# Prévision des trajectoires d'avions par les méthodes d'apprentissage automatique : Approche par CART et forêts aléatoires

Norbert Fouemkeu*,*** Nour-Eddin El Faouzi*
Jacques Sau**, Rémy Fondacci*

*LICIT - Laboratoire d'Ingénierie Circulation Transports,
Unité mixte IFSTTAR/ENTPE
25, avenue François Mitterrand - Case 24 - F- 69675 Bron
Cedex - France
{norbert.fouemkeu, elfaouzi, remy.fondacci}@ifsttar.fr
http://www.ifsttar.fr
**LMFA - Laboratoire de Mécanique des Fluides et d'Acoustique,
UMR 5509, Université Lyon 1,
43, blvd du 11 Novembre 1918,
F - 69622, Villeurbanne Cedex - France
sau@univ-lyon1.fr
http://www.univ-lyon1.fr
***Ecole Polytechnique Universitaire de Lyon 1,
MAM - Mathématiques Appliquées et Modélisation,
15, Boulevard Latarjet, 69622 Villeurbanne Cedex - France
norbert.fouemkeu@univ-lyon1.fr
http://www.istil-epu-lyon1.fr

**Résumé.** La forte croissance du trafic aérien dans l'espace européen ($5\%$ par an) indique que le contrôle aérien à l'avenir devra faire face à un nombre d'avions de plus en plus important. La prévision de l'incertitude sur les trajectoires des avions s'impose alors comme une nécessité opérationnelle. Cet article vise un double objectif : ($i$) A partir des données réelles du trafic aérien, nous montrons que les méthodes d'apprentissage automatique CART et les forêts aléatoires permettent de réaliser les prévisions efficaces des instants de passage des avions en des points de leur trajectoire. Nous quantifions le pouvoir prédictif du modèle construit par la méthode CART et de celui des forêts aléatoires. Ces modèles sont évalués sur les données test. ($ii$) Nous montrons que sous certaines conditions, le modèle des forêts aléatoires présente un risque de surapprentissage.

## 1 Introduction

Le niveau important d'incertitude sur les trajectoires de vols est un facteur de baisse de capacité dans l'espace aérien, et donc du dysfonctionnement de la gestion du trafic aérien

en général. Nous proposons dans cet article des modèles d'apprentissage automatique pour la prévision des trajectoires $4D$ d'avions en fonction des données opérationnelles du trafic aérien. Autrement dit, pour un avion en vol, observé à un instant $t_o$ à une position $M(t_o)$ de sa trajectoire prévue dans son plan de vol, les modèles développés doivent permettre de prévoir l'instant $t$ de passage de cet aéronef au point futur $M(t)$ de cette trajectoire. Un tel modèle peut être formulé par l'équation (1) :

$$t = t^p + \Phi(\mathbf{X}) + \epsilon. \tag{1}$$

où :

- $t$ est l'instant de passage d'un avion au point $M$ de sa trajectoire prévue.
- $t^p$ est l'instant prévu dans le plan de vol. Cet instant est connu à partir des plans de vols déposés[1].
- $\mathbf{X}$ est un vecteur de variables explicatives de la différence $t - t^p$. Il s'agit : Des plans de vols, des conditions météorologiques et atmosphériques, de la complexité du trafic et les paramètres courants des vols. Ces groupes de variables sont décrits ci-dessous.
- $\Phi$ est une fonctionnelle à ajuster en fonction de $\mathbf{X}$.
- $\epsilon$ est une composante résiduelle.

## 1.1 Plan de vol d'origine et plan de vol simulé par OPERA

Lorsqu'un avion est observé à l'instant $t_0$ à la position courante $M(t_0)$ de sa trajectoire réelle, nous utilisons le simulateur du trafic aérien OPERA développé au LICIT pour simuler la trajectoire de vol restante. $M(t_0)$ est considéré comme le point de départ de la trajectoire à prévoir. Les points balises de la trajectoire simulée sont celles du plan de vol d'origine qui jalonnent la portion de la trajectoire de vol prévue qui reste à parcourir. Ce simulateur prend en entrée les paramètres de performances des avions issues de la base BADA (*Base of Aircraft Data*) d'Eurocontrol[2]. Il s'agit des paramètres nominaux des avions disponibles dans Eurocontrol (2009). Les nouveaux instants prévus pour survoler le reste des points balises ont été calculés en initialisant le temps à $t_0$ sur $M(t_0)$. Pour ces instants nous gardons la notation $t^p$. Dans la suite, lorsque nous parlerons de plan de vol ou de la trajectoire prévue, nous ferons référence à cette trajectoire simulée par OPERA.

## 1.2 Vecteur de variables explicatives X

Le vecteur de variables explicatives $\mathbf{X}$ contient quatre groupes de variables susceptibles d'avoir une influence directe sur le temps de passage des avions sur les points de leur trajectoire prévue :

- *Plans de vols* : Chaque plan de vol contient les points balises à traverser, les heures de survol de ces balises, les secteurs à traverser par le vol, les heures d'entrée et de sortie des secteurs, les aéroports de départ et de destination du vol, le niveau de vol de croisière prévu, le type d'avion utilisé, la compagnie aérienne exploitante du vol, etc.

[1] Un vol commence par le dépôt d'un plan de vol auprès des autorités de gestion des flux de trafic. Ce plan de vol contient un ensemble d'informations sur la route prévue pour le vol.

[2] Organisation européenne pour la sécurité de la circulation aérienne.

- *Conditions météorologiques et atmosphériques* : Il s'agit des prévisions du vent (son intensité et sa direction), la température, la pression, la densité de l'air et l'humidité spécifique sur le reste de la trajectoire prévue à partir du point courant $M(t_o)$.
- *Complexité du trafic* : Il s'agit de la complexité du trafic sur le reste de la trajectoire de vol prévue à partir du point courant $M(t_o)$. Elle dépend de la géométrie des secteurs traversés, de type de flux d'avions (horizontal, vertical) entrant ou sortant des secteurs, des interactions potentielles entre les avions pendant le vol et de l'hétérogénéité des performances entre les avions.
- *Paramètres courants* : Les paramètres courants du vol à l'instant $t_o$ comprennent la vitesse, le taux de montée ou de descente de l'aéronef, le retard du vol par rapport au plan de vol d'origine, la distance sur le plan de vol entre le point $M(t_o)$ et le point $M(t)$, la différence entre l'altitude du point courant $M(t_o)$ et l'altitude du point $M(t)$ de la trajectoire prévue.

Une description plus détaillée de ces variables est présentée en annexe (TAB 1 et TAB 2).

## 1.3 Identification du modèle

L'équation (1) définie précédemment peut de façon équivalente s'écrire :

$$t - t^p = \Phi(\mathbf{X}) + \epsilon. \tag{2}$$

Pour un avion et un point $M$ de sa trajectoire prévue, la différence « $t - t^p$ » est l'écart entre l'instant $t$ de passage réel de cet avion au point $M$ et l'instant $t^p$ prévu dans son plan de vol. La prévision de l'instant $t$ en fonction des variables explicatives du vol est alors équivalente à la prévision de l'écart « $t - t^p$ » en fonction de ces mêmes variables. Dans le reste de l'article, nous parlerons tout simplement de *l'écart temporel* pour faire référence à cette différence que nous notons :

$$\triangle\tau = t - t^p.$$

La modélisation consiste maintenant à trouver un estimateur de $\triangle\tau$ en fonction de l'ensemble des caractéristiques du vol représentées par le vecteur $\mathbf{X}$. En moyenne, cet estimateur que nous notons $\widehat{\triangle\tau}$ vérifie la relation suivante :

$$\widehat{\triangle\tau} = \Phi(\mathbf{X}). \tag{3}$$

En effet, la variable résiduelle issue de la modélisation est supposée de moyenne nulle. En utilisant les données d'archives du trafic réel enregistrées par le système CPR (Correlated Position Report), on peut construire un tel estimateur en minimisant les écarts résiduels par la méthode des moindres carrés.

