans le monde. Le comportement des cyclones dépend de l'endroit où ils se trouvent mais aussi de paramètres dont l'influence est encore mal connue ce qui rend la prédiction de leur trajectoire et de leur intensité, caractérisée par la vitesse de vent maximale, encore plus difficile. La précision de ces prévisions est importante car elle permet aux populations concernées de se protéger du pouvoir destructeur des cyclones et d'éviter des évacuations de populations inutiles.
Les méthodes les plus souvent rencontrées utilisent des modèles dynamiques ou statistiques pour prédire la trajectoire des cyclones et la vitesse de vent. Les modèles dynamiques (Peng et al. (2004); Bender et al. (2007)) reposent sur la résolution d'équations physiques dépendant de paramètres atmosphériques. Ces méthodes sont très couteuses en calculs. Leurs résultats peuvent être utilisés par des méthodes de prévisions par consensus (Krishnamurti et al. (1999); Goerss (2007)) qui combinent les prévisions dynamiques pour donner des prévisions

plus précises. Les modèles statistiques (Knaff et al. (2003); Aberson (1998)) tentent de trouver des modèles de régression à partir d'une base de cyclones antérieurs observés dans le même bassin. Ces méthodes prédisent la position du cyclone à différentes dates indépendamment les unes des autres et donc, ne prennent pas en compte les corrélations entre les variables prédites. D'autre part, la prédiction de la trajectoire des cyclones par méthode statistique est un problème de grandes dimensions. Do et al. (2009, 2010); Simon et al. (2009) proposent d'utiliser des arbres obliques dans des forêts aléatoires afin d'améliorer les performances de méthodes statistiques appliquées à ce type de problème.
Nous avons étudié une méthode statistique qui prédit de manière jointe les déplacements en latitude et en longitude sur un horizon de 120 h avec un pas de 6 h à partir des données relatives à la trajectoire (base IBTrACS) ou des données atmosphériques provenant de la base JRA-25 (Japanese 25-year Reanalysis Project) observées sur une fenêtre de 30 h. Le fait de prédire les déplacements en latitudes et en longitudes conjointement permet de prendre en compte des relations existantes entres ces variables. Contrairement aux travaux cités précédemment (Knaff et al. (2003); Aberson (1998)), nous proposons, ici, d'utiliser des arbres de régression pour prédire la trajectoire des cyclones. Cette méthode a l'avantage d'être simple et d'obtenir des résultats comparables à ceux de l'état de l'art. De plus, jusqu'ici, les méthodes statistiques utilisaient seulement les données relatives à la trajectoire. Dans une première Section (Section 2), nous présentons les données utilisées, ensuite dans la Section 3, nous expliquerons comment les forêts aléatoires à arbres de régression permettent d'obtenir un algorithme « multi-sorties ». Puis nous exposerons le protocole expérimental et les résultats obtenus dans la Section 4. Enfin dans la Section 5, nous donnerons des conclusions et des pistes de futures investigations.

2 Description des données

Dans cette section nous allons tout d'abord exposer quelques généralités sur les cyclones (sous-section 2.1). Ensuite, nous présenterons les données IBTrACS et JRA-25 (sous-sections 2.2 et 2.3) qui sont deux types de données utilisées pour prédire la trajectoire des cyclones. Enfin dans la sous-section 2.4, nous mettrons en forme notre problème.

2.1 Généralités sur les cyclones

Les cyclones sont le plus souvent caractérisés par leur vitesse de vent maximale et le bassin qui les a vu naître (Atlantique Nord, Pacifique Est ...). A titre d'exemple la Figure 1, montre que les trajectoires des cyclones présents dans l'Océan Atlantique (dans le cadre rouge) ont des trajectoires curvilignes alors que les cyclones présents dans le Pacifique Est (cadre vert) suivent des trajectoires quasi rectilignes.

D'autre part, les paramètres atmosphériques tels que la température (Camargo et Zebiak (2002)), l'humidité, la vorticité à différentes altitudes (Sharp et al. (2002)) ainsi que la pression et la température de surface de la mer (Moskaitis et al. (2004)) sont également importants pour déterminer l'évolution d'un cyclone.
Nous voulons prédire la trajectoire des tempêtes tropicales et des cyclones. Une tempête tropicale est un phénomène caractérisé par des vitesses de vents soutenus supérieurs à 34 nœuds. Au delà de 64 nœuds, la tempête est cataloguée comme cyclone (cf. Tableau 1). Leur nombre varie d'une année à l'autre selon les bassins. L'année 2005 a enregistré un nombre record de

