HAL
open science

Modélisation de l'incertitude sur les trajectoires d'avions

Norbert Fouemkeu

To cite this version:

Norbert Fouemkeu. Modélisation de l'incertitude sur les trajectoires d'avions. Mathématiques générales [math.GM]. Université Claude Bernard - Lyon I, 2010. Français. NNT : 2010LYO10217 . tel-00710595

HAL Id: tel-00710595

https://theses.hal.science/tel-00710595v1

Submitted on 21 Jun 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N° d'ordre 217 - 2010 Année 2010

THÈSE DE L'UNIVERSITÉ DE LYON

Délivrée par

L'UNIVERSITÉ CLAUDE BERNARD LYON I

ÉCOLE DOCTORALE : INFORMATIQUE ET MATHEMATIQUES

DIPLÔME DE DOCTORAT
(arrêté du 7 août 2006)

Soutenue publiquement

Le 27 octobre 2010

par

M. Norbert FOUEMKEU

Modélisation de l'incertitude sur les trajectoires d'avions

Préparée au LICIT
Laboratoire d'Ingénierie Circulation Transports
Unité mixte de recherche ENTPE-INRETS

Directeurs de thèse : Jacques SAU
Nour-Eddin El FAOUZI
Rémy FONDACCI

Membres du Jury :

M. Younès BENNANI	Professeur, Université Paris 13	Examinateur
M. Florian DE VUYST	Professeur, Ecole Normale Supérieure de Cachan	Rapporteur
M. Jean-Michel LOUBES	Professeur, Université de Toulouse 3	Rapporteur
Mme Gabriela CIUPERCA	Maître de Conférences (HDR), Université Claude Bernard Lyon 1	Examinatrice
M. Jacques SAU	Professeur Emérite, Université Claude Bernard Lyon 1	Examinateur
M. Nour-Eddin EL-FAOUZI	Directeur de recherche (HDR), INRETS	Examinateur
M. Rémy FONDACCI	Ingénieur principal des études et d'exploitation de l'aviation civile, DSAC Centre-Est	Examinateur

A .....

Remerciements

Mes remerciements vont tout d'abord à mes directeurs de thèse :

- Rémy FONDACCI qui m'a proposé le sujet de cette recherche. En sa qualité du Directeur du LICIT, il a mis à ma disposition tous les moyens matériels nécessaires pour le bon déroulement de cette thèse. Ses multiples lectures de mon travail et des discussions permanentes que nous avons eues dans le domaine de la modélisation du trafic aérien ont été fructueuses et ont permis de donner des meilleures orientations à ma recherche.

- Nour-Eddin EL FAOUZI sur qui j'ai pu compter en permanence pendant ces trois années de thèse. Sa rigueur et ses multiples conseils ont été particulièrement précieux et m'ont permis d'aborder avec plus de sérénité la partie statistique et probabilité de ce travail. Pendant ma thèse, c'est lui qui m'a permis d'allier à la fois la recherche et l'enseignement des cours de statistiques à l'Ecole Nationale des Travaux Publics de l'Etat (ENTPE) et à l'Ecole Polytechnique Universitaire de Lyon (EPUL).

- Jacques SAU pour ses multiples casquettes, d'abord parce qu'il a accepté de diriger cette thèse, ensuite, parce qu'il a été présent tout au long de son exécution.

Je tiens également à remercier M. Florian DE VUYST et M. Jean-Michel LOUBES qui ont bien accepté la lourde tâche de rapporter ma thèse. Je remercie enfin M. Younès BENNANI et Mme Gabriela CIUPERCA d'avoir accepté d'être examinateurs de mon travail.

Je remercie tous les membres du LICIT que j'ai côtoyés tous les jours et plus particulièrement les doctorants : Damien (c'est lui qui m'a initié au langage C++), Aurélie, Romain et David. Je n'oublie pas Sophie dont le bureau était en face du mien.

Je remercie EUROCONTROL et plus particulièrement M. Antoine SIMON d'EUROCONTROL qui nous a facilité l'accès aux données les plus récentes du trafic aérien européen. Je remercie aussi M. Patrick JOSSE de Météo-France qui nous a également permis d'accéder facilement aux données météo pour le trafic aérien.

Pendant cette thèse, j'ai bénéficié du financement de l'ENTPE et du soutien matériel de l'Institut National de Recherche sur les Transports et leur Sécurité (INRETS), qu'ils en soient remerciés.

Je tiens particulièrement à exprimer toute ma gratitude à M. Georges ELANGA OBAM, à sa Majesté FO'O MIATSUET Jean-Gallot JIOTSA, à M. Jean-Paul ZOYEM pour leurs conseils et encouragements pendant mes études.

Mes remerciements vont également à mes deux parents sans qui ce travail n'aurait jamais existé. Que cette thèse soit pour eux une source de bonheur. Merci à mon oncle Albert NKENGNE et sa femme pour le cadre de travail qu'ils m'ont offert pendant mes études. Que cette thèse soit pour eux une raison d'espérance. Je ne peux finir ces remerciements sans penser à tous mes frères et sœurs de la famille FO'O MELOUNG.

Merci enfin à Thérence et à notre petite Celya.

Résumé

Dans cette thèse, nous proposons des modèles probabilistes et statistiques d'analyse de données multidimensionnelles pour la prévision de l'incertitude sur les trajectoires d'aéronefs. En supposant que pendant le vol, chaque aéronef suit sa trajectoire 3D contenue dans son plan de vol déposé, nous avons utilisé l'ensemble des caractéristiques de l'environnement des vols comme variables indépendantes pour expliquer l'heure de passage des aéronefs sur les points de leur trajectoire de vol prévue. Ces caractéristiques sont : les conditions météorologiques et atmosphériques, les paramètres courants des vols, les informations contenues dans les plans de vol déposés et la complexité de trafic. Dans cette étude, la variable dépendante est la différence entre les instants observés pendant le vol et les instants prévus dans les plans de vol pour le passage des aéronefs sur les points de leur trajectoire prévue : c'est la variable *écart temporel*.

En utilisant une technique basée sur le partitionnement récursif d'un échantillon des données, nous avons construit quatre modèles. Le premier que nous avons appelé CART classique est basé sur le principe de la méthode CART de Breiman. Ici, nous utilisons un arbre de régression pour construire une typologie des points des trajectoires des vols en fonction des caractéristiques précédentes et de prévoir les instants de passage des aéronefs sur ces points. Le second modèle appelé CART modifié est une version améliorée du précédent. Ce dernier est construit en remplaçant les prévisions calculées par l'estimation de la moyenne de la variable dépendante dans les nœuds terminaux du modèle CART classique par des nouvelles prévisions données par des régressions multiples à l'intérieur de ces nœuds. Ce nouveau modèle développé en utilisant l'algorithme de sélection et d'élimination des variables explicatives *(Stepwise)* est parcimonieux. En effet, pour chaque nœud terminal, il permet d'expliquer le temps de vol par des variables indépendantes les plus pertinentes pour ce nœud. Le troisième modèle est fondé sur la méthode MARS, modèle de régression multiple par les splines adaptatives. Outre la continuité de l'estimateur de la variable dépendante, ce modèle permet d'évaluer les effets directs des prédicteurs et de ceux de leurs interactions sur le temps de passage des aéronefs sur les points de leur trajectoire de vol prévue. Le quatrième modèle utilise la méthode d'échantillonnage bootstrap. Il s'agit notamment des forêts aléatoires où pour chaque échantillon bootstrap de l'échantillon de données initial, un modèle d'arbre de régression est construit, et la prévision du modèle général est obtenue par une agrégation des prévisions sur l'ensemble de ces arbres. Ce modèle est robuste et constitue une solution au problème d'instabilité des arbres de régression propre à la méthode CART.

Les modèles ainsi construits ont été évalués et validés en utilisant les données test. Leur application au calcul des prévisions de la charge secteur en nombre d'avions entrants a montré qu'un horizon de prévision d'environ 20 minutes pour une fenêtre de temps de largeur supérieure à 20 minutes permettait d'obtenir les prévisions avec des erreurs relatives inférieures à 10%. Ainsi, pour l'autorité régulatrice des courants de trafic aérien, ces modèles constituent un outil d'aide pour une gestion améliorée de la charge des secteurs de l'espace aérien contrôlé.

Mots-clés : Trafic aérien, Gestion de l'espace aérien, Instants de passage, Ecart temporel, CART classique, CART modifié, Méthode MARS, Forêts aléatoires, Prévision, Modèles probabilistes, Statistique, Charge secteur.

Abstract

In this thesis, we analyze multidimensional data using probabilistic and statistics models to forecast uncertainty in aircraft trajectories. Assuming that the aircraft follow the 3D trajectory contained in the initial flight plan, we used the characteristics of the flight environment as predictors to explain the crossing time of aircraft at given points on the planned trajectory. These characteristics are : weather and atmospheric conditions, current flight parameters, information contained in the flight plans and the air traffic complexity. Typically, the dependent variable is the difference between actual time observed during flight and planned time to cross pre-established trajectory points : this variable is called *temporal difference*.

We built four models using a method based on recursive partitioning of the sample. The first called classical CART is based on the Breiman CART method. This model uses regression trees to construct a point typology of aircraft trajectories utilizing previous characteristics and to forecast the crossing time of aircraft at these points. The second model, called amended CART, is an improved model. It replaces forecasting estimated by the mean of the dependent variable inside the terminal nodes of classical CART by new forecasting given by multiple regression inside these nodes. This model, developed using a *Stepwise* algorithm, is efficient, because for each terminal node, it calculates the flight time using most relevant predictors inside the node. The third model is based on the MARS (multivariate adaptive regression splines) method. Besides continuity of the dependent variable estimator, the model assesses the direct and interaction effects of the explanatory variables on the crossing time of flight trajectory points. The fourth model uses a bootstrap sampling method. For each bootstrap sample from the initial data, it constructs a random forest tree regression model as in the CART method. The general model forecasting is obtained by aggregating forecasting on the set of trees. This model is robust and produces a stable solution in contrast to the regression trees obtained with the CART method.

The models described have been assessed and validated using test data. Their use in computing sector load forecasting in terms of aircraft count entering the sector shows that the forecast horizon of about 20 minutes, with interval times larger than 20 minutes, allowed forecasting with relative errors of less than 10%. Hence, these models can be of an assistance to agencies in managing the sector load of the controlled airspace.

Keywords : Air traffic, Airspace management, Crossing time, Time difference, classical CART, amended CART, MARS method, Random Forests, Forecast, Probabilistic models, Statistic, Sector load.

Table des matières

Introduction générale **11**

1 Système de trafic aérien **15**

1.1 Généralités 15

1.1.1 Organisation de l'espace 15

1.1.2 Contrôle de la circulation aérienne 16

1.1.3 Gestion des flux du trafic 17

1.2 Limite du système actuel et évolution envisagée 18

1.3 Contexte de la thèse 20

1.4 Problématique 21

1.4.1 Un exemple 21

2 Position du problème **25**

2.1 La charge d'un secteur de l'espace aérien 25

2.2 Méthodologie 27

2.2.1 Formulation du problème 27

2.2.2 Méthodes utilisées 30

2.3 Modèles de trajectoires existants 30

2.3.1 Le modèle énergie totale 31

2.3.2 Modèles paramétriques et non-paramétriques 31

2.4 Conclusion 34

3 Données utilisées pour l'analyse **35**

3.1 Les données du trafic aérien 35

3.1.1 Plans de vols 35

3.1.2 Données courantes du vol 36

3.2 Informations communes aux données météorologiques et atmosphériques 36

3.3 Données météorologiques 37

3.3.1 Influence du vent 37

3.3.2 Généralisation de l'influence du vent sur la trajectoire de vol prévue 37
3.4 Paramètres de l'air atmosphérique 38
3.4.1 Composition de l'air sec 38
3.4.2 Action de l'air sur les corps en mouvement dans l'atmosphère 38
3.4.3 Altitude densité 39
3.4.4 Ecart de densité de l'air 40
3.4.5 Température et pression de l'air 41
3.5 Paramètres de complexité du trafic 42
3.5.1 Flux total d'aéronefs sur la trajectoire de vol 42
3.5.2 Interactions totales sur la trajectoire de vol 42
3.5.3 Score de complexité de trafic sur la trajectoire prévue 43
3.5.4 Densité du réseau de routes aériennes sur la trajectoire prévue 43
3.6 Ecart temporel : variable à expliquer 44
3.7 Modèle théorique : Le simulateur du trafic aérien OPERA 44
3.7.1 Principe de fonctionnement 44
3.7.2 Lien entre la trajectoire temps réel du vol et le modèle OPERA 45
3.8 Conclusion 46

4 Prévision de l'écart temporel par CART 47
4.1 Introduction 47
4.2 CART : Méthode de régression par arbre 48
4.2.1 Avantages et limites de la méthode 48
4.2.2 Revue de littérature 48
4.2.3 Spécification du modèle de régression par arbre : CART 49
4.2.4 Illustration de la méthode 49
4.2.5 Principe de la méthode CART 51
4.2.6 Etapes de construction d'un arbre de régression 51
4.2.7 Construction de l'arbre maximal $\mathcal{A}_{\max}$ 51
4.2.8 Evaluation du modèle de régression par arbre 53
4.2.9 Erreur de l'arbre 54
4.2.10 Procédure d'élagage 54
4.2.11 Sélection du meilleur arbre 55
4.2.12 Divisions suppléantes 56
4.2.13 Importance d'une variable explicative par CART 57
4.3 Prévision par la méthode CART 57
4.3.1 Arbre de régression retenu 58
4.3.2 Diagnostic des résidus du modèle CART obtenu 58
4.4 Analyse des résultats de la régression 62

4.4.1 Variables actives 62
4.4.2 Conditions météorologiques et atmosphériques 62
4.4.3 Caractéristiques courantes des aéronefs en vol 64
4.4.4 Caractéristiques des vols prévus dans les plans de vols 66
4.4.5 Paramètres de complexité du trafic 66
4.5 Conclusion 67

5 CART modifié par l'algorithme Stepwise 69
5.1 Introduction 69
5.2 Méthodologie 70
5.3 Modèle de régression multiple intra-classe 70
5.3.1 Spécification du modèle de régression intra-classe 70
5.3.2 Principe de fonctionnement de l'algorithme Stepwise 71
5.3.3 Comparaison des statistiques sommaires des résidus des modèles CART classique et CART modifié 72
5.3.4 Test d'ajustement des résidus par une loi normale 73
5.3.5 Dispersion des résidus en fonction de l'horizon temporel de prévision 73
5.3.6 Comparaison de la qualité des deux modèles 75
5.4 Effets des facteurs du modèle *CART modifié* 76
5.4.1 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}1$ 76
5.4.2 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}4$ 76
5.4.3 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}8$ 77
5.4.4 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}9$ 77
5.5 Conclusion 77

6 Modèle MARS 79
6.1 Introduction 79
6.2 Méthode MARS 80
6.2.1 Littérature existante 80
6.2.2 Spécification du modèle MARS 81
6.2.3 La fonction $B_k(\mathbf{X})$ et le nœud d'une spline 81
6.2.4 Algorithme MARS 82
6.2.5 Paramètres pour la mise en œuvre du modèle MARS 86
6.3 Prévision de l'*écart temporel* par la méthode MARS 86
6.3.1 Indicateurs de sélection du modèle 86
6.3.2 Test d'ajustement des résidus du modèle MARS par la loi normale 86
6.3.3 Dispersion du modèle et horizon de prévision 87
6.3.4 Qualité des modèles : MARS, CART classique et CART modifié 88

6.4 Estimation des écarts temporels par le modèle MARS . . . 90
6.4.1 Effets directs . . . 92
6.4.2 Effets d'interactions . . . 95
6.5 Conclusion . . . 99

7 Prévision par les Forêts Aléatoires 101
7.1 Introduction . . . 101
7.2 Forêts aléatoires (FA) . . . 102
7.2.1 Généralités et principe . . . 102
7.2.2 Quelques caractéristiques des FA . . . 104
7.2.3 Représentativité d'un échantillon bootstrap . . . 104
7.2.4 Algorithme . . . 105
7.3 Modélisation de l'*écart temporel* par les forêts aléatoires . . . 106
7.3.1 Choix des paramètres . . . 106
7.3.2 Hiérarchie des variables explicatives du modèle FA . . . 108
7.3.3 Analyse du modèle FA . . . 111
7.4 Comparaison de la qualité prédictive du modèle FA à celles des autres modèles . . . 114
7.5 Conclusion . . . 115

8 Prédictivité des modèles et application 117
8.1 Introduction . . . 117
8.2 Evaluation des modèles . . . 118
8.2.1 Indicateurs d'évaluation des modèles . . . 118
8.2.2 Echantillon de données test pour la validation . . . 119
8.2.3 Erreur d'ajustement et erreur de prévision sur données test . . . 119
8.2.4 Identification du meilleur modèle . . . 120
8.3 Prédictibilité du temps de passage par les modèles . . . 123
8.3.1 Généralités . . . 123
8.3.2 Littérature liée . . . 124
8.3.3 Exemple de contrainte de prédictibilité . . . 125
8.3.4 Métrique de prédictibilité utilisée . . . 126
8.3.5 Tests entre hypothèses multiples unilatérales . . . 126
8.3.6 Ajustement de la distribution de probabilité des erreurs absolues $\mathcal{Q}(h)$ . . . 128
8.3.7 Formulation du test final . . . 131
8.3.8 Construction de la statistique critique K_α . . . 132
8.3.9 Région de prédictivité des modèles . . . 133
8.4 Ajustement des résidus par un mélange de gaussiens . . . 135
8.5 Application des modèles à la prévision de la charge d'un secteur . . . 137

8.5.1 Prévision de la charge du secteur LFFFSUP : journée du 14 septembre 2007 à 7h 00 138
8.6 Conclusion 144

Conclusion générale **145**

Bibliographie **147**

A Maillage de l'espace pour la construction des cellules **161**
A.1 Maillage du domaine d'étude par les cellules de même volume 162
A.1.1 Relation entre deux cellules voisines de même aire 162
A.1.2 Illustration 162

B Compléments aux résultats de l'étude **165**
B.1 Description des variables explicatives 166
B.2 Test asymptotique de normalité de Jarque-Bera 168
B.3 Lois gamma, exponentielle et du khi-2 169
B.4 Divisions suppléantes 170
B.5 Coefficients des modèles de régression linéaire 170

Introduction générale

La croissance du trafic aérien, amorcée vers les années 1990, continue malgré des hauts et des bas de faire son chemin à travers le monde et en particulier en Europe. Elle entraîne avec elle la naissance de nouveaux foyers de concentration du trafic, phénomène observé en Europe de l'Est à cause d'une extension du trafic de plus en plus importante vers cette région (FIG.1.2) et de l'accroissement du niveau de saturation des secteurs de contrôle qui étaient déjà très chargés, particulièrement en Europe occidentale et centrale. L'ampleur de cette croissance est telle que les méthodes de régulation habituellement utilisées par le système ATFM (Air Traffic Flow Management) consistant à protéger les contrôleurs de la surcharge de travail par des retards des vols au sol ou à diviser les secteurs surchargés en des secteurs élémentaires plus petits ont atteint leurs limites capacitaires. La baisse de la capacité de gestion des courants du trafic aérien ATFM couplée au manque de prévisibilité réelle du trafic et à l'utilisation non optimale des ressources disponibles [1] sont des arguments clés avancés par les usagers de l'espace aérien, compagnies aériennes régulières et à la demande, les usagers militaires et l'aviation générale pour réclamer[2] aux autorités régulatrices du trafic et aux prestataires des services de la navigation aérienne (ANPS) de nouveaux outils de gestion pour un écoulement souple et efficace du trafic tout en maintenant le niveau de sécurité actuel (PRR 3, 1999)[14].

Malgré la croissance du trafic limitée à 0.4% en 2008 contre 5% en 2007 (PRR 2008)[17][3], relativement au niveau de 2006, les prévisions de croissance indiquent une augmentation du trafic d'environ 73% à l'horizon 2020. Dans la nécessité d'aligner la capacité de gestion du trafic aérien au moins à ce niveau de croissance à l'horizon 2020 et même au-delà, la commission européenne a lancé l'initiative du « ciel unique européen » avec pour objectif phare de réorganiser et de moderniser la gestion du trafic (ATM) en Europe sous la forme d'un réseau flexible, harmonisé et homogène, indépendant des frontières nationales et constitué de blocs d'espace aérien fonctionnels (FAB).

Cette thèse s'inscrit dans le cadre de la construction d'outils d'aide à la gestion des courants de trafic permettant la régulation de la charge des secteurs ou encore de la quantité de conflits dans l'espace aérien, par action en temps réel sur les aéronefs en vol. De tels outils nécessitent une prévision précise des gran-

[1] Il s'agit du découpage actuel de l'espace aérien en des secteurs de contrôle en fonction les frontières nationales.

[2] Conformément à la Convention de l'OACI, les usagers de l'espace aérien sont en droit d'attendre un écoulement sûr, ordonné et rapide du trafic aérien.

[3] Le PRR 2008 présente la performance des services de la navigation aérienne (ANS) européens en 2008 du point de vue de l'aviation et l'analyse sous l'angle des domaines-clés de performance que sont la sécurité, la ponctualité et la prévisibilité, la capacité/les retards, l'efficacité des vols, l'efficacité économique et l'impact environnemental.

deurs à réguler et des instants d'entrée et de sortie des aéronefs dans les secteurs ou des instants de passage sur des points de croisement de trajectoires. Nous nous proposons donc d'élaborer les modèles probabilistes et statistiques d'analyses de données multidimensionnelles permettant une prévision efficace de l'incertitude sur les trajectoires temporelles des avions pendant leur vol. Ainsi, connaissant la position courante de chaque aéronef en vol, le modèle final doit être capable de prévoir avec la plus faible erreur possible les instants de passage de ces aéronefs en des points futurs de leur trajectoire de vol prévue, en fonction de l'ensemble des caractéristiques de l'environnement du vol. Dans cet environnement, la complexité du trafic, la configuration physique de l'espace (la géométrie des secteurs, le croisement des routes aériennes à l'intérieur de ces secteurs), le comportement des pilotes et la politique des compagnies aériennes, le partage de l'espace aérien entre l'aviation civile et l'aviation militaire interagissent pour influer sur l'évolution globale du trafic. Le problème traité ici est alors un sous-module d'un problème beaucoup plus global et complexe de prévision en temps réel des trajectoires $4D$ des vols afin d'anticiper la charge de secteurs de l'espace et la quantité de conflits entre les aéronefs en vol.

Les modèles proposés sont non-paramétriques, basés sur les méthodes de partitionnement récursif. Il s'agit de partir du modèle classique de classification par arbre de régression appelé CART (Classification And Regression Trees) pour construire un modèle de prévision des instants de passage des aéronefs en des points de leur trajectoire de vol prévue en fonction des données d'archives du trafic (plans de vols), de l'état actuel du trafic, de la complexité du trafic et des conditions météorologiques et atmosphériques prévues. Ensuite, des améliorations sont apportées à ce modèle et ont pour objectifs d'une part d'atténuer les effets de la discontinuité de l'estimateur de la fonction des prévisions entre les sous-régions adjacentes et d'autre part de corriger l'instabilité des arbres inhérente à la méthode CART. Ce mémoire de thèse est organisé comme suit :

Le chapitre 1 présente dans un premier temps l'organisation actuelle du système du trafic aérien et ses limites. Ensuite une synthèse des projets d'amélioration entrepris au niveau européen est exposée. Enfin, nous situons cette thèse dans son contexte et présentons la méthodologie pour le traitement de la problématique définie plus tôt.

Le chapitre 2 présente la formulation probabiliste du problème traité et détaille ensuite la méthodologie pour le résoudre. Il présente également quelques modèles existants dans le cadre de la prévision des trajectoires d'avions.

Le chapitre 3 est consacré aux données utilisées dans cette étude et à la construction des variables explicatives du modèle à élaborer. Dans le paragraphe (3.1), nous présentons les données du trafic aérien où nous définissons le plan de vol et ses évolutions en fonction des filtres ATFM. Le paragraphe (3.2) présente les conditions de recueil des données météorologiques et atmosphériques. Le paragraphe (3.3) décrit l'influence des conditions météorologiques prévues sur les trajectoires des vols ainsi que la construction d'un indicateur d'influence du vent sur le vol. Le paragraphe (3.4) quant à lui se focalise sur les variables explicatives liées à l'air atmosphérique et décrit l'influence de celles-ci sur les aéronefs en vol. Dans le paragraphe (3.5), la complexité du trafic est abordée ; nous construisons un indicateur de complexité du trafic à partir de différents facteurs d'interaction entre les aéronefs en vol dans un volume de l'espace. Le paragraphe

(3.6) s'attache à la définition de la variable à expliquer dans les modèles. Enfin, nous présentons dans le paragraphe (3.7) le simulateur aérien *OPERA* car celui-ci utilise le modèle de performances avion Bada d'EUROCONTROL, qui sera utilisé comme référence par le modèle développé dans cette thèse.

Le chapitre 4 est focalisé sur la méthode de prévision par arbre de régression (CART). Le paragraphe (4.1) expose les motivations du choix de cette méthode pour le problème traité. Le paragraphe (4.2) s'attache à la description des principes fondamentaux de cette méthode. Dans le paragraphe (4.3), nous appliquons la méthode CART à la prévision des instants de passage des aéronefs en des points de leur trajectoire de vol prévue. Ainsi, l'arbre de régression optimal est proposé, le diagnostic des résidus du modèle correspondant est réalisé et nous évaluons l'erreur d'ajustement de ce modèle en fonction de l'horizon temporel de prévision. Le paragraphe (4.4) s'attache à l'analyse des résultats du modèle obtenu. Les effets des variables explicatives actives sont analysés. Le modèle ainsi construit est baptisé CART classique.

Le chapitre 5 consiste en une amélioration de la précision du modèle CART classique. Le paragraphe (5.2) est consacré à la méthodologie. En (5.3), nous utilisons le modèle de régression linéaire généralisé basé sur l'algorithme *Stepwise* pour améliorer les prévisions des écarts temporels dans les nœuds terminaux de l'arbre de régression obtenu au chapitre précédent. Nous comparons la qualité d'ajustement de ce nouveau modèle à celle du modèle CART classique et ce en fonction de l'horizon temporel de prévision. Le paragraphe (5.4) propose une interprétation des effets marginaux des caractéristiques des vols sur les instants de passage des aéronefs pour les nœuds où les modèles de régression présentent des bons indicateurs de qualité d'ajustement. Ce nouveau modèle est baptisé CART modifié.

Dans le chapitre 6, nous appliquons la méthode MARS (Multivariate adaptive regression splines) pour modéliser les instants de passage des aéronefs en des points de leur trajectoire. Le paragraphe (6.2) se focalise à la définition et aux principes de ce type de modèle. Le modèle est alors spécifié et l'algorithme de construction des fonctions splines de la base de l'estimateur de la variable dépendante est présenté. Le paragraphe (6.3) est complètement consacré à la modélisation de la variable explicative par la méthode MARS. Les résultats de ce modèle sont comparés à ceux obtenus aux chapitres 4 et 5. Le paragraphe (6.4) analyse les effets directs (principaux) des facteurs explicatifs et ceux des interactions entre ces facteurs sur la variable à expliquer.

Le chapitre 7 propose un autre modèle basé sur les forêts aléatoires. Nous consacrons le paragraphe (7.2) à la présentation de quelques fondements du modèle des forêts aléatoires et de son algorithme. Dans le paragraphe (7.3), nous appliquons ce modèle à la prévision des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue. Nous tirons également parti de ses possibilités de l'hiérarchisation des variables explicatives pour affiner la première hiérarchie de ces variables obtenue plus tôt par le modèle CART classique. L'incertitude du modèle développé est ensuite étudiée en fonction de l'horizon temporel de prévision. Dans le paragraphe (7.4), nous utilisons le coefficient de Theil basé sur l'erreur de prévision pour comparer la qualité d'ajustement du modèle développé par les forêts aléatoires à celles des modèles CART classique, CART modifié et MARS.

Le chapitre 8 est consacré d'une part à l'évaluation de la prédictibilité du temps de passage des aéronefs par des modèles développés et à l'application de ces modèles à la prévision de la charge secteur. Le para-

graphe (8.2) utilise les données test pour évaluer la capacité de généralisation de chaque modèle en situation nouvelle. Les indicateurs d'évaluation sont présentés et nous comparons ensuite les erreurs d'ajustement des modèles à celles de prévision. Le paragraphe (8.3) définit la notion générale de prédictibilité du trafic en en proposant une métrique pour notre étude. Le théorème du rapport des vraisemblances monotone est rappelé et les conditions de son application sont vérifiées. Nous montrons que ce théorème est applicable pour la recherche des régions de prédictibilité des instants de passage des aéronefs par des modèles. Le paragraphe (8.4) traite enfin du problème d'ajustement de la loi de probabilité des résidus de chacun des quatre modèles proposés. Le paragraphe (8.5) se focalise à l'utilisation des modèles pour la prévision proprement dite de la charge d'un secteur de l'espace. Pour cela, la formulation probabiliste de la charge secteur est complètement définie. Nous proposons un coefficient de correction qui permet de déduire la part de la charge secteur dûe aux aéronefs entrants dans le secteur alors qu'ils n'ont pas prévu y entrer, en fonction de la part de la charge de ce secteur dûe aux aéronefs qui ont prévu de le traverser et qui l'ont effectivement traversé. Finalement, la prédictibilité de la charge secteur par les modèles est évaluée simultanément en fonction de l'horizon temporel de prévision et de la largeur de la fenêtre de prévision.
Nous achèvons notre travail par une conclusion générale où nous donnons quelques perspectives d'utilisation de notre travail.

Chapitre 1

Système de trafic aérien

1.1 Généralités

Avant d'aborder le thème de cette thèse, commençons par préciser pour les lecteurs non familiers du monde de la circulation aérienne quelques notions de base, utiles à la compréhension du reste du document. Les notions présentées dans cette section sont pour l'essentiel extraites de l'ouvrage de Georges MAIGNAN (1991)[47] sur le contrôle de la navigation aérienne. Cet ouvrage, bien que relativement anciens est encore pertinent pour la présentation des principes généraux de la circulation aérienne.

Le système de trafic aérien repose sur une organisation de l'espace ainsi que sur des méthodes de gestion des flux du trafic et de contrôle de la circulation aérienne.

1.1.1 Organisation de l'espace

Pour faciliter la surveillance et le contrôle des aéronefs, l'espace aérien est découpé sous l'égide de l'OACI (Organisation de l'Aviation Civile Internationale) en deux parties : l'espace aérien inférieur (FIR)[1] et l'espace aérien supérieur (UIR)[2]. Chacun de ces espaces est à son tour découpé en secteurs de contrôle (FIG.1.1), plus simplement appelés secteurs. Un secteur est un contour géographique délimité par un ensemble de points (latitude, longitude), une altitude minimum et une altitude maximum. Plusieurs secteurs dits élémentaires peuvent être regroupés pour ne former qu'un seul secteur de contrôle auquel on attribue une capacité et on affecte une équipe de contrôleurs pour gérer les flux d'avions entrants et sortants. Pendant le vol, un aéronef passe d'un secteur à un autre en suivant des routes aériennes qui sont matérialisées par une succession de tronçons limités par les points appelés balises. Un tel point équipé d'un radiophare[3] est

[1]Flight Information Region. Il correspond à des altitudes inférieures à 19500 pieds ($FL195$). L'espace aérien inférieur n'est contrôlé qu'autour des voies aériennes (AWY) et dans les régions terminales autour des aérodromes. A l'extérieur de ces zones, l'avion ne bénéficie que du service d'information, mais pas du service d'anticollision.

[2]Upper Information Region. C'est tout l'espace compris entre 19500 et 66000 pieds (entre $FL195$ et $FL660$). Au-dessus du niveau de vol 195, l'espace est contrôlé partout. Il n'y a pas de voies aériennes mais des routes prédéterminées (PDR). Cet espace est partagé entre l'aviation générale et l'aviation militaire.

[3]C'est une installation radio-électrique de navigation appelée VOR.

en général le point de croisement avec une ou plusieurs autres routes.

Dans le système ancien, les aéronefs ne pouvaient naviguer commodément que sur les radiales de ces radiophares, en s'en rapprochant ou en s'en éloignant. Grâce aux évolutions technologiques actuelles, les avions de transports modernes sont, pour un grand nombre, équipés des calculateurs de bord couplés au pilote automatique, et sont capables de suivre avec une grande précision n'importe quelle route définie par un point de début et un point de fin. Ce système de navigation de surface s'appelle RNAV[4] et permet aux pilotes de s'affranchir de la navigation point à point bien plus coûteux en carburant et sur le plan environnemental du fait de l'allongement des distances de vol.

FIG. 1.1 – Un extrait de l'espace aérien Européen (2007) : Les traits bleus sont les limites des secteurs aériens, les traits rouges sont des routes aériennes et les points noirs sont les balises ou les points de croisements de plusieurs routes aériennes et les triangles sont les points de report obligatoire de la position des vols.

1.1.2 Contrôle de la circulation aérienne

En circulation aérienne civile, on distingue principalement trois types de contrôle. Le contrôle en route s'effectue entre les régions terminales (TMA : TerMinal control Area) des aérodromes de départ et de destination du vol. Il est assuré à partir de centres spécialisés souvent implantés en dehors des aérodromes. En France, ce sont les « centres régionaux de la navigation aérienne » (CRNA) ; il y en a 5[5], situés à Brest, Athis-Mons, Reims, Bordeaux et Aix-en-Provence. Le contrôle d'approche concerne tout aérodrome où il

[4] Area Navigation.

[5] Aux Etats-Unis, il y a 21 centres de contrôle en-route pour l'ensemble du territoire

existe un important trafic aux instruments (IFR : Instrument Flying Rules). Les contrôleurs d'approche coordonnent le transfert des aéronefs entrants du centre de contrôle en route vers la tour de contrôle (CTR) responsable du contrôle d'aérodrome de destination du vol. De même ils coordonnent le transfert des aéronefs en partance de la tour de contrôle responsable du contrôle d'aérodrome de départ vers le centre de contrôle en route. Enfin, le contrôle d'aérodrome s'occupe des pistes et de la circulation au sol sur la plateforme aéroportuaire. Il est assuré à partir de la tour de contrôle installée sur l'aérodrome.

Le rôle premier du contrôle de la circulation aérienne est la prévention des abordages entre les aéronefs. Le contrôle aérien doit ainsi veiller au respect des normes de séparation verticales[6] et horizontales[7]. Lorsqu'il y'a un risque de perte simultanée de ces seuils d'espacements, on dit qu'il y'a conflit entre les aéronefs impliqués.

1.1.3 Gestion des flux du trafic

Un vol commence par le dépôt d'un plan de vol (PLN) auprès des autorités de gestion des flux de trafic. Ce plan de vol contient un ensemble d'informations sur le vol prévu telles que : le jour du vol, les aéroports d'origine et de destination, l'heure de départ souhaitée, la route aérienne empruntée, l'ensemble des points balises jalonnant cette route, les heures prévues pour le survol des balises, l'ensemble des secteurs à traverser, les heures d'entrée et de sortie des secteurs, le niveau de vol de croisière, le type d'aéronef, le nom de la compagnie aérienne exploitant du vol, le numéro tactique du vol et le numéro d'identification du vol. Il arrive que plusieurs pilotes dans leurs plans de vol, demandent à survoler les mêmes balises et à traverser les mêmes secteurs aux mêmes heures. Afin d'éviter tout risque de conflits entre les avions et la surcharge du travail des contrôleurs aériens dans ces zones de l'espace, il existe tout un mécanisme de régulation qui organise, de façon la plus rationnelle possible, l'utilisation des ressources disponibles (le contrôle) et répartit de façon impartiale les attentes inévitables. En Europe cette mission de régulation des courants de trafic (ATFM : Air Traffic Flow Management) est assurée par la CFMU (Central Flow Management Unit)[8] en collaboration avec les centres nationaux. Pour son efficacité, l'activité d'ATFM visant à organiser la fluidité du trafic dans les conditions de sécurité maximale agit à travers trois *filtres* hiérarchisés ci-après :

- Le filtre ATFM stratégique : Appliqué plus d'un an à 6 mois avant les vols, il consiste à définir un schéma général d'orientation de trafic en fonction des flux. Il a pour but de définir les stratégies de régulations, les schémas d'orientation du trafic, des modalités d'utilisation des routes et itinéraires, de participer à l'aménagement de l'espace par modification de la sectorisation. Le filtre stratégique nécessite donc une parfaite connaissance des flux de trafic du centre de contrôle ainsi que du fonctionnement de la CFMU et des mécanismes de régulation et d'allocation de créneaux. Il permet également la gestion d'événements modifiant sensiblement les caractéristiques du trafic (salon du Bourget, coupe du monde de football).

[6] L'espacement vertical est fixé à 1000 pieds soit 300 m.

[7] La séparation horizontale est fixée à 5 NM (9 Km). Il s'agit de la séparation radar en route. Lorsque la couverture radar est de qualité parfaite, cette norme fixée à 3 NM (5.5 Km) en région terminale.

[8] La CFMU est opérationnelle depuis 1995.

- Le filtre ATFM pré-tactique : L'objectif de ce filtre est de prendre des mesures préventives afin d'éviter toute surcharge de secteurs. Il est appliqué un à deux jours avant le départ du vol et permet la préparation des plans de régulation ATFM. Ces plans de régulation consistent en une synthèse des informations disponibles et concernent les dernières mises à jour des PLN[9], les schémas d'ouverture des centres et le nombre de positions de contrôle actives pour le jour de départ, ce qui permet d'organiser une journée du trafic et d'identifier les secteurs saturés où le trafic sera régulé.
- Le filtre ATFM tactique : L'objectif de ce filtre est d'assurer, en temps réel, une charge de trafic gérable par les contrôleurs. L'ATFM tactique évalue les créneaux de décollage des vols en tenant compte simultanément des dernières mises à jour sur la capacité du système ATC (Air Traffic Control) et de la demande de trafic, des créneaux prévus par les compagnies aériennes et des plans de régulation pré-tactiques. Il est rendu nécessaire à cause de l'incertitude sur l'évolution quantitative et qualitative des flux de trafic. Pour cela, l'ATFM tactique propose des mesures de re-routements, consistant en un changement d'itinéraire du vol pour éviter les secteurs saturés. Rappelons que le système qui réalise l'ensemble de ces calculs au niveau de la CFMU est le système CASA (Computer Assisted Slot Allocation).
- Le filtre ATC agit sur la séparation physique des vols afin de maintenir en permanence les distances de sécurité entre les aéronefs dans les conditions économiques optimales. C'est un service anticollision fourni aux avions par les contrôleurs aériens. Il peut également proposer aux aéronefs un re-routement qui consiste en un changement d'itinéraire du vol pour éviter les secteurs saturés. Le cap[10] et le niveau de vol[11] peuvent ainsi être changés.

1.2 Limite du système actuel et évolution envisagée

Pour la régulation du trafic aérien européen, qui a pour but la gestion de la saturation des secteurs, la CFMU a souvent recours à deux solutions. La première consiste à retarder les vols au sol, ce qui engendre inéluctablement des coûts économiques importants à la fois pour les compagnies aériennes et pour les passagers. La deuxième consiste à diviser un secteur dont la surcapacité est récurrente en deux nouveaux secteurs plus petits et repartir ainsi la charge de travail des contrôleurs du secteur initial. Cette deuxième stratégie, qui sert à protéger les contrôleurs contre la surcharge de travail pénalise fortement les prestataires de services de la circulation aérienne (ANSP)[12] qui doivent investir dans le recrutement et la formation des nouvelles équipes de contrôleurs. Rappelons que dans un secteur de l'espace contrôlé, la charge de travail des contrôleurs est une somme pondérée de trois composantes et définie par l'équation (1) :

$$\mathbf{C}_{secteur} = \alpha.\mathbf{C}_{Surv} + \beta.\mathbf{C}_{Coor} + \gamma.\mathbf{C}_{Conf} \tag{1}$$

[9] Il s'agit des plans de vol.
[10] le cap est l'angle que fait l'axe longitudinal de l'avion avec le Nord.
[11] Le niveau de vol est la hauteur en pieds divisée par 100 de l'avion au dessus de l'altitude pression 1013.25 Hpa.
[12] Air Navigation Service Providers.

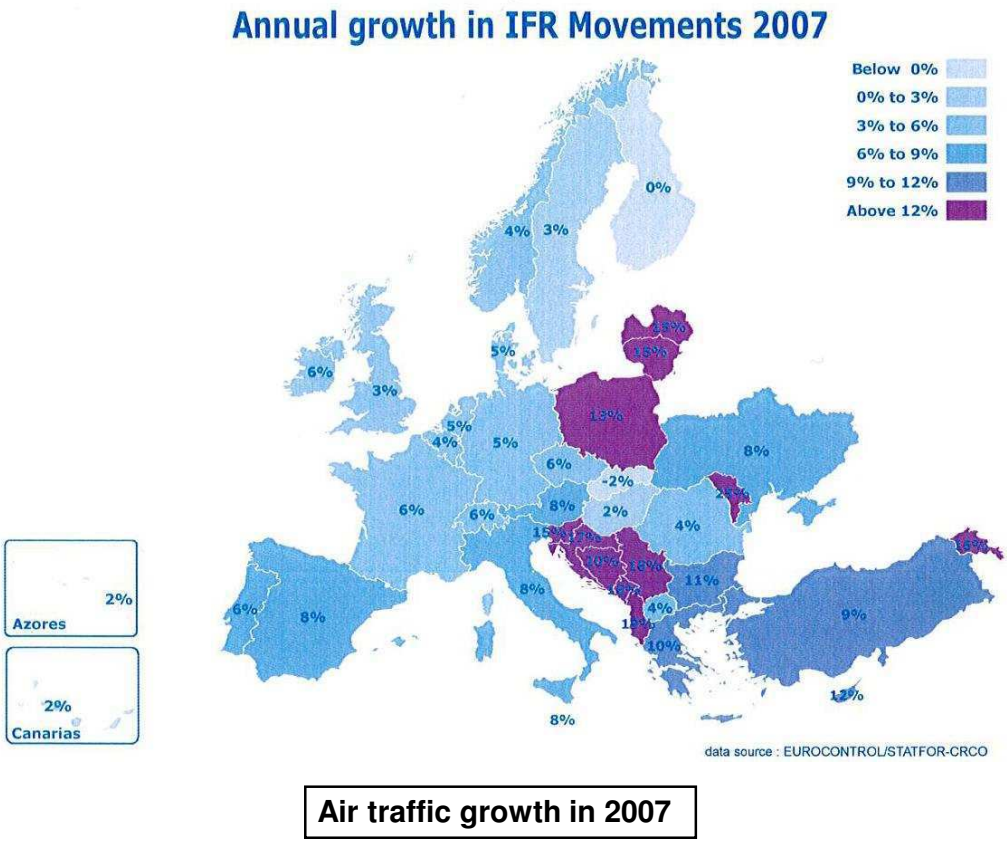

FIG. 1.2 – Illustration de la croissance du trafic aérien en Europe

où

- $\mathbf{C}_{Surv}$ est la charge dite de surveillance, qui est la part de la charge de travail du contrôleur nécessaire pour surveiller les avions qui sont dans son secteur à un moment donné.
- $\mathbf{C}_{Coor}$ est la charge liée à la coordination qui dépend des entrées et des sorties des avions d'un secteur.
- $\mathbf{C}_{Conf}$ est la charge de travail liée à la détection et à la résolution des conflits.
- α, β et γ sont des paramètres qu'il faut définir avec l'aide d'experts contrôleurs[13].

La croissance soutenue du trafic aérien observée (FIG.1.2) pendant cette décennie montre clairement que dans un avenir très proche, ces solutions ne permettront plus d'absorber la demande du trafic de plus en plus élevée. En effet, un secteur de contrôle a des limites de taille au-dessous de laquelle la division n'est plus possible, ce qui à nouveau met les régulateurs du trafic aérien au pieds de ce que l'on appelle le « *mur de la capacité* ». D'où l'urgence de repenser les techniques d'organisation du trafic et d'en proposer de nouvelles qui permettent d'augmenter l'offre de capacité de contrôle en agissant sur les vols en temps réel sans nécessairement recourir aux retards des vols au sol. Pour faire face à cette croissance du trafic, plusieurs

[13]Ils définissent les paramètres lors des simulations du trafic afin que les indicateurs de charge produits soient réalistes.

concepts ont émergé et proposent des changements radicaux dans la façon de gérer le trafic aérien :

- A l'échelle internationale, le programme du minimum de séparation verticale réduit RVSM (2003) [51] a été lancé en 1997 sous les auspices de l'organisation de l'aviation civile internationale (OACI), en coopération avec l'association du transport aérien international (IATA), EUROCONTROL, la fédération internationale des associations de pilotes de ligne (IFALPA), la fédération internationale des associations de contrôleurs de la circulation aérienne (IFATCA) et l'industrie aéronautique. Grâce à ce programme la séparation verticale entre les aéronefs effectuant des vols sur les principales routes reliant l'Asie, le Moyen-Orient et l'Europe est passée de 2000 ft à 1000 ft. Le minimum de séparation verticale réduit (RVSM) augmente l'accès aux niveaux de croisière plus efficaces, ce qui permet de réduire la pollution atmosphérique, les coûts en diminuant la consommation de carburant, et les retards au sol. En Europe, le programme RVSM permet aux transporteurs aériens d'économiser 4 milliards d'euros par année.
- Au niveau de l'Europe le projet majeur lié à la mise en œuvre du ciel unique européen est SESAR (Single European Sky Air Traffic Management Research). Lancé en 2004, il vise à moderniser le système de gestion du trafic aérien en Europe. Son objectif est d'optimiser le trafic aérien et de le rendre plus sûr par l'utilisation de nouvelles technologies de contrôle et de communication entre le sol et les avions. Les chiffres provenant de ce projet indiquent qu'en 2020 le trafic aura augmenté de 73% par rapport au niveau observé en 2006. L'objectif à long terme du projet SESAR, après 2020, est d'essayer de tripler la capacité du trafic afin de permettre d'absorber la croissance du transport aérien et maîtriser ainsi les retards.

1.3 Contexte de la thèse

Pendant la décennie précédente, la croissance rapide du trafic aérien en Europe a causé dans le système de gestion des flux de trafic un certain nombre de problèmes dont le plus important est la baisse de la capacité de gestion des courants du trafic aérien. A cela s'ajoutent le manque de prévisibilité réelle du trafic, l'utilisation non optimale des ressources disponibles et l'augmentation du niveau des retards des vols. Cette faible capacité du système face à la croissance soutenue de la demande des usagers de l'espace aérien interpelle de plus en plus les chercheurs de la communauté aéronautique pour définir et proposer de nouvelles stratégies d'organisation et de gestion du trafic.

Dès la deuxième édition du rapport d'examen des performances d'EUROCONTROL (PRR2, 1999)[14] portant sur les retards, le déficit de capacité du système était indentifié comme l'un des éléments clés de la baisse des performances du système de gestion des flux du trafic. Ce constat est toujours d'actualité, et la situation pourrait s'aggraver si des actions d'envergure n'étaient pas entreprises en vue de réduire durablement le fossé entre une demande de trafic en forte progression et une offre de capacité en stagnation. Le rapport de la commission d'examen des performances (PRR, 2007) [16] d'EUROCONTROL montre que la capacité de gestion du trafic aérien ATM en Europe est bien inférieure à la demande Du trafic, et que le

retard moyen ATFM dû à la gestion du trafic en-route a augmenté au cours de l'année 2007 (1.6 min/vol) par rapport au niveau de 2006 (1.4 min/vol) [16].

Parmi les causes de cette contre performance, l'incapacité à prévoir correctement les niveaux réels du trafic est sans doute la plus importante. Ainsi, les compagnies aériennes, les prestataires de service de navigation aérienne (ANSP), les autorités aéroportuaires et d'autres usagers de l'espace aérien sont unanimes sur l'urgence de mettre en place de nouveaux outils de gestion des flux du trafic. Ces outils ont pour finalité d'accroître la capacité de gestion du trafic avec une meilleure prise en compte de la croissance réelle de la demande. Ils doivent permettre d'assurer une planification et une régulation optimale du trafic sans compromettre les paramètres fondamentaux de performances du système de gestion du trafic en Europe que sont : la sécurité des passagers et des aéronefs, les retards et l'efficacité économique du trafic.

En somme, ces nouveaux outils doivent contribuer efficacement à une organisation du trafic aérien par leur capacité à adapter en temps réel la demande de trafic à la capacité dynamique de gestion des flux du trafic. Pour cela, ils doivent notamment permettre à une meilleure prévision de la charge du trafic et de la quantité de conflits dans chaque zone de l'espace aérien.

Ce sont les motivations qui ont sous-tendu la volonté du LICIT (Laboratoire d'Ingénierie Circulation Transports) à initier de nouvelles recherches dans ce sens. Celles-ci visent notamment, à proposer de nouvelles techniques de planification et de régulation du trafic aérien par l'allocation en temps réel de trajectoires en fonction de la charge de trafic et de la quantité de conflits à gérer par les contrôleurs.

1.4 Problématique

Dans cette thèse, nous utilisons les données opérationnelles de trafic pour la mise en place de techniques pour une prévision probabiliste des instants d'entrée et de sortie des aéronefs des secteurs de l'espace. Pour cela, nous avons :

- Les archives des plans de vol ;
- Les caractéristiques courantes des vols ;
- Les paramètres de complexité du trafic ;
- Enfin, les conditions météorologiques et atmosphériques du trafic.

L'enjeu de ce travail repose sur le fait que les trajectoires $3D$ prévues dans les plans de vol ne sont pas souvent respectées par les pilotes. En effet, si chaque aéronef respectait son plan de vol $3D$, la modélisation de l'incertitude sur les instants d'entrée et de sortie des secteurs ou sur les points de croisement des trajectoires suffirait pour prévoir la présence des aéronefs dans les secteurs d'une part, et la quantité de conflits dans l'espace d'autre part.

1.4.1 Un exemple

La figure FIG.1.3 présente un exemple de comportements souvent rencontrés pendant l'évolution des aéronefs sur leur trajectoire. Nous considérons deux aéronefs (a) et (b) qui se trouvent aux points A et B de

leur trajectoire et qui évoluent vers leurs destinations respectives situées aux points C et D.

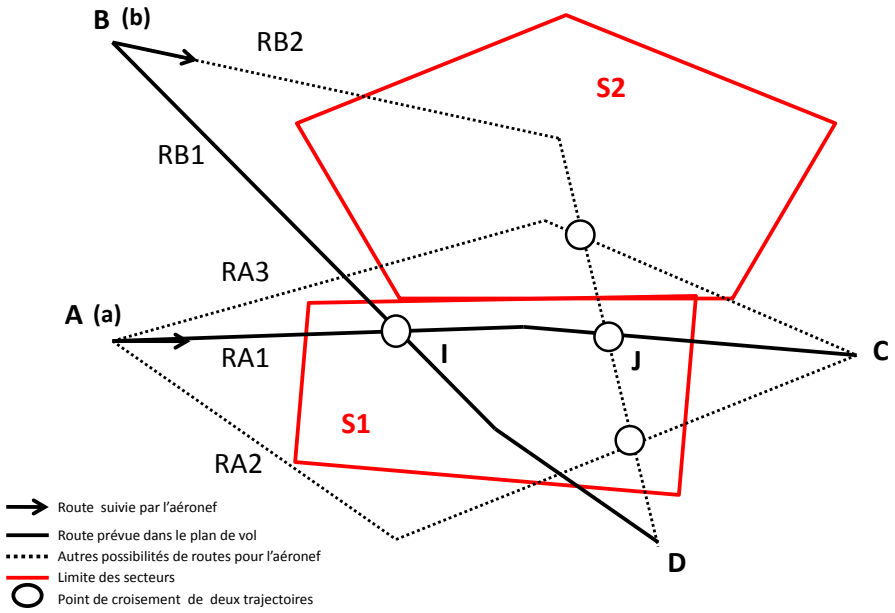

FIG. 1.3 – Illustration de la façon dont les trajectoires sont respectées pendant les vols d'aéronefs

Conformément à son plan de vol prévu, l'aéronef (a) a choisi la route $RA1$ et doit traverser le secteur $S1$ pour atteindre sa destination au point C. En revanche, l'aéronef (b) a choisi la route $RB2$ qui est une route alternative différente de $RB1$ prévue dans son plan de vol. Il traverse successivement les secteurs $S2$ et $S1$ pour atteindre sa destination au point D. J est alors le point de croisement des trajectoires des aéronefs (a) et (b). Ce point est bien différent de I qui est le point de croisement potentiel prévu par les deux plans de vol.

En choisissant la route $RA2$, l'aéronef (a) traverse toujours le secteur $S1$ conformément à son plan de vol, mais y passe moins de temps par rapport à la durée de sa présence dans ce secteur par la route $RA1$. Bien qu'étant toujours à l'intérieur du secteur $S1$, le point de croisement de sa trajectoire avec celle de l'aéronef (b) est modifié. Ce qui modifie également les conditions d'éventuel conflit entre ces deux aéronefs ou la charge de trafic dans ce secteur.

En choisissant la route $RA3$, l'aéronef (a) ne traverse plus le secteur $S1$. Il atteint sa destination en traversant le secteur $S2$ qui n'est pas un secteur de sa trajectoire prévue dans son plan de vol. Le point de croisement de sa trajectoire avec celle de l'aéronef (b) se trouve maintenant à l'intérieur du secteur $S2$. Les conditions d'éventuel conflit entre ces deux aéronefs peuvent également changer, de même la charge de trafic dans le secteur $S2$ augmente (au moins en nombre d'aéronefs entrants).

Cet exemple illustre la complexité du problème de modélisation du temps de présence d'un aéronef dans un secteur de sa trajectoire et des instants de passage des aéronefs sur les points de croisement des trajectoires de vol. Ainsi, l'élaboration des modèles de prévision de la charge du trafic et de la quantité

de conflits dans une zone de l'espace nécessitent une meilleure prise en compte des choix stratégiques des pilotes pour la gestion des vols. Pour un secteur de l'espace donné, les modèles recherchés devraient tenir compte des trois cas suivants :

- Des aéronefs qui ont prévu dans leur plan de vol de traverser ce secteur et qui l'ont effectivement traversé. C'est le cas de l'aéronef (a) avec le secteur $S1$.
- Les aéronefs qui ont prévu dans leur plan de vol de traverser ce secteur et qui ne l'ont pas traversé. C'est le cas de l'aéronef (a) qui a prévu de traverser le secteur $S1$ et qui aurait choisi la route $RA3$ pour atteindre sa destination au point C.
- Les aéronefs qui n'ont pas prévu dans leur plan de vol de traverser ce secteur et qui l'ont finalement traversé. C'est typiquement le cas de l'aéronef (b) qui n'a pas prévu de traverser le secteur (S2), mais qui l'a finalement traversé pour atteindre sa destination au point D.

La difficulté d'agréger ces trois composantes au sein d'un même modèle probabiliste pour la prévision de l'instant d'entrée ou de sortie d'un aéronef des secteurs de sa trajectoire est partiellement dûe au déficit d'information sur les raisons qui orientent les pilotes dans leur choix stratégique de la route à suivre pour atteindre un point de leur trajectoire. Dans le chapitre suivant, nous posons le problème qui est traité dans cette thèse.

Chapitre 2

Position du problème

2.1 La charge d'un secteur de l'espace aérien

Dans cette thèse, nous voulons proposer une méthode d'estimation de la charge des secteurs de l'espace aérien. Il s'agit en effet, pour un secteur donné de l'espace de proposer une estimation probabiliste de sa charge. En utilisant la définition de la charge secteur utilisée par la CFMU, nous nous intéressons ici à la charge d'un secteur qui est le nombre d'avions entrant dans ce secteur pendant une période de temps fixe. Ces périodes sont des intervalles de type [00h ; 01h], [01h ; 02h], [02h ; 03h], etc. Ainsi, une période est définie par une heure de début de période fixe notée T_d et une heure de fin de période fixe notée T_f. Nous notons $N_s(T_d, T_f)$ la charge d'un secteur s pour désigner le nombre d'avions qui entrent dans ce secteur pendant la période définie par l'intervalle de temps $[T_d, T_f]$.

Avant de proposer une estimation de $N_s(T_d, T_f)$, définissons quelques variables et indicateurs importants qui serviront à la calculer. Posons :

- s est un secteur de l'espace aérien.
- i est un aéronef en vol dans l'espace.
- Y_s est un vecteur de variables aléatoires binaires $(Y_{s1}, ..., Y_{si}, ...)$ telles que pour un aéronef i, la variable aléatoire Y_{si} est définie par : $Y_{si} = 1$ si pendant le vol l'aéronef i est entré dans le secteur s et $Y_{si} = 0$ sinon.
- Z_s est un vecteur d'indicateurs binaires $(Z_{s1}, ..., Z_{si}, ...)$ tel que pour un aéronef i qui est entré dans le secteur s, $Z_{si} = 1$ si i a prévu dans son plan de vol d'entrer dans s et $Z_{si} = 0$ sinon.
- T_s est un vecteur de variables aléatoires $(T_{s1}, ..., T_{si}, ...)$ tel que pour chaque aéronef i son instant d'entrée dans s est donné par la variable aléatoire T_{si}.
- $\mathcal{P}_s$ est l'ensemble des aéronefs qui auront pénétré dans s et qui avaient prévu dans leur plan de vol d'y entrer. $\mathcal{P}_s^c$ est l'ensemble des aéronefs qui auront pénétré dans s alors qu'ils n'avaient pas prévu dans leur plan de vol d'y entrer.
- Pour chaque aéronef $i \in \mathcal{P}_s$, on connaît t_{si}^p, l'instant prévu dans son plan de vol pour entrer dans le

secteur s, t_{si} est son instant réel d'entrée dans ce secteur pendant le vol. Ces deux instants sont liés dans le modèle que nous cherchons par la relation (1) suivante :

$$t_{si} = t_{si}^p + \phi_{si} + \epsilon_{si}, \tag{1}$$

où $(\phi_{si})_{(i\in\mathcal{P}_s)}$ est un terme à déterminer en fonction des caractéristiques connues ou prévues des vols, $(\epsilon_{si})_{(i\in\mathcal{P}_s)}$ est le résidu dont nous estimons la loi de probabilité dans ce mémoire.