Comme il est ardu, voire impossible de disposer de tous les points constitutifs de la trajectoire réelle d'un avion, nous construisons les modèles en utilisant les données sur les points d'entrée et sortie des secteurs. Par ailleurs, nous supposons que le point courant $M(t_0)$ est situé après le décollage de l'avion.

Cet article est organisé comme suit : Le paragraphe (2) est consacré d'abord sur la méthodologie et les données utilisées sont ensuite présentées. Le paragraphe (3) consacré à la méthode CART rappelle d'abord la spécification de ce type de modèle avant de présenter le critère optimisé dans le processus de division des nœuds. La modélisation de l'écart temporel est présentée en mettant en évidence les variables explicatives ayant le plus contribué au partitionnement des données. Le paragraphe (4) expose les principes fondateurs du modèle des forêts aléatoires et la modélisation proprement dite par cette méthode est présentée au paragraphe (5). Le paragraphe (6) est dédié à la présentation des résultats. Les pouvoirs prédictifs des modèles sont présentés en fonction de l'horizon de prévision, une quantification du surapprentissage du modèle des forêts aléatoires est ensuite présentée. Une conclusion clôt ce travail.

## 2 Méthodologie et données utilisées

Nous proposons dans un premier temps un modèle fondé sur la méthode CART pour la prévision de l'écart temporel. Ensuite, nous proposons un second modèle fondé sur les forêts aléatoires. En utilisant comme indicateur, le coefficient de Theil (1958), nous comparons ces deux modèles selon leurs performances du point de vue apprentissage des données. Leur pouvoir prédictif est ensuite évalué sur les nouvelles données test et nous en tirons parti pour finalement montrer le risque de surapprentissage de la technique des forêts aléatoires lorsque l'horizon de prévision croît.

Les données utilisées dans cette étude résultent du couplage de données du trafic aérien fournies par Eurocontrol et de données météorologiques et atmosphériques collectées par Météo-France sur la même zone de l'espace aérien. Elles couvrent le mois de septembre 2007. Nous avons extrait de cette masse de données les caractéristiques pertinentes susceptibles de contribuer à la prévision au travers de la modélisation statistique le temps de passage des avions en des points de leur trajectoire, donc de prévoir l'écart temporel. Notre étude porte sur un échantillon de 25000 observations avec 22 variables explicatives. Les avions considérés ont été observés aux intants courants $t_0$ égaux à $8$, $11$, $14$ et $17$ heures.

## 3 Méthode CART

La méthode de régression par CART (Classification And Regression Trees) a été développée par Breiman et al. (1984). Cette méthode qui fait partie de l'ensemble des techniques largement citées dans la littérature de l'apprentissage automatique est inspirée des méthodes de partitionnement récursif initialement initiées par Messenger et Mandell (1972) et Morgan et Messenger (1973) à la suite des travaux de Morgan et Sonquist (1963) et Sonquist et Morgan (1964) portant sur les arbres de décision. La méthode CART a été étendue au traitement des données censurées avec les travaux de Davis et Anderson (1989), Olshen et al. (1990) et de Leblanc et Crowley (1992). En partant de CART, Gelfand et al. (1991) ont proposé une méthode itérative qui construit et élague l'arbre alternativement.

La popularité de la méthode de régression par arbre est en grande partie due aux avantages qu'elle présente et dont les principaux sont :

- Absence ou tout au moins des faibles hypothèses dans la mise en place des modèles par rapport aux méthodes classiques. Par exemple, aucune hypothèse n'est exigée sur la distribution de probabilité, ni pour la variable à expliquer, ni pour les variables explicatives.
- Elle est particulièrement adaptée au cas où les variables explicatives sont nombreuses. Ces variables peuvent être soit qualitatives ou soit quantitatives.
- Elle permet la sélection des variables les plus informatives avec la prise en compte des interactions entre les variables explicatives.
- Grâce aux divisions suppléantes définies comme les divisions les plus semblables à la meilleure division retenue, cette méthode gère l'existence de données manquantes.

En revanche, le principal inconvénient de cette méthode est l'instabilité des arbres de régression. En effet, une légère modification des entrées, c'est-à-dire des mesures des variables explicatives peut conduire aux arbres avec des structures très différentes Breiman (1996b), Ghattas (1999)

Afin de palier cet inconvénient des nouvelles approches innovantes sont apparues et tentent d'apporter des améliorations relatives à la stabilité des arbres. Il s'agit notamment du $Bagging$ Breiman (1996a), du $Boosting$ Freund et Schapire (1997); Schapire (2002) et les forêts aléatoires Breiman (2001). Toutes ces méthodes ont pour principe commun de construire plusieurs arbres de décision ou de régression à partir des données d'apprentissage et de les agréger ensuite. Dans le contexte d'arbre de régression, le modèle agrégé s'obtient par le calcul de la moyenne des prévisions obtenues sur les différents échantillons bootstrap. L'application de ces méthodes par Quinlan (1996) montre une amélioration considérable des performances des modèles agrégés en terme de stabilité et de précision. Le modèle des forêts aléatoires est une variante du bagging avec une deuxième randomisation sur l'ensemble des variables explicatives. Une partie de son efficacité est due au fait que les éventuelles valeurs extrêmes de l'échantillon initial ne se retrouvent que dans certains échantillons bootstrap et que la moyenne des estimations bootstrap fait perdre ou tout au moins diminue les effets de ces valeurs extrêmes dans le modèle agrégé, Tuffery (2006). L'étape de randomisation des variables explicatives est introduite pour rendre les arbres de régression plus indépendants des données d'apprentissage.

Dans le cadre de cette étude nous utilisons la version originale de la méthode CART pour disposer d'un benchmark de la prévision de l'écart temporel. Ensuite, le modèle des forêts aléatoires est utilisé et nous illustrons le gain de performance induit par ce dernier.

## 3.1 Spécification du modèle CART

Dans son principe, la méthode CART construit une partition de l'ensemble des observations à l'aide d'un arbre binaire en minimisant la somme des carrés des erreurs intra-classes. Les arbres de régression divisent l'espace des variables explicatives en un ensemble d'hypercubes. Le modèle de régression par arbre est un modèle simple et non paramétrique où une constante est ajustée sur chaque hypercube. Ce modèle est spécifié par la fonction de régression suivante :

$$M(X) = \sum_{i=1}^{n_0} c_i * \mathbb{1}_{\{X \in K_i\}}$$

Les $c_i$ sont des constantes et les $K_i$ constituent une partition de l'ensemble des individus de l'échantillon. Il s'agit dans ce modèle d'estimer la constante $c_i$ pour chaque classe $K_i$, de déterminer de façon optimale le nombre $n_0$ des classes $K_i$ de cette partition. Dans cette étude, $X$ est le vecteur de variables explicatives. C'est notamment l'ensemble des caractéristiques connues ou prévues des vols. Une fois les $K_i$ déterminés, le meilleur estimateur au sens des moindres carrés des $c_i$ est donné par la moyenne de la variable à expliquer dans la classe $K_i$. L'ouvrage collectif de Celeux et Nakache (1994) rappelle un ensemble de fondements théoriques de la méthode CART.

Parmi les critères utilisés pour sélectionner la meilleure division, Breiman et al. (1984) recommandent l'utilisation des critères de la minimisation de la « *déviance* » lorsque la variable d'intérêt est continue. C'est ce critère[3] qui est utilisé dans cette étude pour la fonction de déviance $\mathbf{D}(X_j, s)$ définie par l'équation suivante :

$$\mathbf{D}(X_j, s) = \min_{c_1, c_2} \left( \sum_{(x_i, y_i) \in a_G(X_j, s)} (y_i - c_1)^2 + \sum_{(x_i, y_i) \in a_D(X_j, s)} (y_i - c_2)^2 \right),$$

où $a_G(X_j, s)$ et $a_D(X_j, s)$ sont respectivement les nœuds gauche et droite obtenus de la division du nœud parent $a$ par la variable explicative $X_j$ au seuil $s$. Les estimateurs de $c_1$ et $c_2$ que nous notons $\widehat{c}_1$ et $\widehat{c}_2$ sont donnés par les relations suivantes :

$$\widehat{c}_1 = \frac{1}{\mathrm{Card}(a_G(X_j, s))} \sum_{(x_i, y_i) \in a_G(X_j, s)} y_i,$$

$$\widehat{c}_2 = \frac{1}{\mathrm{Card}(a_D(X_j, s))} \sum_{(x_i, y_i) \in a_D(X_j, s)} y_i.$$

Ce sont les estimateurs empiriques de la moyenne de la variable dépendante $Y$ à l'intérieur des nœuds descendants gauche et droit respectivement.
La meilleure division notée $d^* = (X_j^*, s^*)$ qui minimise la fonction déviance $\mathbf{D}(X_j, s)$ est donnée par :

$$d^* = \arg \min_{(X_j, s) \in \mathscr{D}} \mathbf{D}(X_j, s)$$

où $\mathscr{D}$ est l'ensemble des divisions admissibles du nœud parent $a$. $X_j^*$ est la variable explicative sur la quelle la division optimale est obtenue avec le seuil $s^*$. On obtient un résultat équivalent en cherchant la division maximisant l'indice de réduction « *d'impureté* » Ghattas (1999). Par ailleurs, un nœud est déclaré terminal s'il est pur ou si le nombre d'observations qu'il contient est inférieur à un nombre minimal fixé.