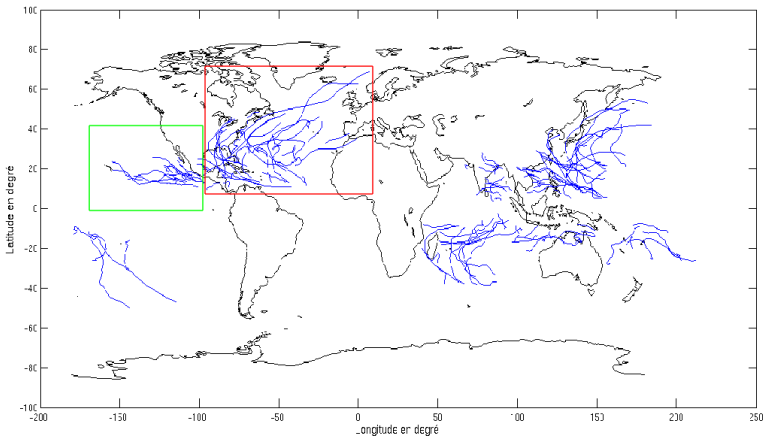


FIG. 1 – *Trajectoires des cyclones pour l'année 2005 tracées à partir des données IBTrACS.*

tempêtes tropicales dans le bassin Atlantique Nord (Figure 1). En effet, 28 tempêtes ont été enregistrées dont 15 ont atteint le stade d'ouragan c'est-à-dire que le vent soutenu maximum enregistré a atteint au moins 64 nœuds soit environ 119 km/h. La vitesse de vent soutenu est le vent mesuré pendant dix minutes à 10 mètres du niveau de la mer. Ces chiffres sont environs trois fois plus importants que le nombre moyen de tempêtes tropicales et de cyclones enregistrés depuis 1965 dans l'Atlantique Nord. C'est pour cette raison que nous avons choisi de prédire la trajectoire des cyclones de l'année 2005 dans ce bassin.

	Vitesse des vents soutenus
Tempête Tropicale	de 34 à 73 nœuds
Ouragan de force 1	de 74 à 82 nœuds
Ouragan de force 2	de 83 à 95 nœuds
Ouragan de force 3	de 96 à 113 nœuds
Ouragan de force 4	de 114 à 135 nœuds
Ouragan de force 5	>135 nœuds

TAB. 1 – *Classification des cyclones selon l'échelle de Simpson (Simpson (1974)).*

2.2 Les données IBTrACS

La base IBTrACS (International Best Track Archive for Climate Stewardship) recense un ensemble d'informations sur les cyclones et les tempêtes tropicales. Les données IBTrACS correspondent aux valeurs médianes, moyennes et maximales de vitesse de vent et de pression enregistrées par 12 stations réparties sur le globe toutes les six heures (Knapp et al. (2010)). La Figure 2 présente par une croix rouge les positions de l'ouragan Wilma enregistrées dans

IBTrACS toutes les 6 h. La flèche indique la position de ce cyclone à la date du 25 octobre 2005 à 00h, ainsi que les mesures de pression et de vitesse de vent soutenu maximale enregistrées.