En utilisant ces variables et indicateurs, nous proposons une estimation $\mathbf{N}_s(T_d, T_f)$ de la charge du secteur s entre les instants T_d et T_f. Cette estimation est définie comme l'espérance mathématique de la variable aléatoire jointe (Y_s, T_s). En supposant l'indépendance de Y_s et T_s, nous avons les égalités suivantes :

$$\begin{aligned}
\mathbf{N}_s(T_d, T_f) &= \sum_i \mathbb{P}\left(y_{si} = 1, T_d < t_{si} < T_f\right) \\
&= \sum_{i\in\mathcal{P}_s} \mathbb{P}\left(y_{si} = 1\right)\mathbb{P}\left(T_f < t_{si} < T_f\right) + \sum_{i\in\mathcal{P}_s^c} \mathbb{P}\left(y_{si} = 1\right)\mathbb{P}\left(T_d < t_{si} < T_f\right).
\end{aligned}$$

On pose les relations :

$$\begin{aligned}
\delta_{i,1} &= \mathbb{P}\left(y_{si} = 1, i \in \mathcal{P}_s\right) \\
\delta_{i,0} &= \mathbb{P}\left(y_{si} = 1, i \in \mathcal{P}_s^c\right)
\end{aligned}$$

où $\delta_{i,1}$ est la probabilité que l'aéronef i entre dans s qui est un secteur prévu de sa trajectoire de vol et $\delta_{i,0}$ est la probabilité que l'aéronef i entre dans s qui n'est pas un secteur prévu de sa trajectoire de vol. Ces deux probabilités sont déterminées par le choix des routes et de leur tenue par les pilotes. Le modèle du choix des routes n'est pas développé dans cette thèse mais fera l'objet d'un modèle spécifique. Ainsi, en utilisant les données historiques des vols, nous avons estimé $\delta_{i,1}$ et $\delta_{i,0}$ par leur valeur moyenne pour chaque secteur de l'espace. Finalement, le problème à résoudre revient à estimer la fonction de densité f_{ϵ_s} qui permet de calculer la probabilité $\mathbb{P}\left(T_d < t_{si} < T_f\right)$ en utilisant la relation (1) ci-dessus. On écrit alors :

$$\mathbb{P}\left(T_d < t_{si} < T_f\right) = \mathbb{P}\left(T_d - \phi_{si} - t_{si}^p < \epsilon_{si} < T_f - \phi_{si} - t_{si}^p\right) \tag{2}$$

$$= \int_{T_d-\phi_{si}-t_{si}^p}^{T_f-\phi_{si}-t_{si}^p} f_{\epsilon_s}(t)\, dt \tag{3}$$

Il apparaît de l'équation (2) que pour les aéronefs qui respectent leur trajectoire de vol $3D$, l'estimation de la charge d'un secteur s de leur trajectoire est une fonction de densité de probabilité de la variable résiduelle ϵ_s. En revanche, pour les aéronefs qui pénètrent les secteurs non prévus dans leur trajectoire $3D$, l'absence des instants prévus t_s^p dans leur plan de vol rend difficile la prévision de leur entrée dans les secteurs de leur trajectoire effectivement réalisée pendant le vol.

Le nombre d'avions entrant dans un secteur s entre les instants T_d et T_f est décomposé par la relation :

$$\begin{aligned}\mathbf{N}_s(T_d,T_f) &= \sum_{i\in\mathcal{P}_s}\delta_{i,1}\int_{T_d-\phi_{si}-t^p_{si}}^{T_f-\phi_{si}-t^p_{si}} f_{\epsilon_s}(t)\,dt + \sum_{i\in\mathcal{P}_s^c}\delta_{i,0}\mathbb{P}\left(T_d<t_{si}<T_f\right)\\ &= \mathbf{N}_{\mathcal{P}_s}(T_d,T_f)+\mathbf{N}_{\mathcal{P}_s^c}(T_d,T_f).\end{aligned}$$

Ainsi, estimer $\mathbf{N}_s(T_d,T_f)$ revient à estimer chacun des deux termes du membre de droite :

- $\mathbf{N}_{\mathcal{P}_s}(T_d,T_f)$ est le nombre d'avions qui entrent dans un secteur s pendant l'intervalle de temps $[T_d,T_f]$ et qui avaient prévu dans leur plan de vol d'entrer dans ce secteur.
- $\mathbf{N}_{\mathcal{P}_s^c}(T_d,T_f)$ est le nombre d'avions qui entrent dans le secteur s pendant l'intervalle de temps $[T_d,T_f]$ alors qu'ils n'avaient pas prévu dans leur plan de vol d'entrer dans ce secteur.

A partitr de ces équations, nous voyons que pour déterminer la charge d'un secteur pendant un intervalle de temps donné, nous avons besoin des instants prévus dans les plans de vol pour que les aéronefs entrent dans ce secteur. Ainsi, pour les aéronefs qui respectent leur trajectoire $3D$ prévue dans leur plan de vol, la prévision de la charge secteur se fait avec moins de difficultés. En effet, utilisant les données de l'ensemble des plans de vol, nous pouvons estimer la loi de probabilité de ϵ. De même, la probablilité pour qu'un aéronef entre dans un secteur de sa trajectoire prévue peut être calculée. Sachant aussi qu'un avion entre dans un secteur de sa trajectoire prévue, il est possible de déterminer la probabilité qu'il y entre dans le créneau correspondant au calcul de la prévision de la charge. En revanche, il est plus difficile de prévoir la part de la charge secteur dûe aux aéronefs qui n'ont pas prévu d'entrer dans ce secteur et qui finalement le traversent. Par ailleurs, ne disposant d'aucune information sur leurs instants prévus pour entrer dans les secteurs qui n'appartiennent pas à leur trajectoire prévue, il est difficile de déterminer une estimation de la loi de probabilité de ϵ. Notons qu'il s'agit pour certains, des aéronefs qui, pendant leur vol sont soumis aux contraintes de régulation, notamment le re-routement pour un écoulement efficace de trafic. Ce facteur de la charge secteur est attribuable à la complexité de gestion des courants de trafic.

Il apparaît que pour une meilleure prévision de la charge des secteurs, une modélisation fine de la loi de probabilité des résidus ϵ est nécessaire. Pour y parvenir, nous utilisons une approche détaillée dans la méthodologie ci-dessous.

2.2 Méthodologie

2.2.1 Formulation du problème

Nous nous proposons dans cette thèse, de déterminer des modèles probabilistes d'incertitude sur la trajectoire de vol des aéronefs en fonction des données opérationnelles de trafic. C'est-à-dire, pour un aéronef observé à un instant t_o à une position $M(t_o)$ de sa trajectoire réelle, les modèles que nous développons doivent permettre de prévoir au moyen d'une fonction de densité de probabilité, l'instant t de passage de cet

aéronef en un point futur $M(t)$ de sa trajectoire prévue. Il s'agit de proposer une solution plus générale de l'équation (1) précédente. Car, connaissant l'instant d'entrée de chaque aéronef dans un secteur de l'espace, on peut en déduire le nombre total d'aéronefs qui entrent dans ce secteur pendant un intervalle de temps donné. Finalement, notre modèle à résoudre peut être formulé par l'équation (4) :

$$t = t^p + \Phi(\mathbf{X}) + \epsilon. \tag{4}$$

, où :

- t est l'instant de passage d'un aéronef au point M de sa trajectoire prévue.
- t^p est l'instant prévu dans le plan de vol pour le passage de cet aéronef au point M. Cet instant est connu à partir des plans de vols déposés[1].
- $\mathbf{X}$ est un vecteur de variables explicatives de la différence $t - t^p$ à prévoir. 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.
- Φ est une fonctionnelle à ajuster en fonction de $\mathbf{X}$.
- ϵ est une composante résiduelle.

2.2.1.1 Plan de vol d'origine et plan de vol simulé par *OPERA*

Lorsqu'un aéronef 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. C'est en réalité une « balise fictive » de ce plan de vol. Les autres 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 aéronefs de la base BADA (Base of Aircraft Data) d'EUROCONTROL. Il s'agit des paramètres nominaux des aéronefs. Les nouveaux instants prévus pour survoler le reste des points balises ont été calculés en initialisant le temps à t_0 au point $M(t_0)$. Ces instants sont notés t^p. Dans la suite, lorsque nous parlerons de plan de vol ou de trajectoire prévue, nous ferons référence à cette trajectoire simulée par *OPERA*.

2.2.1.2 Variables explicatives $\mathbf{X}$

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

- *Plans de vols* : Il s'agit des points balises à traverser, les heures de survol de ces balises, les secteurs à traverser, les heures d'entrée et de sortie de chaque secteur, 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 le vol prévu.

- *Conditions météorologiques et atmosphériques* : Il s'agit des prévisions du vent (intensité et direction), la température, la pression, la densité de l'air et de l'humidité spécifique sur le reste de la trajectoire prévue à partir du point courant $M(t_o)$.
- *Complexité du trafic* : C'est la complexité du trafic sur le reste de la trajectoire 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 l'hétérogénéité des performances entre les aéronefs.
- *Paramètres courants* : Les caractéristiques courantes du vol à l'instant t_o comprennent la vitesse, le taux de montée ou de descente, 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 explicatives est présentée en annexe de ce mémoire dans les tables (TAB.B.1, TAB.B.2, TAB.B.3, TAB.B.4).

2.2.1.3 Modèle à estimer

L'equation (4) définie précédemment peut de façon équivalente s'écrire :

$$t - t^p = \Phi(\mathbf{X}) + \epsilon. \tag{5}$$

Pour un aéronef et un point M de sa trajectoire de vol prévue, la différence « $t - t^p$ » est l'écart entre l'instant t de passage réel de cet aéronef 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 du mémoire, 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 :

$$\widehat{\triangle\tau} = \Phi(\mathbf{X}). \tag{6}$$

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, on peut construire un tel estimateur en minimisant les écarts résiduels par la méthode des moindres carrés.

Comme il est ardue, voire impossible de disposer de tous les points constitutifs de la trajectoire réelle d'un aéronef, sans nuire à la généralité, 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'aéronef. Nous présentons au paragraphe suivant les méthodes statistiques que nous utilisons.

2.2.2 Méthodes utilisées

Afin de proposer un modèle de prévision de l'écart temporel, nous utilisons une classe de modèles statistiques basée sur l'apprentissage automatique des données. Il s'agit des méthodes de partitionnement récursif.

Dans un premier temps, nous utilisons la méthode de régression par arbre CART. Ensuite, nous proposons une amélioration de la qualité de prévision de cette méthode en proposant une nouvelle version de ce modèle que nous appelons « CART modifiée ». Dans cette version, nous remplaçons l'étape de prévision utilisant la moyenne des classes par une phase où à l'intérieur de chaque classe, un modèle de régression linéaire généralisé est construit pour la prévision de la variable à expliquer en fonction des paramètres pertinents des vols pour la classe. Ensuite, nous utilisons une trosième méthode appelée MARS (Multivariate Adaptative Regression Splines). Elle vise à corriger au moyen des splines de régression, la discontinuité des prévisions du modèle CART entre les nœuds (classes) adjacents. Nous tirons parti de cette méthode pour évaluer les effets des interactions impliquant plusieurs variables explicatives sur l'écart temporel. Enfin, un quatrième modèle basé sur les forêts aléatoires (FA) est construit. Cette technique fondée sur les méthodes bootstrap permettra d'apporter une correction à l'instabilité des arbres de régression inhérente à la méthode CART.

Finalement, la fonction de densité de la loi de probabilité des résidus de chacun des modèles proposés est déterminée en utilisant l'algorithme EM (Expectation-Maximization) de mélange de lois Gaussiennes. Avant de commencer la modélisation preprement dite, présentons d'abord quelques modèles existants dans le cadre de la prévision de la trajectoire des avions.

2.3 Modèles de trajectoires existants

La prévision de la trajectoire d'un avion consiste à déterminer quel sera sa position dans l'espace à un instant futur de son mouvement. Elle utilise un ensemble de paramètres connus sur les types d'avions. A ce jour, plusieurs modèles ont été proposés et les méthodes utilisées sont aussi variées. Par exemple, on utilise les simulations des avions en vol pour déterminer les prévisions de leur trajectoire. Cette approche par simulation a l'avantage de permettre l'évaluation du trafic avec un nombre très important d'avions et présente malheureusement l'inconvénient d'être coûteux car elle nécessite beaucoup de paramètres pour chaque type d'avion intervenant dans le processus de simulation.

2.3.1 Le modèle énergie totale

Le modèle de description de la trajectoire le plus connu est le modèle énergie totale TEM (Total Energy Model). Ce modèle est régi par trois équations fondamentales de la dynamique :

- La première porte sur l'équilibre de l'avion dans le plan vertical où la portance doit compenser le poids, la composante verticale de la force aérodynamique est telle que $F_z = m.g$. Dans cette relation, z fait référence à l'altitude où se trouve l'avion. Cette relation permet de calculer la traînée correspondante.
- La seconde est l'équation de la conservation de l'énergie : la somme de l'énergie cinétique et l'énergie potentielle est égale à l'énergie appliquée à l'avion. L'énergie que ce dernier reçoit est égale à celle fournie par ses moteurs (force de poussée : F) moins celle perdue sous forme de traînée T. L'équation de synthèse peut s'écrire :

$$m.g.z + \frac{1}{2}mV^2 = (F - T).l, \tag{7}$$

  où l est la longueur du déplacement effectué, soit encore

$$m.g.\frac{dz}{dt} + m.V\frac{dV}{dt} = (F - T).V. \tag{8}$$

- La troisième est liée au plan horizontal et permet d'évaluer l'évolution du CAP de manière suivante :

$$\psi_{t+\Delta t} = \psi_t + \frac{d\psi}{dt}.\Delta t, \tag{9}$$

  où ψ_t est le cap de l'avion à l'instant t.

2.3.2 Modèles paramétriques et non-paramétriques

Afin de construire des modèles de prévision de trajectoires moins dépendants des paramètres des aéronefs, les modèles mathématiques paramétriques et d'autres non-paramétriques ont fait l'objet d'importantes études. Dans ce cadre, C. Bontemps (1997) [7] propose dans ses travaux deux modèles de prévision de trajectoires. Son approche consiste à trouver la courbe qui correspond le mieux aux points passés afin d'estimer la position du suivant. Cette méthode exige de disposer d'un ensemble de points observés pour générer cette courbe définie par divers coefficients de pondédration.

- Dans le cadre des modèles paramétriques, on étudie une trajectoire définie par la suite $(Y_i)_{i=1,\ldots,N}$ des points par lesquelles est passé un mobile au cours du temps. Ces méthodes ont pour but de trouver une fonction $f(t, \theta)$, où θ est un paramètre vectoriel à estimer, telle qu'elle corresponde le mieux possible aux données observées Y_i à des instants T_i (Delecroix, 1983) [23]. Il s'agit donc de minimiser au sens des moindres carrés, les écarts entre l'ensemble des points réels et la courbe de $f(t, \theta)$. Cela signifie qu'il faut trouver la valeur du paramètre $\widehat{\theta}_N$ telle que :

$$\widehat{\theta}_N = \arg\min_{\theta} \sum_{i=1}^{N} (Y_i - f(T_i, \theta))^2. \tag{10}$$

En optimisant ce critère, les modèles suivants ont été construits pour l'estimation de la fonction f pour le paramètre $\theta = (\alpha, \beta)$:
- Le modèle linéaire : $f(t, \theta) = \alpha + \beta.t$;
- Le modèle racine : $f(t, \theta) = \alpha + \beta.\sqrt{t}$;
- Le modèle multiplicatif : $f(t, \theta) = \alpha.t^{\beta}$;
- et le modèle logarithmique : $f(t, \theta) = \alpha + t^{\beta}$.

En utilisant ces modèles pour la prévision de la trajectoire de montée des avions sur un horizon temporel inférieur à 600 secondes (10 minutes), Bontemps trouve que les intervalles de confiance ne contiennent pas souvent la trajectoire réelle. Ainsi, il développe une nouvelle méthode de prévision de la trajectoire d'aéronefs basée sur les modèles non-paramétriques.

- En condérant cette fois les couples observés $(Y_i, T_i)_{i=1,\dots,N}$, la méthode paramétrique consiste à trouver un estimateur $\widehat{f}$ de la fonction f tel que pour tout t proche des valeurs T_i, la quantité $\widehat{f}(t)$ soit proche de Y_i. Dans (Courot et *al.*, 1984) [4], la forme générale d'un estimateur non-paramétrique d'une fonction f s'écrit sous la forme :

$$\widehat{f}(t) = \sum_{i=1}^{N} W(T_i, t).Y_i, \tag{11}$$

sous la contrainte

$$\widehat{f}(t) = \sum_{i=1}^{N} W(T_i, t) = 1, \tag{12}$$

où $W(T_i, t)$ est le poids associé à l'observation Y_i à l'instant t. Pour satisfaire à la condition (12), on pose :

$$W(T_i, t) = \frac{w(T_i, t)}{\sum_{k=1}^{N} w(T_k, t)}. \tag{13}$$

La forme de l'estimateur oriente le choix des valeurs de $w(T_k, t)$. Une méthode consiste à prendre en compte tous les points mais en limitant leur influence en fonction de leur distance par rapport au point courant. On utilise les coefficients de pondération du type :

$$w(T_i, t) = \Phi\left(\frac{T_i - t}{h}\right), \tag{14}$$

où Φ est une fonction continue dont les valeurs diminuent lorsque $|T_i - t|$ augmente, h est un paramètre de lissage qui permet de contrôler la façon dont s'effectue cette diminution. Le noyau de Gauss est souvent utilisé pour définir la fonction Φ :

$$\Phi(T_i, t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \tag{15}$$

En utilisant cette fonction de noyau (15) gaussien pour la prévision de trajectoires d'aéronefs, Bontemps propose un estimateur du type :

$$\begin{aligned}
\widehat{f}(t) &= \sum_{i=1}^{N} W(T_i,t).Y_i \\
&= \sum_{i=1}^{N} \frac{w(T_i,t)}{\sum_{k=1}^{N} w(T_k,t)}.Y_i \\
&= \sum_{i=1}^{N} \frac{\Phi\left(\frac{T_i-t}{h}\right)}{\sum_{k=1}^{N} \Phi\left(\frac{T_k-t}{h}\right)}.Y_i \\
&= \frac{\sum_{i=1}^{N} \Phi\left(\frac{T_i-t}{h}\right).Y_i}{\sum_{i=1}^{N} \Phi\left(\frac{T_i-t}{h}\right)}.
\end{aligned}$$

Cet estimateur a l'avantage d'être défini en tout point et en particulier pour les i supérieurs à N. Avec cet estimateur, la prévision en un point est un barycentre des données observées. Ainsi, toute valeur prévue par cette méthode ne peut dépasser la valeur la plus grande, ni être plus faible que la plus petite. Ce qui constitue un réel handicap si la trajectoire à prédire est une trajectoire de montée ou de descente. Pour contourner ces problèmes, Yann Le Fablec (1999) [26] propose d'autres méthodes innovantes de prévision des trajectoires d'aéronefs basées sur les réseuax de neurones.

Dans ses travaux, Yann Le Fablec a développé une méthode de prévision de trajectoires d'aéronefs utilisant les réseaux de neurones. Celle-ci consiste à faire apprendre à ces réseaux un ensemble de trajectoires avant de les utiliser pour en prévoir les nouvelles. Dans son modèle, trois méthodes différentes sont proposées. Elles permettent une prévision à long terme dans le plan vertical et à court et à moyen terme dans le plan horizontal. Il montre que la prévision dans le plan vertical est précise tout en ne nécessitant que très peu de données initiales. L'une de ces méthodes est capable d'intégrer de nouvelles données au fur et à mesure que l'avion vole, ce qui lui permet de réagir aux possibles changements de trajectoire et ainsi d'améliorer la qualité des prévisions. Sur le plan horizontal, il montre que la méthode atteint ses limites dans la mesure où elle ne permet pas de réaliser les prévisions à long terme si on ne connait pas la route suivie par l'avion. L'avantage de sa méthode est qu'elle ne nécessite pas la connaissance d'un plan de vol et permet d'être moins dépendants d'informations non accessibles ou mal connues.

Lymperopoulos et *al.* (2006)[46] ont développé un algorithme de prévision de trajectoires d'aéronefs pendant la phase de décollage. Pour un vol donné, leur algorithme génère plusieurs trajectoires possibles. Il utilise ensuite les données radar du début de la trajectoire du vol pour choisir parmi les trajectoires candidates celle qui tend à prévoir le mieux la trajectoire courante du vol. En utilisant des simulations, ils montrent que cette méthode fournit des prévisions satisfaisantes sur un horizon temporel de 20 minutes après le décollage de l'avion.

2.4 Conclusion

L'ensemble des modèles présentés ci-dessus permettent d'améliorer la prévision des trajectoires d'avions. Nous notons cependant qu'ils ne prennent pas toujours en compte l'environnement des vols en temps réel. Par exemple les conditions météorologiques : le vent, la température, la pression atmosphérique et la densité de l'air ne sont pas considérés alors qu'on sait de façon théorique que ces paramètres peuvent affecter le mouvement des avions pendant leur vol. Par ailleurs, ces modèles ne sont pas toujours évalués sur leur capacité à prévoir la charge de travail des contrôleurs et de la quantité de conflits potentiels dans un secteur de contrôle.

Pour les systèmes d'aide à la gestion et l'organisation des courants de trafic, il est nécessaire de développer les modèles de prévision des trajectoires dynamiques en intègrant à la fois les aspects de la complexité de trafic et de son environnement météorologique. Ainsi, notre approche dans cette thèse est différente des méthodes présentées ci-dessus. Nous nous plaçons dans les conditions où les aéronefs suivent leur trajectoire de vol prévue dans les plans de vols. Notre méthode va consister à utiliser l'ensemble des caractéristiques des aéronefs à un instant donné pendant leur vol pour prévoir le moment de leur passage sur les points futurs de leur trajectoire. Finalement, les modèles développés seront ensuite évalués sur leur capacité à prévoir la charge de l'espace en nombre d'aéronefs entrant dans les secteurs.

Chapitre 3

Données utilisées pour l'analyse

Les données utilisées dans cette étude proviennent de deux sources : Les données de trafic aérien ont été fournies par EUROCONTROL et celles liées aux conditions météorologiques et atmosphériques ont été fournies par Météo-France. L'ensemble de ces données couvrent une même période de mesure (septembre 2007) sur le même domaine de l'espace $20N - 72N$ et $32W - 42E$. L'objectif de ce chapitre consiste à extraire de cette masse de données les caractéristiques pertinentes susceptibles d'expliquer, mais aussi de calculer, au travers des modèles probabilistes, l'écart de temps de passage des aéronefs en des points de leurs trajectoires de vol prévues, par rapport au temps de passage prévu dans les plans de vol. Pour chaque variable explicative construite, nous lui donnerons un « nom de code » qui sera utilisé partout dans ce mémoire de thèse en référence de celle-ci. Les noms de code des variables sont chaque fois écrits en italique.

3.1 Les données du trafic aérien

Ce sont les données d'archives de plans de vols déposés et les données radar dites CPR (Correlated Position Report) pour l'ensemble des aéronefs qui ont utilisé l'espace aérien, recueillies par EUROCONTROL pour le mois de septembre 2007. Elles correspondent aux mesures de la période d'AIRAC_299 (Aeronautical Information Regulation and Control : Contrôle et diffusion des renseignements aéronautiques). Nous détaillons ci-dessous chaque type de données.

3.1.1 Plans de vols

Un plan de vol (PLN ; flight plan [52]) est un ensemble de renseignements spécifiés au sujet d'un vol projeté ou d'une partie de vol, déposé et transmis par les compagnies aériennes ou les pilotes aux services de la circulation aérienne. Une fois le plan de vol déposé, trois états successifs du vol sont enregistrés et archivés par EUROCONTROL :

- FTFM : C'est le plan de vol initial.

- RTFM : C'est le plan de vol régulé comprenant les modifications éventuelles imposées par la CFMU (Central Flow Management Unit).
- CTFM : C'est le plan de vol final suivi par l'aéronef. C'est un plan de vol particulier parce que c'est l'état du vol effectivement observé.

De façon générale, un plan de vol contient des informations telles que : les aéroports de départ et de destination, l'heure de départ du bloc de stationnement, l'heure d'arrivée prévue à l'aéroport de destination, l'indicatif du vol, le numéro tactique du vol, la compagnie aérienne exploitant le vol, le type d'aéronef utilisé, la route prévue définie par un ensemble de balises à survoler, le niveau de vol de croisière demandé, etc... Les caractéristiques des plans de vol ont ainsi été utilisées pour construire certaines variables explicatives des modèles que nous proposons. Ces variables sont consignées dans la table TAB.B.1 en annexes avec leurs noms de code et leurs descriptions respectives.

3.1.2 Données courantes du vol

Il s'agit ici des caractéristiques des vols en temps réel. En effet, les données courantes du vol désignent l'ensemble des paramètres des aéronefs au moment où on veut réaliser les prévisions de leurs instants de passage en des points de leurs trajectoires de vol prévues. Ces paramètres sont disponibles grâce aux données d'archives des trajectoires des vols effectivement réalisées et enregistrées par radar dans le système CPR. Dans ce système, les positions radar $4D$ des vols sont renseignées toutes les minutes. Ainsi, pour un vol donné, les informations suivantes sont disponibles : les aéroports de départ et de destination, la position du vol (latitude, longitude, altitude, temps), la vitesse, le cap (par rapport au nord géographique), le taux de montée ou de descente de l'aéronef, le profil du vol (stable ou en évolution), l'indicatif du vol et le numéro du vol dans le système TACT d'EUROCONTROL. Les variables explicatives construites à partir de ces caractéristiques courantes du vol sont consignées dans TAB.B.2 en annexes.

3.2 Informations communes aux données météorologiques et atmosphériques

Les données météorologiques et atmosphériques utilisées dans cette étude proviennent de Météo-France. Ce sont des prévisions fournies par le modèle ARPEGE. Ces prévisions sont actualisées toutes les 6 heures, soit à 00 heure, 06 heures, 12 heures et 18 heures. Ces prévisions sont disponibles par niveau isobare sur le domaine $20N - 72N$ et $32W - 42E$ maillé avec une résolution de 0.25 degré, soit approximativement 15 NM. C'est la résolution la plus fine utilisée par Météo-France pour la prévision des données météorologiques destinées au trafic aérien. Ces informations météorologiques et atmosphériques sont disponibles sur dix niveaux isobares, notamment sur : 200, 250, 300, 400, 500, 600, 700, 850, 900 et 925 hPa par rapport au niveau de pression standard. Ces niveaux isobares correspondent respectivement aux niveaux de vol $FL390$, $FL340$, $FL300$, $FL240$, $FL180$, $FL140$, $FL100$, $FL050$, $FL030$ et $FL025$.

Les paramètres météorologiques et atmosphériques disponibles sont : la force du vent, la direction du vent, la pression de l'air, la température, la densité de l'air et l'humidité spécifique. Pour chaque point du vol

appartenant exactement à un des dix niveaux de vol ci-dessus, la valeur prévue du paramètre météo est celle prévue sur ce niveau de vol au voisinage de ce point. En revanche, si ce point est situé entre deux niveaux de vol, nous utilisons un modèle d'interpolation linéaire pour déterminer une approximation de la valeur du paramètre en ce point à partir des prévisions du paramètre disponibles sur les niveaux de vol supérieur et inférieur.

3.3 Données météorologiques

3.3.1 Influence du vent

Un aéronef en vol est soumis à des mouvements de la masse d'air par rapport au sol, c'est l'effet du vent sur l'aéronef par rapport au sol. En fonction de la direction et de la vitesse du vent, la vitesse de l'aéronef par rapport au sol peut être plus ou moins affectée selon le cap suivi. La relation entre la vitesse de l'aéronef et la vitesse du vent (toutes par rapport au sol) est exprimée par l'équation : $\vec{V_s} = \vec{V_p} + \vec{V_w}$. Ainsi, en projetant la composante vent effectif sur la trajectoire du vol, on obtient la vitesse du vol par rapport au sol $V_s = V_p + V_w \cos(\alpha)$, où α est l'angle au vent, orienté entre la direction d'où vient le vent et la direction de la route suivie par l'aéronef. V_p est la vitesse propre de l'aéronef (TAS ; True airspeed), c'est-à-dire, sa vitesse par rapport à la masse d'air dans laquelle il évolue.

Connaissant une portion de la trajectoire de vol prévue, posons : *TAV* le temps de vol prévu avec vent (par rapport au sol) en supposant que le pilote ne modifie pas sa vitesse propre pour contrer l'effet du vent et *TSV* le temps de vol prévu sans vent (par rapport à la masse d'air). Nous avons défini l'indicateur d'influence du vent que nous notons dans cette thèse par « *Indur* », comme rapport du temps de vol prévu avec vent sur le temps de vol prévu sans vent :

$$Indur = \frac{TAV}{TSV} \tag{1}$$

Intuitivement, nous voyons que Indur égal à 1 signifie que l'influence du vent sur la portion de la trajectoire concernée est globalement nulle. Si Indur est supérieure à 1, cela signifie que le temps de vol avec vent (TAV) est supérieur au temps de vol sans vent (TSV). Potentiellement, le vent prévu sur la trajectoire de vol concerné est de face et allonge le temps de vol. Enfin, Si Indur est inférieure à 1, cela signifie que le temps de vol avec vent (TAV) est inférieur au temps de vol sans vent (TSV). Potentiellement, le vent prévu sur la trajectoire de vol est de dos et pourrait accélérer l'aéronef dans son mouvement pendant le vol.

3.3.2 Généralisation de l'influence du vent sur la trajectoire de vol prévue

Soit une trajectoire de vol prévue, constituée par une suite de petites portions $(M_i M_{i+1})_{(0 \leq i < n)}$, où M_0 et M_n sont les extrémités de cette trajectoire et M_0 est en particulier le point courant du vol. Nous avons généralisé l'indicateur d'influence du vent prévu sur toute la trajectoire du vol par :

$$Indur = \frac{\sum_{i=0}^{n} TAV_i}{\sum_{i=0}^{n} TSV_i} \tag{2}$$

où TAV_i et TSV_i sont les temps de vol prévus respectivement avec vent et sans vent, sur la portion de la trajectoire élémentaire limitée par les points M_i et M_{i+1}. Dans le cadre des modèles statistiques que nous développons, cet indicateur d'influence du vent est une variable explicative écart temporel entre l'instant de passage d'un aéronef en un point de sa trajectoire prévue et l'instant de passage en ce point prévu dans le plan de vol. Elle mesure l'influence du vent entre la position courant de l'aéronef et un point de sa trajectoire prévue. Elle est consignée dans la TAB.B.3.

3.4 Paramètres de l'air atmosphérique

Dans cette étude, nous voulons déterminer les effets de ces paramètres sur la variable écart temporel. Pour cela, rappelons d'abord la composition de l'air.

3.4.1 Composition de l'air sec

L'air atmosphérique est un mélange de deux gaz, l'air sec d'environ 99% et la vapeur d'eau d'environ 1%. L'air sec lui-même n'est pas un gaz simple, sa composition volumétrique est d'environ : 78.09% d'azote, 20.95% d'oxygène, 0.93% d'argon, 0.026% d'oxyde de carbone et les traces d'autres gaz rares. Pour plus de détails, voir Compléments de météorologie de J. LEPAS (1973) [44]. Elle est fonction de la température, de la pression et de la quantité d'eau en vapeur qu'il contient. C'est une grandeur sans unité de mesure. Les effets de l'air sur les paramètres de performance des aéronefs varient selon la densité de l'air (BADA ; 2009/003) [1].

3.4.2 Action de l'air sur les corps en mouvement dans l'atmosphère

De façon générale, un corps en mouvement dans l'air est soumis à la résistance de l'air. L'intensité de cette résistance est fortement liée à la densité de l'air. Plus l'air est dense, plus il va imposer une force de résistance à l'avancement d'une grande intensité à tout objet en mouvement à travers lui. Pour un aéronef en vol, cette résistance constitue la composante horizontale de la résultante aérodynamique. Elle s'appelle traînée, et s'oppose à l'avancement de l'aéronef sur sa trajectoire. Si l'air est peu dense, la traînée devient faible, ce qui améliore la résultante de la force de traction. Cependant, Jack Williams (2005) [63] montre que la perte des performances des aéronefs due à la baisse de la densité de l'air a plus d'importance et ne peut être compensée par le gain que cette baisse induit sur la réduction de la traînée. Pour plus de détails sur la densité de l'air et les performances d'un aéronef, le lecteur peut consulter Jack Williams (2003) [62].

3.4.3 Altitude densité

L'altitude densité est un indicateur souvent utilisé par les pilotes pour établir le lien entre la densité de l'air et les performances aérodynamiques des aéronefs (Jack Williams ; 2003).

3.4.3.1 Définition et description de l'altitude densité

Pour un aéronef dont la densité de l'air à sa position courante de vol est ρ, son altitude densité en ce point est la valeur de l'altitude par rapport à l'atmosphère standard où la densité de l'air est aussi ρ.

- Supposons qu'un aéronef en vol se trouve à une altitude où il fait chaud. Potentiellement, il y'a peu de molécules de gaz par unité de volume constituant l'air. Le comportement aérodynamique de cet aéronef est semblable à celui d'un aéronef en vol à une altitude supérieure relativement à l'atmosphère standard. Ainsi, l'aéronef vole en conditions de haute altitude densité. Donc, dans les conditions de faible densité de l'air.
- Si en revanche, l'aéronef en vol se trouve à une altitude où il fait froid, potentiellement, il y'a plus de molécules de gaz par unité de volume constituant l'air. Le comportement aérodynamique de cet aéronef est semblable à celui d'un aéronef en vol à une altitude inférieure relativement à l'atmosphère standard. Ainsi, l'aéronef vole en conditions de basse altitude densité. Donc dans les conditions de forte densité d'air.

Ainsi, les performances aérodynamiques d'un aéronef varient selon que l'altitude densité est élevée (l'air est moins dense) ou basse (l'air est plus dense).

3.4.3.2 Effets théoriques de la densité de l'air sur les performances d'un aéronef en vol

Les conditions de haute altitude densité réduisent les performances des aéronefs. En effet, l'air est moins chargé des molécules des gaz. Il est plus léger par rapport aux conditions de l'atmosphère standard et affecte les performances des aéronefs de façon suivante :

- Le moteur d'aéronef a un déficit d'air pour entretenir la combustion, sa puissance s'en trouve ainsi réduite.
- Les hélices ont peu d'air pour soutenir la propulsion de l'aéronef dans son déplacement, la poussée est ainsi réduite par rapport aux conditions normales.
- Pour les jets, avions à réaction, les réacteurs éjectent une faible quantité de gaz et la poussée s'en trouve réduite.
- A cause de l'insuffisance des molécules de gaz dans l'air, la force exercée par l'air sur les ailes perd de son intensité, et il en résulte la diminution de la force de portance.
- Pour les aéronefs en croisières, la faible densité de l'air diminue la trainée et donc permet de voler à de grandes vitesses pour une consommation raisonnable.
- Pour les avions au décollage ou à l'atterrissage, la réduction de la poussée et de la portance exige de la part des pilotes et des contrôleurs une attention particulière. En effet, pour compenser le déficit

d'air sur les performances des aéronefs, la distance parcourue sur la piste au moment du décollage ou de l'atterrissage est plus longue. Ainsi, les pilotes doivent connaître la longueur des pistes, la hauteur des arbres et le paysage immédiatement proches de ces pistes pour un meilleur contrôle du taux de montée ou de descente (Jack Williams, 2003).

Connaissant les effets théoriques de l'altitude densité de l'air sur les paramètres de performances des aéronefs, nous définissons ci-dessous un indicateur équivalent pour l'évaluation statistique de l'impact de la variation de la densité de l'air sur les performances des aéronefs. Nous le baptisons : écart de densité de l'air.

3.4.4 Ecart de densité de l'air

Dans le cadre des modèles statistiques proposés dans cette thèse, l'écart de densité de l'air est utilisé en lieu et place de l'altitude densité. Nous le définissons comme la différence entre la densité de l'air observée à l'altitude courante du vol, et la densité de l'air à la même altitude relativement aux conditions de l'atmosphère standard. Pour l'interprétation de l'écart de densité de l'air, nous nous plaçons ici dans des conditions de chaleur ou de froid :

- Supposons que l'aéronef en vol se trouve à une altitude où il fait chaud. L'air est moins dense. L'aéronef se comporte de la même manière que s'il se trouvait à une altitude supérieure en conditions d'atmosphère standard. A l'altitude réelle de l'aéronef, la densité de l'air observée est inférieure à la densité de l'air en conditions d'atmosphère standard. Ainsi, l'écart de densité de l'air est négatif et l'air est pauvre en molécules de gaz et peut entraîner une diminution de la résultante de la traction de l'aéronef.
- Supposons que l'aéronef en vol se trouve à une altitude où il fait froid. L'air est plus dense. L'aéronef se comporte de la même manière que s'il se trouvait à une altitude inférieure en conditions d'atmosphère standard. A l'altitude réelle de l'aéronef, la densité de l'air observée est supérieure à la densité de l'air en conditions de l'atmosphère standard. Ainsi, l'écart de densité de l'air est positif et l'air est riche en molécules de gaz et peut entraîner une amélioration de la résultante de la force de traction de l'aéronef.

Nous notons Moydensa, la variable explicative construite à partir de l'écart de densité de l'air. Elle est définie dans TAB.B.3 en annexe comme l'écart moyen de densité de l'air sur la trajectoire de vol prévue. Connaissant les effets théoriques de la densité de l'air sur le comportement aérodynamique des aéronefs, nous présentons ci-dessous quelques relations qui existent entre cette densité et les autres paramètres de l'air atmosphérique comme la température et la pression.

3.4.5 Température et pression de l'air

La pression et la température ont des effets opposés sur la densité de l'air. Ces paramètres affectent les aéronefs en vol à travers la variation de l'altitude de vol. Les variables explicatives construites à partir de la pression et de la température sont notées respectivement *Moypres* et *Moytemp* et consignées dans TAB.B.3. *Moypres* est l'écart moyen entre la pression prévue et la pression de l'atmosphère standard sur la trajectoire prévue du vol. De même, *Moytemp* est l'écart moyen entre la température prévue et la température de l'atmosphère standard sur la trajectoire prévue du vol.

3.4.5.1 Effet de la température sur la variation de niveau de vol

Quand la température de l'air est supérieure à la température standard, l'aéronef en vol est en réalité à un niveau de vol supérieur à celui indiqué par l'altimètre à bord par rapport au niveau de pression standard 1013.25 hPa. En revanche, quand la température est inférieure à la température standard, l'aéronef en vol est en réalité à un niveau de vol inférieur à celui indiqué par l'altimètre à bord par rapport au niveau de pression standard. La baisse de la température pendant le vol en route rend la vraie valeur courante du niveau de vol au dessous de celle indiquée par l'altimètre à bord. Ainsi, pour connaître sur quel niveau de vol l'aéronef se trouve réellement, les pilotes doivent procéder au calage de l'altimètre à bord par rapport au niveau de pression standard, 1013.25 hPa.

3.4.5.2 Effet de la pression sur la variation de niveau de vol

Lorsque la pression croît, la densité de l'air croît aussi. Le passage de l'aéronef en vol des hautes pressions vers les basses pressions sans correction de l'altimètre par les pilotes peut entraîner une surestimation de l'altitude du niveau de vol. C'est-à-dire que l'aéronef est en réalité au dessous de la valeur de l'altitude indiquée à bord relativement au niveau de pression standard. En revanche, l'aéronef peut gagner de l'altitude en passant de la zone des basses pressions vers la zone des hautes pressions. C'est-à-dire que l'aéronef est en réalité au dessus de la valeur de l'altitude indiquée par l'altimètre à bord par rapport au niveau de pression standard. Pour que l'altitude indiquée à bord soit la vraie altitude du niveau de vol, les pilotes doivent caler l'altimètre par rapport au niveau de pression standard 1013.25 hPa pour tenir compte des variations des niveaux de pression pendant le vol.

La densité, température et la pression atmosphérique sont des variables liées entre elles par la loi de Mariotte. Leurs effets concernent la mesure de l'altitude des niveaux de vol, les performances des moteurs, la résistance de l'air, donc la portance et la trainée, et en conséquence les taux de montée, la vitesse de croisière, la consommation de carburant, que le pilote veut pouvoir maîtriser. Les effets de ces variables explicatives sur le temps de passage des aéronefs en des points de leur trajectoire seront déterminés dans la section du développement des modèles statistiques.

3.5 Paramètres de complexité du trafic sur la trajectoire de vol prévue

La commission d'évaluation des performances de EUROCONTROL a défini un ensemble d'indicateurs de complexité qui servent de référence pour l'évaluation de la gestion des flux de trafic ATFM par les fournisseurs des services à la navigation aérienne (ANSP) en Europe (PRR ; 2005) [15]. La complexité du trafic est ainsi basée sur le concept d'« interaction ». Il y a interaction lorsque deux aéronefs en vol se trouvent dans une même zone de l'espace au même moment. Dans le cadre de notre étude, une interaction de vol est définie comme la présence simultanée de deux aéronefs dans une cellule de 20x20 Milles Nautiques et 3000 pieds de hauteur. Pour la construction des cellules on se reportera à l'annexe.

Définitions

Le « score de complexité du trafic » est le produit de la densité du trafic et de l'indice structurel du trafic. La « densité du trafic » est la mesure du nombre d'interactions potentielles entre les aéronefs.

3.5.1 Flux total d'aéronefs sur la trajectoire de vol

Le flux journalier de trafic pour une cellule donnée est la moyenne sur une journée de trafic du nombre d'aéronefs traversant cette cellule pendant chaque tranche d'une heure. Le flux total d'aéronefs pour un vol donné est la somme des flux journaliers de trafic sur toutes les cellules de la trajectoire de vol prévue, divisée par la distance de cette trajectoire. Elle est représentée dans TAB.B.4 dans l'annexe par la variable notée *Denscum1*. C'est en réalité un indicateur du niveau trafic rencontré par un vol sur sa trajectoire de vol. Il est exprimé en nombre de vols par heure et par mille nautique.

3.5.2 Interactions totales sur la trajectoire de vol

- L'interaction totale pour une cellule donnée est la moyenne sur une journée de trafic du nombre de couples d'aéronefs traversant simultanément cette cellule pendant chaque tranche d'une heure. Cette variable est notée *Inde*.
- Pour une trajectoire de vol donnée, l'indicateur d'interaction totale prévue est la somme des interactions totales sur toutes les cellules de la portion de la portion de la trajectoire de vol prévue, divisée par la distance de celle-ci. Cet indicateur est une variable explicative que nous notons Indecum1 dans TAB.B.4 en annexe. C'est une variable prévisionnelle.
- L'interaction horizontale pour une cellule donnée est la moyenne sur une journée de trafic du nombre de couples d'aéronefs traversant simultanément cette cellule avec une différence de caps supérieure à 20° pendant chaque tranche d'une heure. Lorsque la différence de caps entre deux aéronefs simultanément présents dans une zone de l'espace est supérieure à 20° sur le même plan horizontal, la PRU considère que ces aéronefs présentent un risque de conflit. Nous notons cette variable *Inho*.
- L'interaction verticale pour une cellule donnée est la moyenne sur une journée de trafic du nombre de couples d'aéronefs traversant simultanément cette cellule et se trouvent à des phases de vol différentes (montée, croisière ou descente) pendant chaque tranche d'une heure. Nous notons cette variable *Inve*.

- L'interaction de vitesse pour une cellule donnée est la moyenne sur une journée de trafic du nombre de couples d'aéronefs traversant simultanément cette cellule avec une différence de vitesses supérieure à 35 kts pendant chaque tranche d'une heure. Nous notons cette variable *Invi*. Lorsque la différence de vitesses entre deux aéronefs simultanément présents dans une zone de l'espace est supérieure à 35 kts sur le même plan horizontal, la PRU considère que ces aéronefs présentent un risque de conflit. Cette variable est notée Invi.
- Pour une cellule donnée l'indice structurel de trafic est la somme des trois types d'interactions : verticales, horizontales et de vitesse (PRU, 2005). C'est un facteur de score de complexité du trafic que nous notons *score*.

3.5.3 Score de complexité de trafic sur la trajectoire prévue

Puisque les interactions totales contiennent chaque type d'interactions, nous décomposons l'indice structurel du trafic dans la relation suivante :

$$score = Inho + Inve + Invi = Inde * [\frac{Inho}{Inde} + \frac{Inve}{Inde} + \frac{Invi}{Inde}].$$

On pose :

$$r_Inho = \frac{Inho}{Inde}, r_Inve = \frac{Inve}{Inde} \text{ et } r_Invi = \frac{Invi}{Inde},$$

les rapports des interactions horizontales, verticales et de vitesse respectives sur les interactions totales de la cellule. On obtient finalement :

$$score = Inde * [r_Inho + r_Inve + r_Invi] = Inde * rscore.$$

Désormais, nous appelerons score d'interaction relative pour une cellule, *rscore*, la somme des interactions relatives dans la cellule :

$$rscore = r_Inho + r_Inve + r_Invi.$$

Le score de complexité du trafic pour la trajectoire du vol que nous notons *Rscorecum1* est la somme des interactions relatives sur toutes les cellules de la trajectoire divisée par la distance de celle-ci. Ce score de complexité qui est une variable explicative des modèles à construire est disponible dans TAB.B.4 en annexe.

3.5.4 Densité du réseau de routes aériennes sur la trajectoire prévue

Ici, nous construisons une variable explicative de l'écart temporel. Ainsi, pour un aéronef donné, la densité du réseau de trafic notée *Moycrois*, est définie dans le cadre de cette étude, comme le nombre moyen des routes connectées sur chaque balise de la trajectoire de vol prévue. FIG.3.1 en est une illustration. Considérons a et b deux aéronefs qui décollent du même point D. L'aéronef a traverse les balises B_1, B_2 et B_3 pour atteindre sa destination A_1. Il traverse ainsi trois balises avec respectivement 3, 4 et 5 connections. La densité du réseau de trafic pour cette trajectoire est $\frac{3+4+5}{3} = 4$. L'aéronef b quant à lui traverse B_1 et B_4 pour atteindre sa destination A_2. La densité du réseau sur sa trajectoire est $\frac{3+4}{2} = 3.5$.

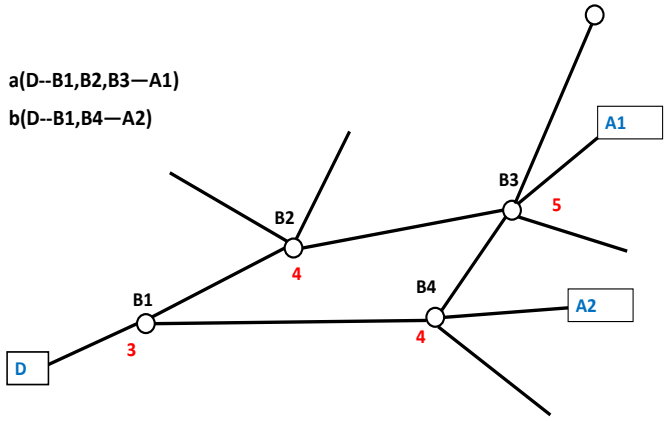

FIG. 3.1 – Illustration de la densité de réseau de trafic

3.6 Ecart temporel : variable à expliquer

Rappelons que la variable à expliquer écart temporel notée « *Ecarttemp* », mesure l'écart entre l'instant observé lors du passage d'un aéronef en un point de sa trajectoire, et l'instant de passage du vol en ce point prévu à l'aide du modèle de performances BADA. Elle est exprimée en secondes. Les modèles statistiques qui seront développés ensuite permettront de déterminer les effets respectifs des variables explicatives précédentes sur l'écart temporel.

3.7 Modèle théorique : Le simulateur du trafic aérien OPERA

3.7.1 Principe de fonctionnement

OPERA est un simulateur du trafic aérien développé au LICIT. Comme variables d'entrée, Il prend les plans de vols déposés par les compagnies aériennes, les données de performances des aéronefs, l'ensemble des balises, l'ensemble des secteurs et l'ensemble des aéroports. Les données de performances sont les données BADA fournies par EUROCONTROL. Elles décrivent notamment les performances des aéronefs à divers niveaux de vol. Les paramètres de performance décrits sont : le niveau de vol de croisière, la vitesse de croisière, les taux de montée et de descente, la consommation en carburant, la masse de l'aéronef au décollage. Comme sortie, le simulateur génère les trajectoires des vols en fonction des paramètres nominaux des types d'aéronefs utilisés. Pour chaque vol, on connait ses positions $3D$ (latitude, longitude, altitude) successives toutes les 10 secondes, le secteur courant où se trouve l'aéronef et la prochaine balise visée sur le reste de sa trajectoire.

La souplesse du simulateur de faire « décoller » un aéronef à partir d'une altitude quelconque de sa trajectoire avec prise en compte de ses paramètres de performances associés lui confère une place centrale dans

cette thèse. En effet, les trajectoires de vol obtenues de cette simulation servent de modèle théorique de référence pour la prévision en situation réelle du trafic, le temps de passage d'un aéronef en un point prévu de sa trajectoire. Ces trajectoires sont construites en utilisant les paramètres nominaux de performances comme la vitesse nominale, le taux de montée/descente, la consommation en carburant, la masse de l'aéronef au décollage.

Le processus de simulation ne prend pas en compte l'influence des conditions météorologique et atmosphérique prévues, ni la complexité du trafic sur la trajectoire de vol prévue.

3.7.2 Lien entre la trajectoire temps réel du vol et le modèle OPERA

3.7.2.1 Balise fictive

A un instant donné t_o, appelé instant courant de prévision, chaque aéronef en vol occupe sur sa trajectoire une position précise $M(t_o)$. Cette position est connue grâce aux enregistrements radar des positions successives des vols. C'est un point réel $4D$ (latitude, longitude, altitude, temps) observé sur la trajectoire au moment où le vol a lieu. Il est utilisé pour initialiser la portion de la trajectoire du plan de vol qui reste à parcourir une fois que l'aéronef s'est rendu en ce point. C'est à partir de ce point que nous utilisons le modèle théorique, c'est-à-dire le modèle implémenté dans le simulateur du trafic aérien OPERA pour simuler le reste de la trajectoire du vol. Nous faisons ainsi l'hypothèse qu'à ce point $M(t_o)$, la trajectoire la plus probable que l'aéronef doit suivre est toujours en cohérence avec les objectifs affichés dans son plan de vol. Dans cette étude, le point $M(t_o)$ est appelé « point courant de prévision ».

Pour simuler un vol à partir du point courant de prévision, on introduit dans OPERA une balise fictive en ce point et on fait partir le vol de cette balise à une altitude donnée. Ce point n'est pas une balise ordinaire du plan de vol d'origine. C'est seulement la position qu'occupe l'aéronef pendant son vol au moment où on réalise les prévisions du temps de son passage en des points futurs sur le reste de sa trajectoire de vol prévue.

3.7.2.2 Plan de vol défini à partir du point courant de prévision

A partir du point courant de prévision, on construit un nouveau plan de vol dont la première balise est une balise fictive introduite dans OPERA en ce point. La deuxième balise est celle du plan de vol initial visée lorsque l'aéronef survole la balise fictive. Le nouveau plan de vol est obtenu en ajoutant les balises du plan de vol d'origine qui suivent la balise visée. Ainsi, on obtient un nouveau plan de vol qui est une entrée du modèle théorique OPERA.

Ce nouveau plan de vol se distingue de celui d'origine par le fait que l'aéronef commence son vol non pas nécessairement à l'altitude de l'aéroport de départ (généralement 0 actuellement), mais à un niveau d'altitude quelconque. En utilisant comme entrée, ces nouveaux plans de vol, le simulateur de trafic aérien OPERA va à son tour générer les trajectoires $4D$ des vols. Ce sont les trajectoires théoriques qui serviront de référence pour la modélisation de l'incertitude sur le temps de passage des aéronefs aux entrées et sorties

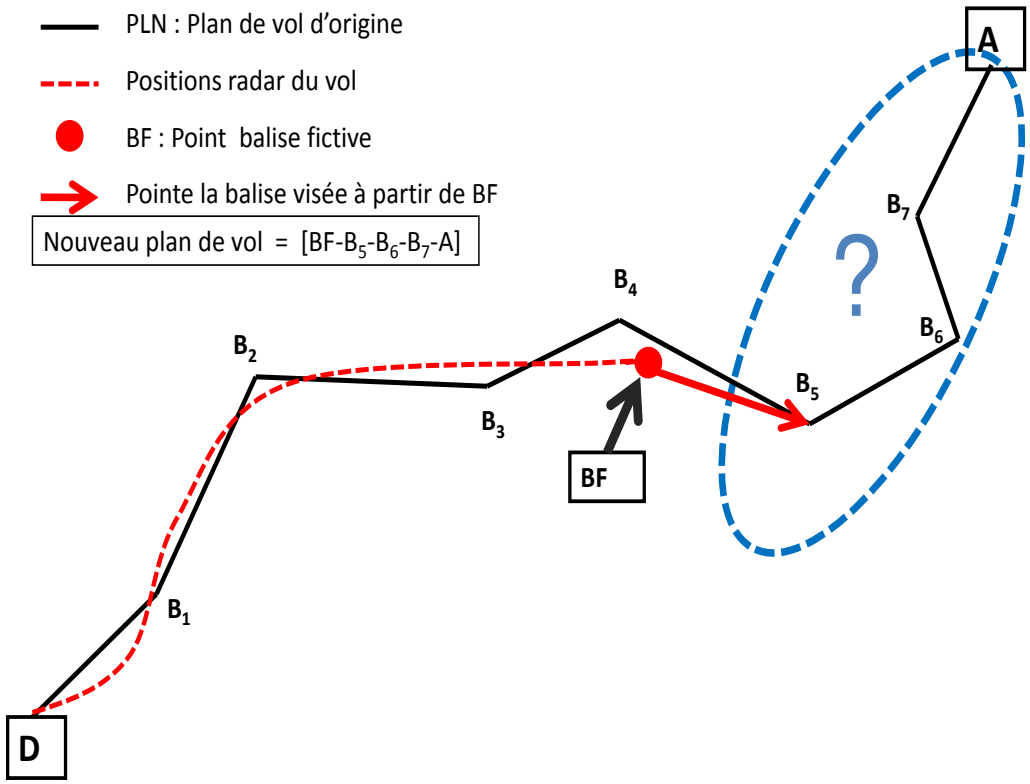

FIG. 3.2 – Illustration de la trajectoire du vol : Le trait plein noir entre les points D et A indique la trajectoire du vol prévue dans le dernier plan de vol déposé avant le départ de l'aéronef. La portion de la trajectoire en pointillés rouges du point D au point BF indique la trajectoire temps réel du vol. A partir du point BF, la portion de la trajectoire délimitée par l'ellipse en pointillés bleus indique les positions des points sur la trajectoire de vol en lesquels nous voulons prévoir le temps de passage de l'aéronef.

des secteurs de l'espace aérien traversés par leur trajectoire de vol prévue. La figure FIG.3.2 est un exemple de la trajectoire du plan de vol prévu, de la trajectoire radar (temps réel) du vol et de la trajectoire du nouveau plan de vol à suivre en temps réel à partir du point courant de prévision.

3.8 Conclusion

Ce chapitre nous a permis de passer en revue toutes les variables explicatives des vols à utiliser dans cette étude. Il s'agit des caractéristiques météorologiques et atmosphériques prévues sur les trajectoires de vol, les paramètres courants de vol (vitesse, taux de montée ou de descente, l'altitude courante, la distance entre le point courant de prévision et un point de la trajectoire prévue, etc...), les paramètres de complexité du trafic et les caractéristiques de performance des aéronefs utilisés tels que le type et le niveau de vol de croisière demandé par la compagnie. L'ensemble de ces variables explicatives sont utilisées dans les chapitres suivants pour construire des modèles probabilistes de prévision du temps de passage des aéronefs en des points des trajectoires de vol prévues.

Chapitre 4

Utilisation de la méthode CART pour la prévision de l'écart temporel de passage des aéronefs sur les points de leur trajectoire de vol prévue

4.1 Introduction

Dans ce chapitre, nous traitons du problème conjoint de classification et de prévision des écarts temporels de passage des aéronefs sur les points de leur trajectoire de vol prévue (écart temporel). Notre objectif principal consistera d'une part, à élaborer une typologie des points prévus sur les trajectoires des aéronefs en fonction des variables explicatives de l'écart temporel de passage des aéronefs en ces points, et d'autre part à déterminer la prévision de cette variable pour chaque groupe de la typologie. La méthode de classification par arbre de régression CART (Breiman et *al.*, 1984) semble la mieux adaptée pour répondre à ces objectifs. En effet, les arbres de régression sont réputés pour le partitionnement des échantillons de données en des groupes disjoints, dans lesquels un modèle de régression constant est ajusté pour la variable à expliquer.

Dans son principe, la méthode CART va permettre de définir une règle décisionnelle pour l'affectation de tout nouveau point de vol dans un groupe homogène des points des trajectoires de vol prévues. Cette affectation s'effectue en fonction des caractéristiques du point sur chacune des variables explicatives. La prévision de l'écart temporel pour ce nouveau point est l'estimation de cette variable par le modèle constant sur son groupe d'affectation.