## 3.2 Modélisation par CART

En appliquant la procédure CART sur les données de trafic aérien, nous observons qu'il y a seulement 5 variables explicatives qui ont contribué activement au partitionnement de l'ensemble des observations. Ces variables actives sont celles qui apparaissent visuellement sur l'arbre de régression, voir la figure 1. Il s'agit de :

[3] Il s'agit de l'inertie intra-classes.

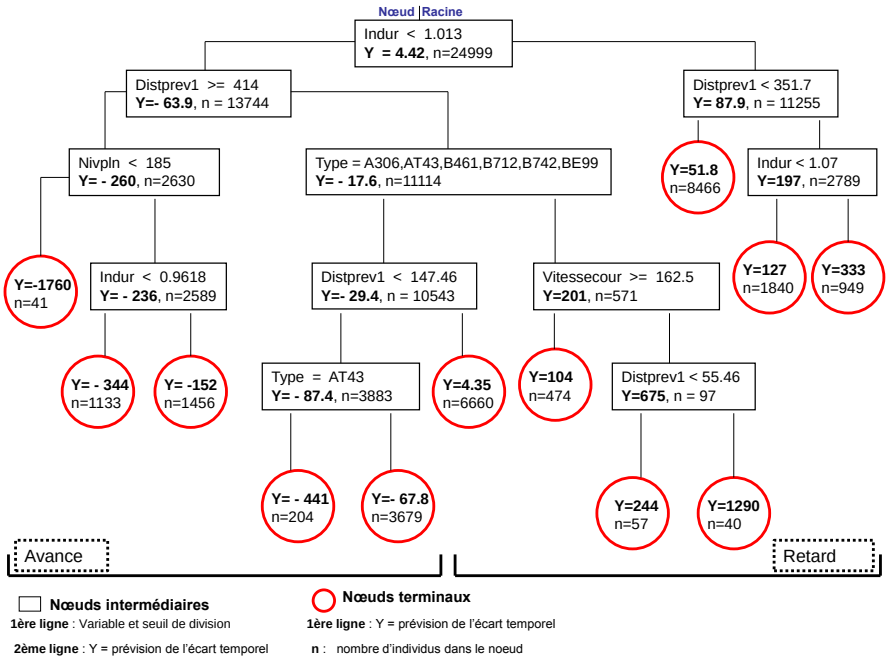


FIG. 1 – *Arbre de régression de la méthode CART pour le modèle de prévision des écarts temporels au passage des avions en des points de leurs trajectoires prévues*

– L'influence du vent prévu sur les trajectoires de vols (*Indur*).
– La distance entre les avions et les points de leurs trajectoires prévues (*Distprev1*).
– Le type d'avion utilisé (*Type*).
– Le niveau de vol prévu dans le plan de vol pour la phase de croisière (*Nivpln*).
– Enfin, la vitesse du vol à l'instant courant de prévision (*Vitessecour*).

Ces variables mettent en évidence l'impact significatif de trois facteurs importants dans la progression des avions sur leurs trajectoires. Il s'agit des conditions météorologiques représentées par l'influence du vent, des paramètres courants des vols représentés par la distance de vol prévue et la vitesse courante, enfin, des caractéristiques des vols définies dans les plans de vols et représentées par le type d'avion et le niveau de vol de croisière. Nous proposons ci-dessous une analyse des effets de ces variables explicatives sur l'écart temporel.

Rappelons qu'une valeur négative de l'écart temporel $\triangle\tau$ s'interprète comme un avion qui passe sur un point de sa trajectoire plus tôt par rapport à l'instant de passage prévu dans le plan de vol (côté gauche de l'arbre, figure 1). Inversement, une valeur positive de l'écart temporel s'interprète comme un retard au passage de l'avion au point de sa trajectoire prévue relativement à l'instant de passage prévu (côté droite de l'arbre, figure 1). Aussi, le vent est de face si l'indicateur *Indur* est supérieur à 1 et de dos sinon. Pour un vent globalement nul sur la trajectoire de vol, *Indur* est égale à 1.

## 3.3 Mise en évidence de l'effet du vent sur le respect du plan de vol

Dans la hiérarchie des variables explicatives, l'influence du vent est une variable de premier niveau et montre que les conditions météorologiques jouent un rôle de premier plan dans le processus de prévision de l'incertitude sur le temps de passage des avions sur les points de leur trajectoire de vol prévue.

La première division sur la variable influence du vent *Indur* de seuil $1.013$ semble bien avoir divisé les données en deux ensembles suivant l'importance du vent auquel les vols ont été soumis. L'estimation de l'écart temporel pour les vols de la branche de droite est d'environ $87.9$ secondes. Celle-ci concerne les vols ayant été soumis à un vent de face dominant et les avions impliqués ont tendance à survoler les points de leur trajectoire après les instants prévus dans leurs plans de vols. Ce résultat est cohérent avec les principes de la dynamique des vols dans la mesure où le vent de face oppose à la progression d'un avion sur sa trajectoire une force de résistance qui réduit sa poussée, donc ralentit son mouvement. En revanche, l'estimation de l'écart temporel pour les vols de la branche gauche est d'environ $-63.9$, soit une avance de $63.9$ secondes. Ce sont des vols ayant été soumis à un vent de dos dominant et les avions impliqués ont tendance à survoler les points de leur trajectoire avant les instants prévus dans leurs plans de vols. Une fois de plus, ce résultat est conforme aux principes de la dynamique des vols dans la mesure où un vent de dos dominant tend à augmenter la poussée de l'avion, donc à accélérer son mouvement.

Par ailleurs, la division sur la variable influence du vent de seuil $1.07$ montre que lorsque le vent de face est très fort ($Indur > 1.07$), les aéronefs concernés survolent les points de leur trajectoire avec des retards encore plus élevés. Il s'agit des vols de la branche droite de ce nœud où la valeur moyenne de l'écart temporel se situe autour de $333$ secondes. En revanche, la division sur la variable influence du vent de seuil $0.96$ montre que lorsque le vent de dos est très fort ($Indur < 0.96$), les aéronefs concernés survolent les points de leur trajectoire avec des avances beaucoup plus importantes. Il s'agit des vols de la branche gauche de ce nœud où la valeur moyenne de l'écart temporel se situe autour de $-344$ secondes, soit une avance d'environ $344$ secondes. Ce qui montre qu'un vent très fort qu'il soit de dos ou de face a pour effet d'amplifier le niveau d'incertitude dans la prévision du temps de passage des aéronefs en des points de leur trajectoire de vol et tend à réduire la capacité de gestion de trafic aérien dans son ensemble.