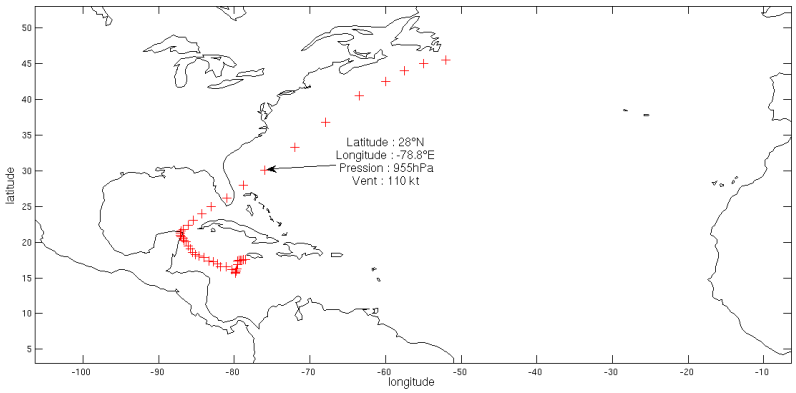
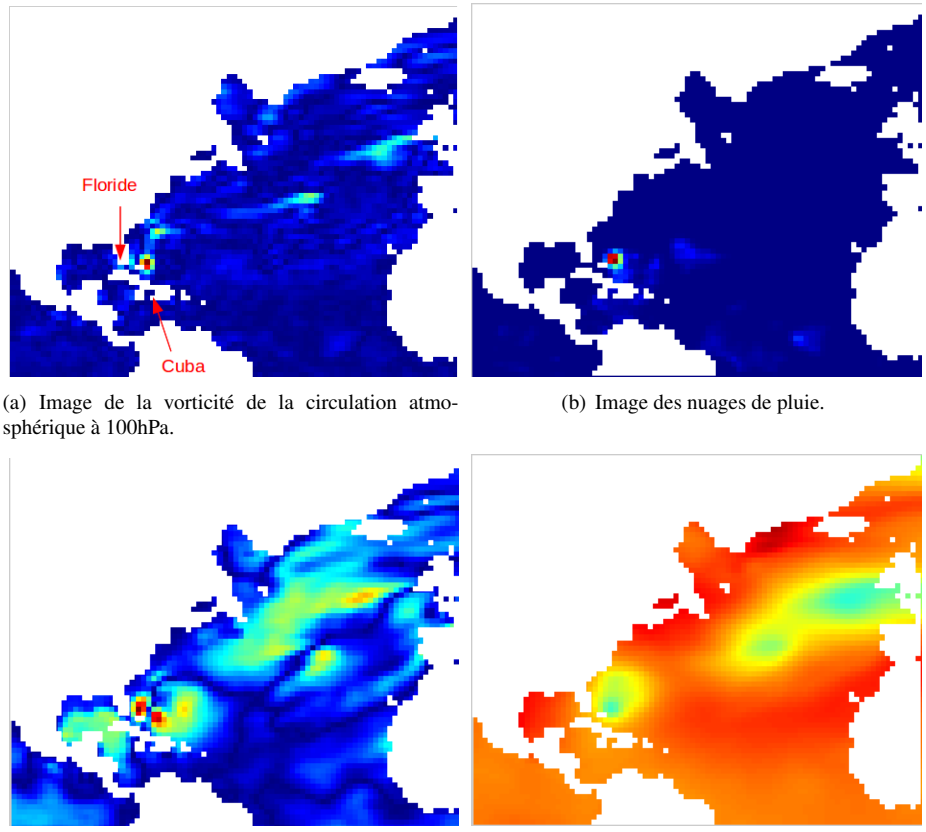


FIG. 2 – *Représentation des données IBTrACS de l'ouragan Wilma.*

2.3 Les données JRA-25

Pour étudier l'impact de l'information contextuelle autour du cyclone, nous avons décidé de prendre en compte des paramètres atmosphériques. Pour cela, nous nous sommes intéressés à la base de données JRA-25. Cette base de données fournit des images satellites du globe toutes les six heures. Ces images satellites sont des mesures de vent, d'humidité, de température, de vorticité et divergence de la circulation atmosphérique à différents niveau de pression. La Figure 3 présente quelques images du bassin Atlantique Nord extraites à partir de cette base de données à une même date. Pour toutes les images, le bleu représente des valeurs faibles et le rouge, les valeurs les plus élevées. Sur toutes ces images, on remarque un point particulier entre la Floride et Cuba. En effet la valeur de la vorticité de la circulation atmosphérique est élevée (Figure 3(a)), des nuages de pluie sont également concentrés sur cette zone (Figure 3(b)), la vitesse de vent selon la latitude est très intense (Figure 3(c)) et la pression est très faible (Figure 3(d)). Toutes ces caractéristiques correspondent au cyclone Wilma. La zone de faible pression et de vent fort située plus près de l'Europe représente une tempête n'ayant pas atteint le niveau d'ouragan.

Ainsi, 250 images satellites sont enregistrées toutes les 6 heures. Chaque pixel d'une image correspond à une surface de $1.25° \times 1.25°$ soit $125km \times 125km$. Pour une image satellite de vent méridional au niveau de la mer, la valeur d'un pixel correspond à la vitesse du vent selon la longitude calculée sur la surface correspondant au pixel par assimilation de données (Onogi et al. (2007)). L'inconvénient de cette base de données est qu'elle ne fournit que des informations à relativement grande échelle.



(a) Image de la vorticité de la circulation atmosphérique à 100hPa.

(b) Image des nuages de pluie.

(c) Image de vent méridional.

(d) Image de pression moyenne au niveau de la mer.