Ce chapitre est organisé comme suit : Le paragraphe (4.2) est dédié aux principes de base de la méthode CART. Dans un premier temps, il fait un passage en revue de la litérature liée, le modèle de régression par arbre est spécifié, la procédure de construction de l'arbre optimal est présentée. Ensuite, nous définirons les concepts des divisions suppléantes et de l'importance des variables explicatives. Dans le paragraphe (4.3),

nous appliquons la méthode CART à la prévision de la variable écarts temporel. Ainsi, l'arbre de régression optimal est proposé, le diagnostic des résidus du modèle correspondant est réalisé et l'étude des erreurs de prévision par ce modèle en fonction de l'horizon temporel de prévision est réalisée. Dans le paragraphe (4.4), nous présentons une analyse des résultats du modèle obtenu. Les effets des variables explicatives actives sont analysés. Le paragraphe (4.5) conclut ce chapitre.

4.2 CART : Méthode de régression par arbre

4.2.1 Avantages et limites de la méthode

La méthode de régression par arbre CART fait partie de la grande famille des méthodes de discrimination par arbre binaire. Ces méthodes présentent plusieurs avantages. En effet, elles nécessitent moins d'hypothèses que les 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. Ces méthodes sont particulièrement adaptées dans les problèmes où les variables explicatives sont nombreuses. Elles sont robustes vis-à-vis des valeurs extrêmes et des données erronées, sélectionnent les variables les plus informatives avec la prise en compte des interactions entre les variables explicatives. Ces dernières peuvent être soit qualitatives ou soit quantitatives. La méthode est invariante à toute transformation monotone des données. Par exemple, si une variable explicative est transformée par la fonction logarithmique, l'arbre de régression obtenu avec la variable transformée est identique à l'arbre avec la variable de départ non transformée. Les algorithmes sont très rapides en phase de construction des arbres et lors du classement de la nouvelle observation. Grâce aux divisions les plus semblables à la meilleure division retenue, la méthode gère l'existence de données manquantes, aussi bien dans la construction de l'arbre et l'estimation de son coût que dans l'application de la règle à de nouveaux individus (Guéguen et *al.*, 1988) [2].

Le principal inconvénient à utiliser ces méthodes est l'instabilité des arbres obtenus. En effet, une légère modification des mesures des variables explicatives peut entraîner d'importantes perturbations dans la structure de l'arbre, voir Breiman (1996a)[8] et Ghattas (1999b) [33].

4.2.2 Revue de littérature

Les méthodes de régression par arbre CART ont été développées grâce aux travaux de Breiman L., Friedman J.H., Olshen R.A. et Stone C.J. (1984) [11]. Ces auteurs ont profondément repensé et redéfini la méthodologie sous-jacente aux arbres de discrimination. Leurs travaux se sont inspirés des méthodes de discrimination par arbre, plus couramment connues sous le nom de partitionnement récursif ou de segmentation, initialement développées par Messenger et Mandell (1972) et Morgan et Messenger (1973) [48] à la suite des travaux de Morgan et Sonquist (1963) [49] et Sonquist et Morgan (1964) [57] portant sur les arbres de régression. Depuis lors, la méthode CART (Classification And Regression Trees) fait référence à la méthode proposée par Breiman et *al.* (1984) [11]. Cette méthode a été étendue aux données censurées avec les travaux de Davis et Anderson (1989) [20], Kwak, Halperm, Olshen et Horning (1990) [40] et Leblanc

et Crowley (1992) [41]. En partant de CART, Gelfand, Ravishankar et Delp (1991) [32] ont proposé une méthode itérative qui construit et élague l'arbre alternativement. Enfin, Chou, Lookabaugh et Gray (1989) [13] ont étendu la méthodologie de CART à d'autres domaines en en donnant une réinterprétation.

Dans le cadre de cette étude nous nous intéressons à la version originale de la méthode CART développée par Breiman et *al.* (1984) [11] où un arbre de régression $\mathcal{A}$ permet de visualiser des variables dites actives qui participent directement à sa construction, et donc à la procédure de discrimination et de prévision correspondante.

4.2.3 Spécification du modèle de régression par arbre : CART

Les arbres de régression divisent l'espace des variables explicatives en un ensemble d'hypercubes (Josse et *al.*)[37]. Le modèle de régression par arbre est un modèle simple, additif et non paramétrique où une constante est ajustée sur chaque hypercube. La fonction de régression est de la forme :

$$y(x) = \sum_{i=1}^{n_0} c_i * \mathbb{1}_{\{x \in K_i\}} + \epsilon. \tag{1}$$

Les c_i sont des constantes et les K_i sont les classes qui constituent la partition de l'ensemble des individus de l'échantillon de données. ϵ est une variable résiduelle. Il s'agit dans ce modèle d'estimer la constante c_i dans chaque classe K_i, de déterminer de façon optimale le nombre n_0 des classes K_i de cette partition. X est un vecteur de variables explicatives du phénomène étudié et x est une réalisation de X. Une fois les K_i déterminés, le meilleur estimateur $\widehat{c}_i$ au sens des moindres carrés des c_i est la moyenne des mesures de la variable dépendante sur K_i. Le modèle final s'écrira :

$$\widehat{y}(x) = \sum_{i=1}^{n_0} \widehat{c}_i * \mathbb{1}_{\{x \in K_i\}}.$$

4.2.4 Illustration de la méthode

Les figures FIG.4.1 et 4.2 donnent une illustration de l'algorithme de construction d'un arbre de régression. Dans cet exemple, nous considérons un échantillon d'individus de taille $n = 60$. On veut prévoir les mesures de la variable dépendante Y en fonction de deux autres variables explicatives X_1 et X_2 en utilisant la méthode CART. Initialement, tous les individus sont placés dans le rectangle a que nous appelons nœud racine.

La première division du rectangle a porte sur la variable X_1. Cette division est définie par l'inégalité $X_1 < 134$. C'est la meilleure division parmi toutes les divisions admissibles sur ce nœud. La valeur 134 est le point de division du rectangle a en deux sous rectangles a_G et a_D suivant la variable X_1. Appelée seuil de division, elle est déterminée selon un critère optimisé qui est présenté plus loin dans ce chapitre. Cette inégalité est en particulier la meilleure division de toutes les divisions admissibles sur cette variable. Les individus pour lesquels cette inégalité est vraie sont affectés au sous-rectangle a_G (nœud fils gauche) alors que ceux pour lesquels on a l'inégalité contraire sont affectés au sous-rectangle a_D (nœud fils droite). Plus

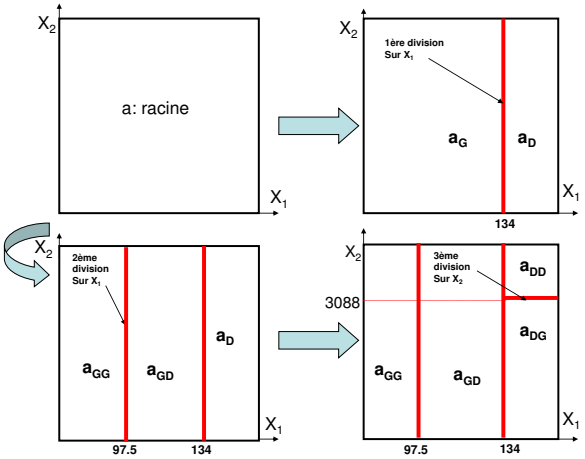

FIG. 4.1 – Un exemple de la division d'un ensemble en plusieurs groupes disjoints

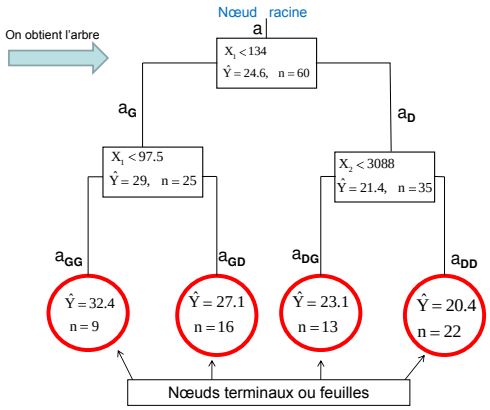

FIG. 4.2 – L'arbre de régression résultant de cette division

généralement, les individus pour lesquels la condition de division du nœud parent est vérifiée sont toujours affectés au nœud fils gauche alors que les autres sont toujours affectés au nœud fils droite.
Lorsque la variable de division est qualitative, la condition de la division est définie par une égalité en lieu et place de la condition d'inégalité. Par exemple, sur un nœud, l'égalité X=m_1, m_2, m_3 signifie que les individus pour lesquels la variable X a pour modalité m_1 ou m_2 ou m_3 sont affectés au nœud fils gauche. Ainsi, les individus avec les modalités autres sont affectés au nœud fils droite.

La 2ème division porte sur le rectangle a_G. C'est le nouveau nœud parent avec la meilleure division sur la variable X_1. Cette division est définie par l'inégalité $X_1 < 97.5$ de seuil 97.5. Les individus pour lesquels cette condition est vérifiée sont affectés au sous-rectangle a_{GG} (nœud fils gauche) alors que ceux pour lesquels on a l'inégalité contraire sont affectés au sous-rectangle a_{GD} (nœud fils droite).

La 3ème division porte sur le rectangle a_D. Elle est définie par l'inégalité $X_1 < 3088$ de seuil 3088. Les individus pour lesquels cette condition est vérifiée sont affectés au sous-rectangle a_{DG} (nœud fils gauche) alors que ceux pour lesquels on a l'inégalité contraire sont affectés au sous-rectangle a_{DD} (nœud fils droite).

Nous avons ainsi procédé à un partitionnement récursif de l'ensemble de l'échantillon en 4 sous-rectangles disjoints appelés nœuds terminaux : a_{GG}, a_{GD}, a_{DG} et a_{DD} de tailles respectives 9, 16, 13 et 22 et les prévisions de la variable à expliquer sont respectivement $\widehat{y}_{GG} = 32.4$, $\widehat{y}_{GD} = 27.1$, $\widehat{y}_{DG} = 23.1$ et $\widehat{y}_{DD} = 20.4$. Pour chaque nœud terminal, cette prévision est l'estimation de la moyenne de la variable à expliquer calculée sur les mesures portées par les individus affectés dans ce nœud.

Cette illustration suscite cependant un certain nombre de questions : Comment choisit-on la meilleure division ? Quel critère doit-on optimiser pour déterminer le meilleur arbre ? Que signifie division admissible pour un nœud ? Qu'est ce qu'un nœud terminal ? Quel est le lien entre la variable à expliquer et le processus de construction des groupes d'individus disjoints ? Nous proposons des réponses à ces questions dans les paragraphes suivants.

4.2.5 Principe de la méthode CART

La méthode CART a pour principe de partitionner l'ensemble des individus en sous-ensembles à l'intérieur desquels la variable dépendante est de variance minimale. Il s'agit de construire les groupes disjoints à l'intérieur desquels les individus sont les plus semblables possible. Un arbre $\mathcal{A}$ est ainsi construit par divisions successives de l'ensemble des données en sous-ensembles appelés nœuds. Une division d d'un nœud a s'effectue en séparant à l'aide d'une variable explicative le nœud a en deux nœuds descendants gauche et droite notés respectivement a_G et a_D. Au sommet de l'arbre se trouve le nœud racine contenant l'ensemble des individus de l'échantillon. Ce nœud racine est un mélange « impur » ou « hétérogène » composé des groupes d'individus avec un niveau élevé de dissemblance. Un nœud est dit intermédiaire s'il est divisé, et terminal sinon. La construction d'un arbre de régression repose alors sur trois principes :

- Établir pour chaque nœud, l'ensemble des divisions admissibles et définir un critère permettant de sélectionner la « meilleure », c'est-à-dire la division optimale.
- Définir une règle permettant de déclarer un nœud comme intermédiaire ou terminal.
- Pour chaque nœud terminal, déterminer la prévision de la variable à expliquer.

4.2.6 Etapes de construction d'un arbre de régression

Un arbre de régression est construit à l'aide d'une procédure itérative. La détermination d'une règle d'arrêt est plus délicate et la solution proposée par Breiman et $al.$ (1984) constitue une des spécificités de la méthode de discrimination par arbre. Cette solution comprend deux étapes fondamentales :

- La première consiste à construire un arbre maximal que nous notons $\mathcal{A}_{\max}$, puis à l'élaguer afin d'obtenir une séquence de sous-arbres emboîtés.
- La deuxième consiste à rechercher parmi ces arbres un sous arbre optimal selon un critère pénalisé sur la déviance (erreur quadratique du modèle). Pour y parvenir, deux approches sont possibles, on utilise soit un échantillon test (ou témoin) si l'échantillon total est de taille suffisamment grande, soit la procédure de la validation croisée lorsque l'échantillon total est de taille plus modeste. Ces deux approches sont exposées dans Celeux et $al.$ (1994). La procédure utilisée dans cette étude et concernant la régression est synthétisée dans Ghattas (1999a).

4.2.7 Construction de l'arbre maximal $\mathcal{A}_{\max}$

On dispose d'un échantillon E décrit par un ensemble de p variables aléatoires $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ et une variable Y. Un individu i de l'échantillon est défini par le couple $(\mathbf{x}_i, y_i)$ où $\mathbf{x}_i$=$(x_{i1}, x_{i2}, \ldots, x_{ip})$. L'échantillon de données $E = (x_i, y_i)_{(1 \leq i \leq N)}$ est de taille N et $\mathbf{X}$ est un vecteur des variables explicatives mixtes (quantitatives ou qualitatives) et Y est la variable dépendante (à expliquer) continue.

Dans la procédure itérative de construction d'un arbre, on commence par placer l'ensemble des individus au nœud racine notée a_0. On détermine pour chaque variable explicative toutes les divisions admissibles de l'échantillon au moyen des inégalités : $X_j \leq s$ où X_j est la $j^{\text{ième}}$ variable explicative continue et s est un

nombre réel appelé seuil de la division. Les individus pour lesquels cette inégalité est vraie sont dirigés au nœud fils gauche a_G et ceux pour lesquels $X_j > s$ sont dirigés au nœud fils droite a_D. a_G et a_D sont des nœuds descendants du nœud racine a_0. Si la variable X_j est qualitative de modalités 0 et 1, la division du nœud a_0 consiste à diriger vers le nœud descendant gauche a_G, les individus du nœud a_0 qui présentent la modalité 0, et vers le nœud descendant droite a_D, l'ensemble des individus du nœud a_0 qui présentent la modalité 1. Le processus de division recommence à partir des nœuds fils a_G et a_D. Il est nécessaire de définir le critère de sélection de la meilleure division pour chaque nœud.

4.2.7.1 Critère de division d'un nœud

Parmi les critères utilisés pour sélectionner la meilleure division, Breiman et *al.* (1984) préconisent l'utilisation de celui fondé sur la notion d'« *impureté* » du nœud parent. Dans le cadre de cette analyse, la variable à expliquer est continue. Le critère adapté est celui qui minimise la fonction déviance $\mathbf{D}(X_j, s)$, autrement dit, l'inertie intra-classes des nœuds descendants définie par la relation :

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

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 sont donnés respectivement dans les relations suivantes :

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

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

$\widehat{c}_1$ et $\widehat{c}_2$ sont les estimateurs empiriques de la moyenne de la variable dépendante Y à l'intérieur des nœuds descendants gauche et droit. Parmi toutes les divisions admissibles explorées sur toutes les variables explicatives et tous les seuils associés, la meilleure division $d^* = (X_j^*, s^*)$ qui minimise la fonction déviance $\mathbf{D}(X_j, s)$ est donnée par :

$$d^* = \arg \min_{(X_j, s) \in \mathscr{D}} \mathbf{D}(X_j, s), \tag{3}$$

où $\mathscr{D}$ est l'ensemble des divisions admissibles du nœud parent a et X_j^* est la variable sur laquelle la division optimale sur le nœud a est réalisée. Le couple (X_j^*, s^*) correspond à un seuil s^* de partage de E à partir de la variable X_j^*. Posons maintenant $a_G = a_G(X_j^*, s^*)$ et $a_D = a_D(X_j^*, s^*)$ où (a_G, a_D) est une partition de E optimale vis-à-vis du critère (3). Notons $a_1 = E$, $a_2 = a_G$ et $a_3 = a_D$. a_G et a_D sont les descendants de a_1. En réalisant sur a_2 et sur a_3 la même procédure que sur a_1, on obtient une nouvelle partition optimale de E en quatre classes (a_4, a_5, a_6, a_7) et ainsi de suite.

La procédure de construction de partitions s'arrête lorsque'il n'y a plus qu'un seul élément par classe, ou lorsque dans chaque classe les mesures des observations sont identiques. La procédure peut aussi être itérée jusqu'à l'obtention des sous-ensembles avec très peu d'individus, souvent compris entre 1 et 5 [31]. L'arbre ainsi obtenu est maximal et noté $\mathcal{A}_{max}$, car il contient le nombre maximum de feuilles. Un tel arbre ne présente aucun intérêt pratique, en effet, il peut avoir autant de feuilles que d'observations.

On peut ainsi présenter la suite emboîtée de partitions par un arbre binaire $\mathcal{A}$ avec les conventions suivantes : $a_1 = E$ est la racine, les a_i sont les nœuds, les nœuds extrêmes, c'est-à-dire, ceux de la partition la plus fine sont les feuilles, $\tilde{\mathcal{A}}$ est l'ensemble des feuilles de $\mathcal{A}$ et $\mathcal{A}^a$ est la branche de $\mathcal{A}$ issue du nœud a, c'est-à-dire, l'ensemble de tous les descendants du nœud a. Les feuilles correspondent aux K_i du modèle (4.2.3).

Lorsqu'un arbre est obtenu, il peut être utilisé sur un plan descriptif ou sur un plan prédictif. Dans le premier cas, chaque nœud est étiqueté par une variable et un seuil. Dans le second cas, supposons que l'on dispose de l'observation des variables explicatives utilisées dans le modèle. Cette observation va cheminer dans l'arbre à l'aide des règles précédentes pour se retrouver dans une feuille K_i. La valeur prévue pour y est alors $\widehat{c}_i$ de cette feuille. Une fois le modèle construit, il faut l'évaluer.

4.2.8 Evaluation du modèle de régression par arbre

Supposons d'abord construit un estimateur $\widehat{y}(.)$ de $y(.)$. Il faut évaluer sa capacité à classer les nouvelles observations dans les nœuds terminaux de l'arbre et à prévoir la valeur de chacune de ces observations pour la variable dépendante. Le coût de la régression est évalué par son erreur quadratique définie par :

$$\mathbf{R}(y) = \mathbb{E}\left(Y - y(x)\right)^2 \tag{4}$$

L'estimation de cette erreur quadratique est une étape essentielle à l'estimation du modèle et de sa qualité prédictive. En général, deux méthodes sont utilisées pour l'évaluation de la qualité du modèle : la méthode de l'échantillon test (ou témoin) et la méthode de validation croisée.

- **L'échantillon test pour l'évaluation** :
  Le premier estimateur de $\mathbf{R}$ généralement utilisé pour la validation est celui basé sur un échantillon test. Supposons que l'on dispose d'un second échantillon $\mathbf{F} = (x_i, y_j)_{(1 \leq i \leq N')}$. Une fois $\widehat{y}$ obtenu à partir de l'échantillon de départ, on estime $\mathbf{R}(y)$ par :
  $$\widehat{\mathbf{R}}^{et}(y) = \frac{1}{\mathrm{Card}(\mathbf{F})} \sum_{(x_i, y_i) \in \mathbf{F}} \left(y_i - \widehat{y}(x_i)\right)^2 . \tag{5}$$
  Cet estimateur permet d'éliminer le biais produit par l'estimateur par substitution, mais nécessite un volume important de données.
- **Validation croisée pour l'évaluation** :
  Le second estimateur du risque utilisé est celui obtenu par validation croisée. Il limite aussi le biais d'optimisme et ne nécessite pas d'échantillon test. Cette technique s'appuie sur les méthodes de ré-échantillonnage lorsqu'on ne peut réserver une partie des données pour l'évaluation des modèles.

Ainsi, l'échantillon total $\mathbf{E}$ est divisé aléatoirement en K sous-échantillons de tailles quasiment égales et mutuellement exclusifs $\mathbf{E}_k$, $k = 1, ..., K$. L'arbre $\mathcal{A}_{max}$ est construit à partir de $\mathbf{E}$. Pour chaque k, on définit $\mathbf{E}^k = \mathbf{E} - \mathbf{E}_k$ le complémentaire de $\mathbf{E}_k$ dans $\mathbf{E}$. Sur $\mathbf{E}^k$ on estime $y(.)$ par $\widehat{y}(., \mathbf{E}^k)$ et l'échantillon test dans ce cas est $\mathbf{E}_k$. L'estimateur de $\mathbf{R}(y)$ par validation croisée est alors

$$\widehat{\mathbf{R}}^{vc}(y) = \frac{1}{N} \sum_{k=1}^{K} \sum_{(x_i, y_i) \in \mathbf{E}_k} \left(y_i - \widehat{y}(x_i, \mathbf{E}^k) \right)^2 . \tag{6}$$

Une présentation de cette cette procédure de test par validation croisée est synthétisée dans Celeux et *al.* (1994).

4.2.9 Erreur de l'arbre

Tenant compte de l'introduction de la notion d'arbre, nous introduisons aussi la notion l'erreur quadratique liée à l'arbre obtenu par le modèle. Si $\mathcal{A}$ est un tel arbre, l'erreur de l'arbre que nous notons $\widehat{\mathbf{R}}(\mathcal{A})$ est définie par :

$$\widehat{\mathbf{R}}(\mathcal{A}) = \frac{1}{N} \sum_{t \in \tilde{\mathcal{A}}} \widehat{\mathbf{R}}(a) \tag{7}$$

avec $\widehat{\mathbf{R}}(a) = \sum_{x_i \in a} (y_i - \widehat{y}(x_i))^2$. La variation de la qualité de l'arbre dûe au partage d'un nœud a avec le seuil s est donnée par :

$$\triangle \widehat{\mathbf{R}}(s, a) = \frac{1}{N} \left[\widehat{\mathbf{R}}(a) - \widehat{\mathbf{R}}(a_G) - \widehat{\mathbf{R}}(a_D) \right] . \tag{8}$$

Par construction, $\triangle \widehat{\mathbf{R}}(s, a) > 0$. Donc $\widehat{\mathbf{R}}(\mathcal{A})$ décroît au fur et à mesure que le nombre de feuilles croît, et en particulier $\widehat{\mathbf{R}}(\mathcal{A}_{max}) = 0$. Finalement, on peut réduire l'arbre maximal en arrêtant la procédure séquentielle :

- lorsque pour un nœud $\triangle \widehat{\mathbf{R}}(s, a) \leq \lambda$.
- lorsque le nombre d'observations de chaque nœud devient inférieur à ν.

Si λ et ν sont petits, le nombre de feuilles sera grand, et petit dans le cas contraire. Dans les deux cas l'estimation de $y(x)$ ne sera pas satisfaisante. Breiman et *al.* proposent de conjuguer validation croisée et procédure d'élagage des arbres pour résoudre le problème de la recherche de l'arbre optimal.

4.2.10 Procédure d'élagage

Considérons la branche $\mathcal{A}^a$ d'un arbre $\mathcal{A}$. Cette branche a pour racine le nœud intermédiaire a de l'arbre $\mathcal{A}$. Il est constitué du nœud a lui même et de tous ses nœuds descendants. *Elaguer* une branche $\mathcal{A}^a$ de l'arbre consiste à supprimer de l'arbre de $\mathcal{A}$ tous les nœuds descendants du nœud a, c'est-à-dire tous les éléments de la branche $\mathcal{A}^a$ excepté le nœud a lui-même. L'arbre ainsi obtenu est noté $\mathcal{A} - \mathcal{A}^a$. Si l'arbre $\mathcal{A}'$ est obtenu à partir de l'arbre $\mathcal{A}$ par élagages successifs, alors l'arbre $\mathcal{A}'$ est un *sous-arbre* de l'arbre $\mathcal{A}$. On note $\mathcal{A}' < \mathcal{A}$,

et $\mathcal{A}'$ est dit emboîté dans $\mathcal{A}$.

La recherche d'un sous-arbre optimal pourrait consister à considérer tous les sous-arbres de $\mathcal{A}_{max}$ et à les comparer à l'aide d'un échantillon test. Le nombre important de sous-arbres rend difficiles les calculs. Aussi, l'arbre optimal ainsi obtenu serait optimal vis-à-vis d'un échantillon test, ce qui constitue une deuxième difficultée.

Rappelons que si $\mathcal{A}' < \mathcal{A}$ alors $\widehat{\mathbf{R}}(\mathcal{A}) \leq \widehat{\mathbf{R}}(\mathcal{A}')$. Une stratégie consiste à pénaliser un trop grand nombre de feuilles par l'introduction d'un facteur coût et d'un nouveau critère :

$$\widehat{\mathbf{R}}_\alpha(\mathcal{A}) = \widehat{\mathbf{R}}(\mathcal{A}) + \alpha \left|\tilde{\mathcal{A}}\right|, \tag{9}$$

où $\alpha \geq 0$ et $\left|\tilde{\mathcal{A}}\right|$ est le nombre de feuilles de l'arbre $\mathcal{A}$. Le terme $\alpha \left|\tilde{\mathcal{A}}\right|$ est interprété comme un coût de complexité et $\widehat{\mathbf{R}}_\alpha(\mathcal{A})$ est l'erreur quadratique pénalisée. Si $\alpha = 0$, alors $\widehat{\mathbf{R}}_0(\mathcal{A}) = \widehat{\mathbf{R}}(\mathcal{A})$. Notons aussi que, si a est un nœud $\widehat{\mathbf{R}}_\alpha(a) = \widehat{\mathbf{R}}(a) + \alpha$.

4.2.11 Sélection du meilleur arbre

La procédure d'élagage permet de contruire une séquence finie de L sous-arbres emboîtés $\mathcal{A}_0 > \mathcal{A}_1 > \ldots > \mathcal{A}_L$ où $\mathcal{A}_L$ est la racine de l'arbre maximal $\mathcal{A}_{max} = \mathcal{A}_0$, et une suite croissante $(\alpha_l)_{(1 \leq l \leq L)}$ de coefficients de pénalisation. Pour la construction de ces suites, voir (Celeux et *al.*, 1994) Pour sélectionner l'arbre optimal, une fois de plus on peut utiliser l'échantillon test ou la procédure de validation croisée :

- **Echantillon test pour l'élagage**
  Une fois la suite d'arbres $(\mathcal{A}_l)_{(0 \leq l \leq L)}$ construite, nous notons $\widehat{y}^l$ les estimateurs de y associés à chacun des $\mathcal{A}_l$ et $\mathcal{A}_l = \mathcal{A}(\alpha_l)$. On choisit l'arbre $\mathcal{A}_{l_0}$ tel que :

$$\alpha_{l_0} = \arg\min_l \frac{1}{N'} \sum_{(x_i, y_i) \in \mathbf{F}} \left(y_i - \widehat{y}^l(x_i)\right)^2, \tag{10}$$

  où $\mathbf{F}$ est l'échantillon test de taille N'.
- **Par validation croisée pour l'élagage**
  Considérons à nouveau les K sous-échantillons $\mathbf{E}_k$ ($k = 1, ..., K$) et leur complémentaire dans $\mathbf{E}$, $\mathbf{E}^k$ (ie. $\mathbf{E} - \mathbf{E}_k$) : nous construisons sur le sous échantillon $\mathbf{E}^k$ une suite d'arbres $\mathcal{A}^k(\alpha_l^k)$ de coût pénalisé minimum pour α. Nous procédons à la même construction sur la totalité de l'échantillon d'apprentissage E. Si on note $(\mathcal{A}_l)$ la suite d'arbres obtenue et (α_l) les coûts de complexité correspondants, l'estimateur de l'erreur pour l'arbre $\mathcal{A}(\alpha_l)$ obtenu par validation croisée est donné par :

$$\widehat{\mathbf{R}}^{vc}(\mathcal{A}(\alpha_l)) = \frac{1}{N} \sum_{k=1}^{K} \sum_{(x_i, y_i) \in \mathbf{E}_k} \left(y_i - \widehat{y}^l(x_i, \mathbf{E}^k)\right)^2 \tag{11}$$

  où $\widehat{y}^l(x)$ est l'estimateur de $y(x)$ associé à l'arbre $\mathcal{A}^k(\alpha_l)$. L'arbre optimal vis-à-vis du critère avec pénalisation est $\mathcal{A}(\alpha_{opt})$ tel que :

$$\widehat{\mathbf{R}}^{vc}(\mathcal{A}(\alpha_{opt})) = \min_{\alpha_l} \widehat{\mathbf{R}}^{vc}(\mathcal{A}(\alpha_l)).$$

4.2.12 Divisions suppléantes

Certaines variables explicatives peuvent être assez importantes sans être actives. En effet, ces variables qui ne jouent plus aucun rôle lorsque l'arbre est construit ont pu être pour plusieurs nœuds « conurrentes » des variables actives. La connaissance de ces variables concurrentes est utile et permet d'établir une hiérarchie de l'ensemble des variables explicatives. Cette hiérarchie peut servir pour mettre en œuvre d'autres méthodes statistiques avec un nombre de variables réduit. Ainsi, afin de faciliter l'interprétation des résultats produits par le modèle de régression par arbre, nous définissons ci-dessous quelques terminologies souvent utilisées dans la pratique de la méthode CART :

- *Division concurrente*
  Si d^* est la division d'un nœud a qui minimise le critère de déviance ou de l'inertie intra-classes, la division qui réalise le deuxième minimum de ce critère est la première division concurrente à d^*. Elle peut porter sur une autre variable explicative ou sur la même variable explicative que d^* avec un autre seuil.
- **Division de substitution**
  Soient a le nœud d'un arbre $\mathcal{A}$, d^* la meilleure division de a en des nœuds descendants a_G et a_D, X_j est une variable explicative quelconque avec $1 \leq j \leq p$, $\mathscr{D}_j$ l'ensemble de ses divisions admissibles selon l'inégalité de $X_j \leq s$, et $\mathscr{D}_j^c$ l'ensemble des divisions complémentaires $(X_j > s)$.
  - Posons $P(d_j, d^*)$: la probabilité pour que la division d_j prédise correctement le partage du nœud a par la meilleure division d^*. Autrement dit, la probabilité pour que le partage donné par la division d_j soit le plus proche possible de celui donné par la division optimale d^* (Ghattas, 1999b).
  - Une division $\tilde{d}_j \in \mathscr{D}_j \bigcup \mathscr{D}_j^c$ basée sur la variable explicative X_j est appelée division de substitution pour d^* si :
  $$P(\tilde{d}_j, d^*) = \max_{d_j \in \mathscr{D}_j \bigcup \mathscr{D}_j^c} P(d_j, d^*), \tag{12}$$

  où $\tilde{d}_j$ est parmi l'ensemble des divisions admissibles basées sur la variable X_j, celle qui peut prédire au mieux la partition effectuée par d^*.
- **Association entre les divisions $\widetilde{d}_j$ et d^* et interprétation**

  Soient d^* la meilleure division du nœud a. Elle achemine les observations vers a_G avec la probabilité P_G et vers a_D avec la probabilité P_D. Si la variable sur laquelle est basée la division optimale d^* est manquante pour une nouvelle observation tombant dans le nœud a, il existe alors deux possibilités pour acheminer cette observation vers un des deux nœuds descendants de a. On peut utiliser la règle naturelle fondée sur la probabilité $max(P_G, P_D)$[1] ou bien une division de substitution $\widetilde{d}_j$ de la divi-

[1] Si $P_G < P_D$, l'observation est acheminée vers le a_D et vers a_G dans le cas contraire.

sion optimale. La comparaison relative entre ces deux méthodes d'acheminement est évaluée par la mesure de l'*association* entre la division de substitution $\widetilde{d}_j$ et la division optimale d^* :

$$Asso(\widetilde{d}_j, d^*) = \frac{err_1 - err_2}{err_1}, \tag{13}$$

où $err_1 = min(P_G, P_D)$ est l'erreur de mal acheminer la nouvelle observation par la règle naturelle, et $err_2 = 1 - P(\widetilde{d}_j, d^*)$ est l'erreur de mal acheminer la nouvelle observation par la division de substitution $\widetilde{d}_j$. L'indice d'association $Asso(\widetilde{d}_j, d^*)$ mesure la capacité de la division de substitution à produire une partition la plus proche possible à celle obtenue par la division optimale. Il s'interprète comme suit :

- Si $Asso(\widetilde{d}_j, d^*)$ est petite, mais positive, la division de substitution $\widetilde{d}_j$ contribue peu dans la réduction de l'erreur d'acheminement.
- Si $Asso(\widetilde{d}_j, d^*) = 1$, la division de substitution $\widetilde{d}_j$ prévoit parfaitement la partition obtenue par la division optimale d^*.
- Si $Asso(\widetilde{d}_j, d^*) < 0$, la division de substitution $\widetilde{d}_j$ n'a pas d'intérêt. En effet, la probabilité de mauvais acheminement est plus élevée suite à son utilisation que celle obtenue par la règle naturelle.

4.2.13 Importance d'une variable explicative par CART

L'importance d'une variable explicative X_j pour un arbre de régression $\mathcal{A}$ est définie par la relation :

$$I(X_j) = \sum_{a \in \mathcal{A}} \Delta \widehat{R}(\tilde{d}_j(a), a), \tag{14}$$

où $\Delta \widehat{R}(\tilde{d}_j(a), a)$ se calcule comme dans la relation (8) vue précédemment. C'est la réduction de l'impureté ou de l'inertie intra-classes dûe à la division du nœud a par la variable X_j. Elle s'écrit :

$$\Delta \widehat{R}(\tilde{d}_j(a), a) = \widehat{R}(a) - \widehat{R}(a_G) - \widehat{R}(a_D),$$

$\tilde{d}_j(a)$ est la division de substitution au nœud a basée sur la $j^{ième}$ variable X_j. La mesure de $I(X_j)$ dépend de l'arbre à partir duquel elle est calculée. C'est la somme des diminutions de la déviance provoquée à chaque nœud a de l'arbre, si l'on remplaçait pour chaque nœud la division optimale par la division de substitution sur la variable X_j (Ghattas, 1999b)[33]. Généralement, on ramène l'importance des variables à l'intervalle [0 ; 100]. La variable la plus importante aura pour indice 100, en effet, les importances obtenues en terme de déviance sont divisées par leur maximum et multipliées par cent.

4.3 Application de la méthode CART à la prévision du temps de passage d'un aéronef en un point de sa trajectoire de vol prévue

Nous utilisons la méthode de régression par arbre CART pour prévoir l'écart temporel de passage d'un aéronef en un point de sa trajectoire de vol prévue. Les données utilisées portent sur 25000 individus dé-

crits par 22 variables explicatives. Chaque individu est un point de la trajectoire prévue pour un vol. Notre modèle a pour but de prévoir en chacun de ces points l'écart temporel de passage d'un aéronef conditionnellement aux caractéristiques connues de l'aéronef au point courant de prévision et à celles prévues. Toutes les variables explicatives ont été décrites au chapitre précédent.

4.3.1 Arbre de régression retenu

Après élagage de l'arbre maximal, le sous-arbre ayant le plus petit coût-complexité par la procédure de la validation croisée possède 12 feuilles ou nœuds terminaux. C'est cet sous-arbre que nous avons retenu pour la régression. L'illustration des coûts est fournie par le graphique FIG.4.3 qui représente la fonction des coûts complexité des sous-arbres en fonction du nombre de nœuds terminaux.

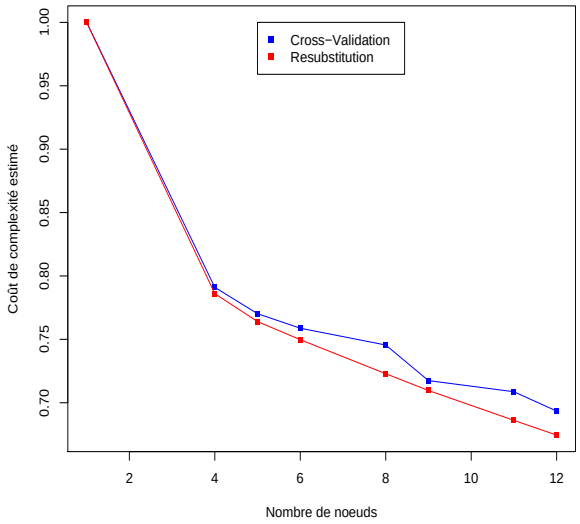

FIG. 4.3 – Coût-complexité des sous-arbres en fonction du nombre de nœuds terminaux

4.3.2 Diagnostic des résidus du modèle CART obtenu

Le diagnostic des résidus du modèle obtenu repose d'une part, sur le test d'ajustement de la loi de probabilité de la variable résiduelle à la loi normale, et d'autre part, sur l'étude de l'ampleur des erreurs du modèle à travers un examen de la dispersion des résidus en fonction de l'horizon de prévision.

4.3.2.1 Test d'ajustement des résidus par une loi normale

Les résidus du modèle obtenu semblent symétriques et concentrés autour de sa moyenne qui est 0 (FIG.4.4). Nous avons réalisé deux tests d'hypothèse de normalité sur cette variable résiduelle :

Le premier et le plus courant est celui de Kolmogorov-Smirnov où la statistique de test D est égale à 0.509 et la p-value est égale à 0. L'hypothèse nulle de normalité est ainsi rejétée au risque de 5%. Selon ce

test, les résidus obtenus du modèle *CART classique* ne suivent pas une loi normale.

Le second est un test asymptotique connu sous le nom de test de Jarque-Bera (1980)[38] et utilisé pour les échantillons de très grande taille. La statistique de test *JBSTAT* suit une loi $\chi^2_{(2)}$ du Khi-deux à 2 degrés de liberté (Une description du test de Jarque-Bera est faite en annexe). Sa valeur observée sur les données d'apprentissage est égale à 568370 et la valeur critique au seuil de signification de 5% est égale à 5.9915. La p-value du test est égale à 0. Ainsi, ce test beaucoup plus performant que le précédent et plus adapté aux données rejette à nouveau l'hypothèse nulle. On conclut finalement que la distribution de probabilité des résidus obtenus du modèle *CART classique* ne peut être ajustée directement par une loi normale. Au regard de la forme de l'histogramme ci-dessus, on peut conjecturer que la distribution de probabilité des résidus de ce modèle est un mélange de lois gaussiennes dont l'estimation des paramètres est présentée plus loin dans ce document.

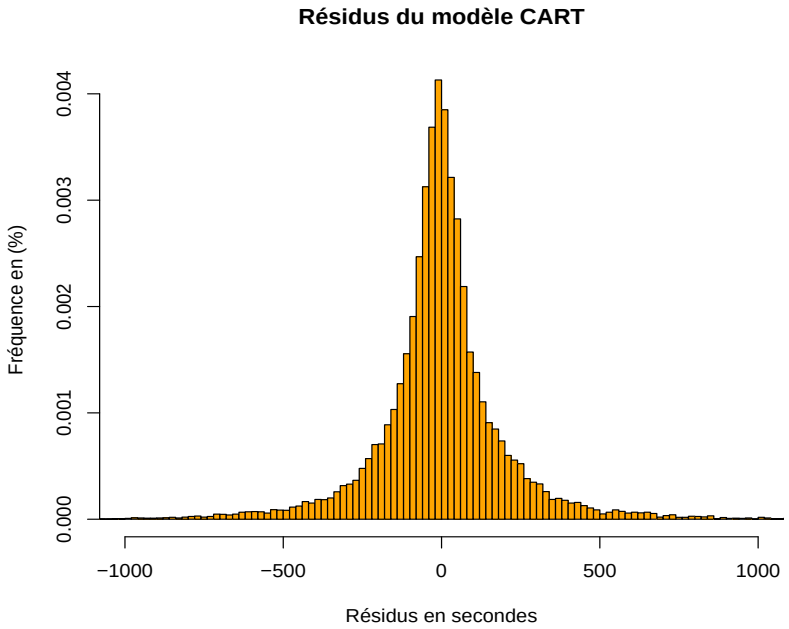

FIG. 4.4 – Histogramme de distribution des résidus du modèle CART classique

4.3.2.2 Horizon de prévision

Dans le cadre de cette étude, un horizon de prévision est un intervalle de temps pendant lequel il est possible d'utiliser un modèle pour calculer les prévisions avec les erreurs bornées. C'est un intervalle de temps compris entre l'instant courant (ici, ramené à 0) et une limite que l'on fixe en fonction des objectifs de la prévision. Pour évaluer l'évolution de l'incertitude des modèles développés en fonction de l'horizon de prévision, nous avons considéré une fenêtre de temps de prévision de longueur supérieur à 60 minutes (3600 secondes) que nous avons ensuite divisé en 12 petits intervalles de 5 minutes (300 secondes) chacun. Le 13ème intervalle contient les points des trajectoires de vol prévues pour lesquels l'horizon de prévision

est supérieur ou égal à 60 minutes. Sur les représentations des boxplots (FIG.4.5), 00-03 indique la forme simplifiée du petit intervalle de limites 0 et 300 secondes et $36et+$ désigne 3600 secondes et plus. Ainsi, sur chacun de ces intervalles, la qualité d'ajustement de chaque modèle proposé est évaluée, ce qui permet d'obtenir finalement une approximation de l'évolution de l'incertitude dans les prévisions en fonction de l'horizon temporel de prévision des modèles.

Compte tenu de la corrélation entre la durée de vol prévue sur une portion de la trajectoire de vol prévue et la longueur de celle-ci, nous avons étendu cette notion d'horizon temporel à l'horizon spatial de prévision. Ainsi, en considérant un horizon de distance inférieure à 400 NM, nous avons recouvert ce domaine de la distance de vol prévue par 10 petits intervalles de même longueur égale à 40 NM. Toutes les distances de vol prévues supérieures ou égales à 400 NM sont représentées par $400et+$, c'est-à-dire 400 NM et plus.

Rappelons que ce découpage est arbitraire. L'objectif ici consiste à avoir une tendance de l'évolution de l'incertitude dans la prévision des écarts temporels en fonction de l'horizon temporel ou de l'horizon spatial.

4.3.2.3 Dispersion des résidus du modèle CART en fonction de l'horizon de prévision

Les résidus sont d'espérance nulle et d'écart-type 240 secondes (4 minutes). La figure FIG.4.5 met en évidence une dépendance de l'ampleur de leur dispersion en fonction des horizons de prévision (temporel et spatial). Ainsi, plus l'horizon temporel du temps de passage en un point est éloigné, plus la dispersion des résidus du modèle est importante. Il en est même que lorsqu'un point de la trajectoire de vol prévue est éloigné du point courant du vol. Il en résulte que la qualité de prévision du modèle CART développé dépend fortement de l'horizon des prévisions. Si la dispersion des résidus est faible pour les horizons inférieurs à 1200 secondes, en revanche, la plus forte dispersion des résidus est observée sur les horizons supérieurs à 2100 secondes (boxplots, FIG.4.5). De même, si la dispersion des résidus est plus faible pour les horizons de distance inférieurs à 200 NM, il apparaît en revanche que, les résidus les plus dispersés sont observés pour les horizons de distance supérieurs à 240 NM (boxplots, FIG.4.5). Finalement le modèle proposé fournit des meilleures prévisions avec une faible incertitude si les points de la trajectoire de vol prévue sont proches du point courant. Mais, la qualité de prévision de ce modèle se dégrade lorsque l'horizon de prévision devient important. L'étude de la qualité de prévision du modèle en fonction de l'horizon de prévision au chapitre 7 permettra d'estimer la profondeur de l'horizon temporel de prévision borné pour une utilisation opérationnelle du modèle.

4.3.2.4 Evaluation des erreurs du modèle CART en fonction de l'horizon de prévision

Theil (1958)[59] a introduit un certain nombre de statistiques qui servent de critères pour évaluer la qualité des prévisions. Nous en donnons ici deux statistiques fondamentales :

- L'erreur absolue moyenne (MAE : mean absolute error) définie par la relation :

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \widehat{y}_i|}{n}, \tag{15}$$

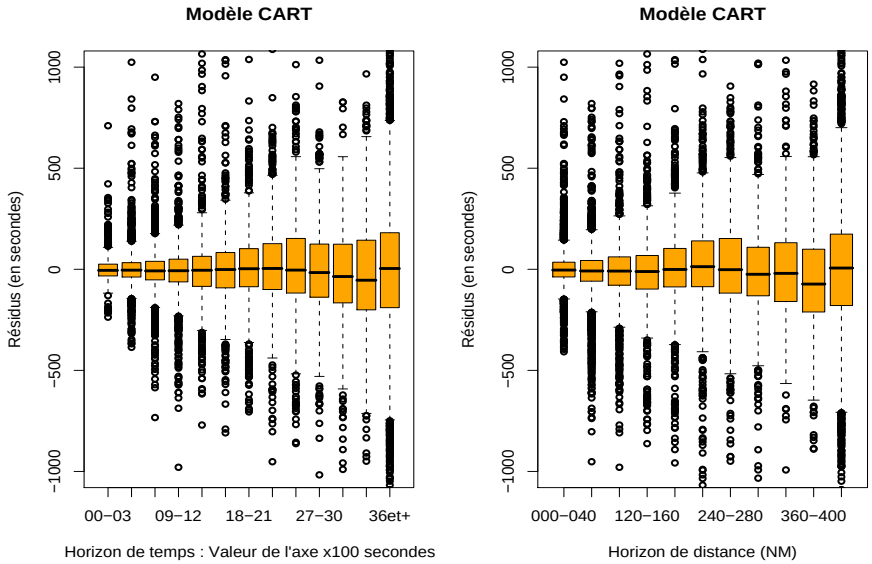

FIG. 4.5 – Dispersion des résidus en fonction de l'horizon temporel (à gauche) et spatial (à droite) de prévision

où $\widehat{y}_i$ est la prévision par le modèle de l'écart temporel et y_i est la mesure de cette variable pour l'observation i. La valeur absolue indique que l'on pénalise les erreurs positives autant que celles négatives.

- La racine carrée de l'erreur quadratique moyenne (RMSE : root mean squared error) est l'indicateur généralement utilisé pour évaluer la précision des modèles en fonction de l'horizon de prévision. Elle est définie par la relation :

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n}}. \tag{16}$$

Cette statistique pénalise plus fortement les erreurs importantes.

Pour l'évaluation de la qualité de notre modèle, nous étudions la variation du RMSE[2] en fonction de l'horizon de prévision. En utilisant cet indicateur, Alexandre M. et *al.* (1994)[3] montrent que de façon générale, au fur et à mesure que l'horizon de prévision se rapproche, les valeurs du RMSE s'améliorent (c'est-à-dire, deviennent faibles). La figure FIG.4.6 indique l'évolution de l'incertitude du modèle en fonction de l'horizon temporel de prévision alors que la figure FIG.4.7 indique l'évolution de l'incertitude du modèle en fonction de l'horizon spatial (distance) de prévision. De ces deux courbes, la précision du modèle est bien meilleure pour les horizons proches. Par ailleurs, les deux figures mettent en évidence la corrélation entre les deux horizons de prévision. Ainsi, pour une analyse approfondie de l'évolution de l'incertitude du modèle, on peut utiliser l'un des deux horizons et aboutir quasiment aux mêmes conclusions. Dans les chapitres suivants, nous utiliserons pour l'évaluation des différents modèles, l'horizon temporel de prévision.

[2]C'est la racine carrée de la somme des carrés des erreurs de prévisions.

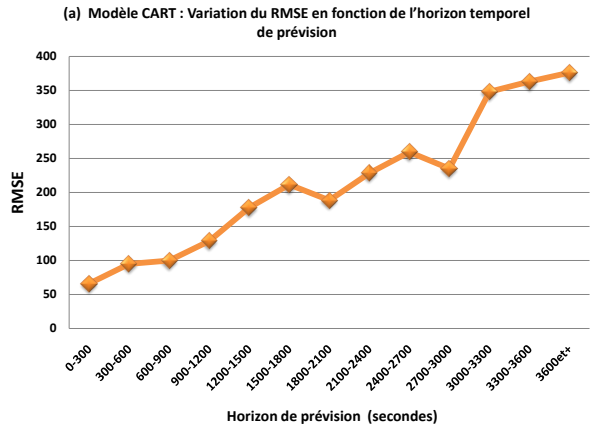

FIG. 4.6 – Qualité de prévision du modèle en fonction de l'horizon temporel de prévision

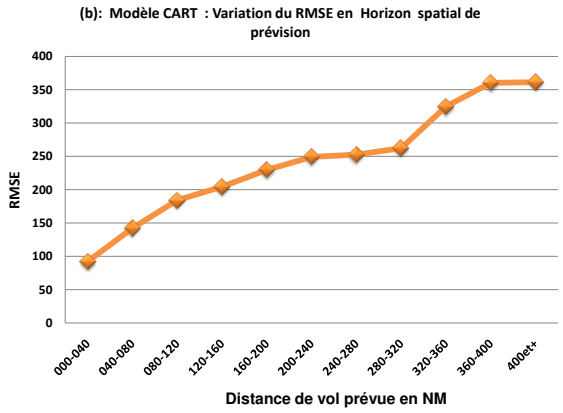

FIG. 4.7 – Qualité de prévision du modèle en fonction de l'horizon spatial de prévision

4.4 Analyse des résultats de la régression

4.4.1 Variables actives

Parmi les 22 variables d'entrée du modèle de régression, seules 5 ont contribué activement au partitionnement de l'échantillon. Ces variables actives sont celles qui apparaissent sur l'arbre lorsqu'il est construit (FIG. 4.8). Elles sont : l'influence du vent prévu sur les trajectoires des vols « *Indur* », la distance sur le plan de vol entre les points courants des aéronefs et les points de leur trajectoire de vol prévue *Distprev1*, le type d'aéronef utilisé « *Type* », le niveau de vol prévu pour la phase de croisière « *Nivpln* » et la vitesse des aéronefs à l'instant courant de prévision « *Vitessecour* » .

Ces variables mettent en évidence l'impact significatif des trois facteurs importants dans le processus de prévision des écarts temporels au passage des aéronefs sur les points de leur trajectoire de vol prévue. Il s'agit des conditions météorologiques qui agissent à travers la variable *Indur*, des paramètres courants des aéronefs à l'instant courant de prévision représentés par les variables *Distprev1* et *Vitessecour*, enfin, les caractéristiques des plans de vol prévus consituées du type d'aéronef (*Type*) et du niveau de vol prévu pour la phase de croisière *Nivpln*. Nous analysons dans les paragraphes ci-dessous les effets spécifiques de chacun de ces facteurs sur la précision des prévisions.

Pour la lecture et l'interprétation de l'arbre de régression (FIG. 4.8) : Sur les nœuds intermédiaires ou terminaux, y est l'estimation moyenne de la valeur dépendante écart temporel sur le nœud, n est le nombre d'individus sur ce nœud.

4.4.2 Conditions météorologiques et atmosphériques

4.4.2.1 Influence du vent : *Indur*

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

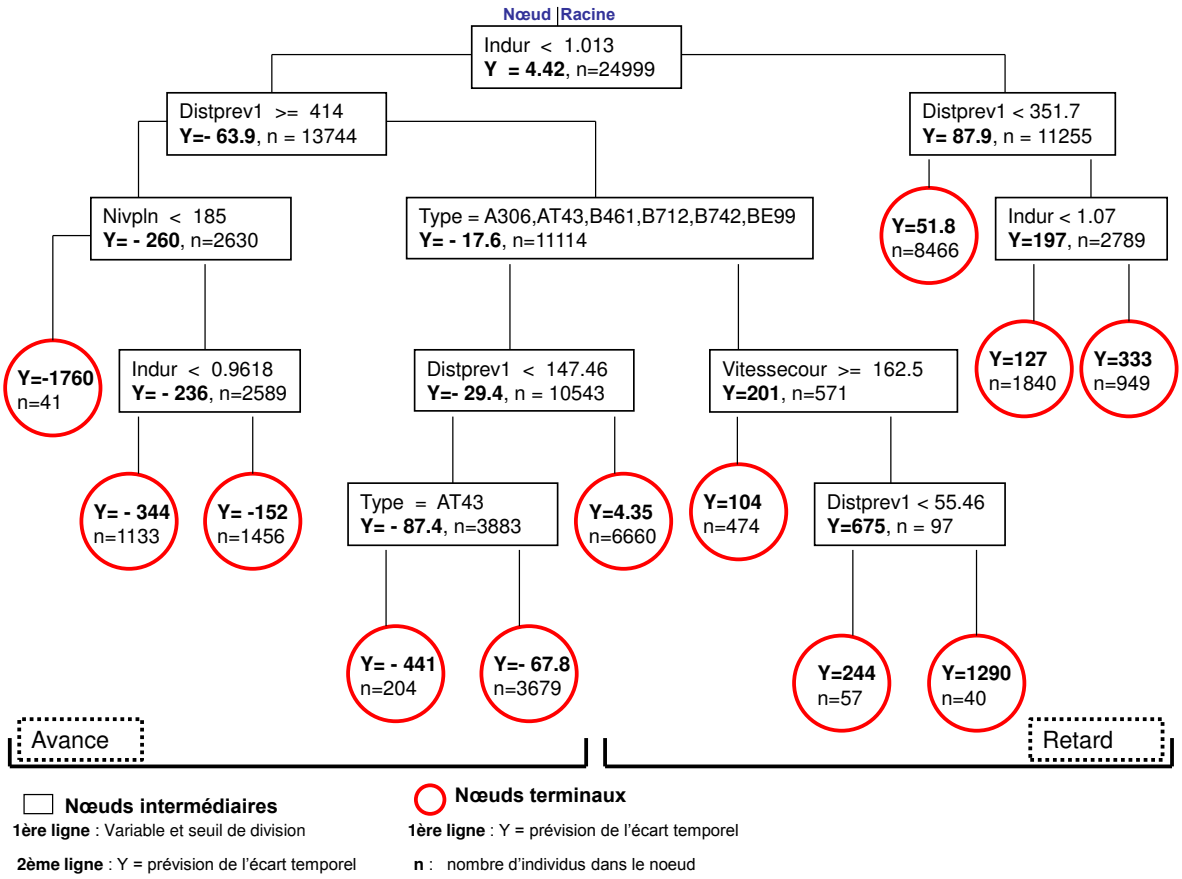

FIG. 4.8 – Arbre de régression de la méthode CART pour le modèle de prévision des écarts temporels

l'incertitude sur le temps de passage des avions en des 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 l'échantillon des données en deux ensembles suivant l'importance du vent auquel les aéronefs ont été soumis. L'estimation de l'*écart temporel* pour les vols de la branche de droite est d'environ 87.9 secondes. Ce sont des vols qui ont été soumis à un vent de face prépondérant et les aéronefs impliqués ont tendance à survoler les points de leur trajectoire prévue après les instants prévus dans leur plan de vol. 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 prépondérant et les aéronefs impliqués ont tendance à survoler les points de leur trajectoire prévue avant les instants prévus dans leur plan de vol. Ces résultats sont cohérents avec les principes de la dynamique des vols dans la mesure où le vent de face oppose au mouvement d'un aéronef une force de résistance qui réduit la résultante de la traction.

Un examen des divisions ($Indur > 1.07$) et ($Indur < 0.96$) sur cette même variable influence du vent montre que les retards et les avances peuvent connaître une certaine amplification lorsque l'intensité du vent augmente et que le pilote ne prend aucune mesure visant à le compenser. Dans le premier cas le retard moyen est d'environ 333 secondes alors que pour le second, l'avance moyenne est de 344 secondes. Finalement, un

vent très fort qu'il soit de dos ou de face a pour effet d'amplifier le niveau de l'incertitude dans le système de prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue et tend à réduire la capacité de gestion de trafic aérien dans son ensemble.

4.4.2.2 Caractéristiques atmosphériques : *Moydensa, Moytemp et Moypres*

Si les conditions météorologiques ont joué un rôle de premier plan dans la construction de l'arbre de régression, il apparaît en revanche que les paramètres atmosphériques ne sont pas intervenus directement dans le processus de partitionnement de l'ensemble des individus. C'est à travers les divisions suppléantes (ie. concurrente ou de substitution) que ces paramètres sont manifestement importants pour expliquer la variabilité de la variable dépendante *écart temporel*. Ainsi, au 4ème nœud de l'arbre optimal sur la variable *Nivpln* (FIG.4.8), l'écart moyen de densité de l'air *(Moydensa)* est la variable de la première division de substitution avec un indice d'association de 98%. Cet indice représente la probabilité que la division de substitution *Moydensa < 0.17* (Annexe ; Tab.B.5) fournit des nœuds descendants gauche et droit qui ressemblent à ceux fournis par la meilleure division sur ce nœud pour la variable *Nivpln*. Notons que la même variable *Moydensa* est également utilisée dans la 3ème et la 2ème divisions de substitution pour les divisions optimales utilisant les variables vitesse courante ($Vitessecour \geq 162.5$) et le type d'aéronef *($Type = AT43$)* respectivement. Les indices d'association sont respectivement 86% et 95%. Les indices d'associations élevés et la compétition remarquable entre la variable *Moydensa* et les variables de performance des aéronefs telles que : le type d'aéronef utilisé *(Type)*, la vitesse courante de l'aéronef *(Vitessecour)* et le niveau de vol de croisière prévu *(Nivpln)* semblent bien confirmer l'existence du lien annoncé au chapitre précédent entre les paramètres de performance aérodynamique des aéronefs et les conditions atmosphériques, notamment l'altitude densité de l'air. Même s'il semble difficile de proposer dans le cadre de cette étude une interprétation des effets de la température *(Moytemp)* et de la pression *(Moypres)*, il est à noter que : les conditions météorologiques et les paramètres atmosphériques sont identifiés comme déterminants pour la prévision de l'*écart temporel*. Les chapitres suivants devront permettre de raffiner l'interprétation de ces facteurs.

4.4.3 Caractéristiques courantes des aéronefs en vol

Les caractéristiques courantes des aéronefs en vol ont participé activement à la construction de l'arbre de régression (FIG.4.8) à travers les variables telles que la distance de vol prévue *(Distprev1)* et la vitesse des aéronefs à l'instant courant de prévision *(Vitessecour)*.