## 3.4 Effets des autres variables sur le respect du plan de vol

On observe de la figure 1 que la deuxième variable active la plus importante est la distance de vol prévue (*Distprev1*). Son interaction avec l'indicateur du vent montre au travers les divisions de seuils respectifs ($Distprev1 \geq 414$) et ($Distprev1 < 351.7$) que l'influence du vent qu'il soit de dos ou de face n'est réellement significative sur la progression d'un avion que si ce dernier est soumis à cette influence de vent sur une distance importante. Par exemple, pour une distance supérieure à $414$ mille nautiques dans les conditions de vent de dos dominant, les avions survolent les points de leur trajectoire avec des avances en moyenne de $260$ secondes contre $17.6$ secondes pour les autres. Pour une distance supérieure à $351.7$ mille nautiques dans les conditions de vent de face dominant, les avions survolent les points de leur trajectoire avec des retards en moyenne $197$ secondes contre $52$ secondes pour les autres.

Le type d'avion et le niveau de vol[4] apparaissent comme des variables actives de troisième niveau. L'interprétation des effets liés à ces paramètres ne semble pas évidente. Par ailleurs, la vitesse de vol à l'instant courant qui est une variable active de quatrième niveau sur l'arbre a des effets conformes aux principes de la dynamique des vols. En effet, au regard de la division sur le nœud $(Vitessecour \geq 162.5)$ où le retard moyen est de $201$ secondes, on observe que ce retard se situe à $104$ secondes pour les avions avec une vitesse à l'instant courant supérieure $162.5$ knots[5]. En revanche, ce retard est six fois plus important pour les avions dont la vitesse de vol à l'instant courante est inférieure à $162.5$ knots.

# 4 Forêts aléatoires

Les forêts aléatoires développées par Breiman (2001) sont une collection d'arbres binaires de décisions $(\mathcal{A}_k)_{(k=1,\ldots,K)}$. Ces arbres $\mathcal{A}_k$ sont construits sur des échantillons bootstrap de l'échantillon d'apprentissage initial. Chaque échantillon bootstrap est obtenu par tirage aléatoire avec remise dans l'échantillon de données de départ. Ces techniques d'apprentissage automatique par agrégation de modèles sont très répandues et présentent quelques spécificités dont les plus importantes sont :

- Dans la procédure de construction des arbres, à chaque nœud, un faible nombre de variables est tiré aléatoirement et la recherche de la division optimale est basée uniquement sur ce sous-ensemble de variables.
- Les arbres construits sur les échantillons bootstrap ne sont pas optimisés, ils ne sont pas élagués, donc maximaux.
- Pour chaque arbre, une partie de l'échantillon est mise de côté. Elle est appelée *Out-of-bag sample* notée OOB. Cette partie de l'échantillon d'apprentissage non utilisée pour la construction de l'arbre sert en effet à l'évaluation de l'importance des variables utilisées.
- L'introduction du tirage aléatoire sur l'ensemble des variables explicatives permet d'éviter l'apparition des mêmes variables.

Selon la règle de décision, il existe deux versions des forêts aléatoires : Le *Random Input* où la règle de décision porte sur une seule des variables explicatives tirées au hasard, et l'autre connu sous *Random Features* qui utilise une combinaison linéaire des variables sélectionnées à chaque nœud, avec des coefficients tirés aussi aléatoirement.

Cette procédure présente un certain nombre d'avantages. Elle nécessite peu de paramètres à régler et utilisent les variables explicatives continues et discrètes pour des problèmes de classification et de régression. Les arbres obtenus ne sont pas instables comme ceux fournis par la méthode CART. La procédure de choix aléatoire des variables explicatives à tester lors de la division en chaque nœud de l'arbre bootstrap permet de donner aux variables importantes cachées dans la méthode CART, des rôles plus actifs dans la construction des arbres individuels issus des échantillons bootstrap. Par ailleurs, deux propriétés essentielles expliquent les performances des forêts aléatoires :

- La bonne performance des arbres individuels qui ont un faible biais mais une forte variance, et la faible corrélation entre les arbres de la forêt. La corrélation entre arbres est définie comme celle de leurs prévisions sur les échantillons tests OOB.

[4]Le niveau de vol est la hauteur en pieds divisée par 100 de l'avion au dessus de l'altitude pression 1013.25 Hpa.
[5]Le knot est l'unité de vitesse en navigation. Un knot est la vitesse d'un mille nautique par heure.

– Le fait qu'un faible nombre de variables soit utilisé à chaque nœud des arbres construits, permet de réduire la complexité algorithmique.

Les forêts aléatoires présentent aussi un certain nombre d'inconvénients : Le temps de calcul est important pour évaluer un nombre suffisant d'arbres jusqu'à ce que l'erreur de prévision sur l'échantillon OOB ou sur un autre échantillon de validation se stabilise. La procédure s'arrête si elle tend à augmenter. Il est nécessaire de stocker tous les modèles de la combinaison afin de pouvoir utiliser cet outil de prévision sur les nouvelles données. L'amélioration de la qualité de prévision se fait au détriment de l'interprétabilité, ainsi le modèle finalement obtenu devient une « *boite noire* ».

Les forêts aléatoires dépendent de trois principaux paramètres :

– Le nombre d'arbres générés à partir des échantillons bootstrap que nous notons *ntree*.
– Le nombre de variables testées dans chaque nœud d'un arbre pour la recherche de la division optimale que nous notons *mtry*.
– Enfin, le nombre minimal d'observations dans un nœud terminal.

Breiman (2001) suggère qu'en classification, le nombre de variables testées pour chaque nœud d'un arbre est égal à $\sqrt{p}$, où $p$ est le nombre initial de variables. Cette valeur proposée par Breiman a été confirmée par d'autres auteurs. Liaw et Wiener (2002), Diaz-Uriarte et de Andrés (2006) ont montré l'optimalité de cette valeur en terme de performance des forêts sur les échantillons OOB. Pour la régression, ce nombre est approximativement $\frac{p}{3}$ (Breiman, 2002). Une forte diminution de ce paramètre réduit les chances que des variables importantes soient sélectionnées dans les arbres individuels, et peut ainsi dégrader les performances des forêts. Les forêts aléatoires fournissent également un moyen original de calcul d'un indice d'importance pour l'hiérarchisation des variables explicatives (Ben et Ghattas, 2008))

# 5 Modélisation par les forêts aléatoires (FA)

La version des forêts aléatoires utilisée ici est le *Random Input* où la règle de décision sur chaque nœud porte sur une seule des variables explicatives tirées de façon aléatoire.

## 5.1 Choix des paramètres

Nous avons fixé le nombre minimal d'observations par feuille à 40[6]. Le couple de paramètres *(ntree,mtry)* est déterminé par une procédure de type validation croisée. Rappelons que *ntree* est le nombre d'échantillons bootstrap pour chaque forêt à construire et *mtry* est le nombre de variables explicatives à tirer aléatoirement sur chaque nœud. Pour *mtry*, nous explorons les valeurs entières comprises entre 2 et $\frac{p}{2}$. En effet, cet intervalle contient la valeur optimale qui est approximativement égale à $\frac{p}{3}$. Pour *ntree*, nous avons fixé une valeur maximale du domaine d'exploration à 200. Ainsi, en posant $y_i = \Delta\tau_i$ et $\widehat{y}_i = \widehat{\Delta\tau}_i$, le couple

[6]Il n'existe pas de règle établie pour déterminer le nombre minimal d'observations par nœud.