FIG. 3 – *Images obtenues à partir par JRA 25.*

2.4 Jeux de données exploités

La base de données est considérée comme suit. Pour chaque cyclone (ou tempête tropicale), une série temporelle multivariée échantillonnée avec un pas de 6 h comportant les latitudes, les longitudes, les vitesses de vent maximales ainsi que des variables auxiliaires sont extraites des observations satellites de la base JRA-25. A partir de cette base de données, nous définissons deux jeux de descripteurs. Le premier jeu ne contient que les mesures relatives à la trajectoire (latitude, longitude, vitesse de vent) observées sur une fenêtre de 30 heures données par IBTrACS. Le second jeu de descripteurs est constitué des descripteurs précédents auxquels les mesures atmosphériques, observées également sur une fenêtre temporelle de 30 heures, ont été ajoutées. Les données atmosphériques nécessitent un prétraitement avant d'être utilisées comme descripteurs. A l'aide des données IBTrACS, la position du centre du cyclone est connue précisément. Nous définissons autour de ce centre une fenêtre de 10×10 pixels (Figure 4). Cette valeur a été définie empiriquement et nous permet d'observer l'ensemble du cyclone. Ainsi par exemple, la mesure de l'humidité atmosphérique relative à un cyclone à une date

donnée est représentée par 100 valeurs. Nous avons réduit ce nombre de valeurs en calculant l'histogramme des ces valeurs sur 10 quantiles. Nous avons choisis de représenter ces mesures par des histogrammes car ainsi la représentation des cyclones est invariante par rotation autour du centre du cyclone.
Le premier jeu de descripteurs représente chaque objet par 15 valeurs qui correspondent aux

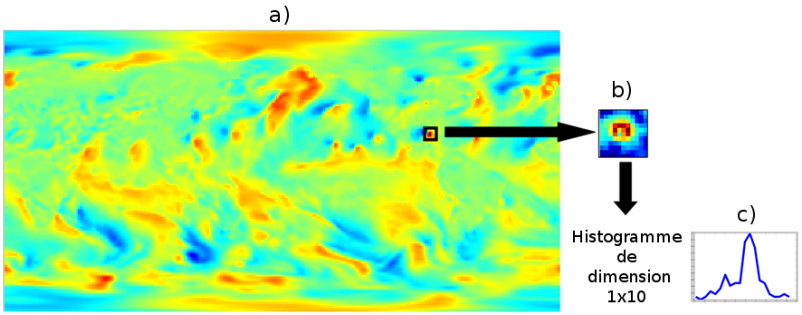


FIG. 4 – *Les trois étapes de l'extraction de descripteurs à partir d'une carte globale d'un paramètre atmosphérique : a) localisation du centre du cyclone b) extraction d'une fenêtre de* 10×10 *pixels centrée sur le centre du cyclone c) calcul de l'histogramme à partir des valeurs du paramètre dans la fenêtre*

latitudes, longitudes et vitesses de vent sur une fenêtre de 30 h échantillonnées avec un pas de 6 h. Avec le second jeu de descripteurs, un objet correspond à 12512 valeurs. Les 12512 valeurs correspondent aux 250 images de paramètres atmosphériques représentées par 10 valeurs prises sur une fenêtre de 30 h échantillonnée avec un pas de 6 h. Les prévisions seront des déplacements en latitudes et en longitudes sur un horizon T avec un pas de 6 h. Ainsi, si nous voulons prédire à un horizon de 120 h, il faudra prévoir 40 valeurs numériques. La prédiction du déplacement du cyclone permet de ne pas être dépendant de sa position courante.

3 Méthodes

Nous avons testé deux méthodes fondées sur les forêts aléatoires pour prédire la trajectoire des cyclones. Les forêts aléatoires ont pour avantages d'être simple à mettre en place et permettent de sélectionner des variables discriminantes automatiquement pour des problèmes contenant un grand nombre de caractéristiques. Avec les deux méthodes utilisées, nous cherchons à prédire la trajectoire future jusqu'à un horizon de prédiction fixé étant donnée la trajectoire passée. La première méthode consiste à prédire les différentes valeurs indépendamment les unes des autres, chacune avec une forêt aléatoire à arbres de régression (sous-section 3.2). La seconde méthode prend en compte la corrélation des valeurs à prédire et détermine de manière jointe les déplacements en latitude et en longitude toutes les 6 h jusqu'à un horizon T donné à l'aide d'une forêt aléatoire (sous-section 3.3).

3.1 Les forêts aléatoires

Dans le cas de la régression, une forêt aléatoire est un régresseur constitué d'un certain nombre d'arbres de régression. La valeur prédite sera la moyenne de celles données par chacun des arbres pour les forêts à arbres de décision et la classe de la plus probable pour les forêts à arbres de classification. Cette méthode utilise le « bagging » et la selection aléatoire de descripteurs qui permet d'obtenir une plus grande variété d'arbres de régression construits à partir d'échantillons d'entrainement obtenus par bootstrapping (Breiman (2001)).