4.4.3.1 Distance de vol prévue : *Distprev1*

La distance de vol prévue est utilisée dans 4 meilleures divisions et apparaît comme l'une des variables les plus actives dans le processus de partitionnement de l'ensemble des données. Placée au deuxième niveau de l'arbre de régression par les divisions respectives *($Distprev1 \geq 414$)* et *($Distprev1 < 351.7$)*, elle permet de raffiner l'interprétation des effets du vent sur le mouvement des aéronefs au moyen de son interaction

avec la variable explicative influence du vent. Nous observons que l'effet du vent qu'il soit de dos ou de face n'est réellement significatif sur la progression d'un aéronef sur sa trajectoire que si la distance de vol prévue est importante. En effet :

- La meilleure division ($Distprev1 \geq 414$) porte sur les aéronefs dont le vent prévu sur les trajectoires est de dos. Elle envoie vers le nœud fils gauche les aéronefs dont la distance de vol prévue est supérieure à 414 NM. Ces aéronefs avec une distance de vol prévue plus importante et un vent prévu de dos prépondérant ont des valeurs élvées (négativement) de la variable *écart temporel*. Ces aéronefs semblent survoler les points de leur trajectoire de vol prévue avec des avances qui se situent autour de 260 secondes. En revanche, les aéronefs affectés au nœud fils droit sont caractérisés par de faibles distances de vol prévues. Ces aéronefs survolent les points de leur trajectoire de vol prévue avec une avance modérée d'environ 17.6 secondes.
- La meilleure division ($Distprev1 \leq 351.7$) porte sur les aéronefs dont le vent prévu sur les trajectoires est de face. Elle envoie vers le nœud fils droit les aéronefs dont la distance de vol prévue est supérieure à 351.7 NM. Ces aéronefs avec une distance de vol prévue plus importante et un vent prévu de face prépondérant ont des valeurs élevées (positivement) de la variable *écart temporel*. Ces aéronefs semblent survoler les points de leur trajectoire de vol prévue avec des retards qui se situent autour de 197 secondes. Les aéronefs affectés au nœud fils gauche sont caractérisés par la distance de vol prévue inférieure à 351.7 NM. Dans les conditions de vent de face dominant, ces ces aéronefs avec une faible distance de vol prévue survolent les points de leur trajectoire de vol prévue avec un retard modéré d'environ 51.8 secondes.
- La variable *Distprev1* intervient encore plus en profondeur de l'arbre comme facteur explicatif des écarts temporels, notamment aux quatrième et cinquième niveaux de l'arbre par les divisions ($Distprev1 < 147.45$) et ($Distprev1 < 55.46$) respectivement. Ces divisions portent toutes sur les vols pour lesquels le vent prévu sur leur trajectoire est de dos. La tendance à survoler les points des trajectoires de vol prévues se confirme.
- La distance totale de vol prévue dans le plan de vol initial *(Distpln)*, la différence d'altitude entre le point courant de prévision et le point de la trajectoire de vol prévue (*Difalti*) et l'altitude courante du vol *Alticour* sont des variables utilisées dans les trois premières divisions de substitution pour les divisions optimales ($Distprev1 \geq 414$) et ($Distprev1 < 351.7$). Pour la première, les indices d'association sont respectivement de 82%, 81% et 81% pour les divisions sur *Distpln*, *Difalti* et *Alticour*. Pour la deuxième, les indices comparables sont 77%, 75% et 75%. Ces variables dans leur majorité sont des caractéristiques courantes des aéronefs en vol.

4.4.3.2 Vitesse de vol au point courant de prévision : *Vitessecour*

La vitesse courante des aéronefs (*Vitessecour*) apparaît au quatrième niveau de l'arbre de régression par la division ($Vitessecour \geq 162.5$). Cette dernière envoie vers le nœud fils gauche les aéronefs avec la vitesse courante supérieure à 162.5 kts. Ces aéronefs survolent les points de leur trajectoire de vol prévue

avec un retard estimé en moyenne à 104 secondes contre 675 secondes pour les aéronefs affectés au nœud fils droit caractérisés par la vitesse courante de vol inférieure à 162.5 kts. Ce qui montre que la vitesse est un facteur déterminant qui permet à un aéronef de réaliser un gain de temps pendant le vol. La vitesse courante de vol est est également utilisée dans la deuxième division de substitution pour la division optimale ($Distprev1 < 147.46$) avec un indice d'association d'environ 63%.

Remarque : En somme, la participation des caractéristiques courantes du vol comme variables actives et comme variables utilisées dans les divisions suppléantes montrent l'importance de ces facteurs dans le processus de classification des points des trajectoires des vols et de prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue.

4.4.4 Caractéristiques des vols prévus dans les plans de vols

Dans le partitionnement de l'ensemble des données, l'implication active des variables telles que le niveau de vol demandé pour la croisière (*Nivpln*) et le type d'aéronef utilisé pour le vol (*Type*) montre que les caractéristiques prévues dans les plans de vols ont des effets significatifs dans le processus de prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol. En effet, ces variables apparaissent toutes au troisième niveau de l'arbre et le type d'aéronef apparaît encore plus en profondeur de l'arbre au cinquième niveau.

Même s'il semble difficile de donner une interprétation au niveau de vol demandé pour la croisière, cette variable est la plus discriminante dans le groupe des aéronefs qui survolent les points de leur trajectoire en avance par rapport aux instants prévus dans les plans de vols.

Le type d'aéronef contribue au processus de partitionnement de l'échantillon de données au troisième et cinquième niveaux de l'arbre de régression (Fig.4.8). La première division sur cette variable montre que les aéronefs dont le type de référence appartient à $\{A306$, $AT43$, $B461$, $B712$, $B742$, $BE99\}$ ont des écarts temporels très faibles, soit une avance moyenne de 29.4 secondes pour le survol des points des de leur trajectoire. En revanche, les aéronefs de type de référence « *autres* »sont caractérisés par leur retard important au passage des points de leur trajectoire de vol prévue. Ce retard est estimé à 201 secondes. Par ailleurs, la variable type d'aéronef est utilisée dans la deuxième division de substitution de la division optimale sur le niveau de vol prévu ($Nivpln < 185$) avec un indice d'association de 98%. Ainsi, la participation du type d'aéronef et du niveau de vol demandé pour la phase de croisière comme variables actives et parfois comme variables utilisées dans les divisions suppléantes montrent que les caractéristiques de performances des aéronefs sont des déterminants significatifs dans le processus de classification des points des trajectoires de vols et de la prévision des écarts temporels correspondants.

4.4.5 Paramètres de complexité du trafic

L'arbre de régression obtenu montre que les paramètres de complexité n'ont pas contribué activement à la procédure de classification des points des trajectoires de vol prévues. Cependant, la variable densité

du réseau de routes aériennes (*Moycrois*) est utilisée dans la deuxième division de substitution pour les divisions optimales ($Type$=$A306$, $AT43$, $B461$, $B712$, $B742$, $BE99$) et ($Indur < 1.07$) avec des indices d'association respectifs 95% et 67%. Ce caractère compétitif de *Moycrois* montre bien que n'étant pas actif dans le processus de partitionnement de l'ensemble des individus et de construction de l'arbre de régression, les paramètres de complexité ne sont pas neutres dans la prévision des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue.

4.5 Conclusion

Nous avons proposé dans ce chapitre un modèle de prévision des instants de passage des aéronefs en des points de leur trajectoire de vol prévue. Le modèle ainsi développé est basé sur la méthode CART (Breiman et *al.*, 1984). En utilisant l'ensemble des variables explicatives, nous avons construit un arbre de régression pour proposer d'une part, une classification des points des trajectoires de vol prévues, et d'autre part, pour prévoir les écarts temporels de passage des aéronefs sur ces points pour chaque classe correspondante.

Afin d'évaluer la qualité du modèle d'ajustement et l'amplitude de l'incertitude du modèle ainsi construit, nous avons étudié la variabilité du RMSE (Theil, 1958) en fonction des horizons temporel et spatial de prévision. Il est apparu que la précision du modèle s'améliore au fur et à mesure que l'horizon de prévision se rapproche de l'origine (FIG.4.6 et FIG.4.7).

Dans le processus de construction de l'arbre de régression, les variables explicatives les plus significatives sont : l'influence du vent *(Indur)*, la distance de vol prévue par rapport au point courant *(Distprev1)*, la vitesse du vol au point courant de prévision *(Vitessecour)*, le niveau de vol prévu pour la croisière *(Nivpln)* et le type d'aéronef utilisé *(Type)*. Ainsi, les conditions météorologiques, l'environnement courant des vols et les caractéristiques des plans de vol sont des facteurs les plus explicatifs du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue. Par ailleurs, à travers les divisions suppléantes, la densité de l'air atmosphérique *(Moydensa)* et la densité du réseau de routes aériennes *(Moycrois)* apparaissent aussi comme des facteurs déterminants susceptibles d'avoir des effets sur la prévision de la variable dépendante *écart temporel*.

Rappelons toutefois que le modèle obtenu souffre du problème d'instabilité de l'arbre qui est une défaillance inhérente aux méthodes CART. Nous traitons ce problème au chapitre 6 de ce mémoire.

Dans chaque nœud de l'arbre, l'estimation de la variable explicative par la moyenne ne permet pas toujours de prendre en compte l'ensemble des caractéristiques propres aux différents profils des trajectoires de vol prévues. Au chapitre suivant, nous généralisons le modèle construit ici au modèle de régression multiple intra-classe. Un nouveau modèle est alors construit et permet d'élaborer à l'intérieur de chaque groupe homogène de la typologie fournie dans ce chapitre, un modèle de régression linéaire généralisé. Celui-ci a pour but d'améliorer les prévisions du modèle CART, en proposant des estimations à l'intérieur de chaque groupe en fonction de l'ensemble des caractéristiques les plus pertinentes des vols. Dans toute la suite de ce mémoire, nous parlerons du modèle *CART classique* pour désigner le modèle CART que nous avons proposé dans ce chapitre.

Chapitre 5

Prévisions intra-classes de l'écart temporel de passage des aéronefs sur les points de leur trajectoire par un modèle de régression multiple basé sur l'algorithme Stepwise

5.1 Introduction

En optimisant le critère sur la déviance (l'inertie intra-classe) selon la variable *écart temporel*, nous avons établi au chapitre précédent une typologie des points des trajectoires des vols. Ainsi, pour un vol donné, en fonction des caractéristiques météorologiques et atmosphériques prévues, ses paramètres courants, la complexité du trafic prévue, chaque point de la trajectoire prévue peut être affecté dans un et un seul des 12 nœuds terminaux. Or, dans chacun de ces nœuds, la prévision de l'écart temporel est déterminée par l'estimation de la moyenne de cette variable sur ses mesures à l'intérieur de ce nœud. Pour prendre en compte les caractéristiques spécifiques aux individus à l'intérieur de chaque groupe, nous proposons dans le présent chapitre, un modèle CART modifié où la phase de prévision classique par l'estimation de la valeur moyenne est remplacée par une étape de régression linéaire généralisée. Pour cela, à l'intérieur de chaque sous-groupe homogène obtenu par la méthode *CART classique*, nous proposons un modèle basé sur l'algorithme séquentiel de sélection et de l'élimination automatique des variables explicatives (*Stepwise*). Ce nouveau modèle est parcimonieux, car construit uniquement à partir des seules variables explicatives pertinentes pour le sous-groupe. Il constitue sans doute une première amélioration du modèle précédent, mais sans aucune prétention de résoudre le problème de l'instabilité de l'arbre.

Le chapitre est organisé comme suit : Le paragraphe (5.2) est consacré à la méthodologie. Dans le paragraphe (5.3), nous utilisons le modèle de régression linéaire généralisé pour améliorer les prévisions de l'écart temporel dans les nœuds de l'arbre de régression. En fonction de l'horizon temporel de prévision,

nous réalisons une comparaison de la qualité de ce nouveau modèle avec celle du modèle fourni par *CART classique*. Dans le paragraphe (5.4), nous proposons une interprétation des coefficients des variables explicatives. Le paragraphe (5.5) conclut.

5.2 Méthodologie

La méthodologie adoptée dans ce chapitre repose sur deux étapes essentielles : une pour la classification des points des trajectoires des vols et l'autre pour la prévision de l'écart temporel. La procédure de classification est exactement la même que celle utilisée dans le modèle par arbre de régression CART. En effet, l'élaboration de ce modèle reposait sur deux phases : La première consistait en la classification des individus en des sous-groupes homogènes et la deuxième consistait à prévoir la variable dépendante par l'estimation de sa valeur moyenne à l'intérieur de chaque sous-groupe (nœud terminal).

Dans le présent chapitre, nous proposons une modification dans la phase de prévision du modèle CART classique. Pour cela, dans chaque sous-groupe, la prévision de la variable dépendante est déterminée par un modèle de régression linéaire généralisé utilisant l'algorithme « *Stepwise* ». Dans tout le reste du mémoire, le nouveau modèle proposé dans ce chapitre sera appelé *CART modifié*. La qualité d'ajustement de ce modèle est ensuite comparée à celle du modèle CART classique en utilisant le coefficient de Theil (1958).

5.3 Modèle de régression multiple intra-classe

Rappelons qu'il s'agit de l'élaboration des modèles de prévision de la variable *écart temporel* en fonction des variables explicatives des vols. Ainsi, pour chaque nœud terminal obtenu précédemment, un modèle de régression linéaire multiple par l'algorithme *Stepwise* est déterminé en fonction des paramètres pertinents caractérisant les points des trajectoires de vol prévues appartenant à ce nœud.

5.3.1 Spécification du modèle de régression intra-classe

Nous considérons $\mathbf{E}$, l'échantillon de données présenté au chapitre précédent. Maintenant, notre objectif consiste à exprimer les valeurs de la variable dépendante $\mathbf{Y}$ en fonction du vecteur des variables explicatives $\mathbf{X}$. Supposons construit la partition $(K_i)_{(1\leq i\leq n_0)}$ de l'ensemble $\mathbf{E}$. Nous pouvons formuler le modèle par la relation :

$$\mathbf{Y} = \sum_{i=1}^{n_0} \mathbf{Y}_{K_i}.\mathbb{1}_{K_i}, \tag{1}$$

avec

$$\mathbf{Y}_{K_i} = \beta_{K_i}\mathbf{X}_{K_i} + \epsilon_{K_i}, \tag{2}$$

où $\mathbf{Y}_{K_i}$ est la variable dépendante pour les individus de l'échantillon appartenant au nœud terminal K_i, $\mathbf{X}_{K_i}$ est un vecteur de variables explicatives pour les observations dans K_i, $\mathbb{1}_{K_i}$ est la fonction indicatrice

de l'ensemble K_i, β_{K_i} est un vecteur de paramètres à déterminer dans le modèle de l'équation (2) et ϵ_{K_i} est la variable des résidus du modèle sur K_i.

Si nous notons $\widehat{y}_{K_i}(.)$ un estimateur $y_{K_i}(.)$, alors la prévision de la variable dépendante par l'équation (1) devient :

$$\widehat{y}(x) = \sum_{i=1}^{n_0} \widehat{y}_{K_i}(x).\mathbb{1}_{\{x \in K_i\}}, \tag{3}$$

où x et y sont des réalisations des variables aléatoires $\mathbf{X}$ et $\mathbf{Y}$ respectivement, en supposant que ϵ_{K_i} est d'espérance nulle, on écrit :

$$\widehat{y}_{K_i}(x) = \widehat{\beta}_{K_i} x_{K_i}.$$

Finalement, pour trouver une solution à l'équation (1), il suffit pour chaque nœud terminal K_i de déterminer un vecteur d'estimateurs $\widehat{\beta}_{K_i}$ des coefficients de régression linéaire multiple défini par la relation (2). Pour cela, nous avons utilisé la méthode de sélection de variables explicatives fondée sur l'algorithme *stepwise*. Avant de présenter les statistiques sommaires sur les résidus des modèles CART (classique, modifié), rappelons d'abord quelques principes de fonctionnement du modèle de régression par l'algorithme *Stepwise*.

5.3.2 Principe de fonctionnement de l'algorithme Stepwise

Plutôt que de chercher à expliquer la variable y par toutes les p variables explicatives, on peut chercher seulement un ensemble de q variables parmi les p qui donne une reconstitution presque aussi satisfaisante de y. Les objectifs visés par une telle démarche sont nombreux : économiser le nombre de prédicteurs (variables explicatives dans un modèle), obtenir les formules d'un bon pouvoir prédictif en éliminant des variables redondantes qui augmentent le biais, obtenir un modèle plus facile à interpréter. Pour cela plusieurs techniques de sélection des variables ont été proposées : La recherche exhaustive et les méthodes de pas à pas en particulier l'algorithme stepwise.

Les méthodes pas à pas sont utilisées lorsque p est élevé et qu'il n'est pas possible de procéder à une recherche exhaustive. Elles procèdent par élimination successive ou ajout successif de variables :

- La méthode descendante consiste à éliminer la variable la moins significative parmi les p, en général celle qui provoque la diminution la plus faible des R^2 coefficient de détermination multiple. On recalcule alors la régression et on recommence jusqu'à élimination de $p-1$ variables.
- La méthode ascendante procède en sens inverse : on part de la meilleure régression à une variable et on ajoute celle qui fait progresser le plus R^2.
- La méthode stepwise est un perfectionnement de l'algorithme précédent. Elle consiste à effectuer en plus à chaque pas des tests de signification du type Student ou de Fisher (F) pour ne pas introduire une variable non significative et pour éliminer éventuellement des variables déjà introduites qui ne seraient plus informatives compte tenu de la dernière variable sélectionnée. L'algorithme s'arrête quand

on ne peut plus ajouter ni retrancher de variables.

Ces méthodes ne mettent cependant pas à l'abri de l'élimination intempestive de variables réellement significatives, ce qui risque de biaiser les résultats.

5.3.3 Comparaison des statistiques sommaires des résidus des modèles CART classique et CART modifié

Les statistiques sommaires des résidus pour les deux modèles sont consignées dans la table TAB.5.1.

N' Nœuds		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Ecart temporel moyen		**-1760**	-344	-152	**-441**	-67.8	4.35	104	**244**	**1290**	51.8	127	333
Taille		41	1133	1456	204	3679	6660	474	57	40	8466	1840	949
CART classique avec prévision par la moyenne	Min	**-2830**	-2007	-1235	**-1706**	-1648	-285	-595.10	**-308.8**	**-1477**	-1790	-1834	-1873
	1^{er}Qu.	**-1226**	-167.2	-172.1	**-361.70**	-117.7	-42.35	-148.8	**-209.8**	**-601.9**	-65.84	-192.9	-199.4
	Mediane	**835.1**	35.82	-0.37	**55.01**	-12.19	8.65	-85.09	**-101.8**	**-229.4**	-14.84	-10.1	9.57
	Moyenne	**0**	0	0	**0**	0	0	0	**0**	**0**	0	0	0
	3^{ème}Qu	**1266**	181.8	147.4	**297.8**	103.8	48.65	17.66	**140.2**	**634.6**	53.16	184.4	160.6
	Max	**1728**	1386	2846	**1546**	1394	1013	2149	**1190**	**2161**	3064	2783	3440
CART modifié (algorithme (Stepwise)	Min	**-267.3**	-1515	-810.7	**-1714**	-1370	-2643	-684.3	**-375.3**	**-526.4**	-1961	-1716	-1833
	1^{er}Qu.	**-62.41**	-147.4	-165.1	**-189.7**	-110.3	-44.34	-140.1	**-89.82**	**-156.5**	-63.92	-190.1	-147.5
	Mediane	**-14.03**	-1.51	-4.1	**36.76**	-9.73	-0.47	-18.28	**-11.11**	**12.18**	-2.08	-12.32	-0.02
	Moyenne	**0**	0	0	**0**	0	0	0	**0**	**0**	0	0	0
	3^{ème}Qu.	**62.41**	145	144.7	**176.9**	91.70	49.21	86.38	**90.58**	**207.9**	60.64	171.9	152.1
	Max	**330.30**	1393	2493	**766**	1422	1041	1885	**669.9**	**574.6**	2762	2624	3239
	R^2**-ajusté**	**0.99**	0.32	0.17	**0.57**	0.18	0.10	0.25	**0.55**	**0.90**	0.07	0.12	0.20

TAB. 5.1 – Synthèse des résidus des modèle *CART classique* (prévision par la moyenne) et *CART modifié* (prévision par régression linéaire multiple). Chaque colonne représente les statistiques sur chaque nœud de l'arbre de régression CART correspondant.

Au regard des résultats des statistiques sommaires sur les résidus respectifs des deux modèles, les écarts résiduels semblent plus faibles pour le modèle *CART modifié* (voir nœuds $n°1$ et $n°9$). Il s'agit notamment des nœuds où la proportion de la variance expliquée (coefficient de détermination) de l'écart temporel est la plus importante avec des valeurs respectives de 99% et 90%. Ces nœuds semblent se distinguer des autres par des valeurs algébriques élevées de la variable dépendante. En effet, pour le nœud $n°1$, la valeur moyenne de cette variable se situe autour de 1760 secondes en avance par rapport aux instants prévus dans les plans de vol alors que la statistique comparable pour le nœud $n°9$ est de 1290 secondes, mais en retard. Le troisième et le quatrième coefficient de détermination ajustés les plus élevés sont observés sur les nœuds terminaux $n°4$ (57%) et $n°8$ (55%) qui eux aussi se distinguent des autres par les valeurs algébriques fortes de l'écart temporel. Ces valeurs sont respectivement 441 secondes en avance et 244 secondes en retard. Sur ces nœuds, les résidus du modèle CART modifié sont beaucoup moins dispersés relativement à ceux obtenus à l'aide du modèle CART classique. La proportion de la variance expliquée de l'écart temporel dans les autres nœuds semblent trop faibles pour permettre une meilleure interprétation des résultats du modèle dans ces classes. Pour ces nœuds, le modèle CART modifié ne semble pas avoir apporté une amélioration significative par rapport aux prévisions obtenues à l'aide du modèle CART classique. Toutefois, une analyse comparative de la dispersion des résidus des deux modèles en fonction de l'horizon temporel de prévision permettra sans doute de quantifier le gain de performance apporté par le modèle CART modifié.

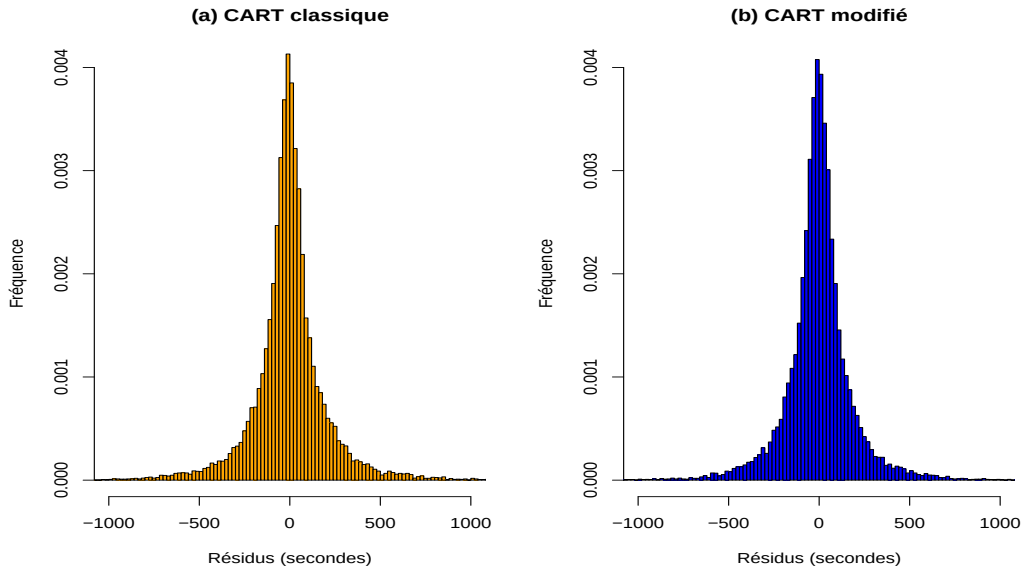

FIG. 5.1 – Histogrammes de distribution des résidus des modèles CART classique et CART modifié (Stepwise)

5.3.4 Test d'ajustement des résidus par une loi normale

L'histogramme de la distribution de fréquence de ces résidus est illustré par la figure FIG.5.1. Dès le chapitre précédent, nous avions utilisé le test de normalité de Kolmogorov-Smirnov pour réfuter l'hypothèse selon laquelle la distribution des résidus est ajustée par une loi normale. En appliquant le même test sur les résidus du modèle CART modifié, on observe une très légère amélioration de la statistique de Kolmogorov-Smirnov. Elle est d'environ 0.49 contre 0.50 pour le modèle CART classique. On rejette encore l'hypothèse de normalité pour la distribution des résidus de ce nouveau modèle. Le test asymptotique de Jarque-Bera fournit une statistique égale à 535350 avec une p-value égale à 0. On conclut à nouveau au rejet de l'hypothèse de la normalité des résidus. Comme pour le premier modèle, une loi d'ajustement de la densité des résidus de ce modèle sera déterminée plus loin dans ce mémoire à l'aide des méthodes d'estimation des composantes de mélange de lois gaussiennes.

5.3.5 Dispersion des résidus en fonction de l'horizon temporel de prévision

C'est dans les nœuds $n^{\circ}1$ et $n^{\circ}9$ que nous observons une compression remarquable des résidus pour le nouveau modèle. En effet, si dans le nœud n° 1, le modèle CART classique a des résidus sont compris entre -2830 et 1728, les limites comparables pour le modèle CART modifié sont de -267.3 et 330.30. De même, pour le nœud $n^{\circ}9$, le premier modèle a des résidus variant entre -1477 et 2161 contre -526.4 et 574.6 pour le second. L'illustration des boxplots dans la figure FIG.5.2 montre qu'il n'est pas évident d'identifier d'autres différences entre les deux modèles. Pour cela, nous avons réalisé une étude comparée de la variation du RMSE de ces deux modèles ajustés, en fonction de l'horizon temporel de prévision. La synthèse des résultats est illustrée dans la figure FIG.5.3. Une lecture de ce graphique montre que lorsque

l'horizon de prévision est éloigné de l'origine, la dispersion des résidus du modèle CART modifié croît moins vite que celle des résidus du modèle CART classique. Ce qui traduit une amélioration de l'incertitude dans les prévisions fournies par ce nouveau modèle.

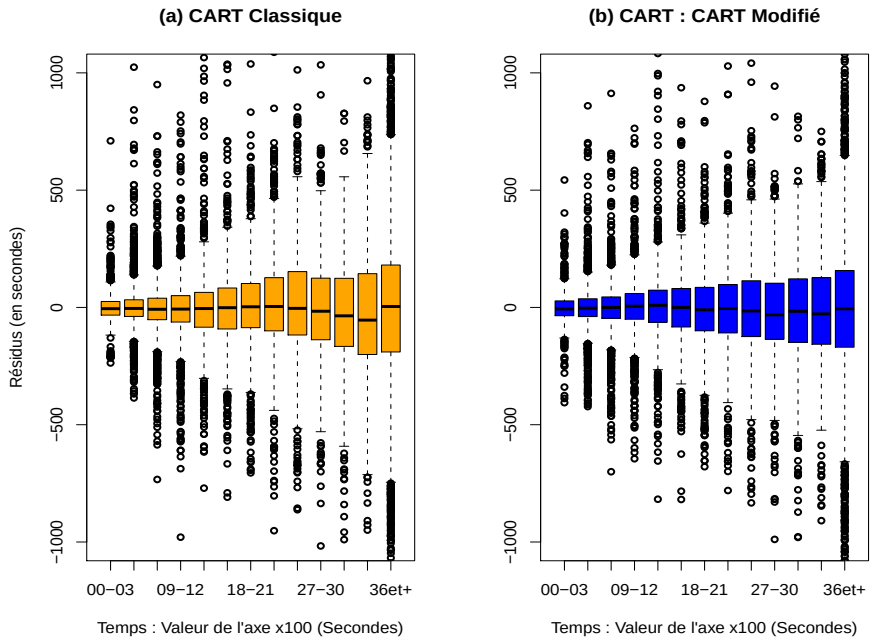

FIG. 5.2 – Dispersion des résidus en fonction de l'horizon temporel de prévision pour les modèles : CART classique et CART modifié (Stepwise). 00-03 indique par exemple un horizon de prévision compris entre entre 0 et 300 secondes.

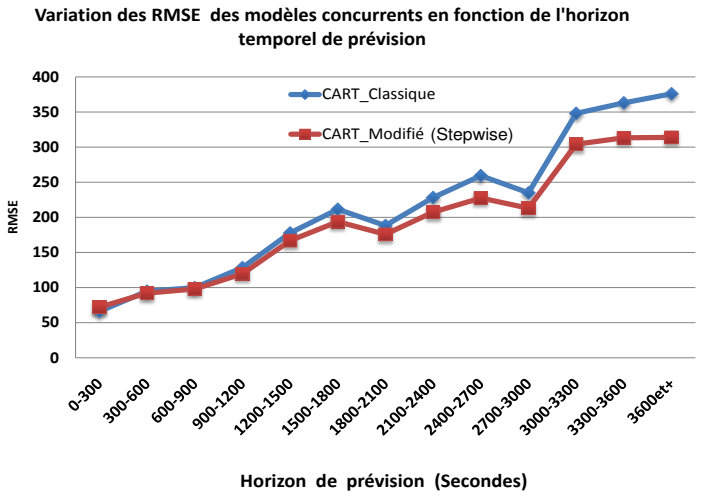

FIG. 5.3 – Evolution comparée de l'incertitude en fonction de l'horizon temporel de prévision du modèle CART classique à celle de CART modifié (Stepwise)

5.3.6 Comparaison de la qualité des deux modèles

Nous utilisons le coefficient de Theil (1958) pour comparer une prévision avec celle issue d'une méthode alternative. Ce coefficient est constitué du RMSE du modèle de référence normé par celui du modèle alternatif (Mathis A. et *al.*, 1994)[3]. Il est donné par la relation :

$$Theil = \frac{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i^{ref})^2}{n}}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i^{alt})^2}{n}}}. \tag{4}$$

Si le coefficient de Theil est égal à un, les prévisions du modèle ne sont pas meilleures que celles données par le modèle alternatif. Si le coefficient de Theil est supérieur à un, les prévisions du modèle de référence sont plus mauvaises que celles données par le modèle alternatif. Dans notre étude, nous considérons comme modèle alternatif, le modèle CART modifié. Ainsi, le coefficient de Theil est égal à 1.15 supérieur à 1. Ainsi, de façon globale, les prévisions du modèle CART modifié sont bien meilleures que celles données par le modèle CART classique.

Nous avons étudié ce coefficient en fontion de l'horizon temporel de prévision. Le calcul numérique de cette variation est résumé dans la figure FIG.5.4 ci-dessous. Il apparaît que pour un horizon temporel de prévision inférieur à 600 (10 minutes), le modèle CART classique fournit des prévisions bien meilleures que celles données par CART modifié. En revanche, au dela de cet horizon, cette tendance s'inverse et le gain apporté dans la qualité des prévisions de l'écart temporel par le modèle CART modifié croît progressivement à mesure que l'horizon de prévision augmente. Nous nous intéressons dans le paragraphe suivant à déterminer les effets des variables explicatives ayant contribué de façon significative à la construction des modèles de régression multiple intra-classes.

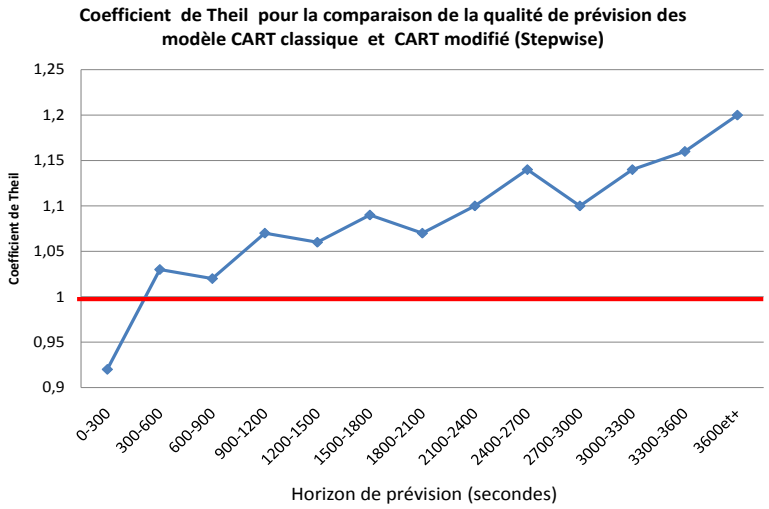

FIG. 5.4 – Coefficient de Theil pour la comparaison de la qualité des modèles CART classique et CART modifié (Stepwise). Ce coefficient est égal au rapport du RMSE du modèle CART classique sur le RMSE du modèle CART modifié. Ref = référence et alt = alternatif.

5.4 Effets des facteurs explicatifs du modèle *CART modifié*

Dans ce paragraphe, nous présentons les effets des facteurs explicatifs des écarts temporels de passage des aéronefs en des points de leur trajectoire de vol prévue. Pour cela, à l'intérieur des nœuds obtenus par la méthode CART classique, nous proposons une interprétation des coefficients du modèle de régression linéaire obtenu. Cette interprétation sera limitée uniquement aux nœuds dans lesquels la proportion de la variance de l'écart temporel est supérieure à 50%. Les nœuds terminaux concernés sont $n^{\circ}1$, $n^{\circ}4$, $n^{\circ}8$ et $n^{\circ}9$ avec les coefficients de détermination multiples ajustés respectivement de 99%, 57%, 55% et 90%. Ces coefficients sont disponibles en annexe respectivement dans les tables TAB.B.6, TAB.B.7, TAB.B.9 et TAB.B.10. D'après l'arbre de régression obtenu dans le chapitre précédent, tous ces nœuds sont en particulier des sous-ensembles du nœud fils gauche de la division optimale ($Indur < 1.013$).

5.4.1 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}1$

Ce nœud regroupe les aéronefs dont la distance de vol prévue est supérieure à 414 NM et dont le niveau de vol demandé dans la phase de croisière est inférieur à FL185. Une lecture des coefficients de la régression multiple dans ce nœud (TAB.B.6, annexe) montre que la pression atmosphérique, la distance de vol prévue, le taux de montée/descente, la vitesse courante du vol ont des effets significatifs sur les instants de passage des aéronefs en des points de leur trajectoire de vol. Le signe négatif de leur coefficient de régression indique que plus ces variables ont des mesures élevées par rapport à leur moyenne, plus les aéronefs sont accélérés dans leur mouvement. Ce qui se traduit par une amplification de l'incertitude dans les prévisions. En effet, les aéronefs survolent les points de leur trajectoire très en avance par rapport aux instants prévus dans leur plan de vol.

La complexité du trafic, notamment la densité du réseau des routes aériennes présente des effets opposés à ceux des variables précédentes sur l'écart temporel. Le signe positif de son coefficient dans la régression multiple indique que lorsque ses mesures augmentent par rapport à sa moyenne, elle contribue à retarder les aéronefs dans leur mouvement. Notons cependant que malgré la valeur du coefficient de détermination élevée (99%), la faible taille (41) de l'échantillon rend difficile l'interprétation des coefficients de ce modèle.

5.4.2 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}4$

Ce nœud à la particularité de ne contenir que les aéronefs d'un seul type de référence ($AT43$). Selon les données de performances BADA, il s'agit des aéronefs dont le niveau de vol maximal à charge nominal se situe autour de $FL240$ avec une vitesse de croisière correspondante de 272 kts. A l'intérieur de ce nœud, la densité de l'air et la température apparaissent respectivement avec les coefficients de régression négatif et positif (TAB.B.7, annexe). L'augmentation de la densité par rapport à sa moyenne semble accélérer les aéronefs dans leur mouvement. En revanche, l'accroissement de la température par rapport à la même référence a tendance à réduire la force de traction et peut ainsi retarder les aéronefs pendant leur vol.

Comme sur le nœud $n^{\circ}1$, la densité du réseau des routes aériennes et le score de complexité du trafic

ont des effets significatifs sur l'écart temporel et contribue particulièrement à l'amplification de l'incertitude lorsque le mouvement des aéronefs devient très lent. On observe aussi que la vitesse courante et le taux de montée/descente ont des effets similaires à ceux du nœud $n^{\circ}1$.

5.4.3 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}8$

Ce nœud regroupe les vols dont la distance de vol prévue est inférieure à 55.46 NM, la vitesse courante de vol est inférieure à 162.5 kts et l'influence du vent est inférieure à 1.013. Nous pensons que ces sont dans la zone d'approche à l'arrivée.

Le signe positif (TAB.B.9, annexe) du coefficient de la distance de vol prévue indique que plus un aéronef est loin du point prévu de sa trajectoire, plus l'incertitude sur son instant de passage en ce point est élevée, car les aéronefs sont plus retardés dans leur mouvement. En revanche, la variable altitude courante du vol est de signe négatif, ce qui indique que plus les aéronefs volent haut, plus ils ont tendance à être en avance.

Le retard bloc des aéronefs est significatif dans le modèle et montre que le passé d'un vol peut influer sur le comportement du pilote pendant le reste du vol. Son coefficient dans la régression linéaire multiple étant positif, donc les aéronefs qui ont fait l'objet d'un retard dans le passé peuvent voir ce retard se propager sur le reste de leur vol.

5.4.4 Facteurs explicatifs de l'écart temporel pour le nœud $n^{\circ}9$

Par rapport à la distance de vol prévue, les aéronefs de ce groupe sont opposés à ceux du groupe $n^{\circ}8$ et ont donc la distance de vol prévue plus longue, supérieure à 55.46 NM. Les effets des variables telles que l'influence du vent, la distance du vol prévue et le retard bloc sont significatifs sur la variable écart temporel et dans le même sens que ceux observés dans le nœud $n^{\circ}8$ (TAB.B.10, annexe). Par ailleurs la faible taille (40) de l'échantillon ne permet pas réellement de tirer une règle générale sur le comportement des aéronefs dans ce nœud.

5.5 Conclusion

Nous avons proposé dans ce chapitre une amélioration du modèle de régression par arbre CART (Breiman et *al.*, 1984). Celle-ci a consisté à remplacer les prévisions par la moyenne de la variable dépendante de la méthode CART classique, par une étape d'estimation de prévisions par la régression linéaire multiple à l'intérieur des nœuds terminaux obtenus par CART. Le modèle de régression linéaire ainsi développé à l'intérieur de chaque nœud est parcimonieux, car il utilise l'algorithme séquentiel de sélection et d'élimination automatique de variables explicatives « *stepwise* ». Cette méthode permet de ne retenir dans le modèle que les variables significatives et pertinentes pour expliquer la variable dépendante.

Ce nouveau modèle que nous avons appelé *CART modifié* améliore la précision des prévisions de l'écart temporel, donc celles des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue. En

particulier, il fournit des prévisions bien meilleures que celles données par CART classique pour les horizons de prévision éloignés de l'origine. Il apparaît que l'utilisation du modèle de régression linéaire multiple pour l'estimation de la variable dépendante au lieu de la moyenne est une première tâche accomplie dans le but d'améliorer la qualité des prévisions une fois que l'arbre de régression est construit.

Nous proposons au chapitre suivant un autre type de modèle utilisant une procédure adaptative qui combine à la fois le partitionnement récursif dont CART est un cas particulier et la régression de la variable dépendante sur les bases des fonctions splines construites à partir des variables explicatives.

Chapitre 6

Prévision de l'écart temporel de passage des aéronefs sur les points de leur trajectoire par le modèle MARS

6.1 Introduction

L'objectif de ce chapitre est de proposer un modèle qui permette d'améliorer la précision des prévisions de l'écart temporel fournies par les deux méthodes précédentes. Le modèle que nous proposons maintenant est une généralisation des méthodes basées sur le partitionnement récursif de l'échantillon des données. C'est une procédure de régression multivariée par les splines adaptatives (MARS) pour construire des modèles de régression non-paramétriques dont l'estimation des valeurs de la variable d'intérêt est une somme de produits de fonctions polynômes par morceaux, c'est-à-dire des splines. Ces fonctions peuvent être non-linéaires des p variables explicatives $X_1, ..., X_p$.

La méthode MARS a l'avantage de permettre la modélisation des interactions entre les variables explicatives, car la condition hiérarchique[1] des prédicteurs dans les modèles est moins contraignante. Un prédicteur peut ainsi intervenir dans une interaction sans être au préalable une variable explicative significative dans le modèle. Inspiré sur les méthodes de régression par partitionnement récursif, le modèle MARS se distingue de la méthode CART classique par la continuité de sa fonction de régression.

Nous avons organisé le chapitre comme suit : Le paragraphe (6.2) fait d'abord une revue de la littérature liée à la méthode MARS. Il se focalise ensuite sur la spécification du modèle et définit la base de fonctions splines dans laquelle une estimation de la variable à expliquer sera exprimée. Il s'intéresse également à la construction et au rappel des propriétés des fonctions qui forment cette base. Enfin, la procédure d'estimation de chaque fonction de la base est synthétisée au travers d'un algorithme. Dans le paragraphe (6.3), nous

[1] La condition hiérarchique exige que pour qu'une variable explicative soit facteur d'un terme d'interaction dans un modèle, il faut que cette variable prise séparément soit déjà un terme additif significatif dans le modèle.

appliquons cette approche pour construire un modèle de prévision de la variable écart temporel. Le paragraphe (6.4) dédié aux résultats, analyse les effets directs des variables explicatives et de leurs interactions sur la variable à expliquer. Enfin, le paragraphe (6.5) conclut le chapitre.

6.2 Méthode MARS

6.2.1 Littérature existante

La méthode MARS developpée par Friedman (1991)[30] est une généralisation des méthodes de régression par partitionnement récursif. Beaucoup de travaux ont été publiés sur les modèles relatifs au partitionnement récursif dont le plus populaire est la méthode CART qui produit les arbres de régression où la fonction de régression est constante sur chaque feuille de l'arbre. Certains auteurs l'ont étendu à la régression sur les données censurées. Dans cette lignée, on cite les travaux de Davis et *al.* (1989) [20], de Kwak et *al.* (1990) [40] et de Leblanc et *al.* (1992) [41]. Dans leurs travaux, Kooperberg et *al.* (1995) [39] ont développé une méthode adaptative pour la régression de risque (HARE[2]). Cette dernière utilise les régressions linéaires par morceaux sur les splines où les nœuds et les variables de division sont sélectionnés en fonction de leur contribution dans la fonction de risque. Les résultats de cette technique sont très proches de ceux obtenus par la méthode MARS. Etudiant le modèle à risque proportionnel, Gray (1992) [34] utilise une technique fixant les nœuds des splines (La notion de nœud de spline est définie plus loin dans ce chapitre). Dans ce cas, la procédure de recherche des points de division n'est pas optimisée. Buja et *al.* (1991) [12] et Stone (1991) [58] ont proposé une méthode basée sur les moindres-carrés pondérés pour étendre la technique de régression par la méthode MARS aux modèles de régression généralisés. Leblanc et *al.* [42] ont développé une méthode pour construire des modèles adaptatifs de régression par spline pour l'exploration des données de survie. Leur méthode combine le modèle de régression de Cox (1972) [18] avec une version des moindres carrés pondérés de la technique de régression multivariée adaptative par splines (MARS) pour sélectionner de façon adaptative les variables qui interviennent dans les termes des produits des splines dans le modèle et les nœuds correspondants.

La méthode MARS se présente comme une extension de la méthode CART car elle construit un estimateur de la variable à expliquer au moyen de partitionnement récursif, mais avec la contrainte de la continuité des prévisions. Ce modèle est utilisé dans ce chapitre pour proposer un estimateur de prévision de la variable dépendante *écart temporel*. Cette approche permet ainsi de prendre en compte d'une part, la continuité des prévisions entre les sous-domaines adjacents et d'autre part, pour tenir compte dans le modèle des effets non-linéaires et de ceux des interactions entre les variables explicatives sur le temps de passage des aéronefs sur les points de leur trajectoire de vol prévue.

[2] Adaptative Hazard Regression

6.2.2 Spécification du modèle MARS

Posons : $\mathbf{X} = (X_1, \ldots, X_p)$ un vecteur de variables explicatives, $\mathbf{Y}$ est la variable dépendante. Le modèle d'estimation de $\mathbf{Y}$ en fonction de $\mathbf{X}$ par méthode la MARS peut être formulé par une équation de la forme :

$$Y = f(\mathbf{X}) + \epsilon, \tag{1}$$

où ϵ est le résidu du modèle, $f(\mathbf{X})$ est une fonction de $\mathbf{X}$ définie par l'équation :

$$f(\mathbf{X}) = \beta_o + \sum_{k=1}^{m} \beta_k B_k(\mathbf{X}), \tag{2}$$

$\mathscr{B} = (B_0, B_1, \ldots, B_m)$ une base de fonctions de $\mathbf{X}$, m est le nombre de termes différents de la fonction constante dans l'expression de $f(\mathbf{X})$, β_o est le coefficient de la fonction constante $B_0(\mathbf{X}) = 1$, pour $k \in \{1, \ldots, m\}$, β_k est le coefficient de la fonction $B_k(\mathbf{X})$ différente de la fonction constante.

Le but de la méthode de régression multivariée par les splines adaptatives est de construire $f(\mathbf{X})$ qui est un estimateur de la variable $\mathbf{Y}$. Cet estimateur est une somme de produits des fonctions linéaires par morceaux sur les variables explicatives. Il sera entièrement construit par la détermination de chaque fonction non constante de la base $\mathscr{B}$. Ces fonctions sont construites en utilisant une procédure itérative sur k. Avant de présenter cette procédure, nous nous intéressons d'abord à la définition d'une fonction B_k de la base $\mathscr{B}$.

6.2.3 La fonction $B_k(\mathbf{X})$ et le nœud d'une spline

Avant toute chose, rappelons qu'une spline est une fonction définie par morceaux et par des polynômes. Dans l'équation (2), la fonction $B_k(\mathbf{X})$ est une spline construite à l'itération k. Elle peut être linéaire par morceau ou un produit de plusieurs fonctions linéaires par morceaux. Notons K_k le nombre de fonctions linéaires par morceaux qui interviennent dans la définition de la spline $B_k(\mathbf{X})$ à l'itération k, on écrit :

$$B_k(\mathbf{X}) = \prod_{j=1}^{K_k} b_{jk}(\mathbf{X}). \tag{3}$$

Posons $\mathbf{X}_{(k,j)}$, la variable explicative $\mathbf{X}_j$ utilisée à l'itération k pour construire la spline $B_k(\mathbf{X})$. Pour une valeur c_{kj} de cette variable, l'échantillon des données est divisé en deux sous-domaines gauche et droite. On parle de la division de l'échantillon des données. Celle-ci permet de construire une paire de fonctions splines linéaires $b_{jk}(X)$ définie sur la variable $\mathbf{X}_j$ à l'itération k par :

$$\begin{aligned} b_{jk}(X) &= (X_{(k,j)} - c_{kj})^+ \\ b_{j,k+1}(X) &= (c_{kj} - X_{(k,j)})^+, \end{aligned}$$

où nous avons par définition :

$$\begin{aligned} (X_{(k,j)} - c_{kj})^+ &= \max(0, (X_{(k,j)} - c_{kj})), \\ (c_{kj} - X_{(k,j)})^+ &= \max(0, -(X_{(k,j)} - c_{kj})). \end{aligned}$$

Dans ces fonctions, c_{kj} est le seuil de division de l'échantillon de données sur la variable explicative $\mathbf{X}_{(k,j)}$. On l'appelle le nœud des splines linéaires $b_{jk}(X)$ et $b_{j,k+1}(X)$. Dans la méthode MARS, le seuil de division ou encore nœud de spline est une mesure observée de la variable explicative sur laquelle la division a lieu alors que dans la méthode CART, ce seuil n'est pas nécessairement une mesure de la variable explicative utilisée dans la division.

Dans le cas où la fonction $B_k(X)$ est une seule fonction linéaire par morceau, elle permet d'évaluer dans le modèle, les effets principaux ou directs de la variable explicative correspondante sur la variable à expliquer Y. En revanche, lorsque c'est un produit de plusieurs fonctions linéaires par morceaux, elle permet au modèle d'évaluer les effets des interactions entre les variables explicatives intervenant dans ce produit sur Y. Ainsi, dans l'élaboration du modèle, il sera nécessaire de spécifier le degré d'interactions désiré qui est le nombre maximum de variables explicatives devant intervenir dans chaque fonction de la base $\mathscr{B}$.

Les fonctions linéaires $b_{jk}(X)$ étant construites de façon itérative. A chaque itération, deux termes potentiels du produit définissant la $k^{ème}$ fonction $B_k(\mathbf{X})$ sont construits par une des variables explicatives du vecteur $\mathbf{X}$. Ces termes sont des fonctions linéaires par morceaux $b_{jk}(X)$ et $b_{j,k+1}(X)$. La variable à faire entrer dans la définition de $B_k(\mathbf{X})$ doit être celle qui n'est pas encore présente (comme facteur du produit) dans $B_k(\mathbf{X})$. Toutes les paires des fonctions linéaires par morceaux doivent être considérées et évaluées à chaque étape courante de construction de la fonction $B_k(\mathbf{X})$. La paire qui entre finalement dans la définition de cette dernière est celle qui minimise le critère des moindres carrés. K_k est finalement le nombre de variables explicatives utilisées pour définir la fonction $B_k(\mathbf{X})$. Avant de présenter l'algorithme qui synthétise la construction des fonctions de la base $\mathscr{B}$, rappelons quelques unes de leurs propriétés.

A partir de l'échantillon d'étude, la méthode MARS détermine le nombre de nœuds et leur position pour les différents prédicteurs. Elle utilise les propriétés suivantes pour construire rigoureusement les fonctions à faire entrer dans la base des splines :

- La base $B_0(\mathbf{X}) = 1$ est la première à entrer dans le modèle et doit toujours y rester.
- Les bases participant à la définition de l'estimateur $f(\mathbf{X})$ de Y sont des fonctions linéaires par morceaux. Elles sont de la forme $(X_{(k,j)} - c_{kj})^+$ ou $(c_{kj} - X_{(k,j)})^+$ et de leurs produits. Rappelons que les c_{kj} sont les nœuds des fonctions splines linéaires.
- L'ensemble des nœuds potentiels pour la division de l'échantillon de données sont des valeurs observées $X_{j(i)}$ où $j = 1, ..., p$ et $i = 1, ..., n^*$. Il s'agit de l'ensemble de toutes les statistiques d'ordre pour chaque variable explicative.
- Les produits des splines doivent toujours être formés de facteurs dont chacun contient une et une seule variable explicative.

6.2.4 Algorithme MARS

L'algorithme MARS comprend trois étapes essentielles. La première est une phase de construction et d'addition des fonctions de la base. Ce processus d'addition des fonctions splines est repété jusqu'à la

construction du plus grand modèle (au sens du nombre de fonctions dans la base). Ce modèle maximal contient en général plus de fonctions splines qu'il en faut pour un modèle optimal. La deuxième étape de l'algorithme est une phase d'élimination. A chaque étape, elle consiste à éliminer par la procédure Backward Stepwise le terme qui augmente la somme des carrés des résidus (critère des moindres carrés). La troisième étape porte sur la sélection du modèle optimal. En général, le modèle qui minimise le score de la Cross-validation généralisée (GCV : Cross-validation generalized) est sélectionné (Craven et *al.*, 1979) [19].

6.2.4.1 Etape 1 : Addition

Les fonctions de la base $\mathscr{B}$ étant construites de façon itérative. L'algorithme MARS commence par ajuster le modèle avec pour seule fonction de base, la fonction constante $B_0(\mathbf{X}) = 1$. On obtient ainsi l'estimateur $f(\mathbf{X})$ de Y tel que :

$$f(\mathbf{X}) = \widehat{\beta}_0 B_0(\mathbf{X}).$$

Pour faire le premier pas d'addition, on passe à la deuxième itération. On cherche alors la variable explicative et un point de cette dernière qui permettent d'obtenir une meilleure division de l'échantillon de données en deux sous-groupes disjoints. Ce point de division doit permettre d'améliorer l'estimation de $f(\mathbf{X})$ en utilisant deux splines linéaires. On impose un nombre minimal de points entre les nœuds pour s'assurer d'obtenir un estimateur convergent. Parmi tous ces points, le choix de la division optimale s'effectue de la manière suivante :

- Pour chaque variable explicative X_j, $j = 1, ..., p$, on considère $X_{j(i)}$ la valeur de cette variable portée par l'individu $i = 1, ..., n$ telle que $c_{ji} = X_{j(i)}$ est un nœud potentiel de division pour cette variable. On ajuste le modèle

$$\begin{aligned} Y(X) &= a_0 B_0(X) + a_1 B_0(X)(X_j - c_{ji})^+ + a_2 B_0(X)[-(X_j - c_{ji})]^+ \\ &= a_0 + a_1 (X_j - c_{ji})^+ + a_2 [-(X_j - c_{ji})]^+ \end{aligned}$$

  où a_0, a_1 et a_2 sont des coefficients à estimer par un modèle de régression.
- Pour chacun de ces modèles, on calcule la somme des carrés des résidus. Le meilleur point de division est celui qui permet de minimiser cet indicateur.
- Notons X_{j^*} la variable d'indice j^* qui entre dans le modèle et c_{j^*} le meilleur point de division pour cette variable parmi toutes les $X_{j(i)}$ explorées. On construit deux nouvelles fonctions pour le modèle de la façon suivante :

$$B_1(X) = B_0(X)(X_j - c_{j^*})^+ = (X_j - c_{j^*})^+ \tag{4}$$

$$B_2(X) = B_0(X)[-(X_j - c_{j^*})]^+ = [-(X_j - c_{j^*})]^+ \tag{5}$$

  $B_1(X)$ et $B_2(X)$ sont deux fonctions de la base construites à la 2ème itération de l'algorithme MARS.

De façon analogue, on construit des fonctions pour les itérations suivantes. A chaque itération deux splines linéaires sont construites à la fois. Supposons que le modèle de dimension $2K-1$ de l'itération $K-1$ ($K>1$) a été sélectionné et comprend les fonctions de la base $B_0(X), B_1(X), \ldots, B_{2K-1}(X)$. Le but de l'étape K est de trouver la meilleure division de l'échantillon parmi tous les cas possibles avec prise en compte des splines déjà présentes dans le modèle. Ainsi, on ajuste tous les modèles de la forme

$$Y(X) = \sum_{i=0}^{2K-1} a_i B_i(X) + a_{2K} B_l(X)(X_j - c_j)^+ + a_{2K+1} B_l(X)[-(X_j - c_j)]^+ \tag{6}$$

sachant que les produits des splines doivent toujours être formés de variables explicatives différentes. Notons que : $l = 0, ..., 2K-1$, $j = 1, ..., p$, et c_j prend ses valeurs dans l'ensemble des nœuds potentiels de la variable X_j (il s'agit des statistiques d'ordre de cette variable). On contraint le modèle à ne pas introduire un terme de la base qui fait déjà partie de l'estimateur. On choisit ensuite le modèle à $2K+1$ termes qui minimise la somme des carrés des résidus et les deux fonctions de la base nouvellement construites sont définies par les relations suivantes :

$$B_{2K}(X) = B_{l^*}(X)(X_{j*} - c_{j*})^+ \tag{7}$$

$$B_{2K+1}(X) = B_{l^*}(X)[-(X_{j*} - c_{j*})]^+ \tag{8}$$

La procédure continue en incrémentant K d'une unité à chaque pas jusqu'à l'obtention du modèle de taille maximale qui par défaut est donné par $max(21, 2p+1)$. Si p est le nombre de variables explicatives, en revanche, nous n'avons pu trouver une explication sur l'origine de la valeur 21 proposée par Friedman, le pionnier de cette technique.

6.2.4.2 Etape 2 : Elimination

La deuxième étape de l'algorithme MARS est une phase d'élimination du modèle de dimension maximale. A chaque itération de cette phase, une procédure d'élimination par la méthode Backward Stepwise est appliquée. Toutes les fonctions de la base sauf la constante $B_0(X) = 1$ sont candidates à l'élimination. Pour déterminer celle qui doit être éliminée, on utilise le critère de score de la Cross-validation généralisée. A ce critère, on associe un paramètre de lissage ξ qui pénalise la complexité[3] du modèle. Considérant un modèle de taille maximale K, le nombre de termes additifs non constant dans le modèle, la fonction de pénalisation de la complexité du modèle est donnée par :

$$C(K) = K + dK, \tag{9}$$

[3]La complexité est liée au nombre important de variables dans un modèle.

où d défini par $d = \xi(1 - \frac{1}{K})$ est le coût de complexité fixé pour l'optimisation de chaque fonction dans le modèle.

Posons : $f_K(\mathbf{X})$ l'estimateur de Y avec K termes de fonctions splines non constantes, y_i est la mesure de Y pour l'individu d'indice i, $\widehat{y}_K(\mathbf{x_i})$ est la prévision de la variable Y pour le même individu en utilisant l'estimateur $f_K(\mathbf{X})$. On définit le score de la cross-validation généralisée par :

$$GCV(f_K(\mathbf{x}), \xi) = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_K(\mathbf{x_i}))^2}{[1 - \frac{C(K)}{n}]^2} \tag{10}$$

où n est le nombre d'individus dans l'échantillon de données, $\sum_{i=1}^{n}(y_i - \widehat{y}_K(\mathbf{x_i}))^2$ est la somme des résidus au carré pour l'estimateur $f_K(\mathbf{X})$, $[1 - \frac{C(K)}{n}]^2$ est le terme de pénalisation du modèle défini avec K termes non constants.

A chaque itération de cette procédure d'élimination, la fonction spline avec la plus faible contribution au modèle est retirée. C'est en effet, la fonction de base qui diminue la valeur du score GCV lorsqu'elle est extraite de l'estimateur $f_K(\mathbf{X})$. Par ailleurs, l'indicateur GCV peut aussi être utilisé pour évaluer l'importance des variables explicatives dans le modèle MARS final. Ainsi, une variable est d'autant plus importante que sa suppression du modèle provoque l'augmentation du score de la cross-validation généralisée GCV. Le processus d'élimination continue jusqu'à ce que l'on obtienne le modèle de départ.