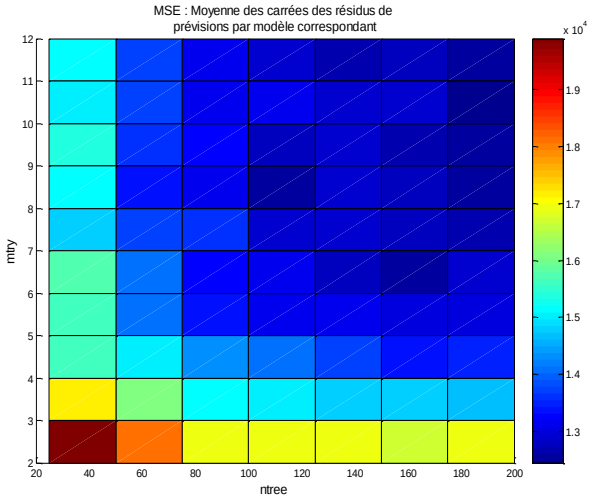


FIG. 2 – *Somme des carrés des erreurs de prévision sur les données d'apprentissage en fonction du nombre d'échantillons bootstrap et du nombre de variables de randomisation sur les nœuds.*

optimal est celui qui minimise donc le critère suivant :

$$\mathcal{E}(ntree, mtry) = \sum_{i=1}^{n} \left( y_i - \widehat{y}_i^{(ntree,mtry)} \right)^2$$

autrement dit :

$$(ntree^*, mtry^*) = \arg \min_{(ntree,mtry)} \mathcal{E}(ntree, mtry) \tag{4}$$

Les valeurs du critère $\mathcal{E}$ en fonction de *ntree* et *mtry* sont représentées dans la figure 2. Un examen de cette figure montre que les performances du modèle de chaque forêt dépend simultanément du nombre d'arbres et du nombre de variables de randomisation. Ainsi, si on privilégie le nombre de variables explicatives au nombre d'échantillonnages bootstrap, le premier couple permettant d'obtenir un bon ajustement des forêts aléatoires est $(120, 9)$ où $120$ est le nombre d'échantillonnages bootstrap et $9$ est le nombre de variables explicatives de randomisation. En revanche, si l'on privilégie le nombre d'échantillonnages bootstrap au nombre de variables de randomisation, le premier couple permettant d'obtenir le meilleur modèle d'ajustement des forêts aléatoires est approximativement $(150, 7)$. C'est ce dernier cas que nous avons retenu pour notre étude. En effet, le nombre de variables de randomisation dans chaque nœud est proche de la valeur préconisée par Breiman (2001), c'est-à-dire proche de la valeur correspondante à la partie entière de la fraction $\frac{22}{3}$. En effet, nous avons $22$ variables explicatives. A partir de ce couple $(150, 7)$, nous avons obtenu un modèle des forêts aléatoires avec un coefficient de détermination $R^2 = 84.52\%$. La figure 3 montre qu'au delà du nombre d'échantillons bootstrap *ntree*=150, le gain de performance en terme de réduction de l'erreur de prévision est négligeable.

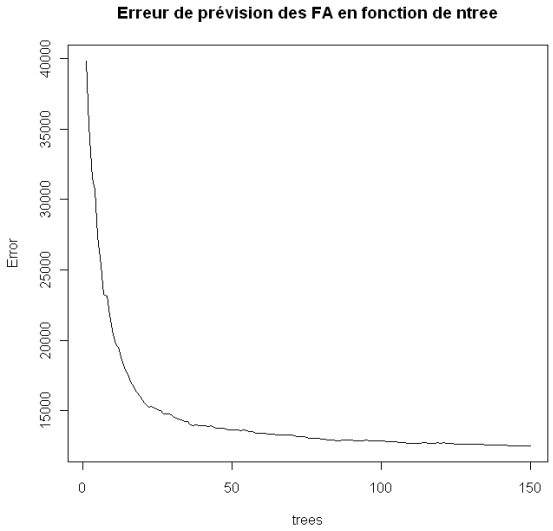


FIG. 3 – *Somme des carrés des erreurs de prévision en fonction du nombre d'échantillons bootstrap (FA=forêts aléatoires)*

## 5.2 Hiérarchie des variables explicatives

Les forêts aléatoires sont un modèle construit par agrégation de plusieurs arbres de régression, donc il n'y a pas d'interprétation directe. Néanmoins, des informations pertinentes peuvent être obtenues par le calcul et la représentation graphique d'indices proportionnels à l'importance de chaque variable dans le modèle et donc de sa participation à la régression. Cela est d'autant plus utile que les variables sont très nombreuses. Plusieurs critères sont ainsi proposés pour évaluer l'importance de la $j^{ème}$ variable explicative.

Concernant les modèles non paramétriques, il n'existe que très peu d'outils permettant d'établir une hiérarchie des variables. Notons cependant que les arbres de régression de la méthode CART et les forêts aléatoires (Breiman, 2001) offrent la possibilité d'établir une hiérarchie des variables explicatives. Dans le cadre des modèles récents de type Séparateurs à Vastes Marges (SVM), Barnhill et al. (2002) et Rakotomamonjy (2003) ont proposé des scores pour chaque variable explicative utilisée. Ces scores permettent d'établir une hiérarchie des variables. En comparant différentes méthodes de sélection des variables sur données réelles et simulées, Ben et Ghattas (2008) montrent que les forêts aléatoires fournissent une hiérarchie plus stable des variables que les autres méthodes. Nous synthétisons dans la figure 4, la hiérarchie des variables explicatives de la variable dépendante *écart temporel* obtenue par le modèle des forêts aléatoires. L'axe des abscisses représente les variables explicatives et l'axe des ordonnées représente l'importance des variables.

Cette hiérarchie a été réalisée en optimisant le critère du *Gain de Pureté du Nœud* (GPN) défini comme suit :

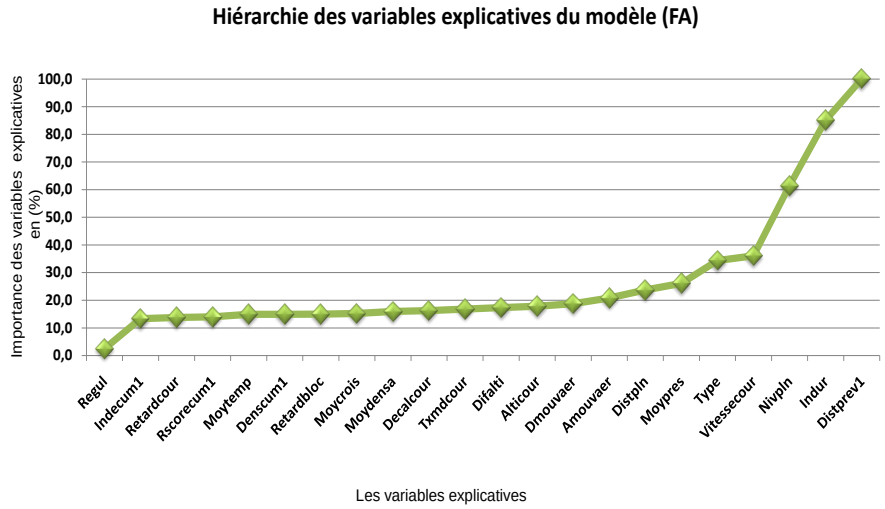


FIG. 4 – *Importance des variables explicatives du modèle des forêts aléatoires. Le modèle est construit à partir de 150 échantillons bootstrap de l'échantillon d'apprentissage initial avec 7 tirages aléatoires des variables explicatives sur les nœuds.*

$$GPN(X_j) = \left| \sum_{q=1}^{ntree} \left[ (y_q - \widehat{y}_q(X_j))^2 \right] - \sum_{q=1}^{ntree} \left[ \left( y_q - \widehat{y}_q^{'}(X_j) \right)^2 \right] \right| \tag{5}$$

où :

- $y_q$ est l'estimation de la variable dépendante $Y$ sur le $q^{ème}$ échantillon bootstrap ;
- $\widehat{y}_q(X_j)$ est la prévision de la variable $Y$ par l'arbre du $q^{ème}$ échantillon bootstrap avec les variables $(X_1, \ldots, X_p)$ et sans permutation de $X_j$ ;
- $\widehat{y}_q^{'}(X_j)$ est la prévision de la variable $Y$ par l'arbre du $q^{ème}$ échantillon bootstrap avec les variables $(X_1, \ldots, X_p)$ après permutation des valeurs de la variable $X_j$.

L'importance des variables explicatives a été rapportée à l'échelle [0 ;100] où chaque valeur du GPN a été divisée par la valeur la plus élevée multipliée par 100. Rappelons que le critère optimisé lors de la construction de l'arbre de chaque échantillon bootstrap est la minimisation de la fonction de déviance présentée précédemment.