3.2 Les forêts aléatoires par arbres de régression

Les forêts aléatoires considérées reposent sur des arbres de régression construits par maximisation du gain d'information, en minimisant la variance des éléments à chaque nœud (Equation 1).

$$\begin{cases} \arg\max\limits_{d,V_d} \Delta R(V_d), \\ \Delta R(V_d) = Var(Y) - Var(Y_d) - Var(Y_g) \end{cases} \tag{1}$$

où Y représente les objets dans le nœud parent, Y_d et Y_g représentent les objets se retrouvant dans les nœuds enfants. Les objets seront des déplacements en latitude ou des déplacements en longitude à un instant donné. L'inconvénient de cette méthode est que l'on ne peut prévoir qu'un élément à la fois. Si nous voulons faire des prévisions à un horizon de 120 h avec un pas de 6 h, nous devons prédire 20 déplacements en latitude et 20 déplacement en longitude. Cela signifie que nous devons appeler 40 fois cet algorithme. De plus, les différents déplacements en longitude et en latitude au cours du temps sont fortement corrélés. En les prédisant indépendamment les uns des autres, l'information mutuelle entre les variables n'est pas exploitée.

3.3 Les forêts aléatoires par arbres de régression multi-sorties

Pour tenir compte de la corrélation entre les prédictions, nous proposons d'utiliser des forêts aléatoires permettant de prédire plusieurs variables en même temps comme l'a déjà proposé De'ath (2002). Ainsi nous pouvons prédire les déplacements en latitude et en longitude à différents instants par une seule forêt dont la sortie sera un vecteur. Le principe est le même que celui des forêts aléatoires dont la variable prédite est une quantité scalaire (cf. sous-section 3.2), on cherche à maximiser le gain d'information en minimisant la variance des objets dans chacun des nœuds issus de la séparation (Equation 2).

$$\begin{cases} \arg\max\limits_{d,V_d} \Delta R(V_d), \\ \Delta R(V_d) = \sum\limits_{i=1}^{P} Var(Y_i) - \sum\limits_{i=1}^{P} Var(Y_{di}) - \sum\limits_{i=1}^{P} Var(Y_{gi}) \end{cases} \tag{2}$$

où P équivaut à la dimension du vecteur prédit, c'est-à-dire que si on veut prévoir à un horizon de 120 h avec un pas de 6 h, alors $P = 2 \times 20$ prévisions. Ainsi en minimisant la somme des variances, la variance de toutes les variables expliquées est optimisées de manière jointe. Dans les forêts aléatoires classiques, les variances des variables sont optimisées séparément. La méthode multi-sorties s'inspire de méthodes développées dans le multi-task learning (Caruana (1997); Evgeniou et al. (2006)). Les travaux conduits dans le multi-task learning ont montré

que le fait d'apprendre plusieurs tâches en même temps rend un prédicteur plus robuste au sur-apprentissage surtout lorsque la taille de l'ensemble d'apprentissage est limité.

4 Expérimentation

Les résultats sont présentés pour les tempêtes tropicales survenues pendant l'année 2005 dans le bassin Atlantique Nord. Cette année compte 28 évènements de ce type dont 15 ont été catalogués comme cyclones. La sous-section 4.1 présente le protocole expérimental mis en place pour déterminer les paramètres optimaux. Ensuite, dans la sous-section 4.2, nous exposerons les résultats obtenus par des forêts aléatoires « classiques » et ceux obtenus pour des forêts aléatoires « multi-sorties » selon le type de descripteurs utilisé.

4.1 Protocole Expérimental

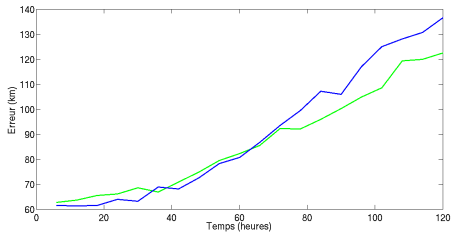
Nous prédisons le déplacement en latitude ($dlat$) et le déplacement en longitude ($dlong$) (cf. Equation 3) des cyclones car cela nous permet d'être indépendant par rapport à la position courante du cyclone tout en ayant sa trajectoire. De manière plus formelle, notre but est de construire une fonction de prédiction définie par :

$$\begin{array}{ll} Y = F(X) & \text{avec } Y = [dlat_{T+6}\ dlat_{T+12} \dots\ dlong_{T+6}\ dlong_{T+12} \dots] \\ X & \text{: descripteurs IBTrACS ou IBTrACS + JRA-25} \\ F & \text{: forêt aléatoire} \\ T & \text{: dernière date d'observation} \end{array} \quad (3)$$

