

Établissement d'inscription: ENSEEIHT DEA Informatique Fondamentale et
Parallélisme
Année: 1996 - 1997

Laboratoire d'Optimisation Globale CENA/ENAC

Dirigé par Jean-Marc ALLIOT

Prévision stochastique de trajectoires : Procédures paramétriques et non-paramétriques

par

Christophe BONTEMPS

bontemps@recherche.enac.fr

Directeur de recherche: Jean-Marc ALLIOT

Responsable du stage: Nicolas DURAND

Septembre 1997

École Nationale de l'Aviation Civile
7, avenue Édouard Belin
31055 Toulouse



Remerciements

Si je ne devais remercier qu'une seule personne, et relater ici ses actions, cette page serait remplie du nom de Jean-Marc ALLIOT, et sa modestie pourrait en souffrir. J'associerai donc à ces remerciements toute l'équipe du Laboratoire d'Optimisation Globale (LOG), pour m'avoir supporté, au sens propre comme au figuré, tout au long de cette difficile année.

Que soient donc remerciés Pascal BRISSET pour son entrain et sa patience, Nicolas DURAND, qui a réussi à me faire aimer le Badminton, (le jour ou je l'ai battu), et a contribué à rétablir mon équilibre physique et mental, Yann LE FABLEC qui a répondu sans rire à toutes mes détresses informatiques. Sans oublier Daniel DELAHAYE, Stéphane PUECHMOREL et Jean-François BOSC qui ont contribué à maintenir une très chaleureuse ambiance ainsi que mes « collègues » Nicolas BARNIER et François HUCHET.

Avec beaucoup de gentillesse, Christelle PIANETTI du CENA a partagé ses sources sur le problème de prévision de trajectoire, je l'en remercie.

Que le département Math-Info trouve ici par l'intermédiaire de ses chefs Eric BLOND et Pierre CREVITZ, ma reconnaissance pour m'avoir hébergé et soutenu au cours de cette année.

Je tiens également à remercier vivement Monsieur le professeur NOAILLES de m'avoir accepté dans ce DEA malgré mon cursus et le Centre d'Études de la Navigation Aérienne (CENA), en la personne de Jean-Marc ALLIOT (encore lui), pour l'avoir financé.

Enfin, Nicole VAYSSIÈRE, et Nathalie LENOIR m'ont permis d'oublier tout, parfois...

Résumé

La prévision de trajectoires fait l'objet d'une attention toute particulière dans le cadre du contrôle aérien. Afin de mieux gérer les flux de trafic, et d'anticiper les conflits, il est nécessaire de prévoir de manière fiable les trajectoires des avions. De nombreux paramètres, pour la plupart inobservables en temps réel ou inconnus, régissent ces trajectoires (type d'appareil, poids, configuration, vent, consigne des compagnies, etc...).

Nous proposons une estimation basée sur des méthodes stochastiques d'estimation. L'idée est d'estimer, sur la base d'observations, la *fonction trajectoire* elle-même plutôt que les nombreux paramètres inconnus régissant son équation. Notre approche sera axée sur *l'inférence* statistique : de l'observation de la trajectoire passée nous essayerons de tirer une estimation du processus (que nous supposons stochastique) ayant engendré cette trajectoire. Nous appliquerons donc le schéma général de la statistique inférentielle que nous pouvons détailler comme :

- *modélisation* du processus inconnu, ayant engendré les données. Nous utiliserons principalement la modélisation « paramétrique » dans laquelle une forme prédéfinie de la fonction est imposée et une modélisation « non-paramétrique » pour laquelle aucune hypothèse sur la forme de la fonction n'est nécessaire.
- *estimation* du processus, ou de son modèle simplifié. C'est la partie calculatoire proprement dite synthétisant les informations apprises dans le passé.
- *prévision* d'une valeur future à l'aide de la modélisation effectuée.
- *test de fiabilité*. L'utilisation de structures stochastiques, sur la nature des erreurs entre modèle approché et réalisations, permet le calcul d'*intervalles de confiance* pour la prévision, donnant la fiabilité de la prévision.

Les méthodes d'estimation seront donc, à l'image des modélisations, paramétriques ou non-paramétriques. Les méthodes paramétriques linéaires et non linéaires seront utilisées sans difficultés et sans grande originalité. Par contre, les méthodes d'estimation non-paramétriques seront développées en détail et adaptées à notre problème. Elles sont de la famille des méthodes à noyau de convolution développées en statistique non-paramétrique (Bosq et Lecoutre (1987), Hardle (1990),...).