Les cinq premières variables les mieux classées de la hiérarchie sont exactement celles ayant été actives lors de la construction de l'arbre de régression du modèle CART. Il s'agit notamment : la distance de vol prévue *(Distprev1)*, l'indicateur d'influence du vent prévu sur la trajectoire de vol *(Indur)*, le niveau de vol prévu pour la phase de croisière *(Nivpln)*, la vitesse du vol au point courant *(Vitessecour)* et le type d'avion utilisé pour le vol *(Type)*. Néanmoins, nous notons que relativement à la hiérarchie obtenue par le modèle CART, l'ordre d'importance de l'indicateur d'influence du vent et celui de la distance de vol prévue sont permutés dans

la nouvelle hiérarchie produite par les forêts aléatoires. La nature des sorties de ce dernier modèle ne nous permet pas d'interpréter d'avantage. Nous présentons dans la section suivante l'ensemble des résultats obtenus et comparons les pouvoirs prédictifs des deux modèles en fonction de l'horizon temporel de prévision.

# 6 Résultats

## 6.1 Horizon de prévision et modèles d'ajustement

Pour que les modèles présentés dans cet article soient utilisés opérationnellement, ils doivent répondre aux conditions de prédictibilité souhaitées par les autorités de régulation et de gestion de trafic aérien. Ainsi, s'impose la nécessité d'étudier les performances de ces modèles dans la prévision de trafic en fonction de l'horizon de prévision à erreur bornée dans lequel les modèles sont appelés à être utilisés. Pour cela, avant la présentation des résultats, nous nous intéressons d'abord à la définition de ce type d'horizon de prévision et des conditions de sa prise en compte.

Un horizon de prévision à erreur bornée est un intervalle de temps pendant lequel il est possible d'utiliser un modèle pour calculer les prévisions avec un niveau d'erreur borné. Il est compris entre l'instant courant (ici, ramené à $0$) et une limite supérieure dépendante de l'incertitude du modèle utilisé. Pour évaluer l'amplitude de l'incertitude des modèles en fonction de l'horizon de prévision, nous avons considéré une fenêtre de temps de longueur inférieure à $60$ minutes ($3600$ secondes) que nous avons ensuite divisé en $12$ petits intervalles de $5$ minutes ($300$ secondes) chacun. Le $13^{ème}$ intervalle contient les points des trajectoires de vols pour lesquels l'horizon de prévision est égal à $60$ minutes ou plus. Sur les représentations graphiques (figure 5), $00-03$ indique le petit intervalle de limites $0$ et $300$ secondes et $36et+$ désigne $3600$ secondes ou plus. Sur chacun des intervalles, le *boxplot* mesure la dispersion des erreurs d'ajustement du modèle correspondant et permet ainsi d'évaluer l'amplitude de l'incertitude de prévisions au fur et à mesure que l'horizon de prévision augmente. Ainsi, une lecture de la figure 5 montre que l'incertitude de prévision sur les données d'apprentissage par le modèle des forêts aléatoires est beaucoup plus faible que celle obtenue du modèle CART. Pour une quantification du différentiel d'ajustement entre ces deux modèles, nous avons utilisé le coefficient de Theil[7] 1958 défini par l'équation suivante :

$$Theil = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i^{ref})^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i^{alt})^2}{n}}}.$$

C'est le rapport du RMSE (Root Mean Squared Error) d'un modèle de référence sur celui des erreurs de prévision de la méthode alternative. Nous considérons la méthode CART comme modèle de référence tandis que les forêts aléatoires sont le modèle alternatif. L'indicateur de Theil ainsi calculé est approximativement égal à $2.18$ supérieur à $1$. Donc relativement à cet indicateur, le modèle des forêts aléatoires a une capacité d'ajustement sur les données d'ap-

[7] Si le coefficient de Theil est égal à un, les prévisions du modèle de référence *(ref)* ne sont pas meilleures que celles données par le modèle alternatif *(alt)*. Si le coefficient de Theil est supérieur à un, les prévisions du modèle alternatif sont bien meilleures que celles du modèle de référence.

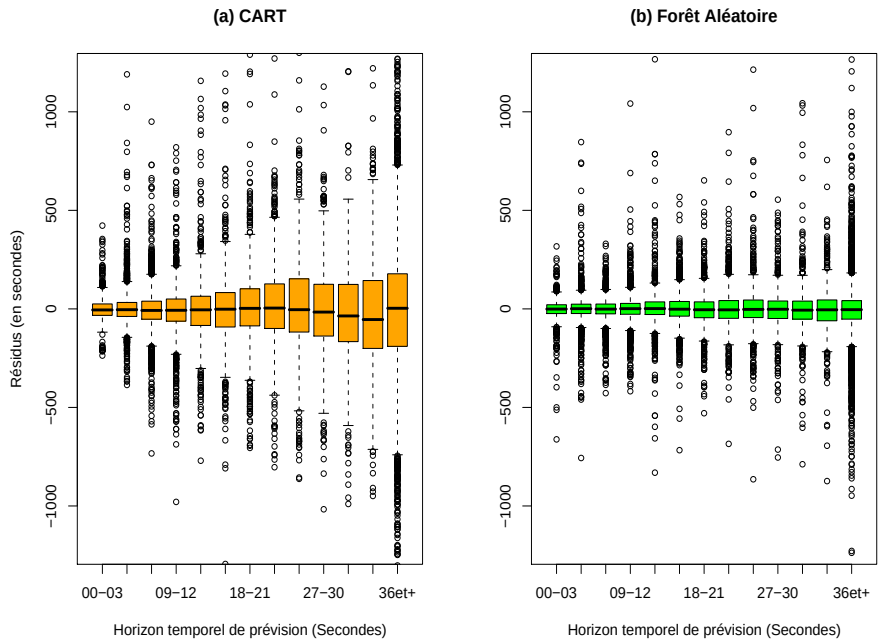


FIG. 5 – *Dispersion des résidus des modèles en fonction de l'horizon temporel de prévision*

prentissage au moins deux fois supérieure à celle du modèle CART.

## 6.2 Pouvoir prédictif des modèles sur données tests

La performance d'un modèle issu d'une méthode d'apprentissage s'évalue par sa *capacité de prévision* ou *capacité de généralisation* en situation nouvelle. Trois types de stratégies sont souvent utilisés pour la mesure de cette performance :

- Un partage de données en un échantillon d'apprentissage et un échantillon test afin de distinguer l'estimation du modèle et les estimations de l'erreur de prévision.
- Une pénalisation de l'erreur d'ajustement faisant intervenir la complexité du modèle.
- Un usage intensif du calcul *(computational statistics)* par la mise en œuvre de simulations.

En général, le choix de la meilleure stratégie dépend de plusieurs facteurs dont la taille de l'échantillon initial, la complexité du modèle envisagé, la variance de l'erreur, la complexité des algorithmes liée au volume de calcul admissible. Dans le cadre de cet article, nous avons l'avantage de disposer d'une base importante de données. Ainsi, la première stratégie semble la mieux adaptée et nous l'utiliserons ici avec un seul échantillon test. Rappelons que l'estimation de la qualité de prévision est un élément central de la mise en place de la stratégie de modélisation par apprentissage automatique. Le point important étant que le meilleur modèle au sens du pouvoir prédictif n'est pas nécessairement celui qui ajuste le mieux les données

d'apprentissage (Besse, 2009). En effet, une diminution importante de l'erreur sur les données d'ajustement peut masquer une situation de surapprentissage.

Lors de la comparaison des modèles d'ajustement, nous avons calculé le coefficient de Theil à partir de la moyenne des carrés des erreurs de prévisions. L'estimateur moyen étant sensible à des valeurs extrêmes, une valeur très grande ou très petite même isolée a une influence considérable dans la moyenne. En revanche, la médiane est connue pour être robuste aux valeurs extrêmes, elle est donc un indicateur plus pertinent et permet par ailleurs d'évaluer la situation de la majorité des observations dans l'échantillon. A cette étape d'évaluation des modèles sur les données de l'échantillon de test, nous comparons les modèles entre eux en utilisant l'indicateur médian noté RMEDSEP (Root Median Square Error of Prediction). Cet indicateur est égal à la racine carrée de la médiane des carrés des erreurs de prévisions sur les données test au lieu du RMSEP (Root Mean Square Error of Prediction). On définit de façon similaire l'indicateur médian calculé sur les données d'apprentissage, noté RMEDSE (Root Median Square Error). La figure 6 synthétise l'évolution de ces indicateurs calculés pour les modèles CART et des forêts aléatoires en fonction de l'horizon temporel de prévision. Les données de l'échantillon de test portent sur 7388 observations et concernent les vols observés aux instants courants $t_0$ égaux à 7 et 16 heures.