Remarques :

Friedman (1991) a réalisé des simulations en faisant varier la valeur du paramètre de lissage ξ. Il suggère que ce paramètre devrait prendre les valeurs comprises entre 2 et 4. Il montre que la qualité du modèle dépend de ce paramètre, donc du coût complexité. Plus le paramètre coût complexité d est élevé, plus un nombre important de fonctions de la base sont éliminées du modèle. En général, d augmente pendant la phase d'élimination dans le but d'obtenir les modèles avec le minimum de fonctions de base.

6.2.4.3 Etape 3 : Sélection

La phase de sélection du modèle est basée sur l'évaluation de leur qualité prédictive. Le modèle final est celui qui minimise le score de la cross-validation généralisée parmi tous ces modèles. L'estimateur final $f(\mathbf{X_1}, \ldots, \mathbf{X_p})$ de Y du modèle de l'Equation (2) peut se décomposer sous la forme d'un modèle d'analyse de variance défini par :

$$f(\mathbf{X_1}, \ldots, \mathbf{X_p}) = f_0 + \sum_{i=1}^{p} f_i(X_i) + \sum_{i<j} f_{ij}(X_i, X_j) + \sum_{i<j<k} f_{ijk}(X_i, X_j, X_k) + ... \tag{11}$$

où f_0 est l'estimation de la constante du modèle. Le deuxième terme est la somme de toutes les splines définies par une seule variable explicative. Le troisième terme est la somme des produits des fonctions de base impliquant deux variables explicatives, etc. De l'estimation $f(\mathbf{X_1}, ..., \mathbf{X_p})$ de Y, on observe que le degré d'interaction peut s'étendre jusqu'à p. Cependant, l'inconvénient d'avoir un degré d'interaction élevé est que les modèles deviennent souvent difficilement interprétables.

6.2.5 Paramètres pour la mise en œuvre du modèle MARS

Dans le cadre de notre étude, nous avons utilisé la fonction *earth()* du package *mda* de l'environnement *R*. C'est la version la plus optimisée de l'algorithme *mars()*. Grâce à cette fonction, nous avons la possibilité de spécifier certains paramètres tels que la valeur du paramètre de lissage *(penalty)*, le nombre maximum de fonctions de la base *(nk)* et le degré maximal des interactions désiré dans le modèle *(degree)*.

Les méthodes non paramétriques nécessitent toujours la spécification d'un paramètre de lissage pour indiquer la flexibilité souhaitée pour un estimateur. Rappelons que, contrairement à la plupart des méthodes de régression non paramétrique, la variation du paramètre de lissage pour la méthode MARS ne permet pas de faire varier la flexibilité de façon continue. C'est ainsi que plusieurs valeurs de paramètre de lissage vont donner le même estimateur et ce dernier peut changer brusquement lorsque l'effet de la variation du paramètre se fait finalement sentir. Ce paramètre agit particulièrement sur le nombre de termes dans la base de l'estimateur $f(X)$ de Y.

6.3 Prévision de l'*écart temporel* par la méthode MARS

6.3.1 Indicateurs de sélection du modèle

Ces indicateurs sont synthétisés dans la figure FIG.6.1. Ainsi, le modèle obtenu a été sélectionné en étudiant le coefficient de régression multiple R^2 (Rsq) et le nombre de variables explicatives (Number of user predictors) en fonction du nombre de fonctions splines de la base (Number of terms) pour l'estimation du modèle. Au dela de 45 termes dans la base construite à partir de 12 variables les plus importantes, les indicateurs de qualité du modèle se stabilisent et le gain apporté par les nouvelles fonctions de la base et les autres variables n'est plus significatif pour améliorer la qualité des prévisions. Ainsi, le modèle optimal est alors construit sur 45 termes des fonctions splines en utilisant 12 variables explicatives.

6.3.2 Test d'ajustement des résidus du modèle MARS par la loi normale

L'histogramme de la distribution de fréquence de ces résidus est illustré par la figure (a) de FIG.6.2. Les résidus du modèle MARS semblent symétriques autour de la moyenne 0. La distance de Kolmogorov-Smirnov pour le test de normalité est égale à 0.50. Ainsi, l'hypothèse nulle selon laquelle la distribution empirique des résidus du modèle MARS ajuste une loi normale est rejetée. Par ailleurs, le test asymptotique de Jarque-Bera fournit une statistique de test égale 596760 avec une p-value égale à 0. Ce qui permet une fois de plus de conclure au rejet de l'hypothèse de normalité de la variable résiduelle. Toutefois la figure (b) de FIG.6.2 sur la comparaison des fonctions de repartition empiriques des résidus des modèles CART classique, CART modifié et MARS montre une forte proximité entre les lois de probabilité de ces trois résidus. A l'étape de validation de ce modèle, nous utiliserons les méthodes d'ajustement des lois de probabilité par le mélange de lois gaussiennes pour proposer un ajustement de la loi de probabilité des résidus par une combinaison de lois normales.

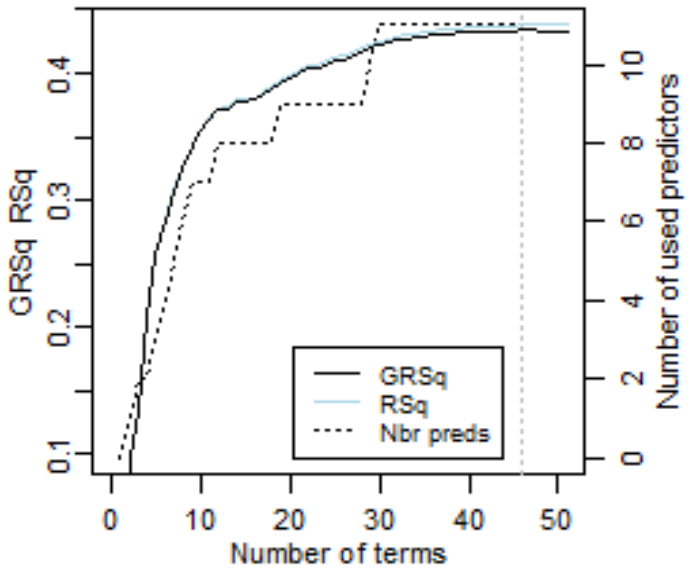

FIG. 6.1 – Indicateur de qualité pour la sélection du modèle optimal en fonction du nombre de fonctions splines et du nombre de variables explicatives.

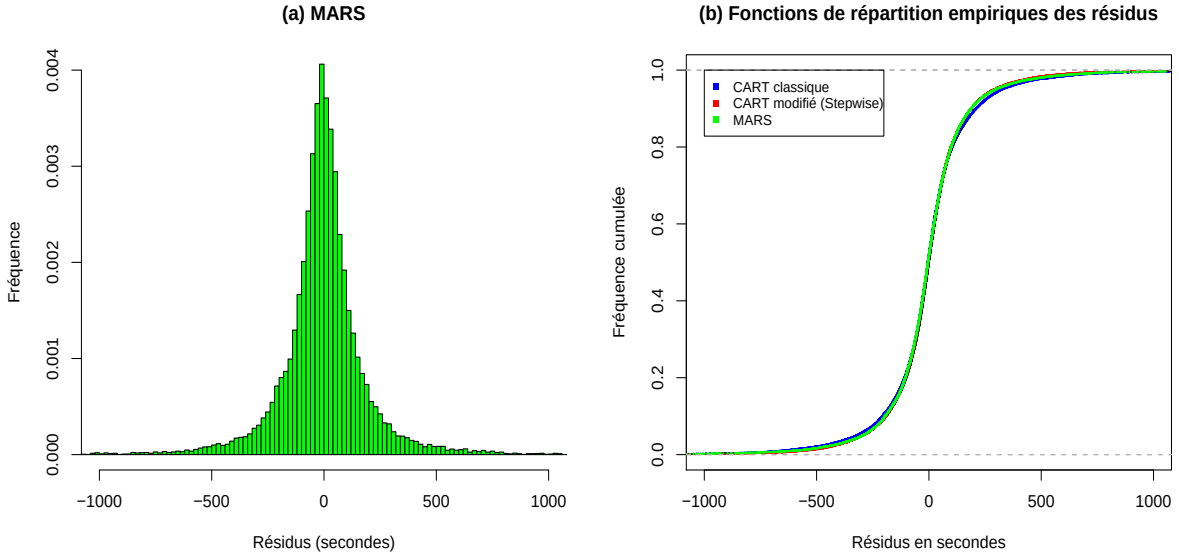

FIG. 6.2 – (a) est l'histogramme de distribution empirique des résidus du modèle MARS, (b) Compare les fonctions de répartition empiriques des résidus des modèles *CART classique*, *CART modifié* (Stepwise) et *MARS*.

6.3.3 Dispersion des résidus du modèle MARS en fonction de l'horizon de prévision

Une lecture des boxplots de la figure (c) de FIG.6.3 montre que la dispersion des résidus du modèle de prévision de la variable écart temporel sont bien concentrés autour de leur moyenne 0 au debut de l'horizon. La variation de l'incertitude du modèle MARS en fonction de l'horizon temporel semble proche de celle des

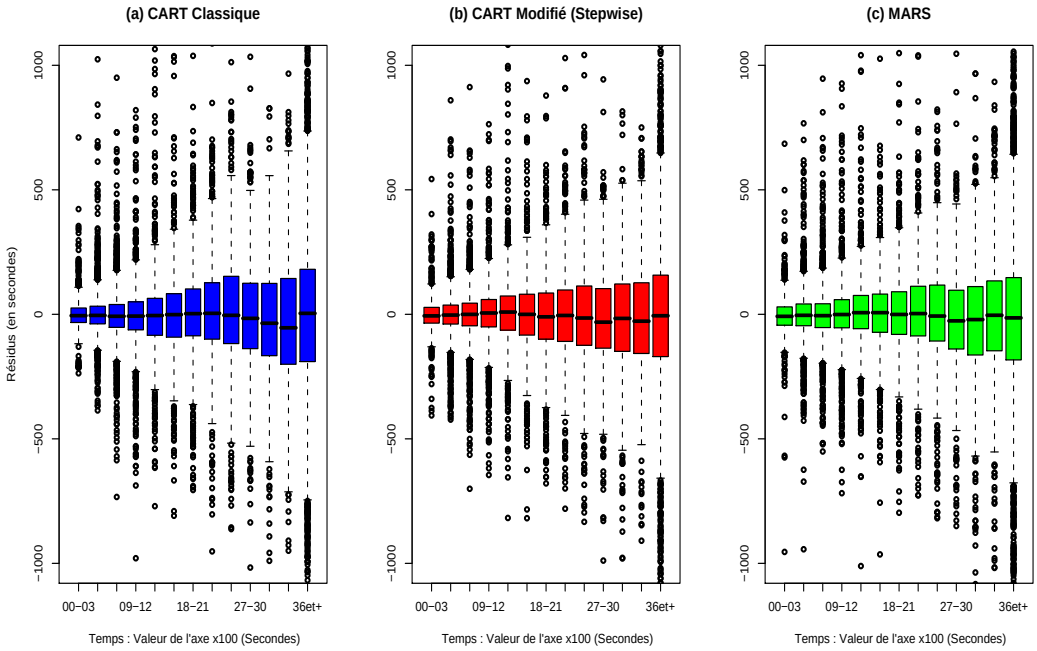

FIG. 6.3 – Dispersion des résidus en fonction de l'horizon temporel de prévision des modèles : CART classique, CART modifié (Stepwise) et MARS. 00-03 indique par exemple l'intervalle de temps compris entre 0 et 300 secondes.

modèles précédents.

En comparant les profils des trois modèles proposés dans la figure FIG.6.4, il apparaît que pour un horizon de prévision inférieur à 2100 secondes (35 minutes), la dispersion des résidus du modèle MARS croît moins vite que celle obtenue des résidus du modèle CART modifié. Par ailleurs, si ce modèle semble founir des mauvaises prévisions au début de l'horizon (0-600 secondes), cela ne suffit pas pour remettre en cause, du moins pour l'instant, sa capacité à fournir de bonnes prévisions. En effet, pour tous ces modèles, 10 minutes apparaissent insuffisantes pour permettre de faire des prévisions sur le trafic en temps réel de façon à prendre des actions correctives en cas de nécessité. Afin de tester l'utilisation opérationnelle de ce modèle, une étape déterminante est nécessaire pour sa validation sur les données n'ayant pas été utilisées dans la phase d'ajustement.

6.3.4 Comparaison de la qualité de prévision du modèle MARS à celles de CART classique et CART modifié

Afin d'affiner la comparaison de la qualité des modèles présentés jusqu'ici, nous utilisons une fois de plus le coefficient de Theil. Nous quantifions la différence entre la qualité prédictive du modèle MARS par rapport au modèle CART classique d'une part, et au modèle CART modifié d'autre part. Le modèle MARS est l'alternatif dans le calcul du coefficient de Theil. Ce coefficient est égal à 1.09 et 1.02 respectivement pour le modèle CART classique et CART modifié. Donc le pouvoir prédictif du modèle ajusté par la méthode MARS permet d'améliorer les prévisions de 9% et 2% relativement aux méthodes CART (classique,

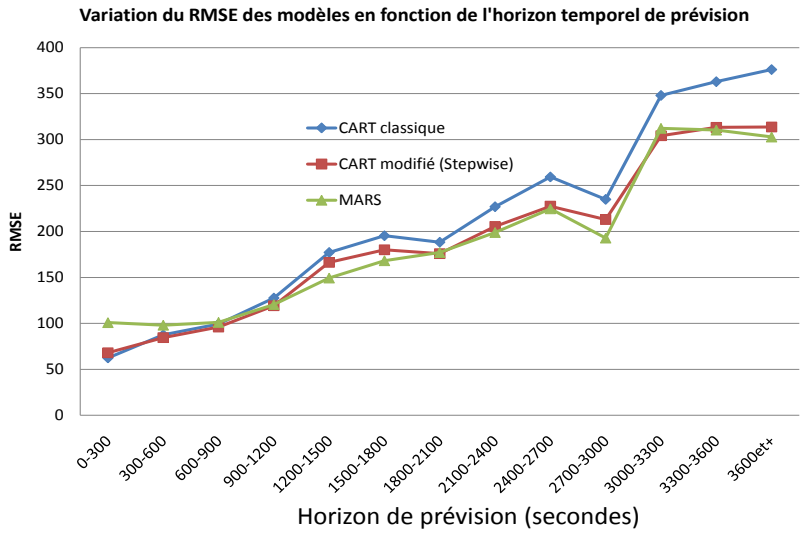

FIG. 6.4 – L'incertitude en fonction de l'horizon temporel de prévision. On compare l'évolution de l'incertitude du modèle MARS à celles des modèles CART classique et CART modifié.

modifié).

Pour les deux modèles de référence, cet indicateur est ensuite évalué en fonction de l'horizon temporel de prévision. Les figures FIG.6.5 et FIG.6.6 présentent la synthèse des résultats obtenus. Ainsi, si l'horizon de prévision est inférieur à 900 secondes, le modèle CART classique est préféré par rapport au modèle MARS. Dans le cas contraire, le modèle MARS est préféré. L'étape indispensable de validation permettra de déterminer la quelle de ces méthodes permet d'obtenir des meilleurs prévisions du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue.

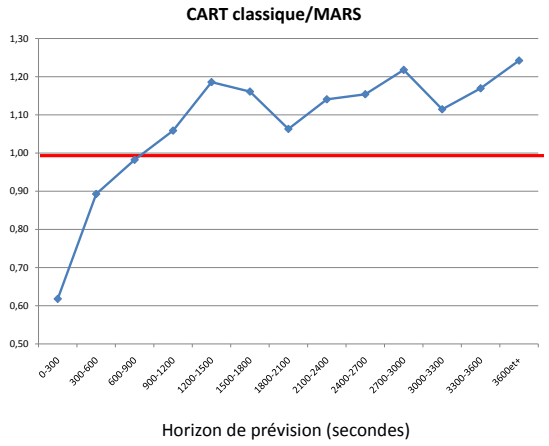

FIG. 6.5 – Indicateur de Theil pour la comparaison de la qualité des modèles *CART classique* et *MARS*. Le coefficient de Theil est égal au rapport du RMSE du modèle *CART classique* sur le RMSE du modèle *MARS*.

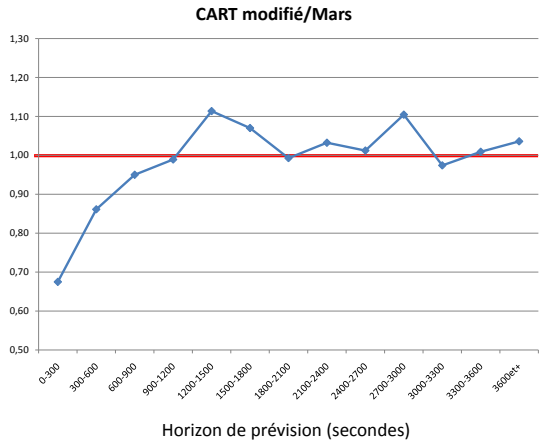

FIG. 6.6 – Indicateur de Theil pour la comparaison de la qualité des modèles CART modifié (Stepwise) et MARS. Le coefficient Theil est égal au rapport du RMSE du modèle CART modifié sur le RMSE du modèle MARS.

6.4 Estimateur de l'écart temporel obtenu par le modèle MARS

Dans cette section, nous proposons une expression analytique de l'estimateur $f_K(\mathbf{x})$ de Y en fonction des caractéristiques des vols. Le paramètre de lissage a été fixé à 2, ensuite, pour éviter la complexité d'interprétation des facteurs d'interactions d'ordre élevé, nous avons limité le degré d'interaction à 2. Ainsi, les fonctions de la base sont des simples fonctions linéaires par morceaux sur une seule variable explicative ou des produits de deux fonctions linéaires par morceaux sur deux variables explicatives.

Les paramètres du modèle de régression sont synthétisés dans la table TAB.6.1. Ainsi, une estimation $f_K(\mathbf{x})$ du modèle est une combinaison linéaire des fonctions splines simples ou des produits de deux fonctions splines définies par les variables explicatives caractéristiques des aéronefs en vol. Les fonctions et les coefficients correspondants se trouvent respectivement dans la première et la deuxième colonne de la table TAB.6.1. Une lecture de cette la table montre que l'estimateur de l'écart temporel sur les instants de passage des aéronefs sur les points de leur trajectoire de vol prévue est une fonction des facteurs directs ou principaux et des interactions de degré 2 entre les variables explicatives. Nous proposons ci-dessous une analyse détaillée des effets des différents facteurs de ce modèle sur la variable réponse.

Paramètres du modèle MARS	
Bases	**Coefficients**
1	30.577
$(Indur - 0.935)^+$	7451.536
$(Distprev1 - 0)^+$	3.056
$(Nivpln - 180)^+$	-0.807
$(Nivpln - 260)^+$	0.847
$(260 - Nivpln)^+$	11.379
$(Vitessecour - 189)^+$	-0.264
$(189 - Vitessecour)^+$	8.296
$(57 - Alticour)^+$	0.679
$(Moypres + 4.477)^+(189 - Vitessecour)^+$	5.135
$[-(4.477 + Moypres)]^+(189 - Vitessecour)^+$	-0.073
$(Moypres + 4.682)^+(57 - Alticour)^+$	3.192
$[-(4.682 + Moypres)]^+(57 - Alticour)^+$	0.018
$(0.935 - Indur)^+(Distprev1 - 78.769)^+$	-4.116
$(Indur - 0.935)^+(Distprev1 - 1057.05)^+$	-13.586
$(Indur - 0.935)^+(1057.05 - Distprev1)^+$	-7.006
$(Indur - 0.935)^+(Vitessecour - 398)^+$	-17.585
$(Distprev1 - 0)^+(Retardcour - 2456)^+$	0.0001
$(Distprev1 - 0)^+(2456 - Retardcour)^+$	0.0001
$(Distprev1 - 0)^+(Txmdcour - 2751)^+$	0.0001
$(Distprev1 - 0)^+(2751 - Txmdcour)^+$	0.0001
$(Distprev1 - 0)^+(Nivpln - 270)^+$	0.039
$(Distprev1 - 0)^+(270 - Nivpln)^+$	-0.102
$(Distprev1 - 0)^+(Nivpln - 180)^+$	-0.042
$(Distprev1 - 147.377)^+(260 - Nivpln)^+$	0.055
$(147.377 - Distprev1)^+(260 - Nivpln)^+$	-0.079
$(Distprev1 - 0)^+(Distpln - 1420)^+$	0.007
$(Distprev1 - 0)^+(1420 - Distpln)^+$	0.0003
$(Distprev1 - 305.225)^+(Vitessecour - 189)^+$	0.0004
$(305.225 - Distprev1)^+(Vitessecour - 189)^+$	0.002
$(Distprev1 - 315.949)^+(189 - Vitessecour)^+$	-0.027
$(315.949 - Distprev1)^+(189 - Vitessecour)^+$	-0.031
$(Distprev1 - 0)^+(Decalcour - 1.089)^+$	0.063
$(Distprev1 - 0)^+(1.089 - Decalcour)^+$	-0.156
$(Distprev1 - 0)^+AT43^+$	-1.026
$(Txmdcour - 2500)^+(260 - Nivpln)^+$	-0.0004
$(2500 - Txmdcour)^+(260 - Nivpln)^+$	0.001
$(260 - Nivpln)^+(Vitessecour - 171)^+$	-0.009
$(260 - Nivpln)^+(171 - Vitessecour)^+$	0.029
$(Nivpln - 180)^+AT43^+$	-3.338
$(Dmouvaer - 46.096)^+(57 - Alticour)^+$	0.099
$(46.096 - Dmouvaer)^+(57 - Alticour)^+$	0.021
$(189 - Vitessecour)^+(Decalcour - 1.467)^+$	21.528
$(189 - Vitessecour)^+(1.467 - Decalcour)^+$	-1.569
$(Vitessecour - 189)^+AT43^+$	-1.295
$R^2 = 0.48$ et $GCV = 41271.57$	

TAB. 6.1 – Les paramètres de l'estimateur $f_K(\mathbf{x})$ de Y. 45 termes ont été sélectionnés sur 49 bases construites dans le modèle maximal. 12 variables explicatives ont également été retenues sur 26.

6.4.1 Effets directs

Dans un modèle de régression, un effet direct[4] désigne la part de la relation causale d'une variable explicative sur une variable dépendante qui n'est pas expliquée par les autres variables présentes dans le modèle. Le coefficient de régression mesure ainsi la part de l'effet d'une variable explicative sur la variable à expliquer qui ne dépend pas des associations que la variable explicative est susceptible d'avoir avec les autres variables explicatives incluses dans le modèle.

Dans notre étude, l'effet direct d'une variable explicative sur la variable *écart temporel* est mesuré par la valeur du coefficient de régression de la fonction spline définie uniquement par cette variable explicative. Sur l'ensemble des variables explicatives utilisées, 5 sont impliquées dans les fonctions splines ayant des effets directs significatifs sur la variabilité de l'écart temporel. On remarque que 4 de ces effets sont dus aux splines construites à partir des variables qui ont été actives dans le processus de construction de l'arbre de régression du modèle CART classique. Il s'agit notamment dans l'ordre d'importance décroissante : l'influence du vent *(Indur)*, de la distance entre le point courant et le point prévue sur la trajectoire de vol prévue *(Distprev1)*, du niveau de vol prévu dans les plans de vols *(Nivpln)* et de la vitesse courante des vols *(Vitessecour)*. Par ailleurs, le modèle MARS classe en 5ème rang la variable altitude courante *(Alticour)* confirmant ainsi son rôle important dans l'explication de la variabilité de la variable *écart temporel*. En effet, dans le modèle CART classique, elle était la première variable de substitution pour les deux premières divisions optimales sur la variable influence du vent avec des mesures d'association respectives de 57% et 67%.

On observe que les nœuds des splines du modèle MARS sont très différents de ceux de la division dans le modèle CART classique. Les effets directs des variables explicatives sur la variable dépendante sont donnés par les signes des coefficients des bases résumés dans la table TAB.6.1. Ce sont notamment des splines définies par une seule fonction linéaire. La figure FIG.6.7 synthétise ces effets pour chacune de ces variables concernées.

Remarque :

Pour les graphiques de la figure FIG.6.7, nous avons utilisé la version robuste de l'algorithme *plotmo()* de Friedman basé sur la médiane. En ordonnée est portée l'estimation en secondes de l'écart temporel des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue. (a) mesure les effets de l'influence du vent, (b) mesure les effets de la distance de vol prévue entre le point courant et le point prévu de la trajectoire de vol, (c) mesure les effets du niveau de vol de croisière prévu dans le plan de vol, (d) mesure les effets de la vitesse courante du vol à l'instant de prévision et (e) mesure les effets de l'altitude courante du vol.

[4] Un modèle de régression estime toujours uniquement l'effet direct d'une variable indépendante, en contrôlant par toutes les autres variables indépendantes incluses dans le modèle.

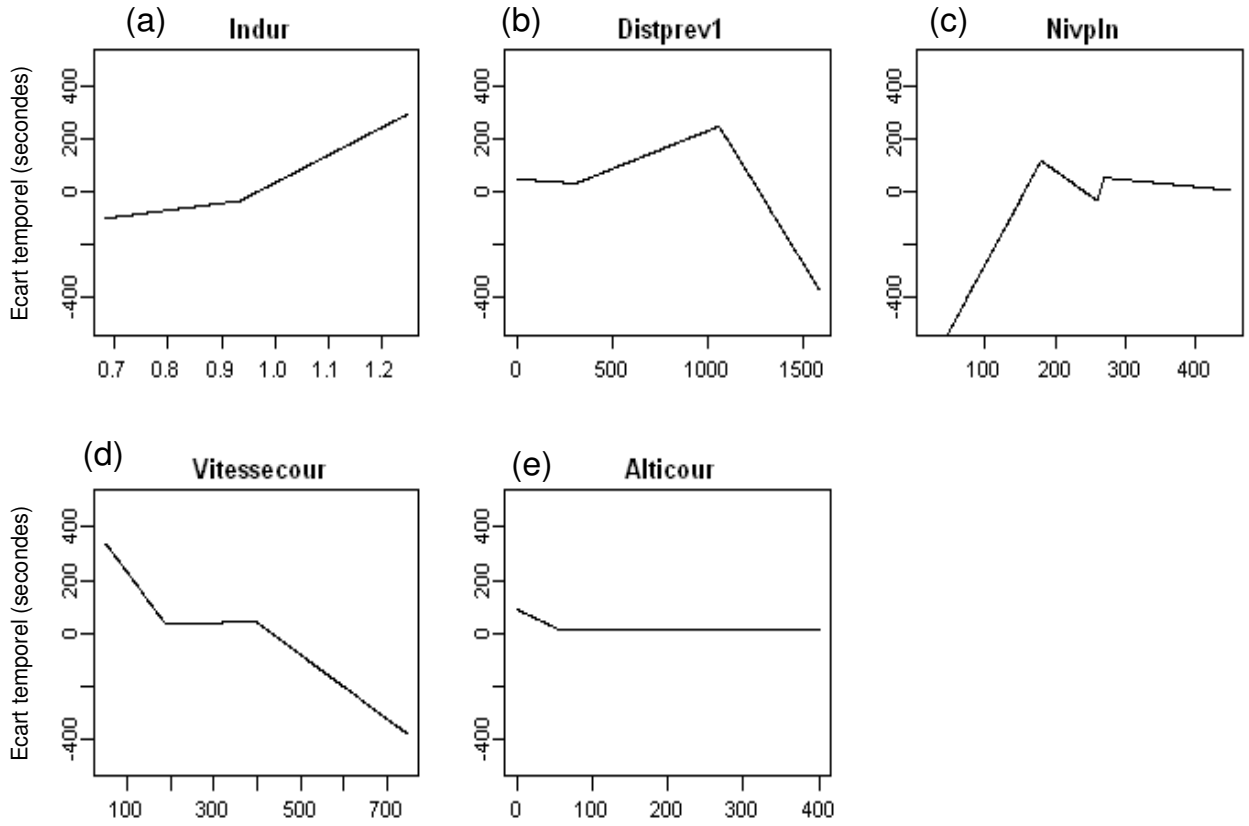

FIG. 6.7 – Effets directs des variables explicatives les plus significatives du modèle. Seulement 5 variables ont des effets directs sur l'écart temporel de passage des aéronefs en des points de leurs trajectoires prévues.

6.4.1.1 Effets directs de l'influence du vent

La première spline du modèle ayant un effet direct[5] sur la variable dépendante est construite à partir de l'influence du vent et notée $(Indur - 0.935)^+$. Son coefficient de régression est égal 7451.536 et consultable dans la table TAB.6.1. Ainsi, au fur et à mesure que l'influence du vent augmente à partir du seuil 0.935, la valeur de l'écart temporel augmente aussi (a) FIG.6.7. Une lecture de cette figure fait apparaître que lorsque le vent prévu sur la trajectoire des vols est de dos, l'écart temporel est algébriquement faible. Il prend en particulier les valeurs négatives et semble monotone croissant au fur et à mesure que le vent de dos diminue d'intensité au profit du vent de face qui augmente d'intensité. Ces résultats confirment et confortent les effets du vent sur la variable écart temporel obtenus du modèle CART classique.

[5]On utilise aussi la terminologie effet principal pour désigner un effet direct. C'est un effet dû à un terme de degré un.

6.4.1.2 Effets directs de la distance de vol prévue

Les effets liés à la distance de vol prévue sur l'écart temporel sont synthétisés dans la figure (b) de FIG.6.7. Ils sont définis par la spline $(Distprev1 - 0)^+$ dont le coefficient dans le modèle MARS est 3.056 (signe positif). Bien que ces effets soient significatifs dans le modèle, il n'est pas pour l'instant évident d'en proposer une réelle interprétation. Nous espérons qu'avec la prise en compte des interactions, la conjonction de cette variable avec d'autres variables explicatives permettront d'affiner les résultats.

6.4.1.3 Effets directs du niveau de vol prévu

Les effets directs du niveau de vol de croisière prévu sont significatifs. Ils se traduisent par la présence de cette variable dans la construction de trois splines de degré un. Celles-ci sont dans l'ordre d'importance décroissante $(Nivpln - 180)^+$, $(Nivpln - 260)^+$ et $(260 - Nivpln)^+$ dont les coefficients de régression respectifs estimés à -0.807, 0.847 et 11.379 (TAB.6.1). Le signe négatif du coefficient de la première spline montre que lorsque le niveau de vol de croisière est inférieur au niveau de vol $FL180$, les mesures de la variable dépendante sont négatives, c'est-à-dire les aéronefs concernés tendent à survoler les points de leur trajectoire prévue en avance. Sur le plan opérationnel, il ne semble pas évident d'enrichir l'interprétation de cette variable.

6.4.1.4 Effets directs de la vitesse courante du vol

La vitesse courante des vols agit sur l'instant de passage des aéronefs sur les points de leur trajectoire par l'intermédiaire des splines $(Vitessecour - 189)^+$ et $(189 - Vitessecour)^+$ qui ont pour coefficients de régression respectifs -0.264 et 8.296. Ces effets sont significatifs dans ce modèle (figure (d), FIG.6.7). Un examen de cette figure montre que l'écart temporel est positif pour les valeurs de la vitesse courante inférieures à 189 kts. Dans ce cas, les retards des vols les plus élevés sont observés lorsque la vitesse courante est plus faible. Intuitivement, il s'agit des aéronefs qui, à l'instant de prévision t_0, se trouvent encore dans la phase de montée ou de descente. A cause de ces retards élevés, l'incertitude de prévision du temps de vol est amplifiée. En revanche, pour les aéronefs avec une vitesse courante élevée[6], on observe deux regimes de variabilité de la variable d'intérêt. Le premier regime concerne la vitesse courante comprise entre 189 et 398 [7] kts. Ici, l'écart temporel moyen est faible et se situe autour de 0. Grâce à la faible valeur de l'écart temporel, l'incertitude dans la prévision du temps de vol est aussi faible. Ainsi, les aéronefs concernés survolent les points de leur trajectoire approximativement aux instants prévus dans leur plan de vol. Dans le deuxième régime caractérisé par la vitesse courante supérieure à 398 kts (la figure (d), FIG.6.7), nous voyons qu'à partir de ce seuil, plus la vitesse courante est élevée, plus les aéronefs concernés survolent les points de leur trajectoire en avance par rapport aux prévisions des plans de vols. Il s'agirait essentiellement des aéronefs qui, à l'instant de prévision t_0, ont déjà amorcé leur phase de croisière. Les résultats obtenus sur cette variable sont bien en accord avec les principes physiques de la dynamique des vols.

[6] Supérieure à 189 kts, seuil de division sur la variable.

[7] Seuil de la vitesse courante dans la spline d'interaction $(Indur - 0.935)^+ : (Vitessecour - 398)^+$.

6.4.1.5 Effets directs de l'altitude courante du vol

Ici, nous avons transformé la variable altitude courante en niveau de vol en divisant ses mesures en pieds par 100 pour obtenir des nouvelles mesures en niveau de vol. Ainsi, l'altitude courante a des effets significatifs dans le modèle par la fonction spline $(57 - Alticour)^+$ qui a pour coefficient de régression égal à 0.679. Ces effets semblent importants lorsque le niveau de vol est inférieur au seuil $FL057$. Ici, l'écart temporel est positif et le retard associé est une source d'importante incertitude sur la prévision du temps de passage des aéronefs en des points de leur trajectoire de vol prévue. En revanche, lorsque le niveau de vol est supérieure à $FL057$, l'écart temporel moyen est très faible et entraîne par conséquent une meilleure précision dans la prévision du temps de vol des aéronefs. Il s'agirait des aéronefs qui, à l'instant courant de prévision, ont armorcé leur phase de croisière.

Au terme de cette première analyse sur les effets directs des variables explicatives les plus importantes sur des écarts temporels de passage des aéronefs en des points de leur trajectoire de vol prévue, nous complétons notre étude dans la section suivante par l'examen des effets liés à la conjonction de plusieurs facteurs. Il s'agira notamment de l'évaluation de l'impact des interactions obtenues du modèle MARS et définies par les produits des splines sur notre variable d'explicative.

6.4.2 Effets d'interactions

Il apparaît des coefficients de la table TAB.6.1 que sur 45 splines ayant des effets significatifs sur la variabilité des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue, 37 sont le fait de la conjonction de plusieurs facteurs. Ce sont des facteurs d'interactions du modèle. Dans leur immense majorité, elles sont construites autour des variables explicatives qui, considérées séparément, ont des effets principaux (ou directs) significativement importants dans le modèle. Nous interprétons les coefficients des interactions en accordant une importance singulière aux facteurs pour lesquels le principe de l'hiérarchie[8] des variables explicatives est vérifié. Par ailleurs, nous tirons parti de cette flexibilité du modèle MARS pour déterminer les effets de certaines variables explicatives qui dans la construction de l'arbre de régression du modèle CART classique n'avaient qu'un rôle suppléant. Il s'agit notamment de la distance de vol prévue dans le plan de vol (*Distpln*), du taux de montée/descente des aéronefs (*Txmdcour*) qui, combinées aux autres variables semblent avoir des effets sur la variabilité de la variable d'intérêt.

[8]Une interaction n'est considérée que si les effets principaux sont significatifs dans le modèle. Le principe des conditions hiérarchiques ne s'applique pas sur le maintien ou non des termes d'interactions dans les modèles MARS.

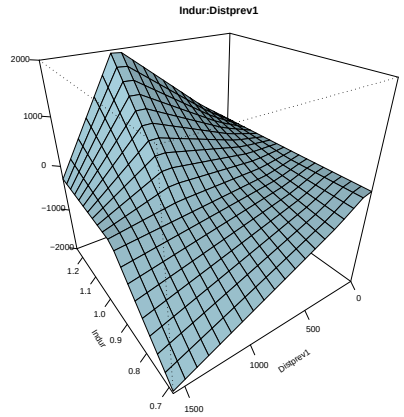

FIG. 6.8 – Interaction entre l'influence du vent prévu sur la trajectoire du vol et la distance de vol prévue. L'axe vertical représente l'écart temporel en secondes.

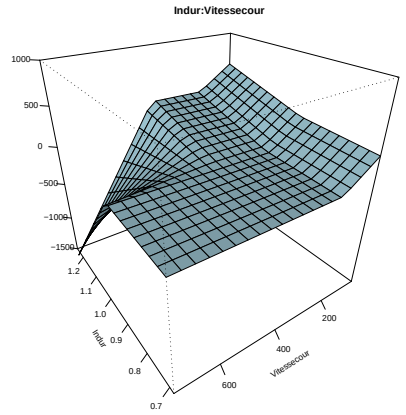

FIG. 6.9 – Interaction entre l'influence du vent prévu sur la trajectoire de vol et la vitesse courante du vol. L'axe vertical représente l'écart temporel en secondes.

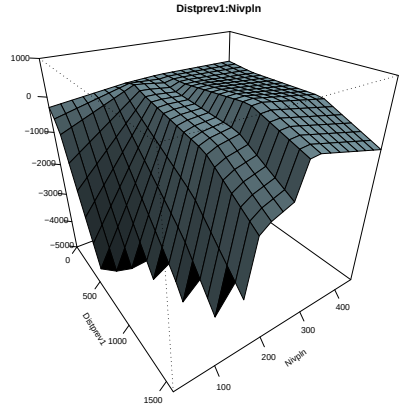

FIG. 6.10 – Interaction entre la distance de vol prévue et le niveau de vol de croisière prévu dans le plan de vol. L'axe vertical représente l'écart temporel en secondes.

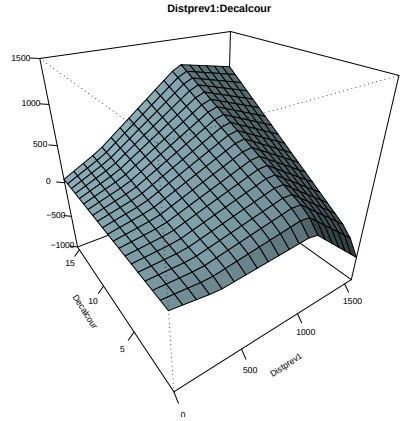

FIG. 6.11 – Interaction entre la distance de vol prévue et le décalage courant du vol par rapport à sa trajectoire prévue. L'axe vertical représente l'écart temporel en secondes.

6.4.2.1 Interactions avec l'influence du vent

L'interaction entre l'influence du vent et la distance de vol prévue est illustrée dans la figure FIG.6.8. Elle est évaluée par les trois fonctions splines $(0.935 - Indur)^+(Distprev1 - 78.769)^+$, $(Indur - 0.935)^+(Distprev1 - 1057.05)^+$ et $(Indur - 0.935)^+(1057.05 - Distprev1)^+$ qui ont pour coefficients de régression respectifs -4.116, -13.586 et -7.006, tous de même signe. Donc, ces fonctions ont des effets de même type sur la variable dépendante. Elles mettent en évidence deux sous-ensembles de l'échantillon de données avec des profils de l'écart temporel bien distincts. Dans le sous-ensemble où le vent de dos est dominant, l'écart temporel est élevé et négatif et se traduit majoritairement les avances importantes par rapport aux plans de vol. De ce fait, l'incertitude de prévision du temps de vol des aéronefs est élevée. En

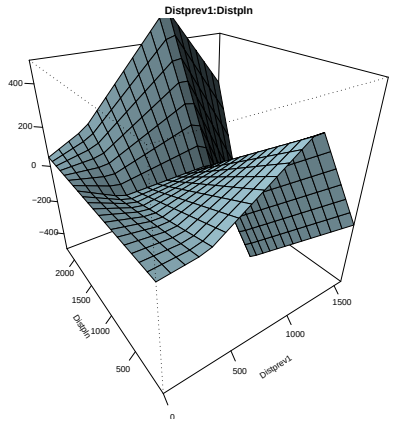

FIG. 6.12 – Interaction entre la distance de vol prévue (entre le point courant et un point la trajectoire prévue) et la distance totale du vol prévue entre l'aéroport de départ et l'aéroport d'arrivée du vol. L'axe vertical représente l'écart temporel en secondes.

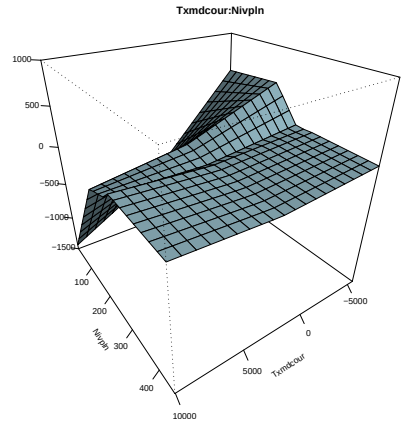

FIG. 6.13 – Interaction entre le taux de montée / descente courant et le niveau de vol de croisière prévue dans le plan de vol. L'axe vertical représente l'écart temporel en secondes.

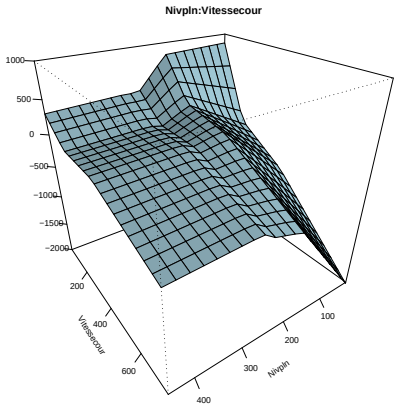

FIG. 6.14 – Interaction entre le niveau de vol de croisière prévu et la vitesse courante du vol au point courant de prévision. L'axe vertical représente l'écart temporel en secondes.

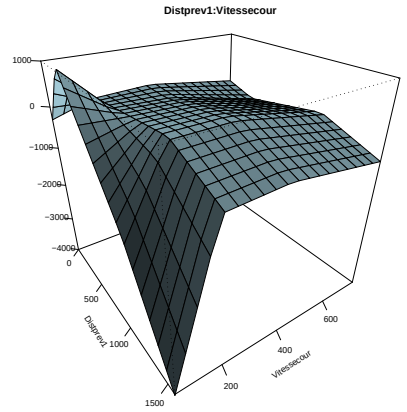

FIG. 6.15 – Interaction entre la distance de vol prévue et la vitesse courante du vol au point courant de prévision. L'axe vertical représente l'écart temporel en secondes.

revanche, avec un vent de face dominant, on observe également une amplification de l'incertitude de prévision du temps de vol sur des longues distances de vol. Cette incertitude se traduit cette fois par d'importants retards.

La figure FIG.6.9 synthétise les effets de l'interaction entre l'influence du vent et la vitesse courante du vol. Les effets de cette interaction sur le temps de vol sont mesurés par la spline $(Indur-0.935)^+(Vitessecour-398)^+$ dont le coefficient de régression est -17.585. Une lecture de ce graphique fait apparaître que la conjonction d'une faible vitesse courante (inférieure à 189 kts) et du vent de face dominant a pour effet de retarder les aéronefs sur leur trajectoire. Par ailleurs, on observe que pour un vent de face dominant, une vitesse courante élevée (supérieure à 398 kts) peut contribuer à atténuer ses effets qui opposent une résistance

contre le mouvement de l'aéronef. Ainsi, les aéronefs concernés peuvent conpenser les retards dûs au vent et l'incertitude associée au retard dans la prévision du temps de vol s'en trouve diminuée.

6.4.2.2 Interactions avec le niveau de vol de croisière prévu dans le plan de vol

Les effets des interactions entre le niveau de vol de croisière prévu et la vitesse courante du vol sur le temps de vol sont évalués dans le modèle MARS par les fonctions splines $(260 - Nivpln)^+(Vitessecour - 171)^+$ et $(260 - Nivpln)^+(171 - Vitessecour)^+$. Elles ont pour coefficients de régression respectifs -0.009 et 0.029. Ainsi, pour les aéronefs ayant des niveaux de croisière dans l'espace aérien supérieur, l'écart temporel est lié à la vitesse courante au moment de la prévision. Une faible vitesse courante amplifie le niveau a tendance au augmenter la valeur des écarts temporels (positifs). Ce qui se traduit par des retards importants qui amplifie le niveau d'incertitude dans la prévision du temps de vol. La figure FIG.6.14 montre que les effets de la vitesse courant sur la variabilité de l'écart temporel sur le temps de vol augmentent lorsque le niveau de croisière de vol est faible. Ces effets gardent les mêmes sens et se traduit par d'importants retards lorsque la vitesse est faible.

Le niveau de vol de croisière interagit aussi avec le taux de montée ou de descente par les fonctions splines $(Txmdcour - 2500)^+(260 - Nivpln)^+$ et $(2500 - Txmdcour)^+(260 - Nivpln)^+$ avec les coefficients de régression respectifs -0.0004 et 0.001. Les résultats de la figure FIG.6.13 montre que les écarts temporels les plus importants concernent les aéronefs qui sont en phase de montée ou de descente et qui ont des faibles niveaux de croisière. Ces aéronefs sont caractérisés par des retards ou des avances très importants qui sont eux aussi source d'incertitudes dans la prévision du temps de vol.

6.4.2.3 Interactions avec la distance de vol prévue

La distance de vol prévue interagit avec plusieurs variables explicatives dans le modèle MARS. Maintenant, nous nous interessons à ses interactions avec la vitesse courante et la distance totale du vol entre l'aéroport de départ et l'aéroport d'arrivée.

Les effets de l'interaction entre la distance prévue et la vitesse courante sont donnés dans le modèle par les splines $(Distprev1 - 305.225)^+(Vitessecour - 189)^+$ et $(305.225 - Distprev1)^+(Vitessecour - 189)^+$. L'incertitude la plus importante dans la prévision de l'*écart temporel* est observée pour la vitesse courante inférieure à 189 kts (FIG6.15). Cette incertitude se traduit par les aéronefs qui survolent les points de leur trajectoire avec d'importants retards.

Les effets d'interactions entre la distance prévue et la distance totale du vol entre l'aéroport de départ et celui d'arrivée sont évalués dans le modèle à travers les splines $(Distprev1 - 0)^+(Distpln - 1420)^+$ et $(Distprev1 - 0)^+(1420 - Distpln)^+$. Il ressort de la figure FIG.6.12 que pour une distance prévue inférieure à 500 NM, l'incertitude de prévision de l'écart temporel est faible et proche de 0 quelle que soit la distance totale de vol prévue dans le plan de vol. En revanche, lorsque la distance prévue pendant le vol est supérieure à 500 NM, écarts temporels sont bien importants pour les valeurs extrêmes de la distance totale de vol. Le survol des points des trajectoires des vols par les aéronefs après les instants prévus traduit

l'importance de l'incertitude dans la prévision du temps de passage sur ces points.

6.5 Conclusion

En utilisant la méthode de régression multiple par les splines adaptatives (MARS), nous avons proposé un modèle non-paramétrique de prévision de de l'écart temporel de passage des aéronefs en des points de leur trajectoire de vol prévue. Pour comparer cette technique aux méthodes CART classique et CART modifié proposés plus tôt, nous avons utilisé le coefficient de Theil calculé à partir des erreurs de prévision des modèles sur les données d'apprentissage. Il ressort que la qualité prédictive du modèle MARS est d'environ 9% et 2% supérieure à celles des modèles CART classique et CART modifié respectivement. Ainsi, l'ajustement du modèle de prévision du temps de passage des aéronefs en des points de leur trajectoire prévue par la méthode MARS permet d'améliorer les deux premiers modèles. Il apparaît toutefois que lorsque l'horizon de prévision est inférieur à 15 minutes (900 secondes), les modèle CART classique et CART modifié ajustent mieux les données que la technique MARS.

Les trois modèles présentés jusqu'ici ont tous leurs avantages et leurs inconvénients. Si le modèle MARS prend en compte la continuité de l'estimateur des prévisions entre les sous-domaines adjacents, il ne résoud pas le problème d'instabilité inhérent aux arbres de régression obtenus avec la méthode CART classique. Rappelons que le modèle CART modifié améliore aussi les prévisions à l'intérieur des feuilles des arbres de régression sans toutefois apporter une réelle réponse à l'instabilité des arbres fournis par la méthode CART classique.

Dans le chapitre suivant, nous présentons les forêts aléatoires qui sont une des méthodes récentes de généralisation des arbres de régression. Une littérature récente montre que cette méthode basée sur l'aggrégation de plusieurs arbres de régression issus des échantillons bootstrap de l'échantillon d'origine est une solution rigoureuse au problème d'instabilité des arbres de régression.

Chapitre 7

Prévision du temps de passage des aéronefs sur les points de leur trajectoire par les forêts aléatoires

7.1 Introduction

Au chapitre 3, nous avons proposé un modèle fondé sur la méthode CART pour la prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue. L'instabilité des arbres de régression inhérente à cette méthode constitue un réel problème pour l'utilisation du modèle ajusté dans des situations nouvelles. En effet, des légères perturbations des données d'apprentissage peuvent provoquer des changements profonds de la structure des arbres. Ce problème a fait l'objet de plusieurs travaux empiriques (Breiman,1996a[8] ; Ghattas,1999b[33]). Ces études montrent que les arbres de régression sont fortement dépendants des échantillons qui ont permis leur estimation. Ce qui pose la question de leur robustesse sur les données nouvelles. Heureusement, des nouvelles approches innovantes sont apparues et tentent d'apporter des solutions d'amélioration. Il s'agit notamment du $Bagging$ (Breiman ; 1996), du $Boosting$ (Freund and Schapire ; 1996a) et les Forêts aléatoires (Breiman ; 2001). Toutes ces méthodes ont pour principe de construire plusieurs arbres à partir des données d'apprentissage et de les agréger ensuite par le calcul de la moyenne des prévisions. L'application de ces méthodes dans la classification par Quinlan (1996) [53] montre une amélioration considérable des performances des modèles de prévisions obtenus.

Le $Bagging$ est une méthode qui consiste à construire une famille d'arbres sur K échantillons bootstrap[1], suivie de l'agrégation de ces arbres pour obtenir un classifieur optimal. Sur chaque échantillon bootstrap, un modèle de prévision est déterminé en utilisant le même principe que dans CART. La prévision optimale est obtenue par agrégation des différentes prévisions par vote lorsqu'il s'agit des arbres de déci-

[1]Un échantillon bootstrap d'un échantillon initial de taille n est un échantillon obtenu par tirage aléatoire avec remise de n individus de l'échantillon initial.

sion. Le bagging est surtout réputé pour améliorer la robustesse des classifieurs instables comme les arbres de décision et les réseaux de neurones. Les travaux les plus récents améliorant la stabilité des arbres de décision sont le $boosting$ de Freund et Schapire (1996a,b)[28, 29] et de Schapire (2002) [56]).

Les forêts aléatoires ou *Random Forests* sont une technique récente de régression qui combine un grand nombre de K arbres binaires construits sur des échantillons bootstrap de l'échantillon d'apprentissage initial. C'est une méthode de sélection des variables explicatives pour des modèles non-paramétriques. Comme le $Bagging$, une partie de son efficacité est dûe 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 de leur nuisance à ces valeurs extrêmes (Tuffery, 2006)[60]. Outre sa grande capacité d'améliorer la stabilité des arbres de régression, les forêts aléatoires offrent la possibilité de hiérarchiser les variables explicatives dans les modèles de prévision. C'est d'ailleurs la méthode que nous avons retenue pour corriger l'instabilité des arbres de régression afin d'assurer la robustesse du modèle de prévision de la variable *écart temporel*, par suite, la prévision de l'instant de passage des aéronefs sur les points de leur trajectoire de vol prévue.

Ce chapitre est organisé comme suit : Dans le paragraphe (7.2), nous présentons quelques fondements du modèle des Forêts aléatoires et son algorithme. Dans le paragraphe (7.3), nous appliquons la méthode des forêts aléatoires pour proposer un modèle de prévision des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue. Ensuite, nous tirons parti de ses possibilités de hiérarchisation des prédicteurs pour affiner la première hiérarchie des variables explicatives obtenue plus tôt sur l'arbre de régression du modèle CART classique. Enfin, l'incertitude du modèle développé est étudiée en fonction de l'horizon temporel de prévision. Nous utilisons au paragraphe (7.4) le coefficient de Theil pour comparer la qualité d'ajustement du modèle par les forêts aléatoires à celles des modèles CART classique, CART modifié et MARS. Enfin, le paragraphe (7.5) conclut.

7.2 Forêts aléatoires (FA)

7.2.1 Généralités et principe

Les forêts aléatoires développées par Breiman (2001)[9] sont une collection d'arbres binaires de régression ou de décisions $\mathcal{A}_k$, $k = 1, ..., K$. Chaque arbre $\mathcal{A}_k$ est construit sur un des K échantillons bootstrap de l'échantillon d'apprentissage initial. Chaque arbre est construit selon les mêmes règles que celles de la méthode CART. L'ensemble des K arbres ainsi obtenus sont agrégés pour former une forêt aléatoire. Le modèle agrégé de prévision optimale est obtenu par le calcul de la moyenne des prévisions de la variable dépendante sur les K arbres des échantillons bootstrap (cas de régression). Les forêts aléatoires 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, en effet, 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 (il s'agit de l'échantillon bootstrap) non utilisée pour la construction de l'arbre sert en effet, à l'évaluation de l'importance des variables utilisées sur cet arbre.
- L'introduction du tirage aléatoire sur l'ensemble des variables explicatives permet d'éviter de voir apparaître toujours les mêmes variables. Il s'agit d'une double randomisation et on parle alors de *forêts aléatoires*.

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.

Les procédures mises en œuvre dans la modélisation par les forêts aléatoires présentent un certain nombre d'avantages. Elles nécessitent 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.
- 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 en revanche, 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 OOB ou sur un échantillon de validation se stabilise et 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 pour la généralisation. L'amélioration de la qualité de prévision se fait au détriment de l'interprétabilité, ainsi le modèle finalement obtenu devient une « *boite noire* ».

7.2.2 Quelques caractéristiques des FA

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 à 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.

Dans les forêts aléatoires, un nœud est déclaré terminal si le nombre d'observations qu'il contient est inférieur à un nombre minimal fixé. 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 explicatives. Cette valeur proposée par Breiman a été confirmée par d'autres travaux. Liaw et *al.* (2002) [45], Diaz-Uriarte et *al.* (2006) [24] ont montré l'optimalité de cette valeur en terme de performance des forêts sur les échantillons tests OOB. Pour la régression, ce nombre est approximativement $\frac{p}{3}$. 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. Ghattas et *al.* (2008)[5] ont observé en étudiant les données de Biopuces que l'importance des variables dans les forêts aléatoires est :

- Insensible à la nature du réechantillonnage utilisé, l'échantillon bootstrap avec ou sans remise,
- Stable en présence de variables explicatives corrélées,
- Invariante vis-à-vis de la normalisation (par exemple : division par l'écart-type),
- Stable vis-à-vis de faibles perturbabtions des données.

Les forêts aléatoires fournissent alors un moyen original de calcul d'un indice d'importance pour la hiérarchisation des variables explicatives.

7.2.3 Représentativité d'un échantillon bootstrap par rapport à l'échantillon initial

Dans une procédure d'échantillonnage bootstrap, il n'est pas certain que chaque observation de l'échantillon initial appartienne à l'échantillon bootstrap. En effet, une observation de l'échantillon initial de taille n appartient à un échantillon bootstrap avec une probabilité estimée à 0.633 :
En effet, à chaque tirage (avec remise), chaque observation a la probabilité $\frac{1}{n}$ d'être sélectionnée. Ainsi, $\left(1-\frac{1}{n}\right)^n$ est la probabilité pour qu'une observation soit sélectionnée zéro fois à l'issue de n tirages aléatoires. Quand n est suffisamment grand, une observation $(obs.i)$ appartient à l'échantillon bootstrap avec la probabilité $\mathbb{P}(obs.i)$ définie par :

$$\mathbb{P}(obs.i) = 1 - \left(1-\frac{1}{n}\right)^n \approx 1 - e^{-1} = 0.633 \tag{1}$$

7.2.4 Algorithme

Pour un échantillon d'apprentissage donné, cet algorithme a pour but de construire par agrégation de plusieurs arbres de régression, une estimation de la variable dépendante en fonction des variables explicatives. Il se présente comme suit :

Considérons un échantillon initial **Z** = (**X**,**Y**) = ((X_i,Y_i),...,(X_p,Y_p)), m_o est la valeur de la variable dépendante **Y** à estimer à partir de **Z** par la méthode des forêts aléatoires. $\widehat{\theta}_Z(m_o)$ est l'estimation de m_o sur l'échantillon Z. Pour $k = 1, ..., K$; $\widehat{\theta}_{Z_k}(m_o)$ est l'estimation de m_o sur l'échantillon Z_k, $\mathcal{A}_k$ est l'arbre de régression de Z_k tel que $\mathcal{A}_k = (F_{kj})_{(1 \leq j \leq J_k)}$. Les F_{kj} sont des nœuds terminaux de $\mathcal{A}_k$ et J_k est le nombre de classes de la partition formée par ces nœuds.