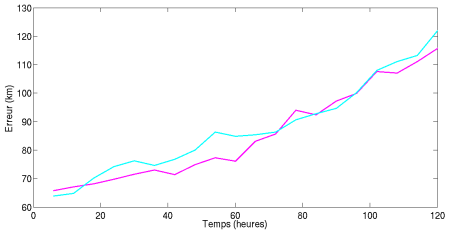
Ces prédictions sont faites sur un horizon de 120 h avec un pas temporel de 6 h ce qui correspond à 40 valeurs numériques. A partir des données décrites dans la section 2, deux groupes de descripteurs sont extraits. Ces descripteurs sont des observations de la trajectoire (IBTrACS) ou des observations atmosphériques (JRA-25) sur une fenêtre temporelle. La taille de cette fenêtre (30 heures) est déterminée par validation croisée à 10 passes (« 10-fold cross validation »). Les données IBTrACS sont les données utilisées par les méthodes statistiques de l'état de l'art (Aberson (1998)). L'algorithme des forêts aléatoires possède comme paramètres le nombre d'arbres (150 arbres) et le pourcentage d'élément d'apprentissage à utiliser (10%) pour le « bagging » (cf. sous-section 3.2). Ces deux paramètres sont également déterminés par validation croisée à 10 passes. Les résultats obtenus sont comparés en calculant l'erreur moyenne exprimée en kilomètre sur la localisation du centre des cyclones.

4.2 Résultats

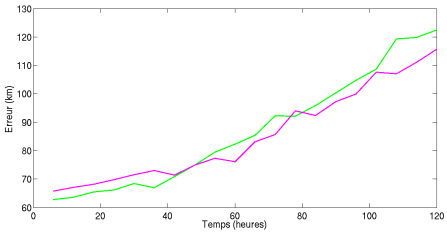
Nous présentons les erreurs moyennes obtenues sur l'estimation de la position du cyclone par les forêts aléatoires classiques et les forêts aléatoires à sorties multiples (Figure 5). Avec les données IBTrACS (Figure 5(a)), les forêts aléatoires classiques (en bleu) semblent réaliser de meilleures performances sur le court terme (72 h), tandis que sur le long terme, la tendance s'inverse en faveur des forêts aléatoires à sorties multiples (en vert, Figure 5(a)). Lorsque les données atmosphériques sont ajoutées aux données IBTrACS (Figure 5(b)) , les forêts aléatoires classiques (courbe bleue claire, Figure 5(b)) et à sorties multiples (courbe rose, Figure 5(b)) réalisent des performances similaires. Les erreurs obtenues avec l'algorithme multi-



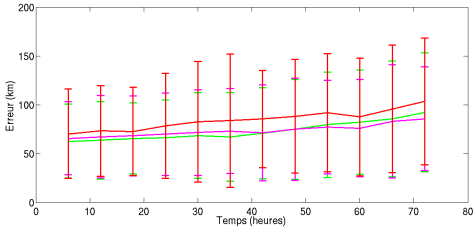
(a) Erreurs moyennes en kilomètres obtenues au cours du temps en heures avec les données IBTrACS. courbe bleue : avec les forêts aléatoires classiques, courbe verte : avec les forêts aléatoires multi-sorties.



(b) Erreurs moyennes en kilomètres au cours du temps en heures avec l'ajout de l'information atmosphérique (IBTrACS+JRA-25). courbe bleue claire : avec les forêts aléatoires classiques, courbe rose : avec les forêts aléatoires multi-sorties.



(c) Comparaison des erreurs moyennes en kilomètres obtenues au cours du temps en heures par les deux jeux de descripteurs avec les forêts aléatoire multi-sorties. courbe verte : données IBTrACS, courbe rose : données JRA-25 + IBTracs.



(d) Comparaison des erreurs moyennes en kilomètre obtenues au cours du temps en heures avec une méthode à vecteur de support (en rouge) et les forêts aléatoires multi-sorties (en magenta) ou classiques (en vert).

FIG. 5 – *Erreur au cours du temps, obtenues par les différents algorithmes.*

sorties pour les deux jeux de prédicteurs sont présentés dans la Figure 5(c). Au delà de 48 h les résultats obtenus avec les données atmosphériques (en rose sur la Figure 5(c)) sont plus précis que ceux obtenus avec les données IBTrACS (courbe verte). Les performances des forêt aléatoires classiques et à sorties multiples (respectivement courbe verte et courbe rose, Figure 5(d)) ont été comparées à celles des méthodes à vecteurs support à noyau Gaussien (SVM, Chang et Lin (2001)), (courbe rouge, Figure 5(d)). Sur cette figure apparaît également la variance des erreurs de prédiction pour les forêts aléatoires indépendantes, les forêts aléatoires conjointes et les SVM. Il peut être noté que les résultats obtenus avec SVM ont une variance plus forte que ceux obtenus par des méthodes utilisant des forêts aléatoires. Les méthodes basées sur les forêts aléatoires, quant à elles, produisent des variances d'erreurs de prédiction similaires. Ces résultats sont concordant avec ceux de Meyer et al. (2003); Do et al. (2009). Ainsi, nous pouvons constater que les forêts aléatoires « multi-sorties » présentent des résultats légèrement plus précis que ceux obtenus par forêts aléatoires classiques. C'est également le cas lorsqu'elles sont comparées à d'autres méthodes classiques de prévisions telles que les SVM. Cependant, nos performances restent à confirmer par une évaluation sur une base d'évaluation plus large que nous sommes en train de constituer. La méthode à sorties multiples permet de gagner considérablement au niveau du coût algorithmique car pour obtenir des prévisions concernant la trajectoire à une horizon de 120 h, il faut appeler 40 fois les forêts aléatoires classiques alors