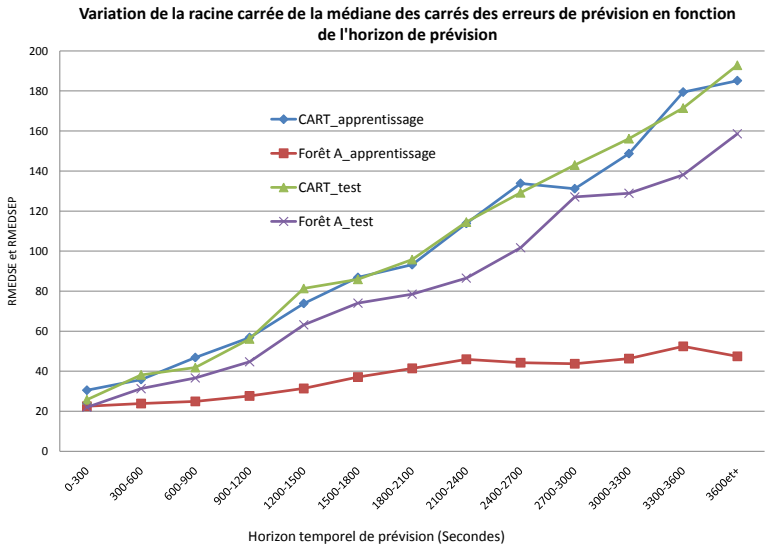


FIG. 6 – *Comparaison des performances des modèles en fonction de l'horizon temporel de prévision. RMEDSE évalue le modèle sur les données d'apprentissage tandis que RMEDSEP l'évalue sur un échantillon test*

A la lumière de cette figure, nous retenons les résultats suivants :

– Le point commun à ces deux modèles : Lorsque l'horizon de prévision est proche de l'instant courant la qualité de prévision des instants de passage des avions en des points de leur trajectoire de vol est meilleure. En revanche, on observe une forte détérioration de la qualité de prévision au fur et à mesure que l'horizon de prévision augmente en

s'éloignant de l'origine. Rappelons qu'un horizon est proche de l'origine ou de l'instant courant de prévision lorsqu'il est inférieur à 40 minutes (2400 secondes).

– Les courbes *ForêtA_apprentissage* et *ForêtA_test* montrent qu'en situation réelle du trafic, le modèle des forêts aléatoires fournit des prévisions dont les erreurs ont une amplitude très élevée par rapport à celles obtenues par ce modèle sur les données d'apprentissage. On observe alors un surapprentissage du modèle des forêts aléatoires qui semble dépendre fortement de l'horizon de prévision.

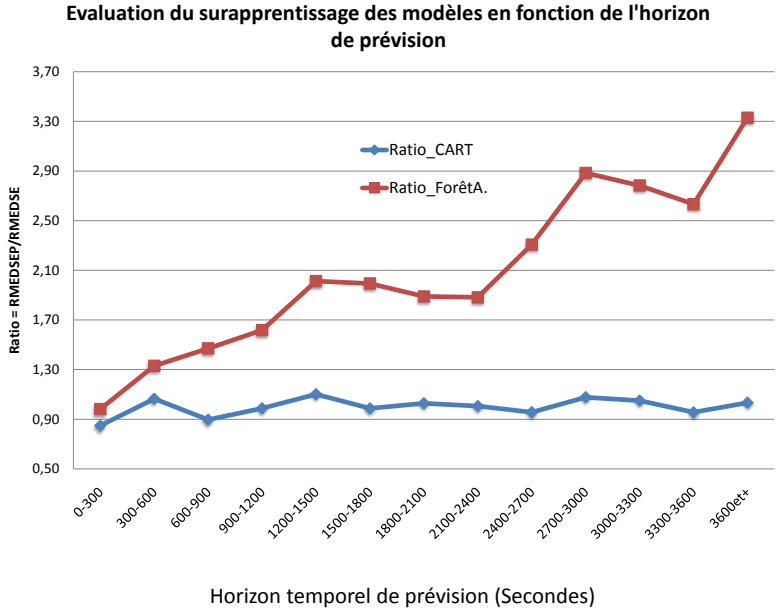


FIG. 7 – *Evaluation du surapprentissage des modèles. Le ratio Ratio_CART mesure le surapprentissage pour le modèle CART en fonction de l'horizon temporel de prévision et Ratio_ForêtA. mesure celui du modèle des forêts aléatoires*

Afin de quantifier l'ampleur du surapprentissage des modèles, nous avons introduit un indicateur du type coefficient de Theil, mais calculé à partir des médianes. Pour chaque modèle, cet indicateur est le ratio du RMEDSEP (Root Median Square Error of Prediction) sur RMEDSE (Root Median Square Error). Ainsi, une valeur élevée de ce ratio traduit l'importance du surapprentissage du modèle ajusté par rapport aux prévisions qu'il fournit en situation nouvelle. A la lumière de la figure 7, nous pouvons en tirer les enseignements suivants :

– Le profile quasi stable autour de la valeur 1 de la courbe *Ratio_CART* montre qu'en situation nouvelle, le modèle obtenu de la méthode CART fournit des prévisions avec des erreurs presque au même niveau que des erreurs d'ajustement sur données d'apprentissage. Ce profile semble ne pas varier en fonction de l'horizon du temps.
– En revanche, à la lumière de la courbe *Ratio_ForêtA.*, il apparaît qu'en situation réelle de trafic, l'utilisation du modèle des forêts aléatoires fournit des erreurs de prévisions très élevées relativement à celles obtenues directement de son modèle d'ajustement. La

figure 7 montre que lorsque l'horizon de prévision se situe autour de 60 minutes, l'erreur de prévision sur données de l'échantillon de test par le modèle des forêts aléatoires peut atteindre 3 fois celle obtenue sur les données d'apprentissage.

Pour une meilleure prévision de l'écart temporel, donc, des instants de passage des avions sur les points de leur trajectoire prévue, le modèle construit par les forêts aléatoires est le plus adapté.

# 7 Conclusion

A travers cet article, les méthodes d'apprentissage automatique apparaissent comme un outil efficace de prévision des instants de passage des avions sur les points de leur trajectoire de vol prévue. Nous avons montré que si le modèle des forêts aléatoires fournit des meilleures prévisions relativement au modèle benchmark CART, son utilisation opérationnelle dans la prévision du trafic aérien nécessite en revanche, un certain niveau de circonspection. En effet, les erreurs de prévisions en situation réelle peuvent très vite atteindre des proportions très importantes (trois fois le niveau des erreurs d'ajustement dans notre cas). Le pouvoir prédictif des forêts aléatoires est toutefois supérieur à celui du modèle CART. Le caractère croissant de l'écart entre les prévisions sur les données test et celles des données d'apprentissage est alors un signal d'alerte sur le recul nécessaire lors de l'utilisation de ce modèle dans le cadre de la prévision des phénomènes dynamiques.

Dans cette étude, nous avons également mis en évidence les effets significatifs des interactions entre l'indicateur d'influence du vent et de la distance de vol sur les avions. Ainsi, le vent qu'il soit de « *dos* » ou de « *face* » n'a d'influence significative sur la progression d'un avion que si ce dernier est soumis dans ce vent sur une distance importante de sa trajectoire.

En fonction du phénomène à prévoir et du niveau d'erreur fixé, une étape complémentaire est indispensable à cette étude. Cette phase vise à compléter les modèles développés en recherchant un horizon temporel optimal pour leur utilisabilité. Dans ce cadre, le modèle des forêts aléatoires développé ici sera comparé aux autres types de modèle non-paramètres tels que, les SVM (Séparateurs à Vates Marges) et le modèle MARS (Multivariate Adaptive Regression by Splines). Le surapprentissage apparent observé sur le modèle des forêts aléatoires sera à nouveau examiné mais cette fois en prenant en compte le nombre d'arbres dans la forêt. Pour l'ensemble de ces modèles, un intervalle de prédictibilité du trafic sera déterminé afin de garantir les prévisions de meilleure qualité.