Pour k=1,...,K, Tirer : un échantillon bootstrap Z_k de l'échantillon **Z**,
Estimer : un arbre de régression sur cet échantillon par la méthode CART,
Initialiser le nombre de nœuds : $N_k = 1$,
Tirer : q variables explicatives avec q $\leq$ p,
Chercher : la division optimale $d^*_{N_k}$ pour ce nœud à partir des q variables,
S'il existe un nœud fils qui admet une division admissible :
$N_k = N_k + 1$,
Tirer à nouveau q variables et chercher la division optimale sur ce nœud fils à partir de ces q variables,
Sinon l'arbre est estimé et s'écrit :
$\mathcal{A}_k = (F_{kj})_{(1 \leq j \leq J_k)}$ tel que $F_{kj} \subset \mathbb{R}^q$,
Calculer $\widehat{\theta}_{Z_k}(m_o)$ sur les nœuds terminaux de l'arbre $\mathcal{A}_k$:
Comme dans la méthode CART, déterminer les coefficients m_{okj} sur l'arbre $\mathcal{A}_k$ par la régression :
$\widehat{m}_{okj} = \frac{1}{Card(F_{kj})} \sum_{y_i \in F_{kj}} y_i$,
$\widehat{\theta}_{Z_k}(m_o) = \left(\widehat{m}_{okj}\right)_{(1 \leq j \leq J_k)}$,
Calculer le modèle de régression pour l'arbre $\mathcal{A}_k$,
$M_k(X) = \sum_{j=1}^{J_k} \widehat{m}_{okj} \mathbb{1}_{\{X \in F_{kj}\}}$,
Calculer l'estimation moyenne de m_o sur les K arbres :
$\widehat{\theta}_Z(m_o) = \frac{1}{K} \sum_{k=1}^{K} \widehat{\theta}_{Z_k}(m_o)$,
Fin pour.

TAB. 7.1 – Algorithme : Forêts Aléatoires

7.3 Modélisation de l'*écart temporel* par les forêts aléatoires

Il s'agit ici, d'utiliser la technique d'apprentissage automatique des données basée sur les forêts aléatoires pour proposer un modèle de prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue. Notre étude porte sur un échantillon de 25000 individus décrits par 22 variables explicatives. En pratique, nous construisons un modèle de prévision de la variable dépendante écart temporel en fonction des variables explicatives qui caractérisent les aéronefs en vol et leur environnement. La version des forêts aléatoires utilisée ici est le Random Input où la règle de décision porte sur une seule des variables explicatives tirées de façon aléatoire.

7.3.1 Choix des paramètres

Dans un contexte de régression, Breiman (2001,2002) [9, 10] propose que le nombre de variables *mtry* de randomisation ou à tirer aléatoirement sur chaque nœud soit proche de la fraction $\frac{p}{3}$, où p est le nombre de variables explicatives impliquées dans le modèle.
Nous avons fixé le nombre minimal d'observations par feuille à 40[2]. 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. Le couple optimal que nous notons $(ntree^*, mtry^*)$ est celui qui minimise le critère suivant :

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

où $\widehat{y}_i^{(ntree,mtry)}$ est la prévision de la variable dépendante pour l'individu i par le modèle des forêts aléatoires construit avec les paramètres $ntree$ et $mtry$. Autrement dit :

$$(ntree^*, mtry^*) = \arg \min_{(ntree,mtry)} \mathcal{E}(ntree, mtry). \tag{2}$$

Les valeurs du critère $\mathcal{E}$ en fonction de *ntree* et *mtry* sont représentées dans la figure FIG.7.1. 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 du modèle 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 est approximativement $(150, 7)$. C'est ce dernier cas que nous avons retenu pour notre

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

étude. En effet, le nombre de variables de randomisation sur 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 FIG.7.2 montre qu'au dela du nombre d'échantillons bootstrap *ntree*=150, le gain de performance en terme de réduction de l'erreur de prévision est négligeable.

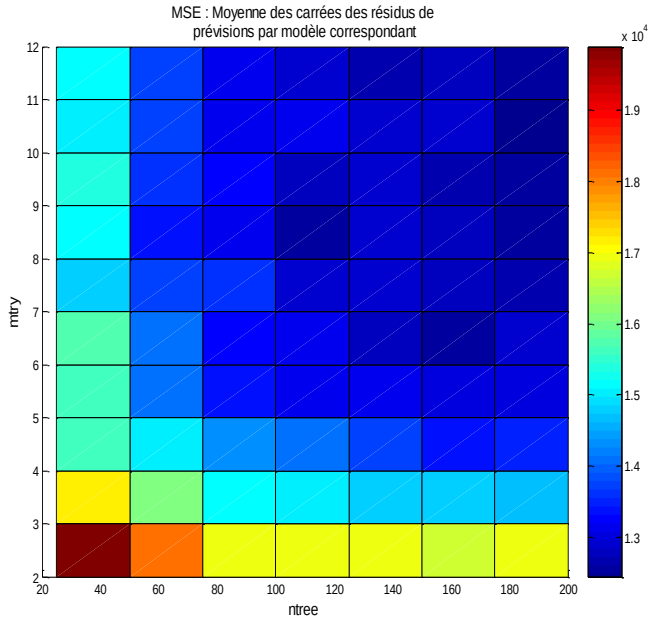

FIG. 7.1 – Somme des carrés des erreurs de prévision sur les données d'apprentissage en fonction du nombre d'échantillons bootstrap et du nombre de variables de randomisation pour la division sur les nœuds.

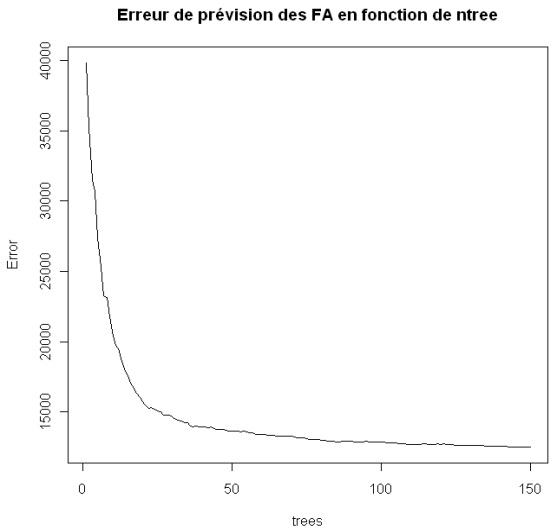

FIG. 7.2 – Somme des carrés des erreurs de prévision en fonction du nombre d'échantillons bootstrap (FA = Forêts aléatoires).

7.3.2 Hiérarchie des variables explicatives du modèle FA

Les Forêts aléatoires sont un modèle construit par agrégation, donc il n'y a pas d'interprétation directe des effets liés aux variables explicatives. Néanmoins des informations pertinentes sont 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.

7.3.2.1 Forêts aléatoires comme méthode de sélection des variables

La hiérarchie des variables explicatives se rapporte aux problèmes de sélection des variables. A ce sujet, plusieurs travaux ont été publiés. La méthode *stepwise* est l'un des premiers algorithmes développés dans le cadre des modèles linéaires ou logistiques. Dans cette lignée, Efron et *al.* (2004) [25] proposent un nouvel modèle de régression *Least angle regression* (LARS) basé sur la sélection des variables explicatives. Ce modèle est plus parcimonieux et plus performant avec un temps d'exécution inférieur à celui de l'algorithme *Stepwise*.

Concernant les modèles non paramétriques, il n'existe que très peu d'outils permettant d'établir une hiérarchie des variables. Néanmoins, les arbres de régression de la méthode CART et les forêts aléatoires offrent la possibilité d'établir une hiérarchie des variables explicatives. Des modèles récents de type Séparateur à Vaste Marge (SVM), Guyon et *al.* (2002) [35] et Rakotomamonjy (2003) [54] ont proposé des scores pour chaque variable explicative utilisée. Ces scores permettent d'établir une hiérarchie des variables. Dans leurs travaux, Ben Ishak et *al.* (2005) [36] suggèrent une procédure de type stepwise. En se basant sur les différents scores estimés par bootstrap, ils montrent que cette procédure est plus fine que la précédente. L'un des derniers développements sur la sélection de variables porte sur les travaux de Park et *al.* (2006). Ils

étudient la sélection de variables par une approche consistant à introduire une pénalité dans le critère d'optimisation utilisé dans la méthode d'estimation des paramètres d'un modèle linéaire. En comparant différentes méthodes de sélection des variables sur données réelles et simulées, Ghattas et *al.* (2008) [5] montrent que les forêts aléatoires fournissent une hiérarchie plus stable des variables que les autres méthodes. Un autre avantage des FA est la possibilité de les utiliser en régression. Voilà dans cette étude, les motivations qui ont guidé le choix de cette méthode pour proposer une hiérarchie des variables explicatives du temps de passage des aéronefs en des points de leur trajectoire de vol prévue.

7.3.2.2 Critères d'importance des variables explicatives

Dans le modèle des forêts aléatoires, plusieurs critères sont proposés pour l'évaluation de l'importance d'une variable explicative X_j sur un arbre :

- Le premier critère (Mean Decrease Accuracy) appelé encore pourcentage de gain de l'erreur quadratique moyenne (%IncMSE) est basé sur une permutation aléatoire des valeurs de la variable X_j, les autres variables restant inchangées. Il consiste à calculer la moyenne sur les observations Out-Of-Bag de la décroissance de la performance d'un arbre lorsque la variable est aléatoirement perturbée :

$$\%IncMSE = \frac{MSE_{OOB}(X_j) - MSE'_{OOB}(X_j)}{\sigma} \tag{3}$$

  où :

  - σ est l'ecart-type de l'erreur de régression ;
  - $MSE_{OOB}(X_j)$ est l'erreur quadratique moyenne de prévision de l'arbre bootstrap sur les données OOB sans permutation de la variable X_j. Cette valeur est identique pour toutes les autres variables explicatives ;
  - $MSE'_{OOB}(X_j)$ est l'erreur quadratique moyenne de prévision de l'arbre bootstrap sur les données OOB après permuation des valeurs de la variable X_j. Cette valeur est spécifique à chaque variable explicative. Plus la prévision est dégradée par la permutation des valeurs d'une variable, plus celle-ci est importante. L'algorithme pour ce critère est détaillé dans Ghattas et *al.*(2008).

- Le second critère Mean Decrease Gini est appelé gain de pureté du nœud (GPN). Il est basé sur la décroissance d'entropie ou encore la décroissance de l'hétérogénéité (pureté ou déviance) définie à partir du critère de Gini. L'importance d'une variable est une somme pondérée des décroissances d'hétérogénéité induites lorsque cette variable est utilisée pour définir la division associée à un nœud. Ce critère illustré dans Ghattas (1999b) utilise les données $In\text{-}Bag$ servant à la construction des arbres. Il est défini par la relation :

$$GPN(X_j) = \sum_{q=1}^{ntree} \left[(y_q - \widehat{y}_q(X_j))^2 \right] - \sum_{q=1}^{ntree} \left[\left(y_q - \widehat{y}_q^{'}(X_j) \right)^2 \right] \tag{4}$$

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.

- Le troisième critère repose simplement sur la fréquence de chacune des variables apparaissant dans les arbres de la forêt. Ce critère caractérisé de rudimentaire n'a pas été retenu par Breiman.

Selon Breiman les deux premiers sont très proches, l'importance d'une variable dépend de sa fréquence d'apparition mais aussi des places qu'elle occupe dans chaque arbre de la forêt.

7.3.2.3 Résultat de la hiérarchie des variables explicatives de l'écart temporel

La hiérarchie des variables explicatives obtenue de cette étude a été réalisée en optimisant le deuxième critère (Mean Decrease Gini). C'est une somme calculée sur les K échantillons bootstrap. La hiérarchie ainsi obtenue est illustrée par la figure FIG.7.3 où l'axe des abscisses représente les variables. L'importance des variables explicatives a été rapportée à l'échelle [0 ;100] en divisant toutes les valeurs obtenues par la plus grande valeur et en multipliant le résultat par 100. Rappelons que le critère optimisé lors de la construction de l'arbre de chaque échantillon bootstrap est la minimisation de l'inertie intra-classes présentée au chapitre 3. Il s'agit d'un critère basé sur la réduction d'impureté lors de la division de chaque nœud.

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 modèle ne nous permet pas d'interpréter d'avantage. Nous présentons dans les paragraphes suivants l'ensemble des résultats obtenus et comparons le pouvoir prédictif de ce modèle à ceux des trois premiers présentés dans les chapitres précédents.

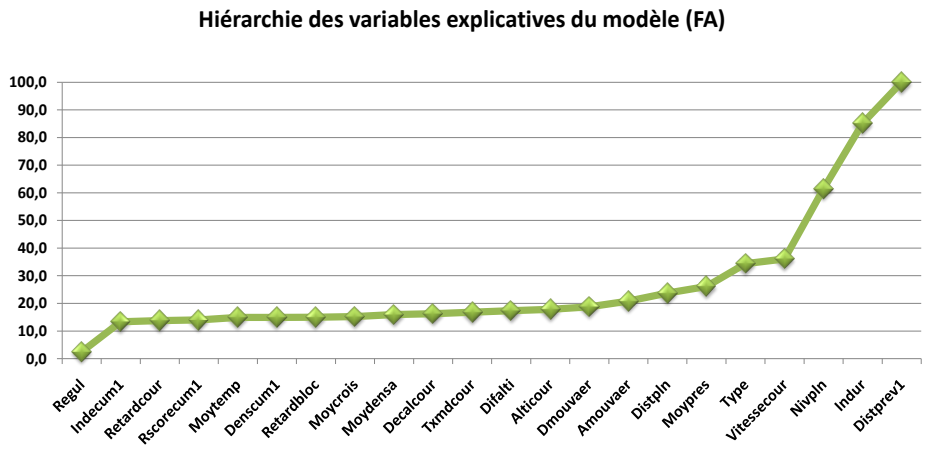

FIG. 7.3 – Importance des variables explicatives du modèle des forêts aléatoires. Le modèle est construit à partir de 150 échantillons bootstrap et 7 tirages aléatoires des variables explicatives sur chaque nœud.

7.3.3 Analyse du modèle FA

Au chapitre précédent, nous avons montré que le modèle CART modifié avait des qualités prédictives bien meilleures que les modèles de régression par arbre basé sur la méthode CART classique. Dans cette section, nous réalisons le diagnostic des résidus du modèle non-paramétrique basé sur les forêts aléatoires. Ce modèle sera comparé au modèle CART modifié.

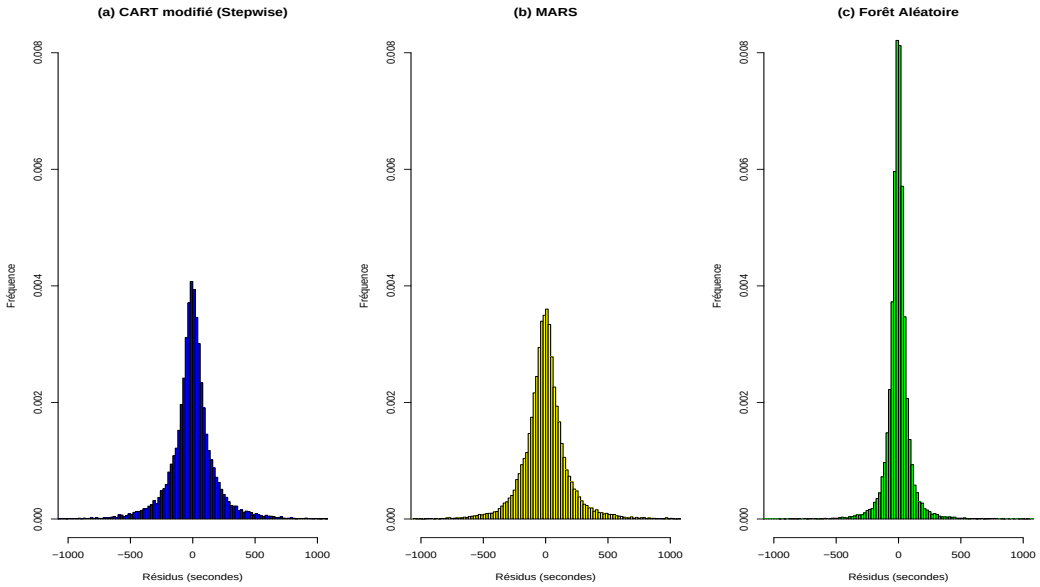

FIG. 7.4 – Histogrammes de distribution des résidus des modèles : CART modifié (Stepwise), MARS et Forêt Aléatoire

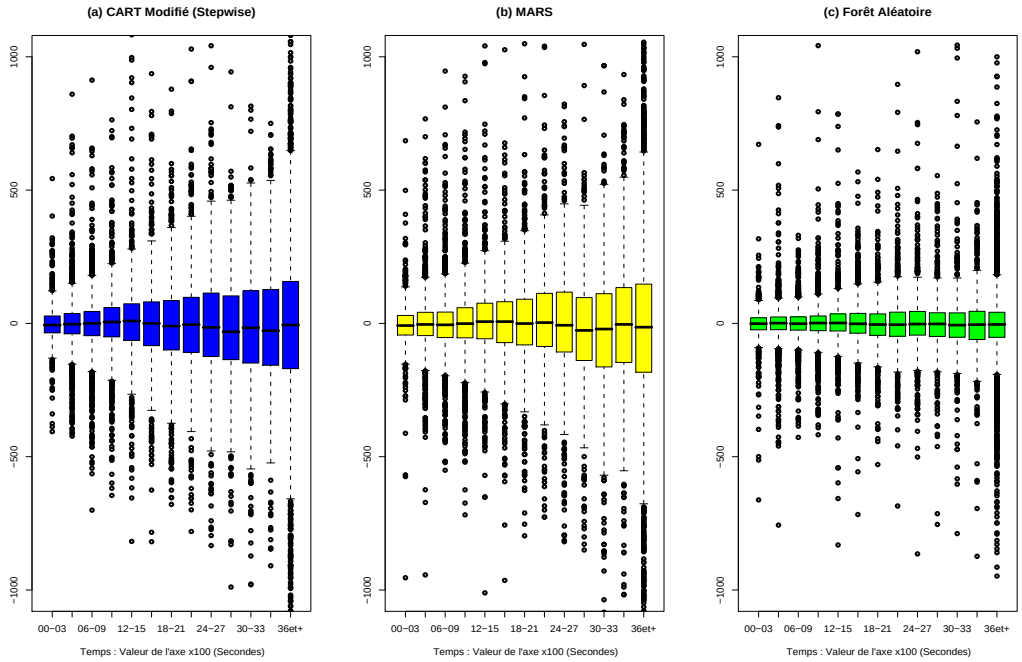

FIG. 7.5 – Dispersion des résidus en fonction de l'horizon temporel de prévision : CART modifié (Stepwise), MARS et Forêt Aléatoire

7.3.3.1 Test d'ajustement des résidus du modèle FA par une loi normale

La distribution des résidus du modèle obtenu par la méthode des forêts aléatoires est illustrée par l'histogramme (c) de la figure FIG.7.4. Les résidus sont très concentrés autour de 0 qui apparaît comme le centre d'une classe modale. La distance de Kolmogorov-Smirnov pour le test de normalité de la variable résiduelle est $D = 0.48$. Le test asymptotique de Jarque Bera a également été réalisé et la statistique de test est gale à 7032400 avec une p-value égale à 0. Ces deux tests conduisent donc à refuter l'hypothèse nulle de normalité de cette variable. Il apparaît que l'on ne peut ajuster la distribution empirique des résidus du modèle des forêts aléatoires par une loi normale de probabilité théorique connue.

7.3.3.2 Dispersion des résidus du modèle des FA en fonction de l'horizon temporel de prévision

A la lecture de la figure (b) de FIG.7.5, il ressort que la dispersion des résidus du modèle obtenu par les forêts aléatoires dépend de l'horizon temporel. De façon générale, le profil de la dispersion obtenue ici est quasiment le même que ceux que nous avons obtenus en étudiant les trois modèles précédents. Toutefois, nous notons que l'accroissement de l'amplitude des résidus obtenus par les forêts aléatoires reste très modéré sur tout l'horizon de prévision.

Afin de comparer l'incertitude de ce modèle à celles des modèles CART classique, CART modifié et MARS, nous avons illustré à travers la figure FIG.7.6 la variation des RMSE de tous ces modèles en fonction de l'horizon temporel de prévision. Comparé au modèle CART modifié, le modèle basé sur les forêts aléatoires semble fournir des erreurs de prévision beaucoup plus faibles sur quasiment tous les horizons

de prévision. Dans le paragraphe suivant, nous voulons quantifier la différence du pouvoir prédictif de ce nouveau modèle par rapport à ceux des modèles précédents.

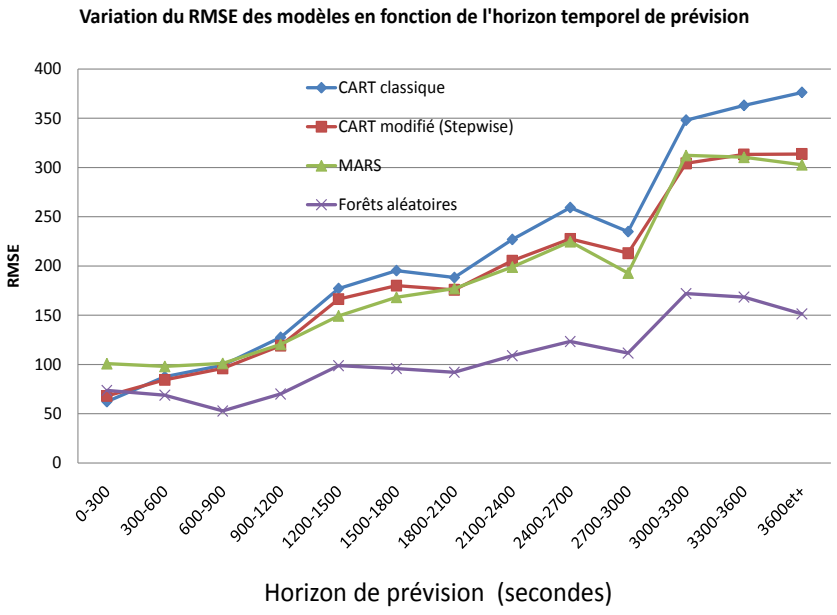

FIG. 7.6 – Variation du RMSE des modèles en fonction de l'horizon temporel de prévision : CART classique, CART modifié (Stepwise), MARS et Forêts aléatoires (FA).

7.4 Comparaison de la qualité prédictive du modèle FA à celles des autres modèles

Afin de quantifier l'apport prédictif du modèle des forêts aléatoires par rapport aux trois premiers modèles développés, nous avons fait recours, une fois de plus au coefficient de Theil. Dans chaque rapport effectué, le modèle alternatif est celui des forêts aléatoires tandis que les autres sont considérés comme des modèles de référence. En fonction de l'horizon temporel de prévision, nous avons présenté l'ensemble des résultats dans la figure FIG.7.7.

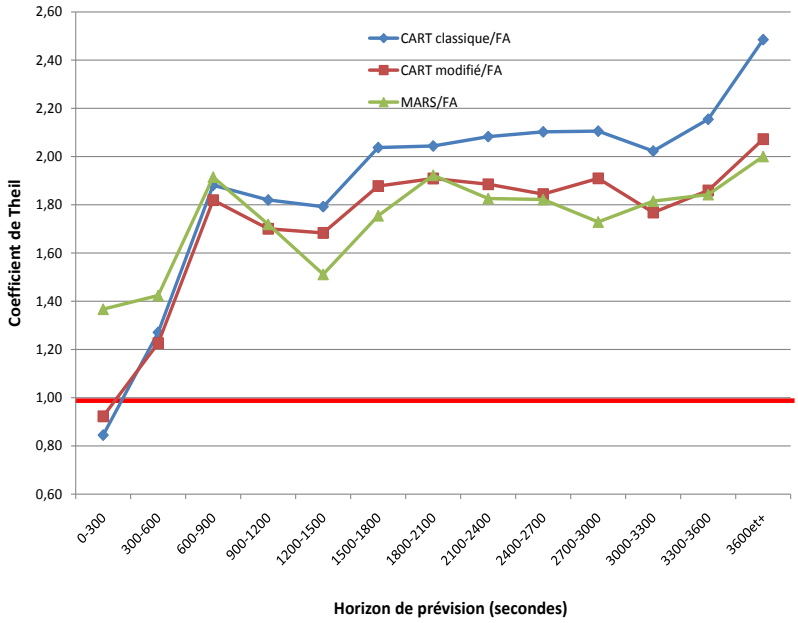

FIG. 7.7 – Indicateur de Theil pour la comparaison du pouvoir prédictif du modèle des forêts aléatoires relativement aux modèles CART classique, CART modifié et MARS. Par exemple, CART classique/FA signifie rapport du RMSE du modèle CART classique sur le RMSE du modèle des forêts aléatoires.

Pour chaque modèle, nous avons évalué le coefficient moyen de Theil sur l'ensemble des horizons. Les valeurs trouvées se situent autour de 2.18, 1.89 et 1.85 respectivement pour les modèles CART classique, CART modifié et MARS. Ainsi, le modèle ajusté par les forêts aléatoires est le plus prédictif et peut permettre d'améliorer la qualité de prévisions du temps de passage des aéronefs en des points de leur trajectoire de près de 118%, 89% et 85% respectivement par rapport aux modèles CART classique, CART modifié et MARS.

Afin de d'affiner l'étude de cette performance d'ajustement du modèle FA, nous observé l'évolution du coefficient de Theil en fonction de l'horizon temporel. Un examen de la figure FIG.7.7 montre que les modèles CART classique et CART modifié peuvent fournir de meilleures prévisions au début de l'horizon de prévision inférieur à 300 secondes (5 minutes). En revanche, cette tendance s'inverse très rapidement à partir de l'horizon de prévision supérieur à 600 secondes (10 minutes). Ce modèle semble plus performant

lorsque l'horizon de prévision devient très éloigné par rapport à l'origine. Quelque soit l'horizon de prévision, le modèle basé sur les forêts aléatoires fournit sur les données d'apprentissage les prévisions de qualité bien meilleures que celles données par les autres modèles. Toutefois, ce modèle semble souffrir comme les autres de la détérioration de son pouvoir prédictif au fur et à mesure que l'horizon temporel de prévision augmente et s'éloigne de l'origine. Le paragraphe suivant est destiné à la conclusion du chapitre.

7.5 Conclusion

Dans ce chapitre, nous avons developpé un modèle de prévision basé sur les forêts aléatoires. Ce modèle, fondé sur la technique d'agrégation d'arbres de régression construits à partir des échantillons bootstrap est une réponse au problème d'instabilité des arbres de régression produits par la méthode CART. Dans cette étude, nous avons montré qu'en outre, le modèle des forêts aléatoires améliore très significativement la qualité des prévisions des modèles CART (classique, modifié) et MARS. A travers le coefficient de Theil, nous avons pu quantifier sur les données d'apprentissage l'apport de cette technique de modélisation par rapport à celle des approches CART et MARS.

L'incertitude de prévision par ce modèle reste fortement dépendante de l'horizon temporel de prévision. La dégradation de ce modèle en fonction de l'horizon de prévision se fait de façon très modérée par rapport à celles observées sur les autres modèles.

Rappelons ici que le pouvoir prédictif réel d'un modèle s'évalue par rapport à son comportement vis-à-vis des nouvelles données. Ainsi, afin de choisir le meilleur modèle pour la prévision de la variable dépendante écart temporel, une étape importante de ce travail est désormais nécessaire. Il s'agit d'une phase de validation des modèles sur les données test différentes de celles d'apprentissage qui ont participé à leur ajustement. C'est au terme de cette ultime étape que nous aurons une idée du comportement de chacun de ces modèles lorsqu'il sera utilisé en situation opérationnelle du trafic. Ainsi, le chapitre suivant va porter sur le concept de prédictivité des modèles et sur la validation des quatre modèles proposés dans cette thèse.

Chapitre 8

Prédictivité des modèles et application à la prévision de la charge secteur

8.1 Introduction

Nous avons proposé quatre modèles non-paramétriques permettant de prévoir en fonction des variables explicatives, les instants de passage des aéronefs en des points de leur trajectoire de vol prévue. A travers ce chapitre, nous visons un objectif double :

- Le premier consiste à évaluer la performance des modèles développés sur leur capacité de prévision en situation réelle du trafic.
- Le deuxième consiste à proposer un horizon temporel de prévision des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue et à calculer la charge des secteurs de l'espace aérien par ces modèles.

Ces modèles ne pourront être utiles que s'ils sont dynamiques et peuvent fonctionner en prenant en compte l'évolution du trafic en temps réel. Ainsi, il est important de bien définir un horizon temporel d'utilisation qui soit compatible avec les objectifs visés et atteignables pour une gestion optimale du trafic. Cet horizon de prévision doit prendre en compte les délais de détection des problèmes dans le trafic et du temps nécessaire pour y apporter efficacement des actions correctives appropriées. Rappelons que le but premier de la prévision du trafic est d'anticiper sur les variations futures des flux avec le plus de précision possible. Au final, la question posée est : En utilisant un ou plusieurs des modèles proposés, peut-on prévoir le trafic ? Si oui, sur quel horizon temporel de prévision ? Telle est la question cruciale à laquelle nous voulons, au travers de ce chapitre, proposer une réponse.

Le chapitre est organisé comme suit. Le paragraphe (8.2) est consacré aux indicateurs d'évaluation des modèles. Le paragraphe (8.3) définit la notion de prédictibilité du trafic en en proposant une métrique pour cette étude. Ainsi, le théorème du rapport des vraisemblances monotone est présenté. Nous détaillons les conditions d'application de ce théorème et construisons ensuite les régions de prédictivité des modèles. Dans le paragraphe (8.4) nous proposons une estimation de la distribution de probabilité des résidus des

différents modèles par un mélange de lois gaussiennes. Le paragraphe (8.5) se focalise à la prévision de la charge d'un secteur de l'espace aérien par les différents modèles proposés. Une approximation de l'horizon de prévision est donnée ainsi que celle de la largeur de la fenêtre de prévision. Enfin, le paragraphe (8.6) conclut le chapitre.

8.2 Evaluation des modèles

La performance du modèle issu d'une méthode d'apprentissage s'évalue par sa *capacité de prévision* dite encore de *capacité de généralisation* en situation nouvelle. Son évaluation est nécessaire dans la mesure où, le meilleur modèle n'est pas nécessairement celui qui ajuste le mieux les données d'apprentissage. Trois méthodes sont souvent utilisées pour la mesure de cette performance (Besse ; 2009)[6] :

- La première consiste en 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 ;
- La deuxième consiste en une pénalisation de l'erreur d'ajustement en faisant intervenir la complexité du modèle, c'est-à-dire du nombre de variables explicatives ;
- La troisième repose sur des simulations.

Le choix de la meilleure méthode dépend de plusieurs facteurs dont la taille de l'échantillon initial, la complexité du modèle envisagé, la variance de l'erreur. Dans le cadre de ce mémoire, nous avons l'avantage de disposer d'une base importante de données. Ainsi, la première méthode serait la mieux adaptée.

8.2.1 Indicateurs d'évaluation des modèles

L'indicateur d'erreur de prévision considéré repose sur le critère minimisé dans la méthode des moindres carrés. C'est une estimation biaisée car trop optimiste de l'erreur de prévision. Elle est liée aux données qui ont servi à l'ajustement du modèle et décroit avec le nombre de variables indépendantes. L'estimation de l'indicateur de qualité ne dépend que de la partie *biais* de l'erreur de prévision et ne prend pas en compte la partie *variance* de la variable à expliquer (Besse ; 2009). Il est donc nécessaire d'éclater l'échantillon en deux parties respectivement appelées échantillon d'apprentissage et échantillon test.

Par ailleurs, le coefficient de Theil tel que nous l'avons utilisé pour la sélection du meilleur modèle d'ajustement semble peu robuste, car il est calculé à partir de la moyenne des carrés des erreurs de prévisions. La moyenne est en effet 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 protégée de l'influence de ces valeurs extrêmes, elle est donc un indicateur plus fiable et représentatif de la situation de la majorité des individus de l'échantillon. Ainsi, dans cette étape d'évaluation des modèles sur les données test, nous comparons les modèles entre eux en utilisant l'indicateur médian noté RMedSEP[1] qui est la racine carrée de la médiane des carrés des erreurs de prévisions au lieu du RMSEP[2]. On définit de façon similaire l'indicateur

[1]Root Median Square Error of Prediction. Cet indicateur est calculé sur les données test.
[2]Root Mean Square Error of Prediction.

médian calculé directement à partir du modèle d'ajustement sur données d'apprentissage noté RMedSE[3].

8.2.2 Echantillon de données test pour la validation

L'échantillon de données test est formé d'aéronefs dont les instants courants de prévision sont 7 heures (matin) ou 16 heures. Les points correpondants sur leur trajectoire de vol sont bien différents de ceux de l'échantillon de données d'apprentissage. En effet, ce dernier est formé d'aéronefs observés pendant leur vol à l'instant $t_0 \in \{8,11,14,17\}$ heures. Si un aéronef est simultanément présent dans l'échantillon de données d'apprentissage et dans l'échantillon de données test, il occupe nécessairement des positions de vol différentes. Ainsi, il présente des caractéristiques de vol différentes, et notamment les caractéristiques aux points courants dans les deux échantillons. L'application des modèles ajustés sur les données test avec 7388 individus va finalement permettre d'évaluer l'erreur de prévision non biaisée de ces modèles lorsqu'ils sont utilisés en situation nouvelle. C'est dans ces conditions que le meilleur modèle au sens prédictif va être déterminé.

8.2.3 Erreur d'ajustement et erreur de prévision sur données test

Pour chacun des quatre modèles, nous avons calculé son indicateur médian de performance RMedSE en fonction de l'horizon temporel de prévision. Cet indicateur obtenu directement du modèle d'ajustement est utilisé comme référence pour évaluer les performances de chaque modèle appliqué sur les données test. Les profils des indicateurs issus de l'ajustement et de ceux calculés sur les données test sont représentés sur le même graphique. Deux profils principaux se distinguent :

- Le premier concerne les modèles CART classique, CART modifié et MARS dont les profils de l'indicateur médian sur les erreurs sont illustrés par les figures respectives FIG.(8.1, 8.2, 8.3). A la lecture de celles-ci, il apparaît que pour un horizon de prévision inférieur à 2400 secondes (40 minutes), chaque modèle fournit sur les données test des prévisions de qualité au moins égale à celle des prévisions de référence obtenue sur l'échantillon de données d'apprentissage.
- Le deuxième profil porte uniquement sur le modèle des forêts aléatoires (FIG.8.4). Il met en évidence le surapprentissage qui se traduit par un très grand écart entre l'erreur de référence issue de l'ajustement et l'erreur de prévision sur l'échantillon des données test. Pour un horizon de prévision rapproché, cet écart est faible et croît au fur et à mesure que cet horizon de prévision augmente. Notons que même si l'écart entre l'erreur d'ajustement sur données d'apprentissage et l'erreur de prévision sur les données test est important, cette dernière reste en revanche inférieure aux erreurs de prévisions données par les trois autres modèles (voir figure FIG.8.5).

[3]Root Median Square Error.

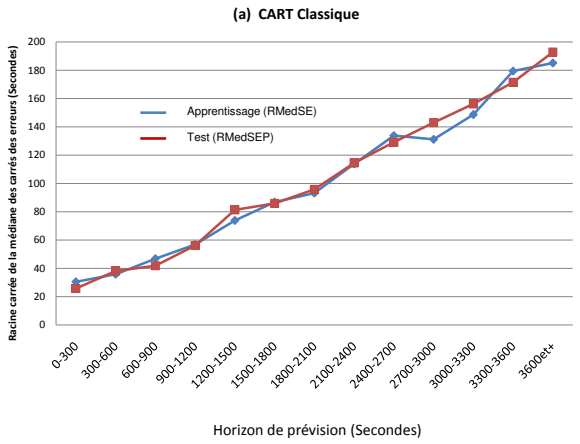

FIG. 8.1 – Comparaison du RMedSEP au RMedSE pour le modèle CART Classique

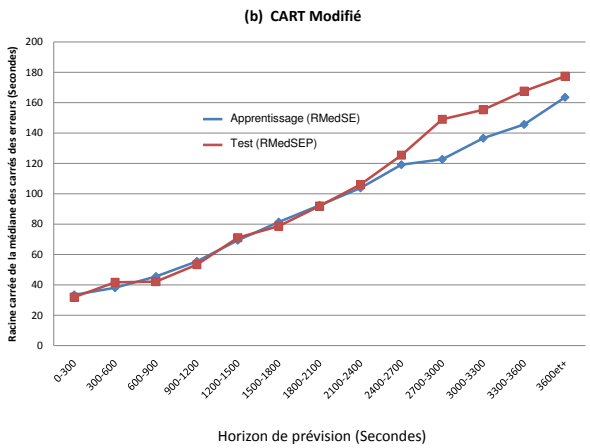

FIG. 8.2 – Comparaison du RMedSEP au RMedSE pour le modèle CART Modifié

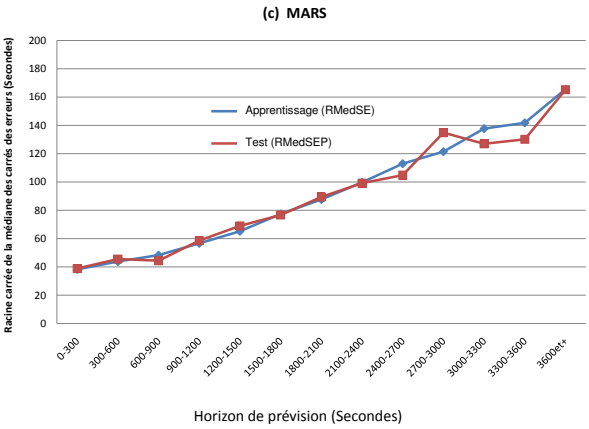

FIG. 8.3 – Comparaison du RMedSEP au RMedSE pour le modèle MARS

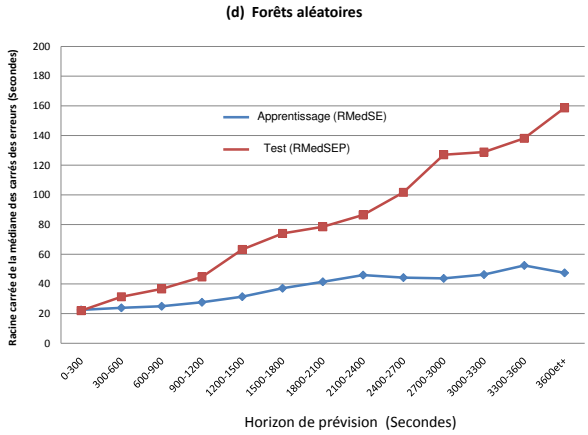

FIG. 8.4 – Comparaison du RMedSEP au RMedSE pour le modèle des Forêts aléatoires

8.2.4 Identification du meilleur modèle

Afin d'identifier le meilleur modèle des quatre proposés dans cette étude, nous avons représenté sur le même graphique FIG.8.5 les profils des RMedSEP pour les différents modèles en fonction de l'horizon temporel de prévision. Dans un sens prédictif, le modèle des forêts aléatoires semble fournir des prévisions bien meilleures sur tous les horizons. Au début de l'horizon de prévision, le modèle MARS est de moins bonne qualité. Notons cependant qu'au fur et à mesure que la profondeur de l'horizon temporel de prévision augmente, cette tendance s'inverse et MARS devient le deuxième modèle le plus prédictif après les forêts aléatoires.

Afin de procéder à la sélection du modèle final, nous allons d'abord quantifier la prédictivité des diffé-

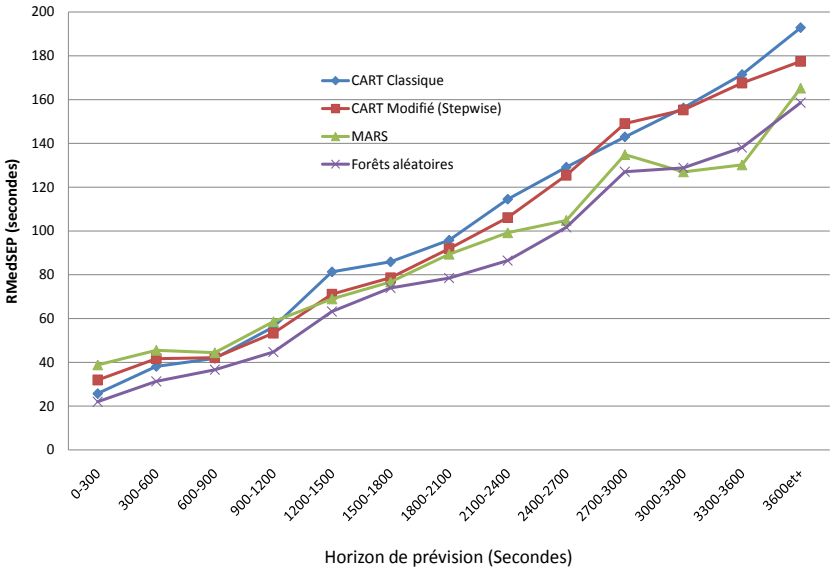

FIG. 8.5 – Profils comparés des RMedSEP des différents modèles sur les données test

rents modèles. L'indicateur généralement utilisé est le rapport de la variabilité SD[4] de la variable dépendante sur la variabilité de l'erreur de prévision du modèle (RMSEP). L'écart-type SD est corrigé de biais et est défini par :

$$SD = \sqrt{\frac{n \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}{n\,(n-1)}} \tag{1}$$

où les $(y_i)_{(1<i\leq n)}$ sont des valeurs de référence des écarts temporels pour l'échantillon des données test. Le rapport $\frac{SD}{RMSEP}$ est souvent utilisé comme meilleur indicateur de mesure de la qualité de prévisions d'un modèle (Williams ; 1990)[64]. Si le RMSEP est proche de l'écart-type (SD)[5] (c'est-à-dire, que ce rapport est proche de un), le modèle ajusté n'est pas performant. En revanche, si ce rapport est supérieur à un, le modèle est plus performant. Ainsi, nous avons calculé cet indicateur pour chaque modèle et les valeurs sont consultables dans la table TAB.8.1. La valeur la plus élevée (1.55) correspond au modèle des forêts aléatoires. Ensuite, pour les différents modèles, nous avons représenté (FIG.8.6) la variation de cet indicateur en fonction de l'horizon temporel de prévision. Il apparaît que les modèles ne sont réellement performants qu'à partir d'une certaine profondeur de l'horizon de prévision (au tour de 1200 secondes, soit 20 minutes).

Par ailleurs, le point commun à tous ces modèles est leur tendance à surestimer les prévisions (FIG.8.7). Rappelons que les valeurs médianes des erreurs de prévisions pour les modèles CART classique, CART modifié, MARS et les forêts aléatoires sont respectivement -9.6, -8.56, -9.44 et -7.53. C'est-à-dire les

[4]Standard deviation.

[5]SD est calculé sur les données initiales. Il s'agit des données d'apprentissage.

Modèles	Indicateur de prédictivité des modèles
CART Classique	1.28
CART modifié (Stepwise)	1.43
MARS	1.44
Forêts aléatoires	1.55

TAB. 8.1 – Comparaison de la prédictivité des modèles proposés. On utilise l'indice SD/RMSEP. Le modèle ayant une forte capacité prédictive a l'indice le plus élevé possible et supérieur à un.

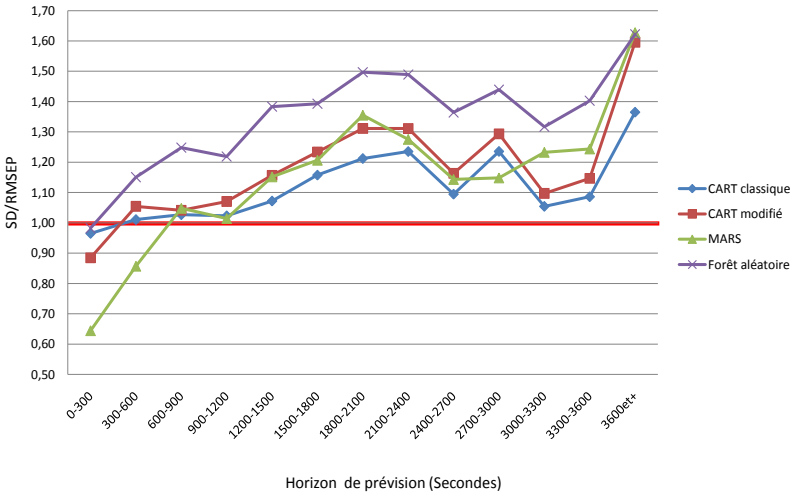

FIG. 8.6 – La capacité prédictive des modèles en fonction de l'horizon de prévision

modèles ont tendance à prévoir les instants de passage des aéronefs sur les points de leur trajectoire plus tôt que les instants de passage réellement observés dans les conditions du trafic. Le modèle des forêts aléatoires est celui qui surestime le moins les prévisions. Il apparaît comme le meilleur modèle avec le pouvoir prédictif le plus élevé lorsqu'il est appliqué sur les données test. Afin de valider ce choix, nous appliquons dans la section suivante les quatre modèles à la prévision du temps de passage des aéronefs sur les points de leur trajectoire d'une part, et en déduire la prévision de la charge des secteurs de l'espace d'autre part.

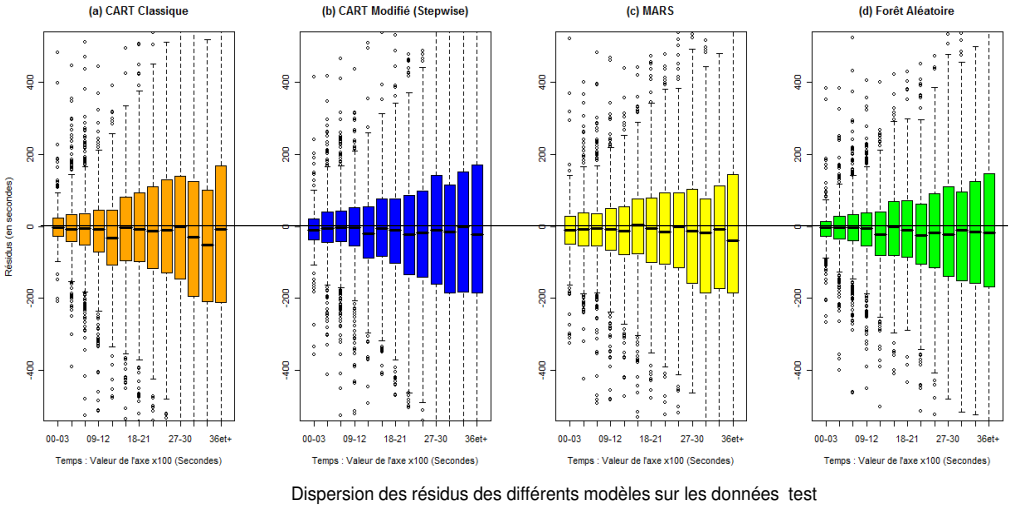

FIG. 8.7 – Zoom de la dispersion des erreurs de prévisions des quatre modèles sur [-500 ; 500]

8.3 Prédictibilité du temps de passage par les modèles

8.3.1 Généralités

Pour une utilisation optimale des modèles proposés en conditions réelles du trafic, il est nécessaire de disposer d'un horizon de temps de prévision suffisant entre le moment où une anomalie est identifiée dans le trafic futur et le moment où une solution y est apportée. Ainsi, un grand intervalle de prévision est nécessaire pour l'évaluation de la situation du trafic et compenser les éventuels retards causés par le prélèvement des données réelles du trafic et pour le calcul des prévisions. Cependant, une faible erreur de prévision du trafic est aussi nécessaire. Car, les actions correctives basées sur des prévisions erronées peuvent entraîner les effets inattendus contre les objectifs de performance et d'efficacité globale du trafic. Ainsi, dans le souci d'une gestion et d'une utilisation optimales des ressources, les autorités de planification et de régulation des flux du trafic ont besoin des prévisions les plus précises possible. Malheureusement, la précision des prévisions se dégrade très rapidement avec l'augmentation de l'intervalle de prévision. Un compromis est alors nécessaire entre une grande fenêtre temporelle de prévision et une faible incertitude de prévision.

Notons que chaque type de trafic obéit à des conditions spécifiques de prédictibilité. Aussi, le même type de trafic peut présenter différentes conditions de prédictibilité si les prélèvements des mesures du trafic en temps réel se font à des échelles de temps différentes, c'est-à-dire sur des intervalles de temps de longueurs différentes. Ainsi, il est important de déterminer des indicateurs statistiques fiables pour la prévision du trafic. Pour cela, nous avons recensé dans la littérature existante un nombre important de travaux déjà réalisés autour du thème de prédictibilité (ou prévisibilité) et horizon de prévision du trafic.

8.3.2 Littérature liée

Parmi les récents travaux au sujet de la prédictibilité et horizon de prévision du trafic, Flener et *al.* (2007) [27] ont introduit la notion d'intervalle de complexité dans un secteur de l'espace aérien. Cet intervalle est utilisé comme moyen pour l'estimation de la charge du travail des contrôleurs (ATC[6] workload) dans ce secteur. Ces auteurs définissent l'intervalle de complexité d'un secteur comme une fenêtre de temps qui est une combinaison linéaire de trois facteurs de complexité suivants : le nombre d'aéronefs évoluant à l'intérieur de ce secteur, le nombre d'aéronefs sur les segments ne correspondant pas à un niveau de vol (c'est-à-dire les aéronefs qui sont en phase de montée ou de descente) et le nombre d'aéronefs dans le secteur et évoluant sur les frontières des secteurs. En effet, les frontières des secteurs de l'espace aérien sont des zones de transmission et de réception des aéronefs qui passent d'un secteur à l'autre. Elles nécessitent à juste titre plus d'attention et plus de surveillance de la part des contrôleurs.

Flener et *al.* proposent un intervalle de complexité sur un horizon temporel compris entre 20 et 90 minutes. L'intervalle de complexité du trafic est alors le temps nécessaire pour prendre les actions appropriées pour la résolution ou la minimisation de la complexité dans la gestion du trafic. Il permet également une meilleure répartition et un équilibrage optimal de la charge du travail due aux différents facteurs de complexité entre les secteurs adjacents d'une région de l'espace. A cet effet, si l'horizon temporel de prévision est inférieur à 20 minutes, le temps n'est plus suffisant pour les calculs et pour la mise en œuvre d'une solution en vue de résoudre un problème de complexité impliquant les aéronefs. Si en revanche, cet horizon est supérieur à 90 minutes, l'incertitude est très importante sur la prévision de la trajectoire du vol. Ce qui une fois de plus, est confirmé par les résultats obtenus dans cette étude et résumé par les figures FIG.(7.6, 7.5).

En modélisant la gestion en temps réel du trafic aérien par des algorithmes génétiques, N. Durand et *al.* [50] fixent l'horizon de prévision entre 10 et 15 minutes. Ils distinguent trois périodes dans l'horizon temporel de prévision. La première est une période verrouillée pendant laquelle aucune modification de la trajectoire ne peut être effectuée. En effet, pendant le temps nécessaire à l'évaluation de la situation réelle du trafic, la résolution des conflits possibles et la transmission des ordres de manœuvres, les avions continuent à voler. Il est donc impossible de modifier leur trajectoire. La période suivante est appelée période définitive, en effet, les ordres de manœuvres donnés dans cette période ne pourront plus être modifiés. La dernière période est celle des manœuvres prévues.

Dans la même lignée que les auteurs précédents, Aimin et *al.*(2002) [55] ont travaillé sur l'horizon temporel de prévision d'un réseau de trafic en général. Ils se sont intéressés aux métriques de prédictibilité du trafic. Leurs études se sont focalisées sur la recherche de l'intervalle maximal de prévision (MPI)[7] du trafic sous la contrainte d'un niveau d'incertitude toléré. Ces auteurs ont étendu leurs travaux, sur l'identification des ressources à mobiliser pour atténuer l'augmentation de l'incertitude dans la prévision du trafic d'une part, et sur les indicateurs pertinents mesurés en temps réel et qui statistiquement sont caractéristiques de la

[6] Air traffic control.
[7] Maximum Prediction Interval.

prédictibilité du trafic d'autre part. Ils montrent que la mise en œuvre du système de prévision en temps réel du trafic souffre énormément de la détérioration de la qualité de prévision au fur et à mesure que l'horizon de prévision croît.

Dans cette section, notre objectif consiste à proposer un intervalle temporel de prévision pour une utilisation opérationnelle des modèles proposés dans cette thèse pour la prévision du temps de passage des aéronefs en des points de leur trajectoire de vol prévue. Pour cela, une méthode de mesure de la prédictibilité appelée « métrique de prédictivité » est nécessaire. En s'appuyant sur le théorème du rapport de vraisemblances monotone, les régions de confiance des prévisions sont construites à partir des objectifs cibles définis par l'autorité de gestion et de régulation du trafic. Ces objectifs sont donnés par un couple de paramètres (λ_0, α) définis ci-dessous.

8.3.3 Exemple de contrainte de prédictibilité

Pendant le processus de prévision du trafic (même dans le cadre du trafic sur les réseaux des télécommunications), les facteurs généralement pris en compte sont : les ressources disponibles pour la gestion et la régulation du trafic, le temps nécessaire à la collecte des données du trafic en temps réel et les statistiques du trafic. Ainsi, la prédictibilité du trafic dépend des objectifs visés par les autorités de gestion du trafic et des moyens dont ils disposent. Dans leurs travaux, Aimin et *al.*(2002) définissent ces objectifs par une contrainte de type erreur des prévisions normalisées donnée par la relation :

$$err(h) = \left| \frac{Y(t_0 + h) - \widehat{Y}(t_0 + h)}{Y(t_0 + h)} \right|. \tag{2}$$

Cette contrainte exige en effet que $err(h)$ ne dépasse pas un pourcentage λ_0 avec une probabilité α fixée. h est l'intervalle de prévision du trafic, $Y(t_0 + h)$ est la valeur observée de la variable dépendante à l'instant $t_0 + h$ et $\widehat{Y}(t_0 + h)$ est la valeur prévue de cette variable dépendante par le modèle à ce même instant. Le couple (λ_0, α) traduit les objectifs visés par le système de prévision et permet finalement de définir une approximation de l'intervalle de confiance de l'horizon de prédictibilité du trafic.

Selon Aimin et *al.*, la prédictibilité du trafic repose sur la détermination de l'intervalle maximal de prévision (MPI). En effet, le trafic est déclaré prédictible si : MPI permet d'avoir suffisamment de marge de temps pour la collecte des données du trafic en temps réel, pour l'identification d'éventuelles zones de conflits et de surchage du travail des contrôleurs dans l'espace. Cette marge de temps doit permettre la prévision des actions de contrôle nécessaires sur l'ensemble du réseau du trafic, et permet également d'atteindre les objectifs du trafic avec le niveau de confiance de prévision exigé et défini par le couple (λ_0, α).

MPI traduit alors les possibilités du trafic d'atteindre un niveau d'efficacité visée. D'une certaine manière, MPI peut être vue comme la borne supérieure d'une fenêtre de temps à l'intérieur de laquelle le modèle proposé prévoit le trafic avec la meilleure précision souhaitée.

8.3.4 Métrique de prédictibilité utilisée

L'erreur de prévision normalisée proposée par Aimin et *al.* comme métrique de prédictibilité pénalise fortement les valeurs extrêmes. Elle est par ailleurs définie pour des variables dépendantes strictement positives. Or, notre variable d'intérêt, les écarts temporels prennent éventuellement les valeurs nulles. Donc, cette métrique n'est pas adaptée. Une alternative à cette métrique est la fonction de perte qui est l'écart absolu de prévision défini par l'équation suivante :

$$\mathcal{Q}(h) = \left| Y(t_0 + h) - \widehat{Y}(t_0 + h) \right|. \tag{3}$$

Cette métrique a l'avantage d'être plus robuste, car moins sensible aux valeurs extrêmes. Elle permet de construire la région de confiance de prédictivité du modèle au moyen des tests composites utilisant la méthode du rapport des vraisemblances. Ces tests sont basés sur la contrainte de prévision du temps de passage d'un aéronef en un point définie par le couple (λ_0, α)[8]. Ainsi, la contrainte définissant les objectifs de gestion et de régulation du trafic exige que l'erreur absolue de prévisions $\mathcal{Q}(h)$ sur les données test ne dépasse pas un seuil λ_0 avec une probabilité fixée α (Aimin et *al.*). Ainsi, en fonction de l'horizon de prévision du modèle, nous allons réaliser aux paragraphes suivants les tests statistiques afin d'estimer la limite de l'horizon de prévision au dela duquel les écarts temporels ne sont plus prévisibles par les modèles avec un niveau de confiance souhaité. Commençons d'abord par rappeler quelques notions de base sur les tests d'hypothèses que nous utilisons.

8.3.5 Tests entre hypothèses multiples unilatérales

8.3.5.1 Définition

On appelle test d'hypothèses multiples unilatérales tout test où les hypothèses sont du type : $\mathbf{H_0} : \theta \leq \theta_0$ contre $\mathbf{H_1} : \theta > \theta_0$ ou $\mathbf{H_0} : \theta \geq \theta_0$ contre $\mathbf{H_1} : \theta < \theta_0$ où θ est un paramètre de dimension 1 ($\theta \in \mathbb{R}$).

8.3.5.2 La vraissemblance

Soit un échantillon $X = (x_1, \ldots, x_n)$ de taille n d'une variable aléatoire distribuée selon la loi de probabilité de densité f de paramètre réel θ (éventuellement vectoriel). La vraissemblance de cet échantillon est la fonction L qui au couple (θ, X), on associe la probabilité :

$$L(\theta, X) = \prod_{i=1}^{n} f(\theta, x_i). \tag{4}$$

Une statistique de X est toute fonction de l'échantillon X à valeur dans $\mathbb{R}$. Une statistique S est dit exhaustive pour le paramètre θ si la distribution de l'échantillon conditionnellement à S ne dépend pas de θ. Autrement dit, la vraissemblance $L(\theta, X|S = s_0)$ est indépendante de θ.

[8]Dans ce contexte λ_0 n'est pas un pourcentage comme c'est fût le cas pour la l'erreur de prévision normalisée, mais il est un seuil de l'erreur absolue du modèle à ne pas dépasser lors de l'utilisation opérationnelle des modèles.

8.3.5.3 Test uniformément plus puissant

Un test est dit uniformément le plus puissant ou UPP si la région critique ou de rejet de l'hypothèse $\mathbf{H_0}$ ne dépend pas du paramètre θ.