que pour les forêts aléatoires à sorties multiples, un appel suffit pour obtenir les prédictions. Dès lors l'utilisation des forets aléatoires à sorties multiples se révèle en terme de coût de calcul beaucoup moins coûteux sachant qu'un appel d'une forêt aléatoire indépendante équivaut à un appel d'une forêt aléatoire multi-sorties.

4.3 Comparaison aux résultats connus

D'après Franklin (2006), la méthode CONU est la méthode donnant les meilleurs résultats, pour l'année 2005, dans l'Atlantique Nord. Elle utilise un consensus de méthodes dynamiques (Goerss (2007)). Ce modèle se base sur les prévisions d'au moins deux modèles dynamiques pour les utiliser dans une équation de régression qui prédit le déplacement du cyclone en latitude et en longitude par rapport à sa position d'origine. Le modèle de régression est construit à partir de deux années précédentes. Nous avons comparé à titre informatif nos résultats à cette méthode (Figure 6) à partir des résultats fournit dans le rapport Franklin (2006) car les informations disponibles dans le papier et le manque d'accès aux données ne nous permettent pas de reproduire exactement le protocole expérimental. La courbe d'erreur moyenne sur l'année

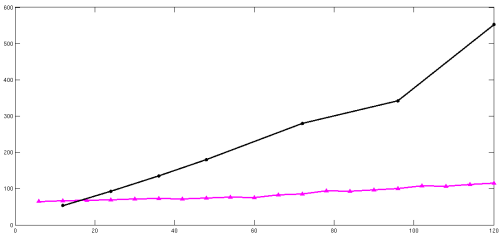


FIG. 6 – *Comparaison des erreurs moyennes en kilomètres au cours du temps en heures de la méthode des forêts aléatoire conjointes (en magenta) à ceux obtenus par la méthode CONU (en noir) présentées dans Franklin (2006).*

2005 obtenue par CONU apparaît en noir sur la Figure 6. Au delà de 18h, la méthode que nous proposons (en rose sur la Figure 6) semble donner de résultats bien meilleurs que la méthode CONU qui prévoit les déplacements en latitudes et en longitudes depuis l'instant initial les uns après les autres. Les mauvaises performances de CONU sur le long terme sont peut-être dues au fait que le modèle de régression est établi à partir des deux années précédentes qui ne sont sans doute pas suffisamment représentatives étant donné la variabilité des saisons cycloniques d'une année à l'autre.

5 Conclusion

Nous avons mené une étude sur la prédiction de la trajectoire de cyclones dans l'Atlantique Nord à l'aide de forêts aléatoire par arbres de régression en comparant, d'une part, l'influence des descripteurs sur les performances, et d'autre part, les performances de forêts aléatoires classiques et multi-sorties. Les résultats obtenus avec des descripteurs issus de données atmosphériques semblent plus précis que ceux relatifs à la trajectoire des cyclones sur le long

terme. Les descripteurs issus de données atmosphériques donnent des résultats assez similaires lorsqu'ils sont utilisés par une forêt aléatoire classique ou une forêt aléatoire multi-sorties. Cependant, lorsque les données relatives aux trajectoires sont utilisées comme descripteurs, une différence de performances entre ces deux méthodes est constatée. Cela nous conforte dans l'idée qu'il faut tenir compte des relations existantes entre les prévisions. Pour étudier plus en détails la différence de performance entre ces deux méthodes lors de l'usage de descripteurs issus de données atmosphériques, il serait envisageable de représenter les données par des méthodes permettant, par exemple, de tenir compte de leur répartition spatiale. De plus, la méthode à sorties multiples permet de réduire considérablement le coût algorithmique. Les résultats obtenus sont encourageant pour continuer sur cette voie puisqu'ils présentent un net avantage sur les méthodes existantes. Cependant la méthode proposée nécessite encore une validation statistique plus exhaustive sur une base de données plus importante, en cours de constitution.