# Références

Barnhill, S., I. Guyon, J. Weston, et V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning 46(3)*, 389–422.

Ben, A. I. et B. Ghattas (April 2008). Sélection de variables pour la classification binaire en grande dimension : Comparaisons et application aux données de biopuces. *Journal de la Société Française de Statistique 145,3*.

Besse, P. (Juillet 2009). *Apprentissage Statistique et Data mining*. Institut de Mathématique de Toulouse : Laboratoire de Statistique et Probabilités UMR CNRS C5583.

Breiman, L. (1996a). Bagging predictors. *Machine Learning 24*, 2350–2383.

Breiman, L. (1996b). Heuristic of instability and stabilization in model selection. *Annals of Statistics 24(6)*, 2350–2383.

Breiman, L. (2001). Random forests. *Machine Learning 45(1)*, 5–32.

Breiman, L. (2002). Manual on setting up, using, and understanding random forests v3.1. *(http://oz.berkeley.edu/users/breiman/Using randomforests V3.1.pdf)*.

Breiman, L., J. H. Friedman, R. A. Olshen, et C. J. Stone (1984). *Classification And Regression Trees*. California: Wadsworth International: Chapman and Hall.

Celeux, G. et J.-P. Nakache (1994). *Analyse discriminante sur variables qualitatives*. Polytechnica edition.

Davis, R. et J. Anderson (1989). Exponential survival trees. *J. Am. Statist. Assoc. 8*, 947–961.

Diaz-Uriarte, R. et S. A. de Andrés (2006). Gene selection and classification of microarray data using random forest. *BMC Bioinformatics 7:3*, 1–13.

Eurocontrol (2009). Project bada eec technical/scientific report 2009/003. Technical report, Eurocontrol.

Freund, Y. et R. Schapire (1997). A decision-theoric generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences 55(1)*, 119–139.

Gelfand, S., C. Ravishankar, et J. Edward (1991). An iterative growing and pruning algorithm for classification tree design. *IEEE transactions on pattern analysis and machine intelligence 13*, 163–174.

Ghattas, B. (1999). Importance des variables dans les méthodes CART. *GREQAM - Université de la Méditerranée*.

Leblanc, M. et J. Crowley (1992). Relative risk trees for censored survival data. biometrics. *Biometrics 48*, 411–425.

Liaw, A. et M. Wiener (2002). Classification and regression by random forest. *Rnews 2*, 18–22.

Messenger, R. et M. Mandell (1972). A model search technique for predictive nominal scale multivariate analysis. *J. Am. Statist. Assoc. 67*, 768–772.

Morgan, J. et J. Sonquist (1963). Problems in the analysis of survey data and a proposal. *J. Am. Statist. Assoc 58*, 415–434.

Morgan, J. N. et R. C. Messenger (1973). Thaid: A sequential serach program for the analysis of nominal scale dependant variables. *Ann Arbor : Institute of social research, university of Michighan*.

Olshen, R., S. Horning, L. Kwak, et J. Halpern (1990). Prognostic significance of actual dose intensity in diffuse large-cell lymphoma : results of a tree-structured survival analysis. *J. of Clinical Oncology 8*, 963–977.

Quinlan, R. (1996). Bagging, boosting and c4.5. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence 13*, 725–730.

Rakotomamonjy, A. (2003). Variable selection using svm-based criteria. *Journal of Machine Learning Research 3*, 1357–1370.

Schapire, R. (2002). The boosting approach to machine learning : An overview. *In MSRI Workshop on Nonlinear Estimation and Classification 55(1)*.

Sonquist, J. et J. Morgan (1964). The detection of interaction effects. *Ann Arbor : Institute for social research, University of Michigan*.

Theil, H. (1958). *Economic Forecasts and Policy*. Amsterdam: North Holland.

Tuffery, S. (April 2006). Data mining et statistique décisionnelle, l'intelligence des données. *Short Book Review, technip edition 26*, Aof 978–2–7108–0888–6.

# Annexe : Les variables explicatives des modèles : TAB 1 et TAB 2

## Summary

The high air traffic demand growth ($5\%$ year) in the european airspace shows that air traffic control will face more and more aircrafts. Hence the uncertainty prediction on aircraft trajectories's become of paramount importance. The goal of our study is twofold: ($i$) First, bsaed on ait traffic operational data we show that the automatic learning based methods as CART and random forests are suitable tools to predict accurately the crossing time on points of the aircraft trajectories. We quantify the predictive effectiveness of both models based CART method and the random forests. These models are assessed using test sample. ($ii$) Second we highlight that, in some case the model based on the random forets can lead to a strong overfitting of data.

| Noms des variables | Description des variables |
|---|---|
| | ***Paramètres prévus dans les plans de vol*** |
| Depart | Aéroport de départ du vol (qualitative). |
| Arriv | Aéroport de destination du vol (qualitative). |
| Droulage | La durée moyenne du roulage dans l'aéroport de départ du vol (secondes). |
| Aroulage | La durée moyenne du roulage dans l'aéroport d'arrivée du vol (secondes). |
| Dmouvaer | Le nombre moyen de mouvements par heure, dans l'aéroport de départ du vol. |
| Amouvaer | Le nombre moyen de mouvements par heure, dans l'aéroport d'arrivée du vol. |
| Type | Le type d'aéronef utilisé par le vol (qualitative). |
| Distpln | La distance de vol prévue dans le plan de vol (NM). |
| Nivpln | Le niveau de vol de croisière prévu dans le plan de vol(FL ; flight level). |
| Regul | Nombre de régulations auxquelles le vol a été soumis avant son départ du bloc de stationnement. |
| Exemdist | Le vol est un long courrier ou un court courrier (qualitative). |
| Retardbloc | Le retard dont l'aéronef a fait l'objet avant son départ du bloc (secondes). |
| | ***Paramètres courants des vols*** |
| Heurecour | L'instant courant du vol (qualitatif). |
| Retardcour | Le retard du vol au point courant (secondes). |
| Txmdcour | Le taux de montée ou de descente de l'aéronef au point courant du vol à (l'instant courant pieds/minutes). |
| Vitessecour | La vitesse de l'aéronef au point courant du vol à l'instant courant (kts). |
| Difalti | La différence d'altitudes entre le point courant du vol à l'instant courant et le point de la trajectoire pévue en lequel on veut prévoir le temps de passage du vol (FL ; pieds divisé par 100). |
| Distprev1 | La distance en projection entre le point courant du vol et un point de sa trajectoire prévue en lequel on veut prévoir l'instant de passage de l'aéronef (NM). |
| Decalcour | L'écart de distance du point courant du vol par rapport à sa trajectoire prévue (NM). |

TAB. 1 – *Les variables explicatives utilisées dans les modèles*

| Noms des variables | Description des variables |
|---|---|
| | ***Variables de complexité et d'infrastructure du trafic***. |
| Secteur | C'est la variable désignant le secteur de l'espace traversé par les trajectoires des vols prévues (qualitatif). |
| Moycrois | Le nombre moyen de routes connectées sur chaque balise de la trajectoire de vol prévue. |
| Denscum1 | Le flux de trafic sur la trajectoire prévue. |
| Indecum1 | La somme cumulée de toutes les interactions potentielles prévues sur la trajectoire du vol (Nombre de vols par NM). |
| Rscorecum1 | Le score des interactions sur la trajectoire de vol prévue. |
| | ***Paramètres météorologiques et atmosphériques*** |
| Indur | Influence du vent sur le vol (sans unité). |
| Moytemp | L'écart moyen entre la température prévue et la température de l'atmosphère standard sur la trajectoire prévue du vol (degré Celsuis). |
| Moytpres | L'écart moyen entre la pression prévue et la pression de l'atmosphère standard sur la trajectoire prévue du vol (hPa). |
| Moydensa | L'écart moyen entre la densité de l'air prévue et la densité de l'air de l'atmosphère standard sur la trajectoire prévue du vol $(kg/m^3)$. |

TAB. 2 – *Les variables explicatives utilisées dans les modèles (suite)*