8.3.5.4 Proposition 1

S'il existe une statistique $S = s(x_1, \ldots, x_n)$ exhaustive minimale à valeurs dans $\mathbb{R}$ et si, pour tout couple (θ_1, θ_2) tel que $\theta_1 < \theta_2$, le rapport des vraisemblances RV défini par :

$$RV(s(x_1, \ldots, x_n)) = \frac{\mathbf{L}(\theta_1, s(x_1, \ldots, x_n))}{\mathbf{L}(\theta_2, s(x_1, \ldots, x_n))} \tag{5}$$

est une fonction monotone de $s(x_1, \ldots, x_n)$, alors il existe un test uniformément le plus puissant (UPP)[9] pour les situations d'hypothèses unilatérales et la région de rejet est soit de la forme : $s(x_1, \ldots, x_n) < K$, soit de la forme $s(x_1, \ldots, x_n) > K$, où K est la valeur critique de la statistique de test. Cette proposition est démontrée dans Lejeune (2004,2005)[43].

8.3.5.5 Proposition 2 (Théorème de Lehmann)

Pour le problème de test : $\mathbf{H_0} : \theta \leq \theta_0$ contre $\mathbf{H_1} : \theta > \theta_0$, on suppose que la loi de probabilité P_θ associée au paramètre θ est à rapport des vraisemblances monotone croissante en la statistique S =$s(x_1, \ldots, x_n)$. Dans ces conditions, il existe un test UPP dont la région critique W est l'ensemble des points $(x_1, \ldots, x_n)$ tels que : $s(x_1, \ldots, x_n) > K$ où la valeur de la constante K est déterminée par le risque fixé $\alpha = P_\theta(W|\theta = \theta_0)$.

8.3.5.6 Proposition 3

Si la loi mère de l'échantillon $X = (x_1, \ldots, x_n)$ est dans une famille appartenant à la classe exponentielle, c'est-à-dire la densité de probabilité du phénomène décrit est de la forme :

$$f(\theta, x) = a(\theta)b(x)exp(c(\theta)d(x)), \tag{6}$$

alors le test du rapport des vraisemblances (RV) a une région de rejet de la forme : $\sum_{i=1}^{n} d(x_i) < K$ si $c(\theta_0) - c(\theta_1) > 0$ ou $\sum_{i=1}^{n} d(x_i) > K$ si $c(\theta_0) - c(\theta_1) < 0$.
En effet, $s(x_1, \ldots, x_n) = \sum_{i=1}^{n} d(x_i)$ est une statistique exhaustive minimale et

$$RV(s(x_1, \ldots, x_n) = \left[\frac{a(\theta_0)}{a(\theta_1)}\right]^n exp\left\{[c(\theta_0) - c(\theta_1)]\sum_{i=1}^{n} d(x_i)\right\}. \tag{7}$$

Le sens de l'inégalité dépend du signe de $c(\theta_0) - c(\theta_1)$. La loi de $\sum_{i=1}^{n} d(x_i)$ est de type connu et K est le quantile d'ordre α de cette loi sous $\mathbf{H_0}$ ou d'ordre $(1 - \alpha)$ selon que le signe est positif ou négatif. En

[9] On dit qu'un test τ^* est uniformément le plus puissant au niveau α s'il est uniformément plus puissant que tout autre test au niveau α.

pratique, on calcule souvent la puissance de ce test optimal pour en déduire son niveau. Après ces quelques rappels, nous allons maintenant montrer qu'il est possible d'approcher la distribution de probabilité des erreurs absolues par une loi de la famille exponentielle.

8.3.6 Ajustement de la distribution de probabilité des erreurs absolues $\mathcal{Q}(h)$

Dans un premier temps, nous procédons aux tests de Kolmogorov-Smirnov pour accepter ou refuter l'hypothèse nulle qui suppose que la distribution de probabilité des erreurs absolues suit une loi exponentielle. Autrement dit, que pour chaque modèle proposé, la loi de probabilité des erreurs absolues est de la forme :

$$f(a, x) = ae^{-ax}, \tag{8}$$

où a est le paramètre à estimer à partir de l'échantillon des données d'apprentissage. En notant $\widehat{a}$ l'estimation de ce paramètre par l'inverse de l'espérance mathématique des erreurs absolues, la fonction de densité correspondante est notée $\widehat{f}$. Les résultats de ces tests sont disponibles dans le tableau (TAB.8.2) :

	K.S : Distance de Kolmogorov-Smirnov	**P-value**	**Estimation du paramètre $\widehat{a}$**	**Décision**
CART classique	0.1071	0	0.0071677	rejet
CART modifié	0.0812	0	0.0079354	rejet
MARS	0.0639	0	0.0079858	rejet
Forêts aléatoires	0.1210	0	0.0168806	rejet

TAB. 8.2 – Synthèse des résultats de Kolmogorov-Smirnov pour le test d'ajustement de la distribution des erreurs absolues à une loi de la forme exponentielle. Rejet signifie que les erreurs absolues ne suivent pas la loi exponentielle pour le modèle developpé. Pour chaque modèle, la fonction de densité théorique ajustée est $\widehat{f}(\widehat{a}, x) = \widehat{a}e^{-\widehat{a}x}$.

L'estimation du paramètre par la méthode de l'espérance mathématique des erreurs absolues est la plus utilisée pour les tests d'ajustement des lois exponentielles. Cette méthode est implantée dans beaucoup de logiciels de statistique. Une lecture des résultats de ces tests montre qu'il est difficile d'ajuster la loi de probabilité des erreurs absolues par une fonction de densité de la famille exponentielle. Ainsi, nous avons procédé à des améliorations des fonctions de densité candidates de cette classe. Pour chaque modèle, la fonction recherchée est toujours de la forme de l'équation (8). La nouvelle approche consiste à utiliser la méthode des *Moindres carrés non-linéaires* pour déterminer de façon optimale le paramètre $\widehat{a}$ qui est une estimation de a dans l'équation (8). La fonction de densité de probabilité ainsi ajustée sera notée $\widehat{f}_{nls}$, où nls désigne Nonlinear Least Squares. Les résultats obtenus sont synthétisés dans le tableau TAB.8.3.

Relativement aux résultats des tests précédents (TAB.8.2), nous observons une nette amélioration de la qualité d'ajustement lorsqu'on utilise la méthode des moindres carrés non-linéaires. Cette méthode a permis de réduire la distance de Kolmogorov-Smirnov ($K.S$) de plus de la moitiée. La réduction la plus importante concerne le modèle $MARS$ où la valeur de $K.S$ est passée de 0.0812 à 0.0187 pour la méthode nls. Cette réduction est sans doute liée à la contrainte de continuité imposée lors de la construction des estimateurs par

le modèle $MARS$. La deuxième grande réduction est observée sur le modèle des forêts aléatoires où $K.S$ est passée de 0.1210 à 0.0499.

Compte tenu de la faible valeur des distances de Kolmogorov-Smirnov obtenues par les moindres carrés non-linéaires, nous pensons que le rejet de l'hypothèse nulle qui dit que la distribution des erreurs absolues suit une loi de forme exponentielle est partiellement expliqué par la puissance du test pour les échantillons de grande taille. Dans la suite de l'étude, nous supposons que la distribution mère de la loi de probabilité des erreurs absolues est de la famille exponentielle. Ce qui nous permet d'utiliser les propriétés du rapport des vraissemblances pour construire les intervalles de prédictivité des modèles.

	K.S : Distance de Kolmogorov-Smirnov	**P-value**	**Estimation du paramètre $\widehat{a}$**	**Décision**
CART classique	0.0552	0	0.00924	rejet
CART modifié	0.0308	0	0.00909	rejet
MARS	0.0187	0	0.00880	rejet
Forêts aléatoires	0.0499	0	0.02264	rejet

TAB. 8.3 – Synthèse des résultats de Kolmogorov-Smirnov pour le test d'ajustement de la distribution des erreurs absolues à une loi de la forme exponentielle. Pour chaque modèle, la fonction de densité théorique ajustée est $\widehat{f}_{nls}(\widehat{a}, x) = \widehat{a}e^{-\widehat{a}x}$.

Pour chaque modèle, les deux fonctions de densité ajustées respectivement par la méthode de calcul de l'inverse de l'espérance mathématique des erreurs absolues et la méthode d'optimisation par les moindres carrés non-linéaires sont représentées dans les figures suivantes. Ces représentations permettent une visualisation de l'amélioration apportée lorsqu'on utilise la méthode nls pour l'estimation du paramètre de la loi exponentielle.

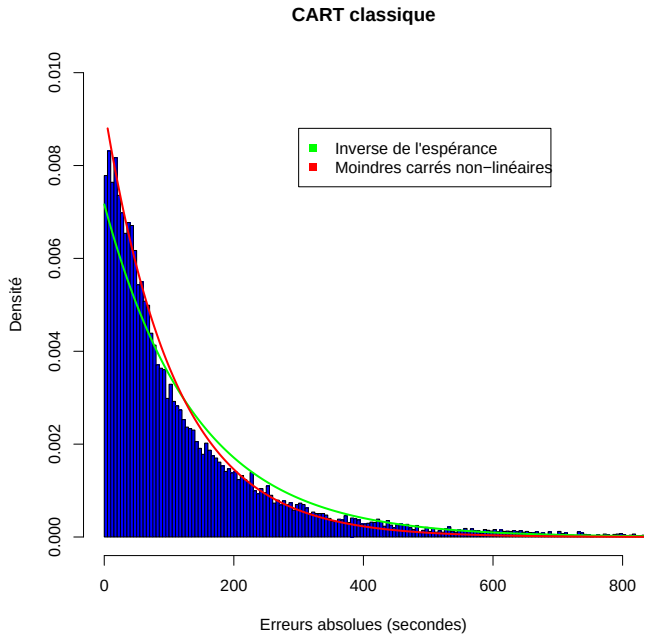

FIG. 8.8 – Ajustement des erreurs absolues du modèle CART classique par une loi de forme exponentielle

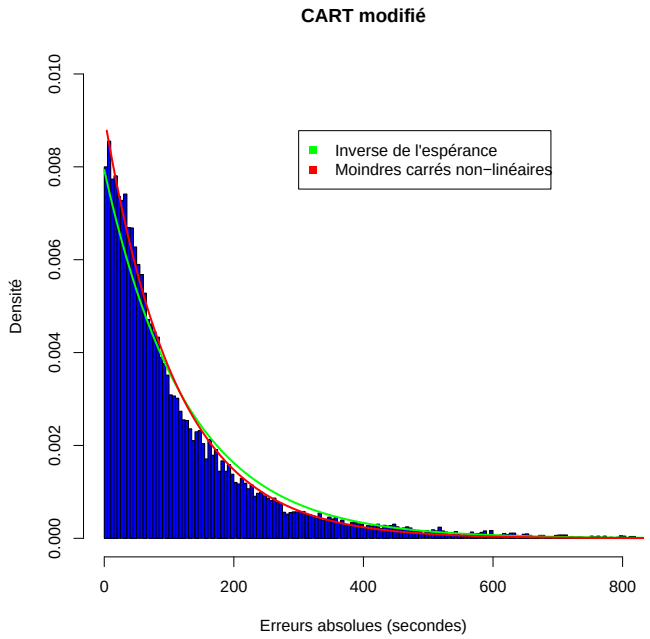

FIG. 8.9 – Ajustement des erreurs absolues du modèle CART modifié par une loi de forme exponentielle

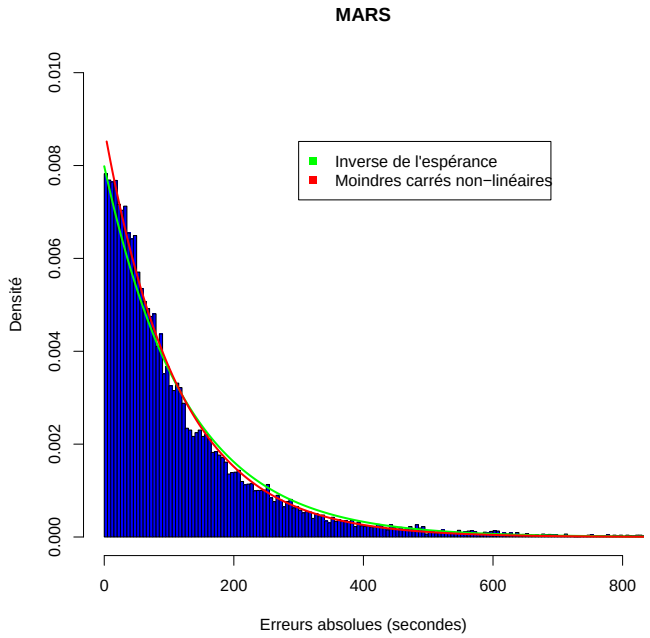

FIG. 8.10 – Ajustement des erreurs absolues du modèle MARS par une loi de forme exponentielle

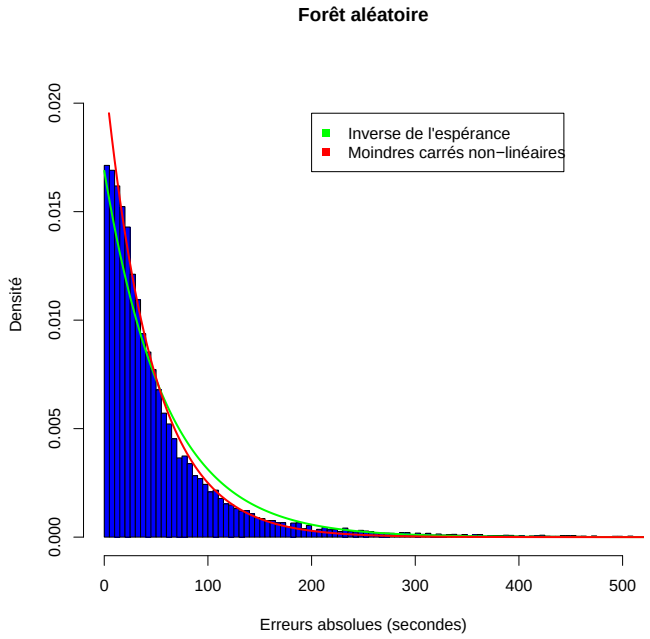

FIG. 8.11 – Ajustement des erreurs absolues du modèle des forêts aléatoires par une loi de forme exponentielle

8.3.7 Formulation du test final

Nous considérons désormais le couple (λ_0, α) qui définit les objectifs visés pour une meilleure circulation du trafic. Ainsi, sur une fenêtre de l'horizon temporel de prévision, le trafic sera dit prédictible avec

un modèle, si en l'utilisant, l'erreur absolue moyenne de la prévision de la variable *écart temporel*, notée λ sur les données réelles du trafic est inférieure au seuil λ_0 avec un risque α. Dans cette étude, nous réalisons les tests avec un niveau de confiance de 95% (un risque de 5%). Ainsi, dans chacun des 12 intervalles de 5 minutes couvrant un horizon de prévision de 60 minutes, nous allons procéder aux tests d'hypothèses suivantes :

$$\mathbf{H}_0 : \lambda \leq \lambda_0 \text{ contre } \mathbf{H}_1 : \lambda > \lambda_0$$

D'après le théorème de Lehmann, il existe un test UPP où la région critique W est l'ensemble des points $(x_1, \ldots, x_n)$ tels que $s(x_1, \ldots, x_n) > K_\alpha$ où $S = s(x_1, \ldots, x_n)$ est une statistique exhaustive du paramètre λ et K_α est la statistique critique déterminée par le niveau de test.

8.3.8 Construction de la statistique critique K_α

La distribution empirique de l'erreur absolue suit une loi de la forme exponentielle de paramètre $\frac{1}{\lambda}$. Si $\lambda_1 < \lambda_2$, le rapport des vraisemblances s'écrit après réduction :

$$RV(s) = \left[\frac{\frac{1}{\lambda_2}}{\frac{1}{\lambda_1}}\right]^n exp\left\{\left[-\frac{1}{\lambda_2} + \frac{1}{\lambda_1}\right] s\right\}, \tag{9}$$

où $s = s(x_1, ..., x_n) = \sum_{i=1}^n x_i$ et $0 < -\frac{1}{\lambda_2} + \frac{1}{\lambda_1}$, donc la vraisemblance est une fonction croissante de s. En appliquant le théorème de la vraisemblance monotone croissante de Lehmann, la région de rejet de l'hypothèse nulle $\mathbf{H}_0 : \lambda \leq \lambda_0$ est de la forme

$$\left\{\sum_{i=1}^n \mathbf{X}_i > \mathbf{K}_\alpha\right\}. \tag{10}$$

La statistique pivotale[10] $\frac{2}{\lambda}\sum_{i=1}^n \mathbf{X}_i$ suit une loi du khi-deux à $2n$ degrés de liberté $\mathcal{X}^2_{(2n)}$. En effet, on utilise le fait que la loi exponentielle appartient à la famille des lois gamma : X suit $\exp(\frac{1}{\lambda}) \equiv \gamma(1, \frac{1}{\lambda})$. En utilisant la propriété d'addition de lois gamma, on conclut que $S = \sum_{i=1}^n \mathbf{X}_i$ suit la loi gamma $\gamma(n, \frac{1}{\lambda})$. Donc, $\frac{2S}{\lambda}$ suit la loi $\gamma(n, 2) = \gamma(\frac{2n}{2}, 2)$ qui est un khi-deux à $2n$ degrés de liberté $\chi^2(2n)$ (voir propriétés en annexe B). Pour appliquer cette propriété, nous supposons que $2n$ est supérieur à 30. Le niveau de test doit vérifier la relation suivante (Théorème de Lehmann) :

[10] Une fonction pivotale pour le paramètre λ est une variable aléatoire $T_n = \phi(X_i, ..., X_n, \lambda)$ dépendante de λ et dont la loi est connue et indépendante de λ.

$$\begin{aligned}
\alpha &= \sup_{\lambda \leq \lambda_0} \mathbf{P}_\lambda(\sum_{i=1}^{n} \mathbf{X}_i > \mathbf{K}_\alpha) \\
&= \mathbf{P}_{\lambda_0}(\sum_{i=1}^{n} \mathbf{X}_i > \mathbf{K}_\alpha) \\
&= 1 - \mathbf{P}_{\lambda_0}(\sum_{i=1}^{n} \mathbf{X}_i \leq \mathbf{K}_\alpha) \\
&= 1 - \mathbf{F}_{\mathcal{X}^2_{(2n)}}(2\frac{1}{\lambda_0}\mathbf{K}_\alpha)
\end{aligned}$$

où $\mathbf{F}$ est la fonction de répartition de la loi du khi-deux $\mathcal{X}^2_{(2n)}$. Par suite, on a :

$$K_\alpha = \frac{\lambda_0}{2}\mathbf{F}^{-1}_{\mathcal{X}^2_{(2n)}}(1-\alpha).$$

Rappelons que pour un échantillon de taille $\nu > 30$, le fractile d'ordre p de la loi du khi-deux à ν degré de liberté est donné par l'approximation suivante :

$$\mathcal{X}^2_{(p)}(\nu) = \nu\left(1 - \frac{2}{9\nu} + u_p\sqrt{\frac{2}{9\nu}}\right)^3,$$

où u_p est le fractile d'ordre p de la loi normale centrée réduite. Ainsi, en posant $\nu = 2n$, la valeur de la statistique critique est :

$$\mathbf{K}_\alpha = n\lambda_0\left(1 - \frac{1}{9n} + u_{(1-\alpha)}\frac{1}{3\sqrt{n}}\right)^3.$$

8.3.9 Région de prédictivité des modèles

Une méthode généralement utilisée pour l'étude de la capacité prédictive des modèles de prévision consiste à comparer la RMSEP[11] à l'écart-type des mesures observées de la variable à expliquer sur les données test ou de validation. Wallach et Goffinet (1987)[61] recommandent l'utilisation du RMSEP en le comparant à un seuil raisonnable. Cependant, il n'est pas aisé de juger un modèle sur la simple valeur de sa RMSEP, car on ne dispose généralement pas de référence de valeur en deçà de laquelle la valeur de cet indicateur est jugée faible, et donc la qualité prédictive jugée bonne.

Dans le cadre de cette étude, la capacité prédictive des modèles développés est établie en s'appuyant sur les tests d'hypothèses basés sur le théorème du rapport des vraisemblances monotone. Ainsi, nous appliquons ce théorème pour construire la région de prédictivité de chaque modèle avec un risque de niveau 5%. Pour chaque $\lambda_0 \in \{30, 40, 50, 60, 70, 80, 90, 95, 100, 105, 110, 115, 120\}$ en secondes et chaque modèle

[11] Root Mean Squared Error of Prediction.

donné, on teste sur chaque fenêtre de prévision de longueur 5 minutes si l'erreur absolue moyenne que nous notons λ est inférieure à λ_0 avec le risque α. λ_0 est considéré comme une borne supérieure de l'incertitude au dela duquel le trafic n'est pas prédictible pour l'horizon de prévision correspondant à la borne supérieure de la fenêtre considérée. Rappelons que l'erreur absolue moyenne de la prévision de l'*écart temporel* par les modèles respectifs est fournie par TAB.8.4 :

	CART classique	**CART modifié**	**MARS**	**Forêt aléatoire**
Erreur absolue moyenne (seconde)	139	126	125	59

TAB. 8.4 – Erreur absolue moyenne des modèles sur les données d'apprentissage

Un examen des résultats des tests d'hypothèses du tableau TAB.8.5 montre qu'aucun modèle ne permet de prévoir le temps de passage des aéronefs avec une incertitude inférieure à 30 secondes sur un horizon de 300 secondes avec un niveau de confiance de 95%. L'erreur absolue moyenne la plus faible est fournie par le modèle des forêts aléatoires. Elle se situe autour de 60 secondes. Pour $\lambda_0 = 60$ secondes, l'horizon de prévision le plus élevé est fourni par le modèle des forêts aléatoires avec une profondeur de 900 secondes. En revanche, cet horizon est estimé à 600 secondes pour les modèles CART classique et CART modifié et à 300 secondes pour le modèle MARS. Finalement, en donnant différents seuils d'incertitude à λ_0 on obtient pour chaque modèle la borne supérieure de l'horizon de prédictibilité du trafic. Plus loin, nous nous intéresserons à l'application des modèles en utilisant d'horizon de prévision et la largeur de la fenêtre de prévision pour le calcul de la charge d'un secteur de l'espace. Le paragraphe suivant propose une estimation de la distribution de probabilité des résidus des modèles développés.

λ_0 **(Secondes)**	30	40	50	60	70	80	90	95	100	105	110	115	120
CART classique		300	300	600	900	1200	1200	1200	1500	1500	1800	1800	1800
CART modifié			300	600	900	1200	1500	1500	1800	1800	1800	2100	2100
MARS				300	900	1200	1500	1500	1800	1800	2100	2100	2100
Forêts aléatoires		300	600	900	1200	1500	1800	1800	2100	2100	2100	2400	2400

TAB. 8.5 – L'horizon de prévision en fonction de la limite de l'erreur de prévision. La valeur de chaque cellule du tableau représente l'horizon de prévision en secondes avec un niveau de confiance de 95%. Le « vide » dans une cellule signifie que le modèle correspondant ne permet pas de réaliser les prévisions du temps de passage des aéronefs avec une erreur absolue moyenne inférieure au seuil d'incertitude λ_0 avec le niveau de confiance de 95%.

8.4 Ajustement des résidus par un mélange de gaussiens

Nous avons vu dans les chapitres précédents qu'il était difficile d'ajuster la distribution de probabilité des résidus des modèles par une seule loi normale. Pour cela, nous avons utilisé l'algorithme EM (Expectation Maximization) implanté dans le logiciel $mixmod$ sous l'interface $Matlab$ pour décomposer la fonction de densité des résidus de chaque modèle en une somme stochastique de deux fonctions de densité toutes normales. Ainsi, si f est la fonction de densité des résidus, alors elle s'écrit sous la forme :

$$f(x) = \alpha_1 * f_1(x) + \alpha_2 * f_2(x), \tag{11}$$

où f_i est une fonction de densité de loi normale $\mathcal{N}(\mu_i, \sigma_i)$ pour $i = 1, 2$. Pour chaque modèle proposé, l'estimation de son mélange de loi gaussienne est illustrée par l'une des figures suivantes :

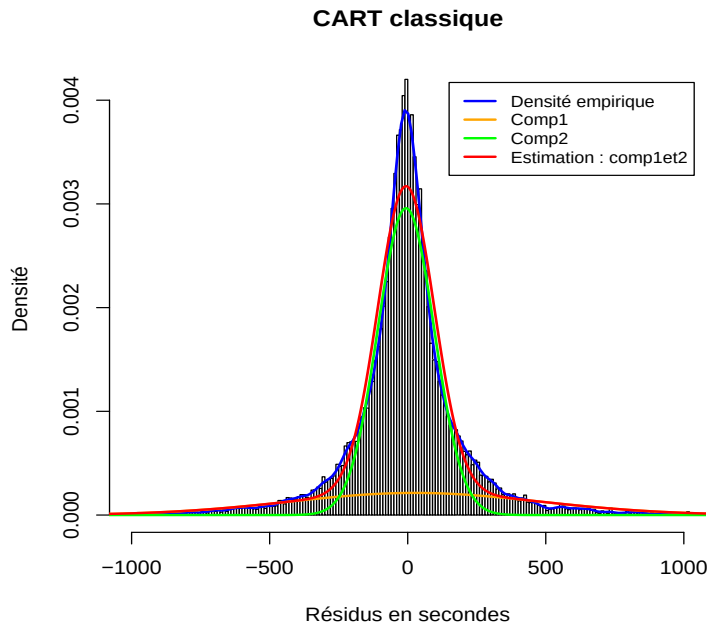

FIG. 8.12 – Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle CART classique. L'estimation du mélange de lois est égale à : $0.14 * \mathcal{N}(16.69, 455.4) + 0.86 * \mathcal{N}(-5.62, 102.64)$.

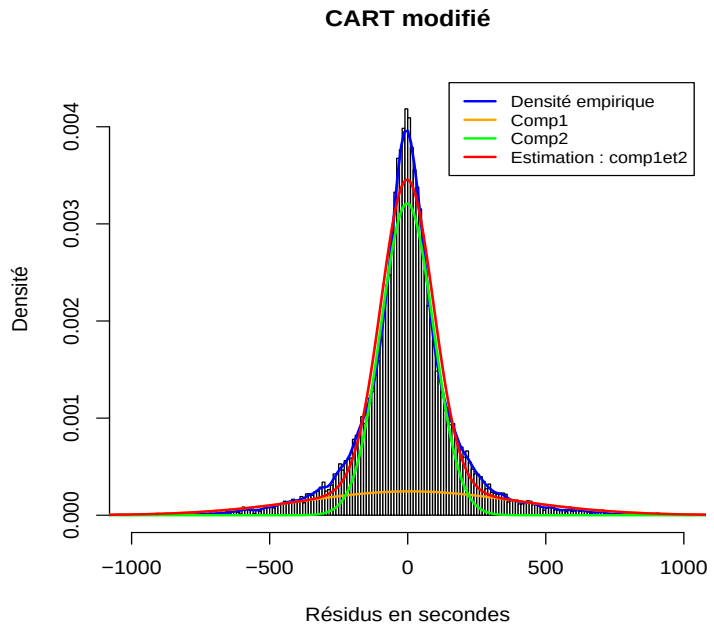

FIG. 8.13 – Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle CART modifié. L'estimation du mélange de lois est égale à : $0.23 * \mathcal{N}(9.84, 390.6) + 0.77 * \mathcal{N}(-2.49, 97.82)$.

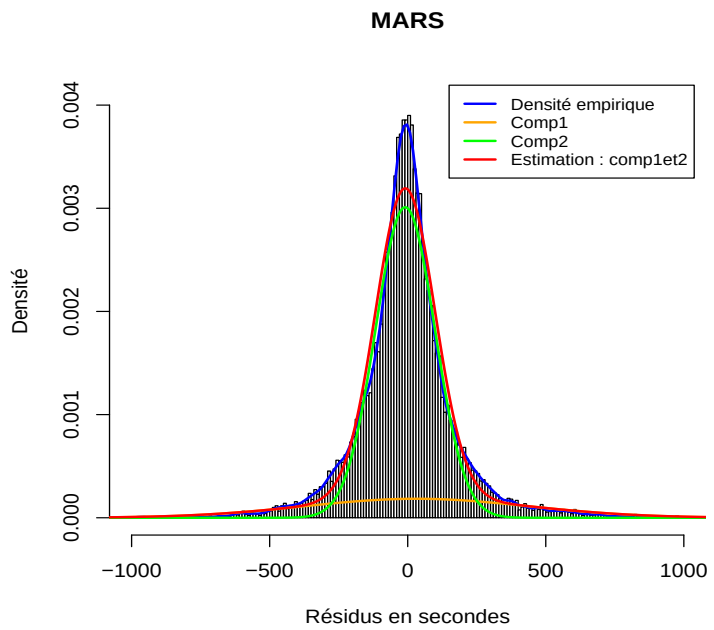

FIG. 8.14 – Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle MARS. L'estimation du mélange de lois est égale à : $0.18 * \mathcal{N}(40.95, 411.3) + 0.82 * \mathcal{N}(-9.17, 108.72)$.

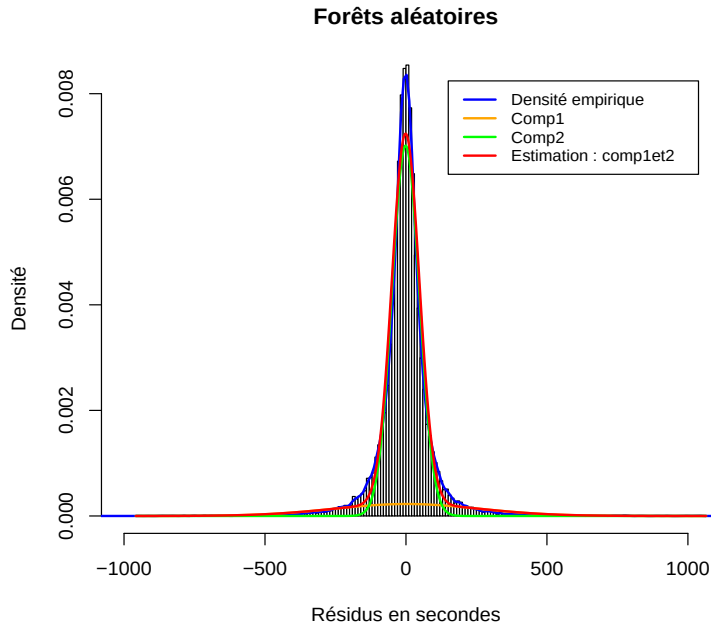

FIG. 8.15 – Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle des forêts aléatoires. L'estimation du mélange de lois est égale à : $0.15 * \mathcal{N}(0.51, 256.63) + 0.85 * \mathcal{N}(-1.52, 48.57)$.

8.5 Application des modèles à la prévision de la charge d'un secteur

Nous avons vu au chapitre 2 que le calcul de la charge d'un secteur s pendant une période de temps définie par l'intervalle $[T_d,\ T_f]$ revenait à trouver une estimation de chacun des termes $\mathbf{N}_{\mathcal{P}_s}(T_d, T_f)$ et $\mathbf{N}_{\mathcal{P}_s^c}(T_d, T_f)$ en fonction des probabilités :
(1) $\delta_{i,1}$,
(2) $\delta_{i,0}$, et
(3) $\mathbb{P}\left(T_d - \phi_{si} - t_{si}^p < \epsilon_{si} < T_f - \phi_{si} - t_{si}^p\right)$

A partir des données d'archive pour le mois de septembre 2007, nous avons estimé les valeurs moyennes des probabilités (1) et (2) pour l'ensemble des secteurs de l'espace : Ainsi $\delta_{i,1} \approx 0.86$ et $\delta_{i,0} \approx 0.14$. Notons par ailleurs que ces valeurs cachent une importante disparité entre les secteurs. Par exemple, pour le secteur LFFFSUP sur lequel nous évaluons la capacité des modèles développés à prévoir la charge, $\delta_{i,1} \approx 0.97$ tandis que $\delta_{i,0} \approx 0.03$[12]. Ainsi la détermination des prévisions des instants d'entrée des aéronefs dans un secteur par chaque modèle permet de calculer complètement la probabilité (3) ci-dessus. Donc, d'en déduire une estimation de $\mathbf{N}_{\mathcal{P}_s}(T_d, T_f)$, la part du nombre d'avions entrant dans le secteur dûe aux aéronefs qui avaient prévu dans leur plan de vol d'entrer dans ce secteur. Puisqu'il est impossible de prévoir les instants d'entrée des aéronefs dans un secteur n'appartenant pas à leur trajectoire de vol prévue, nous propo-

[12]Ce grand écart par rapport à la moyenne pour l'ensemble des secteurs de l'espace est sans doute expliquée par l'étendue du secteur choisi.

sons à partir des données d'archive un *coefficient de correction* qui permet d'estimer $\mathbf{N}_{\mathcal{P}_s^c}(T_d, T_f)$ à partir $\mathbf{N}_{\mathcal{P}_s}(T_d, T_f)$ pour une fenêtre de temps $[T_d, T_f]$ bien définie.

Si pour un secteur s et pour un créneau de temps $[T_d, T_f]$, on pose :

- $\mu^s_{1,1}(T_d, T_f)$: la proportion moyenne d'aéronefs entrants dans le secteur s de leur trajectoire prévue entre les instants T_d et T_f.
- $\mu^s_{1,0}(T_d, T_f)$: la proportion moyenne d'aéronefs entrants dans le secteur s entre les instants T_d et T_f et dont ce secteur n'est pas sur leur trajectoire de vol prévue.

Pour chaque modèle et chaque secteur s, nous estimons la part de la charge de ce secteur dûe aux aéronefs qui y entrent alors que s n'appartient pas à leur trajectoire prévue par la relation :

$$\mathbf{N}_{\mathcal{P}_s^c}(T_d, T_f) = \mathbf{N}_{\mathcal{P}_s}(T_d, T_f) * \frac{\mu^s_{1,0}(T_d, T_f)}{\mu^s_{1,1}(T_d, T_f)}. \tag{12}$$

Finalement, on obtient :

$$\mathbf{N}_s(T_d, T_f) = \mathbf{N}_{\mathcal{P}_s}(T_d, T_f) * \left(1 + \frac{\mu^s_{1,0}(T_d, T_f)}{\mu^s_{1,1}(T_d, T_f)}\right), \tag{13}$$

où la quantité

$$\frac{\mu^s_{1,0}(T_d, T_f)}{\mu^s_{1,1}(T_d, T_f)} \tag{14}$$

est appelé *coefficient de correction* pour la prévision de la part de la charge secteur dûe aux entrées des avions dans les secteurs non prévus de leur trajectoire de vol prévue. Nous pouvons maintenant utiliser les modèles pour la prévision de la charge secteur.

8.5.1 Prévision de la charge du secteur LFFFSUP : journée du 14 septembre 2007 à 7h 00

LFFFSUP est un regroupement de tous les secteurs de l'espace aérien supérieur Français. Même si un tel secteur n'est pas utilisé pour l'écoulement du trafic en période chargée, nous l'avons choisi parce qu'il permet d'avoir suffisamment de vols pour l'évaluation des modèles.

Dans cet exemple d'application, nous nous plaçons à l'instant courant de prévision fixé à t_0=7 h 00 le matin de la journée du 14 septembre 2007. A cet instant précis, on observe 95 aéronefs en vol qui vont entrer dans le secteur LFFFSUP. Notre objectif est de nous placer à l'horizon $t_o + h$ ($h > 0$) et de prévoir le nombre d'aéronefs qui entreront dans ce secteur entre les instants « $t_0 + h$ » et « $t_0 + h + dw$ », où dw est la largeur de la fenêtre de prévision. Nous prenons $h \in \{10, 15, 20, 25, 30, 35, 40, 60\}$ en minutes et $dw \in \{5, 10, 15, 20, 30, 40, 60\}$ en minutes. Nous utilisons dans les modèles la relation (13). Le nombre d'aéronefs entrants prévu par chaque modèle est comparé à celui effectivement observé pendant le vol réel. L'indicateur de comparaison utilisé est l'erreur relative de la prévision par le modèle par rapport au nombre d'aéronefs entrants effectivement observés. La synthèse des résultats est présentée par les figures

(FIG.8.16, FIG.8.17, FIG.8.18, FIG.8.19). Pour chaque horizon, la figure de gauche représente l'erreur relative de chaque modèle en fonction de la largeur de la fenêtre de prévision ((a), FIG.8.17 pour l'horizon 7 h 10 min) et la figure de droite représente le nombre d'aéronefs prévus à l'entrée dans le secteur LFFFSUP en fonction de la largeur de la fenêtre de prévision ((a'), FIG.8.17 pour l'horizon 7 h 10 min). Cette dernière illustre également le nombre d'aéronefs effectivement observés à l'entrée de ce secteur.

Une lecture de ces figures montre que la meilleure prévision de la charge du secteur LFFFSUP est obtenue pour un horizon de prévision d'environ 20 minutes. Pour un horizon de prévision inférieur (FIG.8.16) ou égal à cette valeur, nous voyons que la qualité des prévisions fournies par les différents modèles dépend aussi de la largeur de la fenêtre de prévision. Dans notre exemple, nous observons qu'une fenêtre de largeur inférieure à 15 minutes fournit des mauvaises prévisions de charge avec une erreur relative supérieure à 10%. En revanche, pour des fenêtres de prévision de largeur supérieure à 20 minutes, les prévisions de la charge de ce secteur sont meilleures avec une erreur relative inférieure à 10%. Nous observons que pour un horizon de prévision inférieur à 20 minutes et une fenêtre de largeur supérieure à 20 minutes, les modèles CART classique et les forêts aléatoires fournissent de meilleures prévisions.

Pour les horizons supérieurs à 30 minutes, les meilleures prévisions de charge sont obtenues pour des fenêtres de prévision de largeur supérieure à 40 minutes. Nous notons que la baisse de l'effectif d'aéronefs pour les horizons élevés est expliquée par le fait que les modèles que nous proposons ne prennent pas encore en compte, les aéronefs qui dans leur plan de vol ont prévu de traverser le secteur LFFFSUP mais qui n'ont pas encore décollé de leur aéroport respectif à l'instant courant de prévision t_o = 7 h 00 et donc ne sont pas visibles pour le calcul de la charge future du secteur. Ainsi, les prévisions les plus réalistes dans cette étude sont celles pour lesquelles l'horizon de prévision se situe autour de 20 minutes. Ce seuil est choisi parce que c'est à partir de cet horizon que nous observons une baisse significative du nombre d'aéronefs dans le secteur (voir figures (a'et b', FIG.8.16) et (c' et d', FIG.8.17).

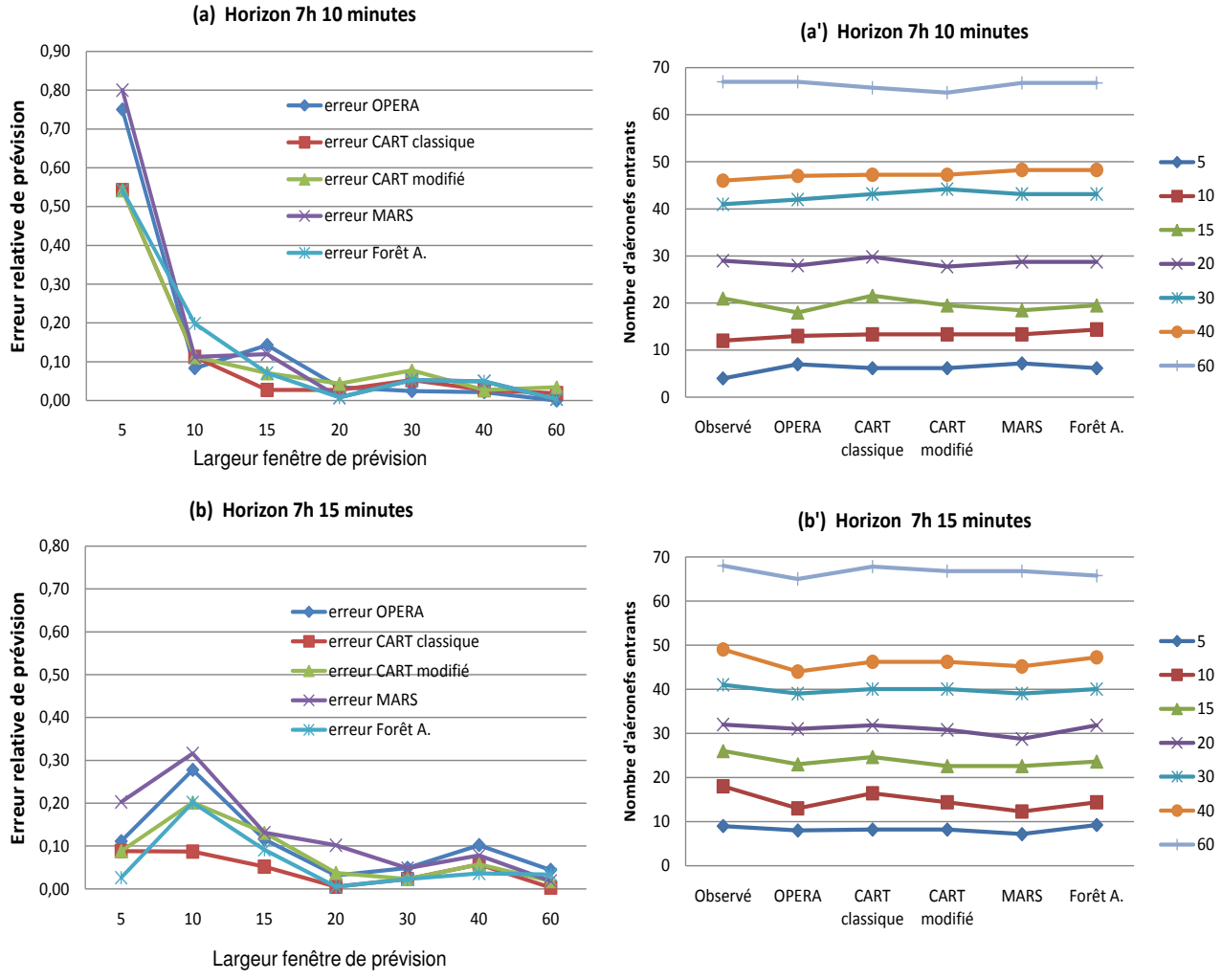

FIG. 8.16 – Prévisions pour les horizons : 7 h 10 minutes et 7 h 15 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision.

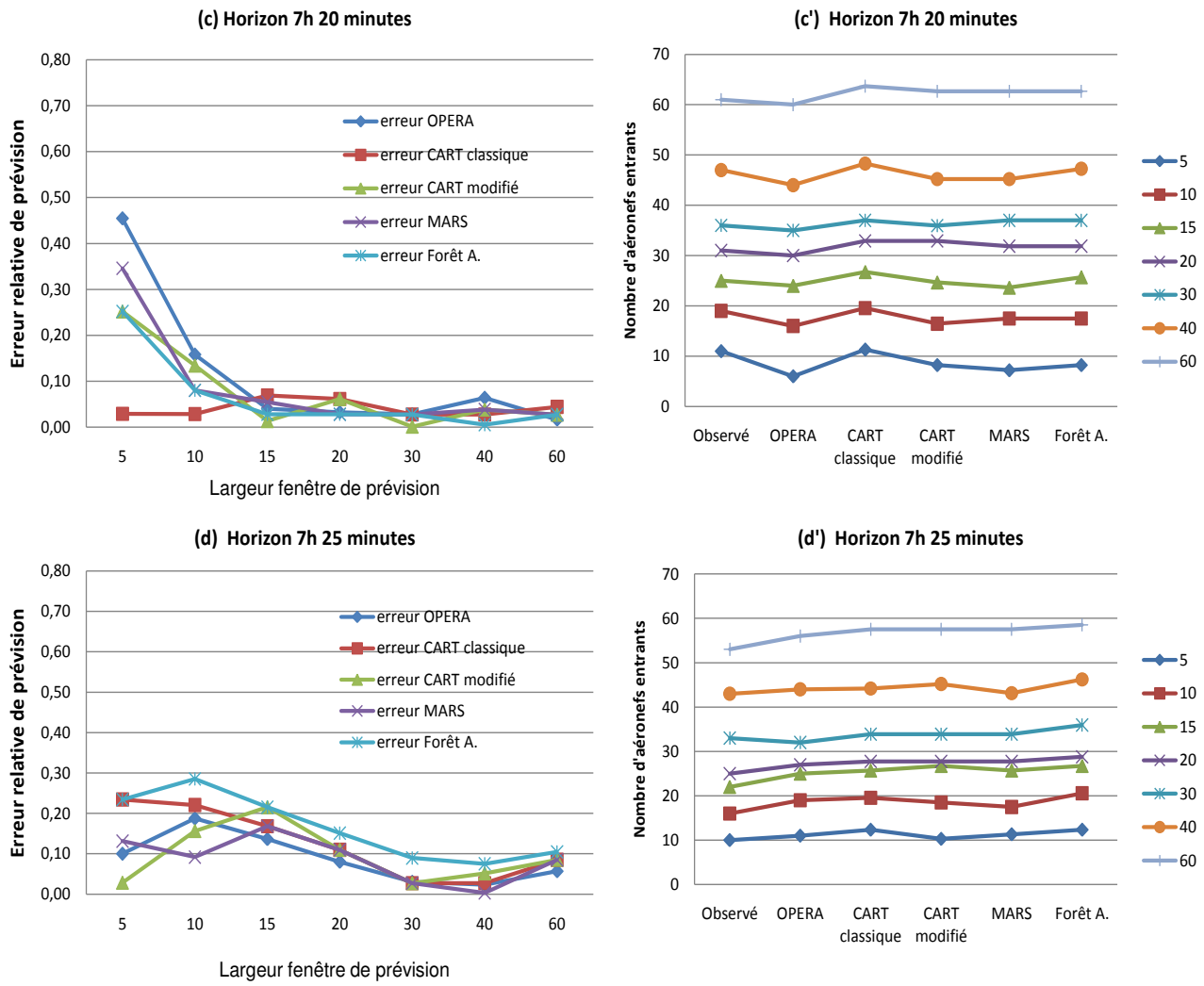

FIG. 8.17 – Prévisions pour les horizons : 7 h 20 minutes et 7 h 25 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision.

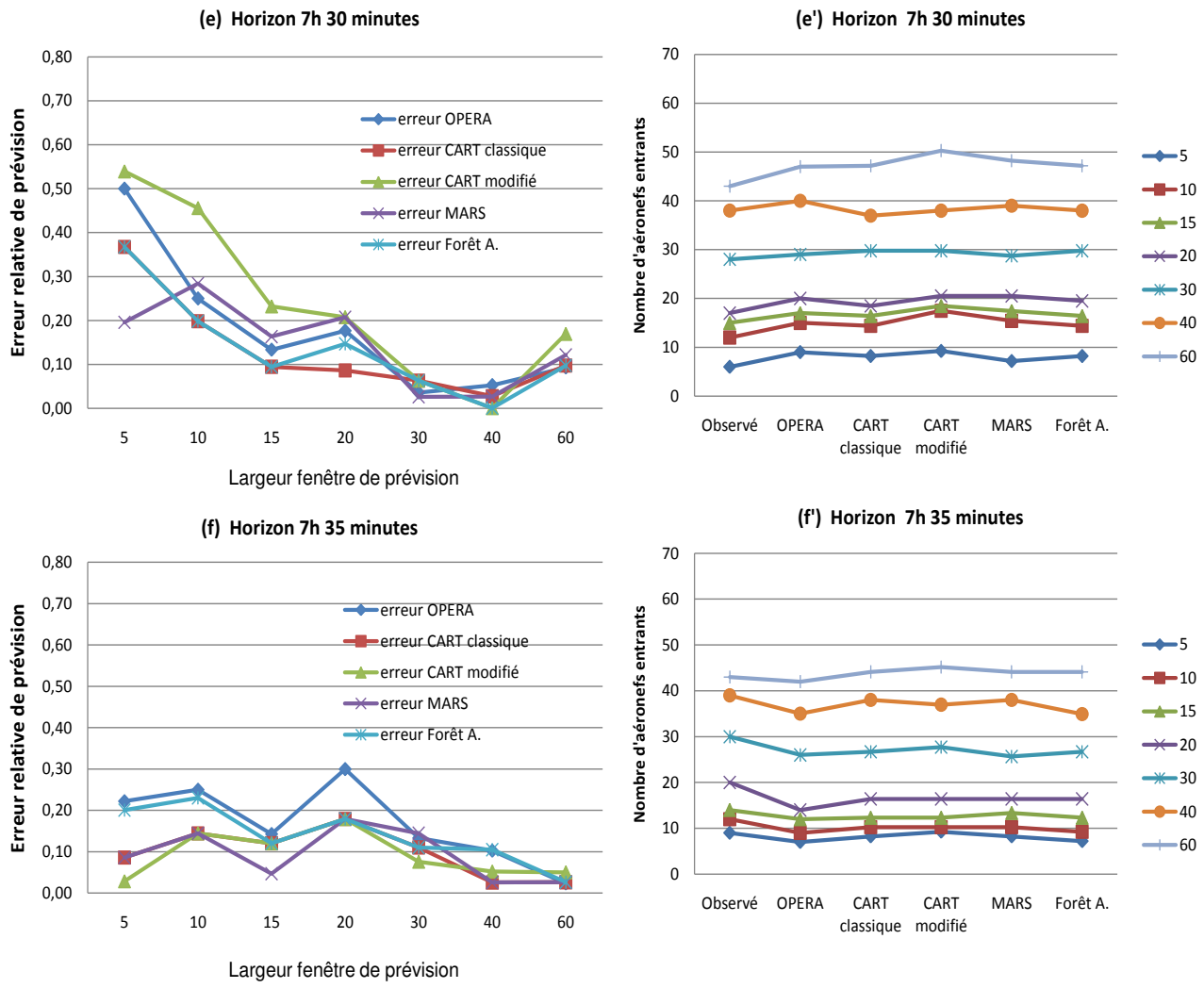

FIG. 8.18 – Prévisions pour les horizons : 7 h 30 minutes et 7 h 35 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision.

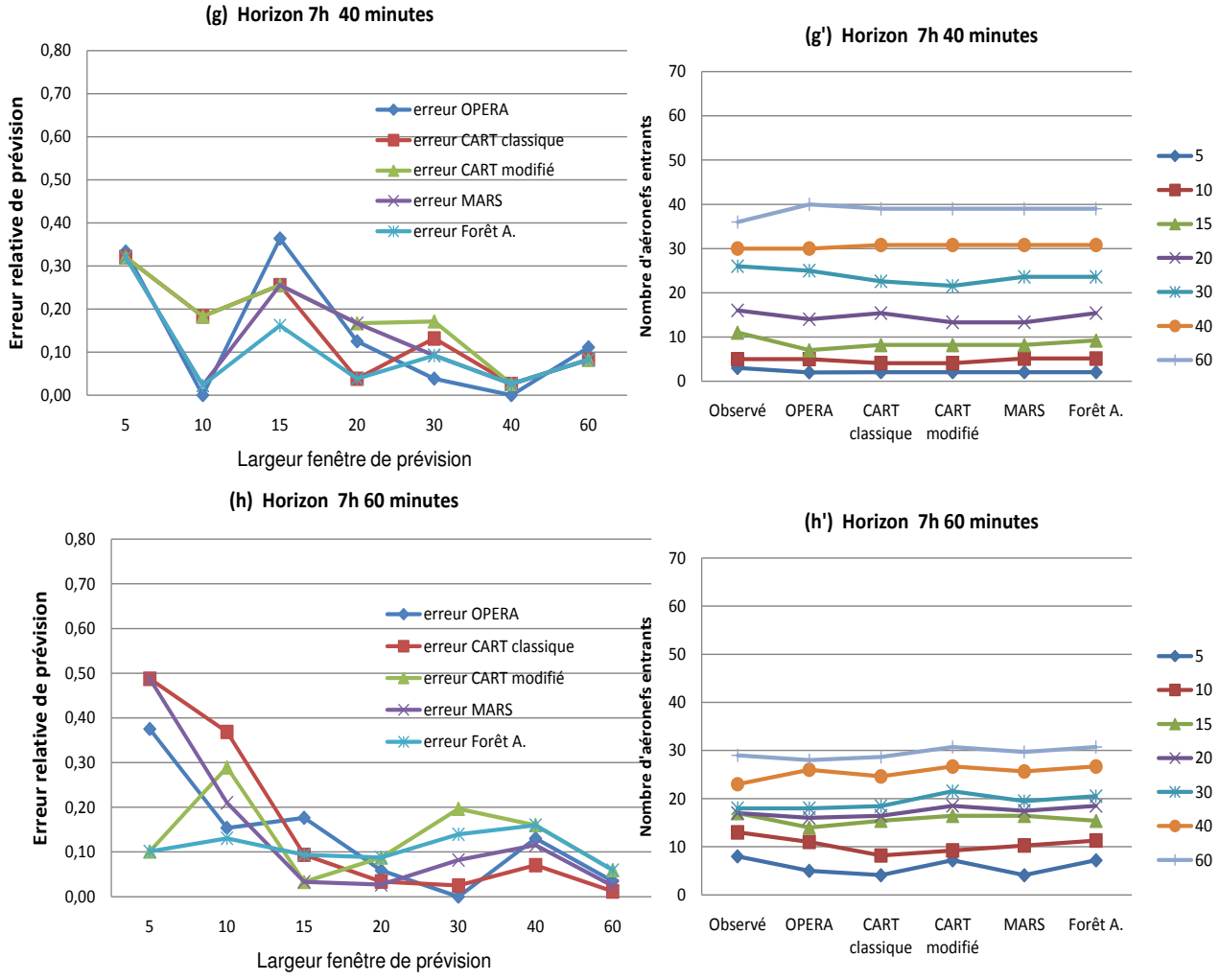

FIG. 8.19 – Prévisions pour les horizons : 7 h 40 minutes et 7 h 60 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision.

8.6 Conclusion

A travers ce chapitre, nous avons évalué les modèles développés par rapport à leur capacité à se comporter vis-à-vis des données nouvelles. Au point de vue erreur de prévision de la variable *écart temporel*, les forêts aléatoires se présentent comme le meilleur modèle, suivi du modèle MARS lorsque l'horizon de prévision est éloigné (FIG.8.5).

Par ailleurs, nous avons évalué la capacité des modèles à la prévision de la charge d'un regroupement de secteurs de l'espace aérien supérieur Européen, notamment la zone LFFFSUP. Il en est ressorti que la prévision de la charge d'un secteur en nombre d'aéronefs entrants est fortement dépendante à la fois de l'horizon temporel de prévision et de la largeur de la fenêtre de prévision. Ainsi, nous avons trouvé que pour l'instant courant de prévision t_o=7 h 00 et un horizon de prévision qui se situe autour de 20 minutes, les meilleures prévisions de la charge secteur en nombre d'aéronefs entrants est obtenue pour une fenêtre de prévision de largeur supérieure à 20 minutes. Ces modèles sont donc utilisables pour les fenêtres de largeurs 30 et 60 minutes qui sont des limites souvent utilisées par la CFMU pour définir la capacité des secteurs de l'espace aérien pour un bon écoulement de trafic. Les modèles des forêts aléatoires et CART classique semblent fournir de meilleures prévisions dans ces conditions.

Même si pour un horizon et une fenêtre de prévision convenablement choisis, les modèles fournissent de meilleures prévisions, nous notons cependant que ces modèles méritent d'être affinés de façon à prendre en compte dans le calcul de la charge du secteur, la part de celle-ci dûe aux aéronefs qui ont prévu de traverser le secteur mais qui, à l'instant courant de prévision ne sont pas encore visibles dans l'espace, car ils n'ont pas encore décollé. Dans le chapitre suivant, nous présentons la conclusion générale à cette thèse avec une synthèse des résultats obtenus, des limites à notre travail et des perspectives envisagées.

Conclusion générale

Dans cette thèse, nous avions pour ambition de proposer un outil d'aide à la régulation et à la planification dynamique du trafic aérien au moyen de la modélisation en s'appuyant sur les méthodes probabilistes et statistiques d'analyses de données multidimensionnelles issues de l'environnement opérationnel des vols. Il s'agissait d'utiliser l'ensemble des informations disponibles à un instant t_o sur les aéronefs en vol pour prévoir les instants de passage de ceux-ci sur les points futurs de leur trajectoire de vol prévue. Pour cela, nous avons développé quatre modèles non-paramétriques basés tous sur le partitionnement récursif de l'échantillon des données. La conjonction des spécificités propres à chacun de ces modèles nous a permis d'aborder notre étude sur les aspects à la fois descriptif et explicatif (ou prédictif). Pour l'évaluation et la validation de ces modèles, nous avons utilisé des données test, c'est-à-dire, celles n'ayant pas été utilisées dans la phase de construction des modèles. Cette approche nous a conduit non seulement à construire des modèles pour la prévision des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue, mais elle a également permis de prévoir la charge future d'un secteur de l'espace à horizon de quelques minutes et sur une fenêtre de temps de largeur définie. Pour terminer, nous rappelons les principaux résultats obtenus, les apports de notre travail, ses limites et les perspectives envisagées.

Résultats

Au cours de cette thèse, nous avons élaboré à partir de la méthode CART et des forêts aléatoires une hiérarchie des variables explicatives de la variable dépendante *écart temporel*, donc, de l'incertitude sur le temps de vol des avions pendant leur progression sur leur trajectoire. Elle montre que parmi les facteurs susceptibles d'influer l'instant de passage des aéronefs sur les points de leur trajectoire, la distance entre le point courant et le point de prévision est le premier facteur important d'incertitude. En effet, plus un point de la trajectoire est éloigné du point courant, plus l'incertitude absolue de prévision du temps de passage de l'aéronef sur ce point est importante. Ensuite, les conditions météorologiques par l'intermédiaire de la variable influence du vent est le deuxième facteur important qui affecte la progression des aéronefs sur leur trajectoire, donc l'incertitude de prévision du temps de vol. Le niveau de vol demandé pour la phase de croisière, la vitesse courante du vol et le type d'avion utilisé sont respectivement troisième, quatrième et cinquième facteurs de cette incertitude.