Références

Aberson, S. (1998). Five-day tropical cyclone track forecasts in the North Atlantic basin. *Weather and Forecasting 13*(4), 1005–1015.

Bender, M., I. Ginis, R. Tuleya, B. Thomas, et T. Marchok (2007). The operational GFDL coupled hurricane-ocean prediction system and a summary of its performance. *Monthly Weather Review 135*(12), 3965–3989.

Breiman, L. (2001). Random Forests. *Machine Learning 45*, 5–32.

Camargo, S. et S. Zebiak (2002). Improving the detection and tracking of tropical cyclones in atmospheric general circulation models. *Weather and forecasting 17*, 1152–1162.

Caruana, R. (1997). Multitask learning. *Machine Learning 28*(1), 41–75.

Chang, C.-C. et C.-J. Lin (2001). *LIBSVM : a library for support vector machines*.

De'ath, G. (2002). Multivariate regression trees : a new technique for modeling species-environment relationships. *Ecology 83*(4), 1105–1117.

Do, T.-N., S. Lallich, N.-K. Pham, et P. Lenca (2009). Un nouvel algorithme de forêts aléatoires d'arbres obliques particulièrement adapté à la classification de données en grandes dimensions. In *EGC*, pp. 79–90.

Do, T.-N., P. Lenca, S. Lallich, et N.-K. Pham (2010). Classifying Very-High-Dimensional Data with Random Forests of Oblique Decision Trees. In *EGC (best of volume)*, pp. 39–55.

Evgeniou, T., C. Micchelli, et M. Pontil (2006). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research 6*(1), 615.

Franklin, J. (2006). 2005 National Hurricane Center Forecast Verification Report.

Goerss, J. (2007). Prediction of consensus tropical cyclone track forecast error. *Monthly Weather Review 135*(5), 1985–1993.

Knaff, J., M. DeMaria, C. Sampson, et J. Gross (2003). Statistical, 5-day tropical cyclone intensity forecasts derived from climatology and persistence. *Weather and Forecasting 18*, 80–92.

Knapp, K., M. Kruk, D. Levinson, H. Diamond, et C. Neumann (2010). The International Best Track Archive for Climate Stewardship (IBTrACS) : Unifying tropical cyclone data. *Bulletin of the American Meteorological Society 91*(3), 363–376.

Krishnamurti, T., C. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C. Williford, S. Gadgil, et S. Surendran (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science 285*(5433), 1548.

Meyer, D., F. Leisch, et K. Hornik (2003). The support vector machine under test. *Neurocomputing 55*(1-2), 169–186.

Moskaitis, J., J. Hansen, et K. Emanuel (2004). An Ensemble Approach to Tropical Cyclone Intensity Forecasting. In *26th Conference on Hurricanes and Tropical Meteorology*.

Onogi, K., J. Tsutsui, H. Koide, M. Sakamoto, S. Kobaayashi, H. Hatsushika, T. Matsumoto, N. Yamazaki, H. Kamahori, K. Takahashi, et al. (2007). The JRA-25 reanalysis. *Journal of the Meteorological Society of Japan 85*(3), 369–432.

Peng, M., J. Ridout, et T. Hogan (2004). Recent modifications of the Emanuel convective scheme in the Navy operational Global Atmospheric Prediction System. *Monthly weather review 132*(5), 1254–1268.

Sharp, R., M. Bourassa, et J. O'Brien (2002). Early detection of tropical cyclones using SeaWinds-derived vorticity. *Bulletin of the American Meteorological Society 83*(6), 879–890.

Simon, C., J. Meessen, et C. De Vleeschouwer (2009). Insertion de proximal SVM dans des arbres aleatoires,mode d'emploi. In *Conférence Francophone sur l'Apprentissage Artificiel*, Hammamet, Tunisie.

Simpson, R. (1974). The hurricane disaster potential scale. *Weatherwise 27*(8), 169.

Summary

This paper presents a study about cyclone tracks forecasting in the North Atlantic basin. We used two types of observations for cyclone track forecasting: wind speed, vorticity, humidity extracted from the JRA database, and 6-hourly latitudes and longitudes localization, and sustained wind extracted from the IBTrACS database. Up to our knowledge, the existing state of the art methods do not take into consideration the correlation between the observation and the forecasted cyclone displacement. To account for the correlation between the observation and the cyclone track during prediction, we propose to use a random forest, which is an ensemble of regression trees, which jointly predicts a whole set of future cyclone displacements consisting in 6-hourly latitude and longitude displacements up to a prediction horizon of 120 hours. The experiments conducted to assess the effectiveness of the proposed method suggest that, for prediction horizons above 18h, our method achieves better performance than state of the art methods.