Le modèle développé à partir de la méthode MARS a permis d'isoler les effets directs de ces facteurs

significatifs et de ceux de leurs interactions sur l'écart temporel de passage des aéronefs sur les points de leur trajectoire de vol. Ainsi, la conjonction des conditions météorologiques et de la position des aéronefs par rapport aux points de leur trajectoire de vol montre que les aéronefs ne sont retardés de façon significative sur leur trajectoire de vol que s'ils ont été soumis au vent de face dominant sur une longue distance de vol. Dans ces conditions, l'incertitude sur la prévision du temps de passage en un point est élevée. Par ailleurs, la vitesse courante de vol affecte l'incertitude sur la trajectoire temporelle des aéronefs lorsqu'elle est très faible ou lorsqu'elle est très élevée. L'interaction entre la vitesse de vol et le niveau de croisière prévu montre que l'incertitude de prévision du temps de passage des aéronefs sur les points de leur trajectoire est plus importante pour les aéronefs qui sont en phase de montée ou de descente à l'instant courant de prévision. En revanche, elle est beaucoup plus faible pour les aéronefs qui à l'instant courant de prévision ont déjà atteint leur phase de croisière prévue. Outre ces résultats descriptifs des modèles développés, nous avons également évalué et validé leur pouvoir prédictif.

L'évaluation et la validation des modèles proposés ont constitué une étape cruciale de cette thèse. Dans cette phase, nous avons trouvé que la RMSEP de chaque modèle sur les données test vérifiait bien la condition de bonne performance prédictive. Ainsi, les modèles CART classique, CART modifié, MARS et les forêts aléatoires sont globalement bons pour la prévision de l'écart temporel, donc de celle des instants de passage des aéronefs sur les points de leur trajectoire de vol prévue. Sur les données test, nous avons calculé l'indicateur médian RMedSEP sur les quatre modèles. Il s'agit de la racine carrée de la médiane des carrés des erreurs de prévision. Il en est ressorti que le modèle des forêts aléatoires est le plus prédictif et permet aussi de construire l'horizon temporel de prévision le plus profond.

Nous avons aussi évalué les modèles sur leur capacité à prévoir la charge des secteurs en nombre d'aéronefs entrants à un horizon de temps et une fenêtre de largeur donnée. Ce test d'évaluation s'est focalisé sur un regroupement de secteurs de l'espace aérien supérieur Européen, LFFFSUP. Ainsi, pour un horizon de prévision d'environ 20 minutes et pour une fenêtre de temps de largeur supérieure à 20 minutes, tous les modèles fournissent des prévisions de charge secteur avec une erreur relative inférieure à 10%.

Apports

Grâce à l'utilisation de méthodes récentes de modélisation non-paramétrique, les apports de cette thèse sont nombreux : Sous un angle prédictif, la validation des modèles sur les données tests a montré que les modèles que nous proposons sont capables de prévoir les instants de passage des aéronefs sur les points de leur trajectoire de vol prévue en fonction des caractéristiques courantes des vols, des paramètres d'archives des plans de vols, de l'environnement météorologique et atmosphérique. Il s'agit d'une démarche nouvelle, basée sur les modèles probabilistes et statistiques d'analyses des données multidimensionnelles. Ces modèles ont la particularité d'être construits à partir des données réelles de trafic. Ce qui n'est pas toujours le cas dans un domaine où la plupart des modèles sont construits à partir des simulations grande échelle. Au delà de l'aspect qualitatif, les principaux résultats présentés dans ce mémoire de thèse sont quantitatifs et permettent d'avoir une idée sur quel horizon temporel de prévision les modèles proposés sont efficaces. Ces

modèles de prévision du temps de passage des aéronefs sur les points de leur trajectoire sont un outil d'aide à la gestion dynamique du trafic, car ils permettent aussi de prévoir la charge des secteurs et la quantité de conflits entre les aéronefs, donc d'anticiper sur le problème de saturation de l'espace aérien contrôlé.

Limites

Si les résultats obtenus dans cette thèse montrent que les modèles proposés permettent de bien prévoir les instants de passage des aéronefs en des points de leur trajectoire prévue, nous pouvons néanmoins relever quelques limites : Notre approche suppose que les aéronefs suivent les trajectoires spatiales $3D$ (latitude, longitude, altitude) prévues dans leur plan de vol ou qu'ils s'en écartent très peu pendant leur déplacement. Ce qui n'est pas toujours le cas dans la mesure où pendant le vol, certains pilotes prennent des « trajectoires plus directes » dès qu'ils sont autorisés à le faire par les contrôleurs aériens. Ainsi, dans cette étude, nous n'avons developpé qu'une composante du modèle nécessaire à la prévision du temps de passage des aéronefs sur les points de leur trajectoire de vol prévue, donc, de la prévision de la charge des secteurs et de la quantité des conflits.

Perspectives

A l'issue de cette thèse, les perspectives qui en découlent sont nombreuses. Elles sont dans leur majorité liées aux travaux complémentaires nécessaires pour l'utilisabilité des modèles proposés. Les trajectoires de référence utilisées dans cette étude ont été créées par le simulateur du trafic aérien OPERA à partir de la position courante des aéronefs et en fonction des plans de vol respectifs :

- La première perspective consiste à implémenter les modèles que nous proposons dans OPERA de sorte que ce simulateur soit capable de prendre en compte dans les simulations, l'environnement des vols tel que les conditions météorologiques.
- La deuxième perspective consiste en la réalisation d'un modèle économétrique du type logistique pour rendre compte du comportement des pilotes sur leurs choix stratégiques des trajectoires en fonction de leur position courante, leurs états (mental), la configuration du trafic dans les zones traversées et de la politique des compagnies aériennes sur la gestion des retards par les pilotes pendant le vol. Autrement dit, Il s'agit de construire la composante du modèle sur le choix et la tenue des routes par les pilotes pendant les vols.
- Dans la perspective d'optimiser la prévision de la charge secteur et la quantité de conflits dans l'espace, la construction d'une composante du modèle est nécessaire pour les aéronefs qui sont encore au sol à l'instant courant de prévision et qui ont prévu de traverser le secteur d'intérêt.

Bibliographie

[1] Project Bada - EEC Technical/Scientific Report. Technical Report 2009/003, EUROCONTROL. Cité p. 38

[2] et J-P. Nakache A. Gueguen. Méthode de discrimination basée sur la construction d'un arbre de décision binaire. *Revue de Statist. Aplli*, 36 :19–38, 1988. Cité p. 48

[3] et A. Brociner A. Mathis. Retour vers le futur. une analyse rétrospective des prévisions de Mosaïque. *Revue de l'OFCE*, 49 :207–228, 1994. Cité p. 61, 75

[4] et F. Droesbeke. B. Courot. Que sais-je ? les méthodes de prévision. *Presses Universitaires de France*, 2157, 1984. Cité p. 32

[5] et B. A. Ishak B. Ghattas. Sélection de variables pour la classification binaire en grande dimension : Comparaisons et application aux données de biopuces. *Journal de la Société Française de Statistique, tome 145*, 3, 2008. Cité p. 104, 109

[6] P. Besse. *Apprentissage Statistique et Data mining*. Institut de M de Toulouse, Laboratoire de Statistique et Probabilités UMR CNRS C5583, Juillet 2009. Cité p. 118

[7] C. Bontemps. Prévision stochastique de trajectoires : procédures paramétriques et non-paramétriques, mémoire de dea informatique fondamentale et parallélisme. master's thesis. *ENAC*, 1997. Cité p. 31

[8] L. Breiman. Heuristic of instability and stabilization in model selection. *Annals of Statistics*, Vol 24(N° 6) :pp 2350 – 2383, 1996a. Cité p. 48, 101

[9] L. Breiman. Random forests. *Machine Learning*, 45(1) :5–32, 2001. Cité p. 102, 106

[10] L. Breiman. Manual On Setting Up, Using, And Understanding Random Forests V3.1. *http ://oz.berkeley.edu/users/breiman/Using randomforests V3.1.pdf*, 2002. Cité p. 106

[11] L. Breiman, J.H. Friedman, R.A. Ohlsen, and C.J. Stone. *Classification and regression trees*. Belmont : Wadsworth edition, 1984. Cité p. 48, 49

[12] A. Buja, D. Duffy, T. Hastie, and T. Tibshirani. Discussion of multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19(1) :93–99, 1991. Cité p. 80

[13] P. A. Chou, T. Lookabaugh, and R. Gray. Optimal pruning with applications to tree-structured source coding and modeling. *IEEE transactions on information theory*, 35 :299–315, 1989. Cité p. 49

[14] Performance Review Commission. *Special performance review report on delays*. EUROCONTROL, November 1999. Cité p. 11, 20

[15] Performance Review Commission. *An Assessment of Air Traffic Management in Europe during the Calendar Year 2004*. Final repport, EUROCONTROL, April 2005. Cité p. 42

[16] Performance Review Commission. *An Assessment of Air Traffic Management in Europe during the Calendar Year 2007*. Final repport, EUROCONTROL, May 2008. Cité p. 20, 21

[17] Performance Review Commission. *An Assessment of Air Traffic Management in Europe during the Calendar Year 2008*. Final repport, EUROCONTROL, May 2009. Cité p. 11

[18] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society*, Series B 34 :187–200, 1972. Cité p. 80

[19] P. Craven and G. Wahba. Smoothing noisy data with spline functions : Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math*, 31 :317–403, 1979. Cité p. 83

[20] R. B. Davis and J. R. Anderson. Exponential survival trees. *Statistics in Medicine*, 8 :947–961, 1989. Cité p. 48, 80

[21] Organisation de l'aviation civile internationale. Assistance météorologique à la navigation aérienne internationale. Technical report, OACI, Juillet 2007. Cité p. 166

[22] Equipe de taxonomie commune. Définitions des phases de vol et notes d'utilisation. Technical report, OACI, October 2002. Cité p. 166

[23] M. Delecroix. Que sais-je ? histogrammes et estimation de la densité. *Presses Universitaires de France*, 2055, 1983. Cité p. 31

[24] R. Diaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7 :3 :pp 1–13, 2006. Cité p. 104

[25] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Ann. Statist.*, 32(2) :407–499, 2004. Cité p. 108

[26] Y. Le Fablec. Prévision de trajectoires d'avions par réseaux de neurones, thèse. *Laboratoire d'Optimisation Globale CENA/ENAC*, 1999. Cité p. 33

[27] P. Flener, J. Pearson, M. Agren, C. Garcia-Avello, M. Celiktin, and S. Dissing. Air-traffic comm-plexity resolution in multi-sector planning using constraint programming. *Journal of Air Transport Management*, 13 :323–328, 2007. Cité p. 124

[28] Y. Freund and R. Schapire. Experiments with a new boosting algorithm. *Machine Learning : Proceedings of the Thirteenth Conference, ed : L. Saitta, Morgan Kaufmann*, pages pp 148–156, 1996a. Cité p. 102

[29] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *http ://www.research.att.com/ yoav*, 1996b. Cité p. 102

[30] J. H. Friedman. Multivariate Adaptative Regression Splines. *The Annals of Statistics*, 19(1) :1–68, 1991. Cité p. 80

[31] et J.-P. Nakache G. Celeux. *Analyse discriminante sur variables qualitatives*. Polytechnica edition, 1994. Cité p. 53

[32] S. B. Gelfand, C.S. Ravishankar, and J.D. Edward. An iterative growing and pruning algorithm for classification tree design. *IEEE transactions on pattern analysis and machine intelligence*, 13 :163–174, 1991. Cité p. 49

[33] B. Ghattas. Agrégation d'arbres de classification. *Revue de Statistique Appliquée*, 1999b. Cité p. 48, 57, 101

[34] R. J. Gray. Flexible methods for analyzing survival data using splines with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87 :942–951, 1992. Cité p. 80

[35] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(-3) :389–422, 2002. Cité p. 108

[36] B. A. Ishak and B. Ghattas. An efficient method for variable selection using svm-based criteria. *Pré-publication de l'Institut de Mathématique de Luminy, Marseille, France*, 2005. Cité p. 108

[37] et E. Matzner-Lober J. Josse. Prévisions des pics d'ozone à Lorient. *Lab. Mathématique Appliquée, Agrocampus Rennes, CS 84215, 35042 RENNES*. Cité p. 49

[38] C. M. Jarque and A. K. Bera. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3) :255–259, 1980. Cité p. 59

[39] C. Kooperberg, C. J. Stone, and Y. K. Truong. Hazard regression. *Journal of the American Statistical Association*, 90 :78–94, 1995. Cité p. 80

[40] L. W. Kwak, J. Halpern, R. A. Olshen, and S. J. Horning. 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, 1990. Cité p. 48, 80

[41] M. Leblanc and J. Crowley. Relative risk trees for censored survival data. *Biometrics*, 48 :411–425, 1992. Cité p. 49, 80

[42] M. Leblanc and J. Crowley. Adaptive regression splines in the cox model. *Biometrics*, 55 :204–213, March 1999. Cité p. 80

[43] M. Lejeune. *Statistique, La théorie et ses applications*. 2004-2005. Cité p. 127

[44] J. Lepas. *Complements de météorologie*. Ecole Nationale de L'Aviation Civile, 2ème edition, 1973. Cité p. 38

[45] A. Liaw and M. Wiener. Classification and regression by random forest. *Rnews*, 2 :18–22, 2002. Cité p. 104

[46] I. Lymperopoulos, J. Lygeros, and A. Lecchini. Model based aircraft trajectory prediction during takeoff. *AIAA Guidance, Navigation, and Control Conference and Exhibit, Keystone, Colorado*, Augost 2006. Cité p. 33

[47] G. Maignan. *Le contrôle de la circulation aérienne*. 108, boulevard Saint-Germain, 75006 Paris, 1er edition, avril 1991. Cité p. 15

[48] J. N. Morgan and R. C. Messenger. *THAID : a sequential serach program for the analysis of nominal scale dependent variables*. Ann Arbor : Intitute of social research, university of Michigan, 1973. Cité p. 48

[49] J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data and a proposal. *J. Am. Statit. Assoc.*, 58 :415–434, 1963. Cité p. 48

[50] et J.-B. Gotteland N. Durand. Algorithmes génétiques appliqués à la gestion du trafic aérien. *J3eA, Journal sur l'enseignement des sciences et technologies de l'information et des systèmes*, Volume 2, Hors-Série 1,6, July 2003. Cité p. 124

[51] OACI. http ://www.icao.int/icao/en/nr/2003/pio200317-f.pdf. Technical report, OACI, Novembre 2003. Cité p. 20

[52] ORA. Ordonnace du DETEC concernant les règles de l'air applicables aux aéronefs 748.121.11. *Département fédéral de l'environnement, des transports, de l'énergie et de la communication*. du 4 mai 1981 (Etat le 28 novembre 2006). Cité p. 35

[53] R. Quinlan. Bagging, boosting and c4.5. *In Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 725 – 730, 1996. Cité p. 101

[54] A. Rakotomamonjy. Variable selection using SVM-based criteria. *Journal of Machine Learning Research*, 3 :1357–1370, 2003. Cité p. 108

[55] A. Sang and San-qi Li. A predictability analysis of network traffic. *Computer Networks*, 39 :329–345, 2002. Cité p. 124

[56] R. Schapire. The boosting approach to machine learning : An overview. *In MSRI Workshop on Nonlinear Estimation and Classification*, 2002. Cité p. 102

[57] J. A. Sonquist and J. N. Morgan. The detection of interaction effects. *Ann Arbor : Institute for social research, University of Michigan*, 1964. Cité p. 48

[58] C. J. Stone. Discussion of multivariate adaptive regression splines (with discussion). *The Annals of Statistics*, 19 :113–115, 1991. Cité p. 80

[59] H. Theil. Economic Forecasts and Policy, Amsterdam : North Holland. 1958. Cité p. 60

[60] S. Tuffery. *Data Mining et statistique décisionnelle, L'intelligence des données*, volume 26 of *978-2-7108-0888-6*. Short Book Review, technip edition, April 2006. Cité p. 102

[61] D. Wallach and B. Goffinet. Mean squared error of prediction in models for studying ecological and agronomic systems. *Biometrics*, 43 :561–573, 1987. Cité p. 133

[62] J. Williams. Why airplanes like cool days better. *Aircraft Owners and Pilots Association's Flight Training Magazine*, July 2003. Cité p. 38

[63] J. Williams. Understanding air density and its effects. *USATODAY.COM*, July 2005. Cité p. 38

[64] P. C. Williams. Variables affecting near-infrared reflectance spectroscopic analysis. in : P. williams and k. norris (editors). *Near-infrared technology in the agricultural and food industries. American association of cereal chemists Inc*, pages 143–167, 1990. Cité p. 121

Table des figures

1.1 Un extrait de l'espace aérien Européen (2007) : Les traits bleus sont les limites des secteurs aériens, les traits rouges sont des routes aériennes et les points noirs sont les balises ou les points de croisements de plusieurs routes aériennes et les triangles sont les points de report obligatoire de la position des vols. 16

1.2 Illustration de la croissance du trafic aérien en Europe 19

1.3 Illustration de la façon dont les trajectoires sont respectées pendant les vols d'aéronefs 22

3.1 Illustration de la densité de réseau de trafic 44

3.2 Illustration de la trajectoire du vol : Le trait plein noir entre les points D et A indique la trajectoire du vol prévue dans le dernier plan de vol déposé avant le départ de l'aéronef. La portion de la trajectoire en pointillés rouges du point D au point BF indique la trajectoire temps réel du vol. A partir du point BF, la portion de la trajectoire délimitée par l'ellipse en pointillés bleus indique les positions des points sur la trajectoire de vol en lesquels nous voulons prévoir le temps de passage de l'aéronef. 46

4.1 Un exemple de la division d'un ensemble en plusieurs groupes disjoints 50

4.2 L'arbre de régression résultant de cette division 50

4.3 Coût-complexité des sous-arbres en fonction du nombre de nœuds terminaux 58

4.4 Histogramme de distribution des résidus du modèle CART classique 59

4.5 Dispersion des résidus en fonction de l'horizon temporel (à gauche) et spatial (à droite) de prévision 61

4.6 Qualité de prévision du modèle en fonction de l'horizon temporel de prévision 62

4.7 Qualité de prévision du modèle en fonction de l'horizon spatial de prévision 62

4.8 Arbre de régression de la méthode CART pour le modèle de prévision des écarts temporels 63

5.1 Histogrammes de distribution des résidus des modèles CART classique et CART modifié (Stepwise) 73

5.2 Dispersion des résidus en fonction de l'horizon temporel de prévision pour les modèles : CART classique et CART modifié (Stepwise). 00-03 indique par exemple un horizon de prévision compris entre entre 0 et 300 secondes. 74

5.3 Evolution comparée de l'incertitude en fonction de l'horizon temporel de prévision du modèle CART classique à celle de CART modifié (Stepwise) 74

5.4 Coefficient de Theil pour la comparaison de la qualité des modèles CART classique et CART modifié (Stepwise). Ce coefficient est égal au rapport du RMSE du modèle CART classique sur le RMSE du modèle CART modifié. Ref = référence et alt = alternatif. 75

6.1 Indicateur de qualité pour la sélection du modèle optimal en fonction du nombre de fonctions splines et du nombre de variables explicatives. 87
6.2 (a) est l'histogramme de distribution empirique des résidus du modèle MARS, (b) Compare les fonctions de répartition empiriques des résidus des modèles *CART classique*, *CART modifié* (Stepwise) et *MARS*. 87
6.3 Dispersion des résidus en fonction de l'horizon temporel de prévision des modèles : CART classique, CART modifié (Stepwise) et MARS. 00-03 indique par exemple l'intervalle de temps compris entre 0 et 300 secondes. 88
6.4 L'incertitude en fonction de l'horizon temporel de prévision. On compare l'évolution de l'incertitude du modèle MARS à celles des modèles CART classique et CART modifié. 89
6.5 Indicateur de Theil pour la comparaison de la qualité des modèles *CART classique* et *MARS*. Le coefficient de Theil est égal au rapport du RMSE du modèle *CART classique* sur le RMSE du modèle *MARS*. 89
6.6 Indicateur de Theil pour la comparaison de la qualité des modèles CART modifié (Stepwise) et MARS. Le coefficient Theil est égal au rapport du RMSE du modèle CART modifié sur le RMSE du modèle MARS. 89
6.7 Effets directs des variables explicatives les plus significatives du modèle. Seulement 5 variables ont des effets directs sur l'écart temporel de passage des aéronefs en des points de leurs trajectoires prévues. 93
6.8 Interaction entre l'influence du vent prévu sur la trajectoire du vol et la distance de vol prévue. L'axe vertical représente l'écart temporel en secondes. 96
6.9 Interaction entre l'influence du vent prévu sur la trajectoire de vol et la vitesse courante du vol. L'axe vertical représente l'écart temporel en secondes. 96
6.10 Interaction entre la distance de vol prévue et le niveau de vol de croisière prévu dans le plan de vol. L'axe vertical représente l'écart temporel en secondes. 96
6.11 Interaction entre la distance de vol prévue et le décalage courant du vol par rapport à sa trajectoire prévue. L'axe vertical représente l'écart temporel en secondes. 96
6.12 Interaction entre la distance de vol prévue (entre le point courant et un point la trajectoire prévue) et la distance totale du vol prévue entre l'aéroport de départ et l'aéroport d'arrivée du vol. L'axe vertical représente l'écart temporel en secondes. 97
6.13 Interaction entre le taux de montée / descente courant et le niveau de vol de croisière prévue dans le plan de vol. L'axe vertical représente l'écart temporel en secondes. 97
6.14 Interaction entre le niveau de vol de croisière prévu et la vitesse courante du vol au point courant de prévision. L'axe vertical représente l'écart temporel en secondes. 97
6.15 Interaction entre la distance de vol prévue et la vitesse courante du vol au point courant de prévision. L'axe vertical représente l'écart temporel en secondes. 97

7.1 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 pour la division sur les nœuds. 107
7.2 Somme des carrés des erreurs de prévision en fonction du nombre d'échantillons bootstrap (FA = Forêts aléatoires). 108
7.3 Importance des variables explicatives du modèle des forêts aléatoires. Le modèle est construit à partir de 150 échantillons bootstrap et 7 tirages aléatoires des variables explicatives sur chaque nœud. 111
7.4 Histogrammes de distribution des résidus des modèles : CART modifié (Stepwise), MARS et Forêt Aléatoire 111
7.5 Dispersion des résidus en fonction de l'horizon temporel de prévision : CART modifié (Stepwise), MARS et Forêt Aléatoire . 112

7.6 Variation du RMSE des modèles en fonction de l'horizon temporel de prévision : CART classique, CART modifié (Stepwise), MARS et Forêts aléatoires (FA). 113
7.7 Indicateur de Theil pour la comparaison du pouvoir prédictif du modèle des forêts aléatoires relativement aux modèles CART classique, CART modifié et MARS. Par exemple, CART classique/FA signifie rapport du RMSE du modèle CART classique sur le RMSE du modèle des forêts aléatoires. 114

8.1 Comparaison du RMedSEP au RMedSE pour le modèle CART Classique 120
8.2 Comparaison du RMedSEP au RMedSE pour le modèle CART Modifié 120
8.3 Comparaison du RMedSEP au RMedSE pour le modèle MARS 120
8.4 Comparaison du RMedSEP au RMedSE pour le modèle des Forêts aléatoires 120
8.5 Profils comparés des RMedSEP des différents modèles sur les données test 121
8.6 La capacité prédictive des modèles en fonction de l'horizon de prévision 122
8.7 Zoom de la dispersion des erreurs de prévisions des quatre modèles sur [-500 ; 500] 123
8.8 Ajustement des erreurs absolues du modèle CART classique par une loi de forme exponentielle 130
8.9 Ajustement des erreurs absolues du modèle CART modifié par une loi de forme exponentielle 130
8.10 Ajustement des erreurs absolues du modèle MARS par une loi de forme exponentielle 131
8.11 Ajustement des erreurs absolues du modèle des forêts aléatoires par une loi de forme exponentielle 131
8.12 Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle CART classique. L'estimation du mélange de lois est égale à : $0.14 * \mathcal{N}(16.69, 455.4) + 0.86 * \mathcal{N}(-5.62, 102.64)$. 135
8.13 Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle CART modifié. L'estimation du mélange de lois est égale à : $0.23 * \mathcal{N}(9.84, 390.6) + 0.77 * \mathcal{N}(-2.49, 97.82)$. 136
8.14 Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle MARS. L'estimation du mélange de lois est égale à : $0.18 * \mathcal{N}(40.95, 411.3) + 0.82 * \mathcal{N}(-9.17, 108.72)$. 136
8.15 Ajustement de la loi des résidus par un mélange de lois gaussiennes - Modèle des forêts aléatoires. L'estimation du mélange de lois est égale à : $0.15 * \mathcal{N}(0.51, 256.63) + 0.85 * \mathcal{N}(-1.52, 48.57)$. 137
8.16 Prévisions pour les horizons : 7 h 10 minutes et 7 h 15 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision. 140
8.17 Prévisions pour les horizons : 7 h 20 minutes et 7 h 25 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision. 141
8.18 Prévisions pour les horizons : 7 h 30 minutes et 7 h 35 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision. 142
8.19 Prévisions pour les horizons : 7 h 40 minutes et 7 h 60 minutes. La legende en marge droite indique la largeur de la fenêtre de prévision. 143

A.1 Maillage du domaine d'étude par les cellules de même volume 163
A.2 Variation de δ en fonction de la latitude 164

Liste des tableaux

5.1 Synthèse des résidus des modèle *CART classique* (prévision par la moyenne) et *CART modifié* (prévision par régression linéaire multiple). Chaque colonne représente les statistiques sur chaque nœud de l'arbre de régression CART correspondant. . 72

6.1 Les paramètres de l'estimateur $f_K(\mathbf{x})$ de Y. 45 termes ont été sélectionnés sur 49 bases construites dans le modèle maximal. 12 variables explicatives ont également été retenues sur 26. 91

7.1 Algorithme : Forêts Aléatoires 105

8.1 Comparaison de la prédictivité des modèles proposés. On utilise l'indice SD/RMSEP. Le modèle ayant une forte capacité prédictive a l'indice le plus élevé possible et supérieur à un. 122

8.2 Synthèse des résultats de Kolmogorov-Smirnov pour le test d'ajustement de la distribution des erreurs absolues à une loi de la forme exponentielle. Rejet signifie que les erreurs absolues ne suivent pas la loi exponentielle pour le modèle developpé. Pour chaque modèle, la fonction de densité théorique ajustée est $\widehat{f}(\widehat{a}, x) = \widehat{a}e^{-\widehat{a}x}$. 128

8.3 Synthèse des résultats de Kolmogorov-Smirnov pour le test d'ajustement de la distribution des erreurs absolues à une loi de la forme exponentielle. Pour chaque modèle, la fonction de densité théorique ajustée est $\widehat{f}_{nls}(\widehat{a}, x) = \widehat{a}e^{-\widehat{a}x}$. 129

8.4 Erreur absolue moyenne des modèles sur les données d'apprentissage 134

8.5 L'horizon de prévision en fonction de la limite de l'erreur de prévision. La valeur de chaque cellule du tableau représente l'horizon de prévision en secondes avec un niveau de confiance de 95%. Le « vide » dans une cellule signifie que le modèle correspondant ne permet pas de réaliser les prévisions du temps de passage des aéronefs avec une erreur absolue moyenne inférieure au seuil d'incertitude λ_0 avec le niveau de confiance de 95%. 134

B.1 Paramètres des plans de vol 166

B.2 Paramètres courants des vols 167

B.3 Paramètres météorologiques et atmosphériques 167

B.4 Les variables de complexité du trafic 168

B.5 Les variables explicatives les plus concurrentes du modèle de régression par arbre CART. Pour chaque nœud, la première ligne représente la variable active de la division optimale et les trois lignes suivantes représentent les trois premières divisions de substitution à la division optimale. 170

B.6 Coefficients du modèle au nœuds n° 1 et 2 171

B.7 Coefficients du modèle au nœud n°3 et 4 171

B.8 Coefficients du modèle au nœud n° 5 et 6 172
B.9 Coefficients du modèle au nœuds n° 7 et 8 172
B.10 Coefficients du modèle au nœuds n° 9 et 10 173
B.11 Coefficients du modèle au nœud n°11 et 12 173

Annexe A

Maillage de l’espace pour la construction des cellules

A.1 Maillage du domaine d'étude par les cellules de même volume

A.1.1 Relation entre deux cellules voisines de même aire

Afin d'identifier les zones de densité et de complexité dans le réseau du trafic aérien Européen, nous avons réalisé un maillage de l'espace couvrant le domaine d'étude 20N-72N et 32W-42E par des parallélépipèdes sphériques de même volume. Dans cette étude, une cellule est un parallélépipède sphérique dont le volume de référence est de 20 NM x 20 NM x 3000 pieds. Notre objectif va consister à recouvrir le domaine d'étude par un ensemble de telles cellules.

Une difficulté dans la réalisation de ce maillage est due à la forme sphéroïde de la terre qui ne permet pas d'obtenir des cellules ayant les mêmes caractéristiques géométriques partout. En effet, supposons fixées deux lignes de méridiens séparées par $0.33^\circ \equiv 20NM$ (Mille nautique), à partir du 20ème parallèle Nord, construisons vers le Nord, deux quadrilatères sphériques consécutifs dont les côtés verticaux portés par les deux méridiens sont des segments $[20^\circ; (20 + 0.33)^\circ]$ et $[(20 + 0.33)^\circ; (20 + 2 * 0.33)^\circ]$ respectivement. La différence de latitude les extrêmités de ces segments est identique égale à 0.33°. A cause de la courbure de la terre qui fait que les deux méridiens concourent au niveau des pôles, les deux domaines n'ont pas la même aire. Le domaine le plus au nord a une surface inférieure à celle de son plus proche voisin au Sud. En effet, plus on s'éloigne vers le Nord, la distance entre les deux méridiens diminue et entraîne la différence de volumes entre ces deux quadrilatères. Or l'objectif ici étant de comparer la densité de trafic d'une cellule à l'autre. Ainsi, pour construire des cellules de même volume partout sur le domaine d'étude, nous avons procédé à un ajustement des côtés des quadrilatères portés par les méridiens suite à la diminution de la longueur des côtés portés par les parallèles.

Ainsi, considérons $D_{i,j}$ et $D_{i+1,j}$ deux domaines consécutifs représentés dans FIG.A.1 ci-dessous. Ils sont limités par les méridiens fixes φ_j et φ_{j+1} et par les parallèles θ_i et θ_{i+1} et θ_{i+1} et θ_{i+2} respectivement. L'aire du domaine $D_{i,j}$ est égale à celle de $D_{i+1,j}$ sous la relation de récurrence définie par l'équation :

$$\theta_{i+2} = \arcsin(2 * \sin(\theta_{i+1}) - \sin(\theta_i)). \tag{1}$$

Ainsi, entre deux méridiens fixes, la connaissance des deux parallèles qui définissent un quadrilatère sphérique suffit pour déterminer la position du parallèle suivant pour obtenir le quadrilatère sphérique vers le nord, adjacent au précédent et ayant la même aire.

A.1.2 Illustration

Pour construire les quadrilatères entre les méridiens fixes φ_j et φ_{j+1}, posons :

- $A_{i,j}(\theta_i, \varphi_j)$, $A_{i+1,j}(\theta_{i+1}, \varphi_j)$, $B_{i,j}(\theta_i, \varphi_{j+1})$ et $B_{i+1,j}(\theta_{i+1}, \varphi_{j+1})$ des points en coordonnées géographiques (en degrés). Ce sont les sommets du domaine $D_{i,j}$ dans FIG.A.1. Entre les méridiens φ_j et φ_{j+1}, on peut définir de cette manière tous les sommets des quadrilatères vers le Nord géographique en faisant varier θ_i.

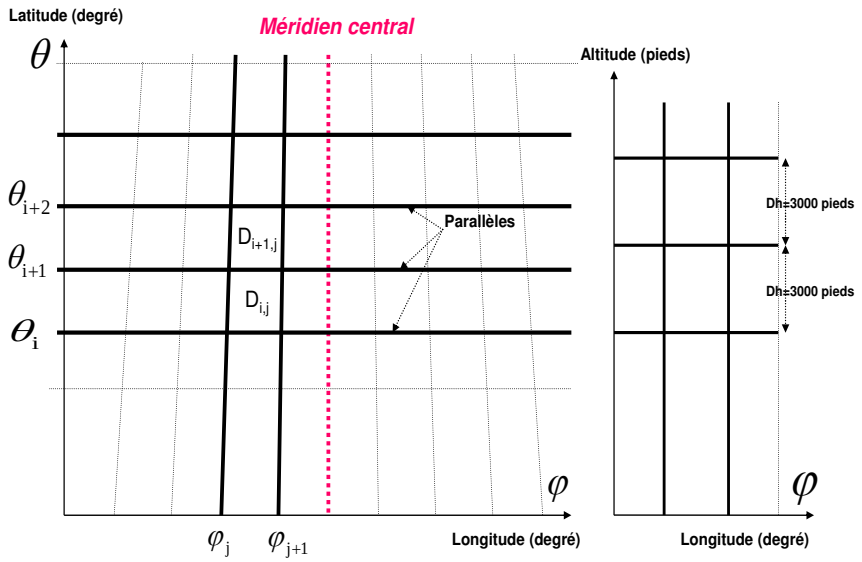

FIG. A.1 – Maillage du domaine d'étude par les cellules de même volume

- $dx_{i,j}$ et $dx_{i+1,j}$ sont des distances orthodromiques entre les points $A_{i,j}(\theta_i, \varphi_j)$ et $B_{i,j}(\theta_i, \varphi_{j+1})$ et entre $A_{i+1,j}(\theta_{i+1}, \varphi_j)$ et $B_{i+1,j}(\theta_{i+1}, \varphi_{j+1})$ respectivement. Ce sont les mesures de distances des côtés portés par les parallèles du domaine $D_{i,j}$.
- $dy_{i,j}$ et $dy_{i,j+1}$ sont des distances orthodromiques entre les points $A_{i,j}(\theta_i, \varphi_j)$ et $A_{i+1,j}(\theta_{i+1}, \varphi_j)$ et entre $B_{i,j}(\theta_i, \varphi_{j+1})$ et $B_{i+1,j}(\theta_{i+1}, \varphi_{j+1})$ respectivement. Ce sont les mesures de distances des côtés portés par les méridiens du domaine $D_{i,j}$. Ces mesures de distances sont identiques.
- $\epsilon_i \geq 0$ et $\delta_i \geq 0$.

A partir du 20ème parallèle, le premier quadrilatère sphérique est construit entre les parallèles $\theta_1 = 20^\circ$ et $\theta_2 = (20 + 0.3333)^\circ$. On initialise les côtés à 0.3333°. Etant proche de l'équateur, les mesures des côtés horizontal (parallèle) et vertical (méridien) sont approximativement $dx_1 = 20NM$ et $dy_1 = 20NM$ resepectivement. Pour construire le quadrilatère dont le côté inférieur est porté par la parallèle θ_2, on détermine θ_3 via l'équation de récurrence précédente (1). Elle est donnée par l'approximation :
$\theta_3 \approx 20.66738^\circ \approx ((20^\circ + 0.3333^\circ) + 0.3333^\circ + 0.0007) \approx (\theta_2 + 0.3333^\circ) + 0.0007^\circ$, où 0.0007° représente la variation de latitude qu'il faut ajouter aux côtés portés par les méridiens φ_j et φ_{j+1} pour compenser la diminution de longueur des côtés portés par les parallèles lors du passage des parallèles θ_1 et θ_2 aux parallèles θ_2 et θ_3 respectivement. Ainsi, on peut exprimer en mille nautique (NM) les mesures de distances des côtés horizontal et vertical du domaine suivant à l'intérieur des mêmes méridiens par les relations : $dx_{2,j} = 20 - \epsilon_2$ et $dy_{2,j} = 20 + \delta_2$. Dans l'exemple précédent, $\delta_1 = 0$ NM et

$\delta_2 = 0.0423$ NM. Il suffit de convertir 0.0007° en NM. De façon générale, entre deux méridiens fixes φ_j et φ_{j+1}, on a les relations : $dx_{i,j} = 20 - \epsilon_i$ (perte en longitude) et $dy_{i,j} = 20 + \delta_i$ (compensation en latitude).

Entre deux méridiens fixes séparés de 0.33°, les variations de $\delta = (\delta_i)$ en fonction de la latitude sont illustrées dans FIG.A.2. La longueur des côtés portés par les méridiens est donc une fonction croissante de la latitude entre 20 et 72 dégrés Nord. Ainsi, on construit l'ensemble des quadrilatères entre les deux méridiens fixes.

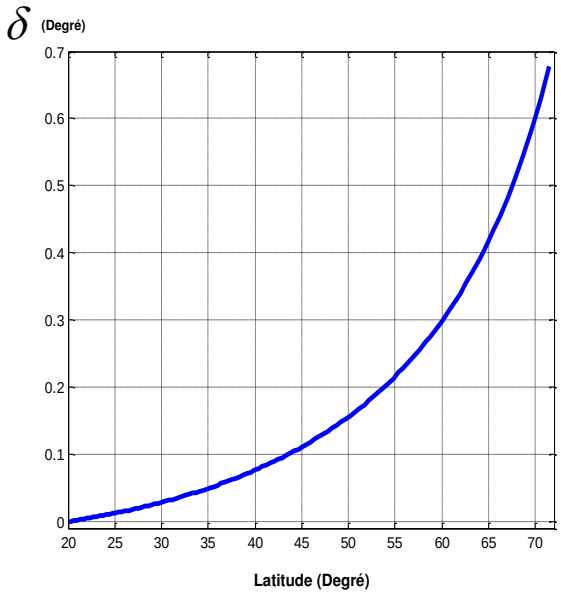

FIG. A.2 – Variation de δ en fonction de la latitude

Pour obtenir l'ensemble des quadrilatères sphériques sur tout le domaine $20N - 72N$ et $32W - 42E$, il suffit d'effectuer des rotations successives d'angle 0.33° des premiers quadrilatères. Ces rotations sont effectuées jusqu'aux méridiens limites du domaine. Après le recouvrement de tout le domaine par des quadrilatères sphériques de même surface, les premières cellules sont obtenues par la translation verticale de la base de chaque quadrilatère par un vecteur de longueur dh = 3000 pieds.

Enfin, les cellules de niveaux supérieurs sont obtenues par des translations successives de même vecteur de longueur dh=3000 pieds. On aboutit ainsi à une couverture de tout le domaine d'étude par un ensemble de cellules de même volume approximativement égal à 20 NM * 20 NM * 3000 pieds.

Annexe B

Compléments aux résultats de l’étude

B.1 Description des variables explicatives

Variables explicatives extraites des plans de vol	
Nom de code	**Description de la variable**
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). A l'aéroport de départ, le roulage comprend deux phases : roulage jusqu'à la piste et le roulage jusqu'à la position de décollage (OACI ; [22]).
Aroulage	La durée moyenne du roulage dans l'aéroport de destination du vol (secondes). A l'aéroport d'arrivée, le roulage comprend une seule phase, le roulage à partir de la piste d'atterrissage et qui se termine à son arrivée à la porte d'embarquement, à l'air de trafic ou à l'aire de stationnement, lorsque l'aéronef cesse de se déplacer par ses propres moyens (OACI ; [22]).
Dmouvaer	Le nombre moyen de mouvements par heure, dans l'aéroport de départ du vol. Cela concerne les vols qui arrivent ou qui partent de cet aéroport (nombre de vols par heure).
Amouvaer	Le nombre moyen de mouvements par heure, dans l'aéroport de destination du vol. Cela concerne les vols qui arrivent ou qui partent de cet aéroport (nombre de vols par heure).
Type	Le type d'aéronef utilisé par le vol (qualitative). Cette variable a été construite au moyen d'une classification ascendante hiérarchique avec le critère d'aggrégation de Ward. Pour cela, nous avons utilisé les données BADA de performances des aéronefs. Ainsi, chaque aéronef est représenté par l'une des modalités suivantes : *A306, AT43, B461, B712, B742, BE99 et AUTR*. Les vols pour lesquels les types ne sont pas disponibles ont pour modalité AUTR (pour dire autre).
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). Le niveau de croisière est le niveau auquel un aéronef se maintient pendant une partie appréciable d'un vol (Annexe 3 de l'OACI ; 2007)[21].
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 de stationnement (secondes). $Retardbloc = AOBT - IOBT$ si le plan de vol initial déposé n'a pas fait l'objet de modifications ultérieures. $Retardbloc = AOBT - COBT$ si le plan de vol initial déposé a fait l'objet de modifications ultérieures où $AOBT$ (Actual Off Block Time) est l'heure de départ bloc réalisée, $IOBT$ (Initial Off Block Time) est l'heure de départ bloc prévue dans le premier plan de vol déposé. $COBT$ (Calculated Off Block Time) est l'heure de départ bloc calculée si le vol s'est vu imposer un retard supplémentaire à cause de la congestion de sa route ou des mauvaises conditions météorologiques.

TAB. B.1 – Paramètres des plans de vol

Variables explicatives extraites des données courantes du vol	
Nom de code	**Description de la variable**
Heurecour	L'instant courant du vol (qualitatif). C'est l'instant à partir du quel on prévoit le temps de passage d'un aéronef à un point à un point de sa trajectoire prévue. Le point correspondant est appelé point courant du vol.
Retardcour	Le retard du vol au point courant (secondes). C'est la différence entre l'intant courant du vol et l'instant prévu dans le plan de vol pour le passage de l'aéronef au point courant.
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. B.2 – Paramètres courants des vols

Variables explicatives extraites des données météorologiques	
Nom de code	**Description de la variable**
Indur	Influence du vent sur le vol. C'est le rapport de la durée de vol prévue avec vent sur la durée de vol prévue sans vent (sans unité).
Moytemp	L'écart moyen entre la température prévue et la température de l'atmosphère standard sur la trajectoire prévue du vol (°C).
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.

TAB. B.3 – Paramètres météorologiques et atmosphériques

Variables explicatives extraites des paramètres de complexité	
Nom de code	**Description des variables de l'étude**
Paramètres 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. C'est la somme cumulée des flux du trafic dans toutes les cellules de la trajectoire prévue du vol, divisée par la distance de cette trajectoire (Nombre d'aéronefs par heure et par NM).
Indecum1	La somme cumulée de toutes les interactions potentielles prévues dans toutes les cellules de la trajectoire de vol prévue, divisée par la distance de cette trajectoire (Nombre de vols par NM).
Rscorecum1	La somme cumulée des scores des différents types d'interactions prévues dans toutes les cellules de la trajectoire de vol prévue , divisée par la distance de cette trajectoire. Il s'agit des scores d'interactions potentielles horizontales, verticales et de vitesse. Par exemple, pour les interactions potentielles horizontales, son score est le nombre d'interactions potentielles horizontales divisé le nombre total de toutes les interactions.

TAB. B.4 – Les variables de complexité du trafic

B.2 Test asymptotique de normalité de Jarque-Bera

Rappelons qu'une loi normale a un coefficient d'asymétrie (*Skewness*) égal à 0 et un coefficient d'applatissement (*Kurtosis*) égal à 3.

Le test de Jarque-Bera est un test de type multiplicateur de Lagrange qui cherche à déterminer si les données suivent une loi normale. Formellement, le test de Jarque-Bera ne teste pas directement si les données ajustent une loi normale, mais plutôt, si les coefficients d'applatissement et les coefficients d'asymétrie des données sont les mêmes que ceux du'une loi normale de même espérance et variance. Le test est formulé comme suit :

$$H_0 : S = 0 \text{ et } K = 3$$
$$H_0 : S \neq 0 \text{ ou } K \neq 3$$

La statistique de ce test est donnée par :

$$JBSTAT = \frac{n-k}{6}\left(S^2 + \frac{(K-3)^2}{4}\right) \tag{1}$$

La statistique $JBSTAT$ suit asymptotiquement une loi χ^2 du khi-deux 2 degrés. n est le nombre d'observations des données. k est le nombre de variables explicatives si les données provienneent des résidus d'une régression linéaire. Sinon, $k = 0$. S est le *Skewness* ou coefficent d'asymétrie. C'est le moment d'ordre 3 d'une variable centrée-réduite. K est le *Kurtosis* ou le coefficient d'applatissement. C'est le moment d'ordre 4 d'une variable centrée-réduite.
Pour la rélaistation de ce test, nous avons utilisé la fonction « *jbtest* »de MATLAB. A partir du package *tseries* du logiciel libre R, on peut aussi calculer le test de Jarque-Bera.

B.3 Lois gamma, exponentielle et du khi-2

X est une variable aléatoire, θ est un réel positif, N un entier positif, k est un réel positif.

- Si X_i suit une loi gamma $\gamma(k_i, \theta)$ pour tout $i \in \{1, ..., N\}$ et si les variables aléatoires X_i sont indépendantes, alors : $\sum_{i=1}^{N} X_i$ suit la loi $\gamma(\sum_{i=1}^{N} k_i, \theta)$.
- Si X suit une loi exponentielle $exp(\theta)$, alors X suit la loi gamma $\gamma(1, \theta)$.
- Si X suit une loi $\gamma(k, \theta)$, alors pour tout réel $t > 0$, la variable aléatoire tX est distribuée selon la loi $\gamma(k, \frac{1}{t}.\theta)$ avec θ qui est un paramètre d'échelle.
- Si X suit une loi $\gamma(\frac{\nu}{2}, 2)$, alors X suit une loi du khi-deux à ν degré de liberté notée $\chi^2(\nu)$.

B.4 Divisions suppléantes du modèle de régression par arbre CART

Synthèse des divisions suppléantes du modèle de régression par arbre CART				
Variable à expliquer :	***écart temporel* : Ecart temporel au passage d'un aéronef à un point de sa trajectoire de vol prévue**			
n° nodes	**Variables explicatives**	**Seuils**	**Association**	**Improvement**
1	**Indur**	**1.013**	**1**	**0.071**
	Alticour	135.5	0.57	
	Dmouvaer	11.35	0.57	
	Txmdcour	639	0.57	
2	**Distprev1**	**413.99**	**1**	**0.112**
	Distpln	918.5	0.82	
	Difalti	-130.99	0.81	
	Alticour	365	0.81	
3	**distprev1**	**351.67**	**1**	**0.058**
	Distpln	912.5	0.77	
	Difalti	-119.65	0.75	
	Alticour	3.267	0.75	
4	**Nivpln**	**185**	**1**	**0.2029**
	Moydensa	0.17	0.98	
	Type	0.225	0.98	
	Decalcour	0.075	0.98	
5	**Type**	**A306,AT43,B461,B712,B742,BE99**	**1**	**0.055**
	Nivpln	35	0.95	
	Moycrois	6.28	0.95	
	Dmouvaer	0.062	0.95	
6	**Indur**	**1.07**	**1**	**0.065**
	Alticour	378	0.67	
	Moycrois	4.82	0.67	
	Distpln	431.5	0.66	
7	**Indur**	**0.96**	**1**	**0.087**
	Vitessecour	487.5	0.58	
	Dmouvaer	64.05	0.58	
	Moycrois	4.602	0.58	
8	**Distprev1**	**147.45**	**1**	**0.058**
	Difalti	288.95	0.69	
	Vitessecour	460.5	0.63	
	Alticour	343.5	0.62	
9	**Vitessecour**	**162.5**	**1**	**0.187**
	Nivpln	115	0.91	
	Moypres	-3.49	0.87	
	Moydensa	0.204	0.86	
10	**Type**	**AT43**	**1**	**0.113**
	Moytemp	20.16	0.95	
	Moydensa	0.315	0.95	
	Distpln	183	0.95	
11	**Distprev1**	**55.46**	**1**	**0.42**
	Difalti	63.89	0.78	
	Distpln	274	0.68	
	Alticour	26.5	0.68	

TAB. B.5 – Les variables explicatives les plus concurrentes du modèle de régression par arbre CART. Pour chaque nœud, la première ligne représente la variable active de la division optimale et les trois lignes suivantes représentent les trois premières divisions de substitution à la division optimale.

B.5 Coefficients des modèles de régression linéaire généralisée par nœud

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°1 ($n = 41$)			
Constante	-3650.34	497.35	0.000
Indecum1	46.45	20.25	0.030
Distprev1	-3.54	0.35	0.000
Nivpln	25.37	1.30	< 0.000
Txmdcour	-0.25	0.04	0.000
Regul	187.10	51.86	0.001
Moycrois	257.64	90.37	0.008
Moypres	-63.23	7.50	0.000
Vitessecour	-2.23	0.55	0.000
réf :TypeA306	–	–	–
TypeAT43	-256.62	252.22	0.318
TypeAUTR	-318.53	178.44	0.085
TypeB461	-478.05	144.93	0.002
TypeB712	-45.28	181.67	0.805
TypeB742	231.62	144.59	0.121
TypeBE99	577.15	211.97	0.011
Adjusted R-squared :		**0.99**	
F-statistic :		**294.3**	
Nœud n°2 ($n = 1133$)			
Contante	-3364	336	< 0.000
Denscum1	25.05	7.92	0.001
Indur	2815	358.7	0.000
Distprev1	-0.65	0.06	< 0.000
Txmdcour	-0.02	0.009	0.049
Nivpln	1.79	0.17	< 0.000
Distpln	0.12	0.04	0.008
Retardbloc	-0.05	0.004	0.000
Decalcour	99.74	14.46	0.000
Alticour	-0.52	0.145	0.000
réf :TypeA306	–	–	–
TypeAT43	415.9	131.8	0.002
TypeAUTR	390.4	47.56	0.000
TypeB461	484.5	56.44	0.000
TypeB712	45.38	195.7	0.020
TypeB742	-40.87	28.18	0.147
Adjusted R-squared :		**0.32**	
F-statistic :		**38.34**	

TAB. B.6 – Coefficients du modèle au nœuds n° 1 et 2

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°3 ($n = 1456$)			
Copnstante	-3815	532.7	0.000
Moycrois	28.47	15.72	0.070
Moytemp	-15.14	4.07	0.000
Moydensa	1571	432.9	0.000
Indur	3407	526.8	0.000
Difalti	-0.42	0.104	0.000
Distprev1	-0.37	0.06	0.000
Retardcour	-0.07	0.01	0.000
Txmdcour	-0.04	0.009	0.000
Nivpln	1.196	0.22	0.000
Distpln	0.17	0.47	0.000
Amouvaer	1.56	0.38	0.000
Decalcour	30.68	11.12	0.006
Alticour	-0.53	0.17	0.002
réf :TypeA306	–	–	–
TypeAT43	-497.1	163.2	0.002
TypeAUTR	472.1	47.42	< 0.000
TypeB461	101.9	59.02	0.084
TypeB712	423.4	20.55	0.039
TypeB742	61.07	32.76	0.063
Adjusted R-squared :		**0.17**	
F-statistic :		**17.17**	
Nœud n°4 ($n = 204$)			
Constante	2456	280.9	0.000
Vitessecour	-5.91	5.20	< 0.000
Nivpln	-4.48	1.08	0.000
Moydensa	-8344	2938	0.005
Distprev1	-1.541	0.52	0.003
Txmdcour	-0.08	0.04	0.018
Moycrois	145.9	35.6	0.000
Amouvaer	-4.69	1.41	0.001
Distpln	-1.16	0.35	0.001
Moytemp	9.28	4.17	0.027
Decalcour	121	54.01	0.026
Rscorecum1	3.35	1.93	0.083
Adjusted R-squared :		**0.57**	
F-statistic :		**25.38**	

TAB. B.7 – Coefficients du modèle au nœud n°3 et 4

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°5 ($n = 3679$)			
Constante	-1190	103.2	< 0.000
Indur	1060	100.5	< 0.000
Vitessecour	-0.30	0.04	0.000
Nivpln	1.43	0.08	< 0.000
Distprev1	-0.52	0.05	< 0.000
Dmouvaer	0.83	0.15	0.000
Amouvaer	0.98	0.17	0.000
Txmdcour	-0.021	0.004	0.000
Alticour	-0.92	0.08	< 0.000
Difalti	-0.55	0.05	< 0.000
Decalcour	22.57	5.31	0.000
réf :TypeA306	–	–	–
TypeB461	195.4	17.08	< 0.000
TypeB712	11.91	7.35	0.105
TypeB742	55.93	12.56	0.000
TypeBE99	120	61.01	0.049
Adjusted R-squared :		**0.18**	
F-statistic :		**58.9**	
Nœud n°6 ($n = 6660$)			
Constante	-102.3	43.68	0.019
Distprev1	-0.51	0.04	0.000
Alticour	-0.38	0.028	< 0.000
Txmdcour	-0.02	0.002	< 0.000
Decalcour	15.22	2.332	0.000
Indur	238.5	42.74	0.000
Moycrois	-11.27	1.772	0.000
Moypres	1.011	0.18	0.000
Moytemp	-1.99	0.41	0.000
Amouvaer	0.18	0.07	0.018
Dmouvaer	0.14	0.07	0.042
réf :TypeA306	–	–	–
TypeAT43	-63.67	5.47	0.000
TypeB461	17.57	6.60	0.008
TypeB712	3.55	3.43	0.299
TypeB742	24.84	5.65	0.000
TypeBE99	89.54	1.65	0.000
Adjusted R-squared :		**0.10**	
F-statistic :		**51.04**	

TAB. B.8 – Coefficients du modèle au nœud n° 5 et 6

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°7 ($n = 474$)			
Constante	218.83	105.13	0.037
Vitessecour	-0.96	0.18	0.000
Txmdcour	-0.05	0.01	0.000
Distprev1	0.67	0.13	0.000
Amouvaer	-1.61	0.69	0.022
Moydensa	2735.56	660.27	0.000
Moycrois	85.44	21.29	0.000
Retardbloc	-0.07	0.02	0.001
Dmouvaer	-1.12	0.59	0.060
Moytemp	-27.62	9.91	0.006
Regul	-41.68	17.88	0.020
Alticour	-0.48	0.23	0.036
Adjusted R-squared :		**0.25**	
F-statistic :		**15.66**	
Nœud n°8 ($n = 57$)			
Constante	-1338	564.2	0.021
Distprev1	11.47	1.714	0.000
Decalcour	251.8	54.36	0.000
Moytemp	-19.57	5.64	0.001
Indur	1534	600.6	0.014
Retardbloc	0.06	271	0.024
Alticour	-2.84	1.61	0.085
Adjusted R-squared :		**0.55**	
F-statistic :		**13.19**	

TAB. B.9 – Coefficients du modèle au nœuds n° 7 et 8

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°9 ($n = 40$)			
Constante	-6187.56	1980.05	0.004
Dmouvaer	22.05	7.02	0.004
Rscorecum1	2922.82	306.16	0.000
Indecum1	-4586.28	553.76	0.000
Distprev1	4.62	0.64	0.000
Moytemp	-100.27	17.04	0.000
Denscum1	-2053.11	654.98	0.004
Retardbloc	0.40	0.09	0.000
Indur	8149.63	2153.29	0.001
Adjusted R-squared :		**0.90**	
F-statistic :		**45.45**	
Nœud n°10 ($n = 8466$)			
Constante	-442.5	50.82	< 0.000
Indur	516.6	43.9	< 0.000
Distprev1	0.23	0.02	< 0.000
Vitessecour	-0.25	0.03	< 0.000
Txmdcour	-0.02	0.001	< 0.000
Alticour	-0.28	0.03	< 0.000
Nivpln	0.21	0.04	0.000
Decalcour	12.32	3.70	0.001
Dmouvaer	0.34	0.09	0.000
Moytemp	-2.45	0.62	0.000
Distpln	-0.05	0.009	0.000
Moycrois	9.18	2.759	0.001
Retardbloc	-0.008	0.003	0.013
Amouvaer	-0.17	0.09	0.074
Adjusted R-squared :		**0.07**	
F-statistic :		**49.17**	

TAB. B.10 – Coefficients du modèle au nœuds n° 9 et 10

Modèle par l'algorithme *Stepwise*			
Variable à expliquer	Ecarttemp : Ecart temporel de passage d'un aéronef en un point		
Variables explicatives	Coefficients	Ecart-type	p-value
Nœud n°11 ($n = 1840$)			
Constante	-3077	528.4	0.000
Vitessecour	-0.64	0.11	0.000
Indur	2507	493.8	0.000
Moydensa	3880	423.1	< 0.000
Nivpln	1.55	0.21	0.000
Regul	-44.98	10.17	0.000
Moytemp	-25.59	4.87	0.000
Moypres	-4.91	1.36	0.000
Distpln	0.18	0.04	0.000
Retardbloc	-0.03	0.01	0.013
Moycrois	35.34	16.38	0.031
Dmouvaer	0.94	0.41	0.023
Indecum1	-37.92	9.71	0.000
Denscum1	213.9	56.20	0.000
Txmdcour	-0.03	0.01	0.014
Alticour	-0.82	0.15	0.000
Decalcour	42.88	14.33	0.003
Difalti	-0.33	0.09	0.001
Adjusted R-squared :		**0.12**	
F-statistic :		**15.11**	
Nœud n°12 ($n = 949$)			
Constante	-1086	299.2	0.000
Vitessecour	-1.29	0.16	0.000
Distprev1	0.49	0.08	0.000
Indur	1282	259.3	0.000
Nivpln	1.43	0.26	0.000
Denscum1	-385.5	83.9	0.000
Rscorecum1	41.58	10.20	0.000
Regul	-37.9	15.52	0.015
Retardcour	0.09	0.02	0.000
Distpln	-0.15	0.07	0.050
Amouvaer	-1.15	0.56	0.039
Retardbloc	-0.07	0.03	0.015
Alticour	-0.85	0.22	0.000
Difalti	-0.42	0.17	0.013
Adjusted R-squared :		**0.20**	
F-statistic :		**18.82**	

TAB. B.11 – Coefficients du modèle au nœud n°11 et 12