Nous envisagerons également différentes approches de la fonction trajectoire, selon qu'on la considère comme une fonction paramétrée du temps $f(t, \theta)$, comme un processus temporel (X_t) de type autorégressif, ou plus généralement comme une fonction du temps $f(t)$.

L'étude réalisée au sein du Laboratoire d'Optimisation Globale CENA-ENAC est programmée en Caml et testée sur la base de données simulées par un générateur de trajectoires développé au CENA. Les résultats sont présentés pour divers horizons de prévision; pour des observations plus ou moins rapprochées et pour les différents modèles de prévision proposés.

Mots Clés : Prévision stochastique, estimation, méthodes non-paramétriques.

Sommaire

1	Problématique	5
1.1	Introduction	5
1.2	Contraintes	7
1.3	Les trajectoires	7
1.4	Fonction réelle ou série temporelle?	8
1.5	Estimation paramétrique ou non-paramétrique?	9
2	Méthodes de régression « classiques »	11
2.1	Un peu de théorie	11
2.2	Régression linéaire simple	12
2.2.1	Méthode	12
2.2.2	Qualité de l'ajustement	14
2.2.3	Tests	14
2.3	Prévisions et intervalles de confiance	15
3	Prévisions « classiques »	19
3.1	Estimation paramétrique	19
3.1.1	Modèle linéaire	19
3.1.2	Modèle « racine »	20
3.1.3	Modèle « multiplicatif »	21
3.1.4	Modèle « Logarithmique »	22
3.2	Prévisions opérationnelles	22
3.3	Bilan	25
4	Méthodes non-paramétriques	27
4.1	Introduction	27
4.2	Définition des estimateurs	28
4.3	Estimateur du noyau de convolution	32
4.3.1	Propriétés générales	34
4.3.2	Propriétés de convergence ponctuelle de l'estimateur \widehat{f}_n :	34
4.4	Du bon choix de la fenêtre	35
4.4.1	Le « <i>plug-in</i> »	37
4.4.2	La « <i>Validation Croisée</i> »	37
4.4.3	L'approche « naïve » de Bosq et Lecoutre	38
4.4.4	Autres méthodes	39
4.5	Conclusion	39
4.6	Annexe au chapitre 3	41
5	Prévisions « non-paramétriques »	45
5.1	Mise en œuvre	45
5.1.1	Prévision et prévisions	45
5.1.2	Méthode utilisée	50

5.1.3	Et les séries chronologiques?	51
5.2	Prévisions à « court » terme	52
5.2.1	Du rôle du « pas »	53
5.2.2	Du choix pratique de la fenêtre	53
5.2.3	Pour en finir avec les fenêtres...	57
5.3	Prévision à « Long » terme	57
5.3.1	Étude de l'erreur initiale	60
6	Bilan et extensions	63
6.1	Raffinement des estimateurs et des procédures	63
6.2	Utilisation d'informations complémentaires	64
6.3	Vers les réseaux de neurones?	64

Table des figures

1.1	Les prévisions dépendent du poids accordé au passé.	6
1.2	Exemple de trajectoire d'un avion en montée vers son niveau de vol	8
2.1	$\hat{Y} = E[Y X]$ est la projection de Y sur W_X	12
2.2	Nuage de points et droite de régression	13
3.1	Modèle de prévision linéaire	20
3.2	Modèle de prévision « racine »	21
3.3	Modèle de prévision multiplicatif	22
3.4	Modèle de prévision logarithmique	23
3.5	Erreurs de prévision pour les différents modèles.	23
3.6	Prévisions pour différentes durées d'observation (Modèle « multiplicatif »).	24
3.7	Prévisions pour différentes durées d'observation (Modèle « racine »).	25
4.1	Exemple de Noyau $K(\frac{u}{h})$ pour 3 valeurs de h	29
4.2	Impact de la fenêtre (h) sur l'estimateur \hat{f} d'une fonction f	32
4.3	Cas limites pour le choix de la fenêtre h dans \hat{f}	33
5.1	Trajectoire d'un A320 en montée	46
5.2	Estimation par filtrage.	47
5.3	Généralisation du filtrage.	48
5.4	Principe de la méthode « D ».	49
5.5	Impact de la fenêtre h et des « poids » sur l'erreur de prévision (Méthode « D »).	50
5.6	Série des innovations: $e_i = Y_i - Y_{i-1}$	51
5.7	Série des innovations « d'ordre 2 »: $u_i = e_i - e_{i-1}$	52
5.8	Estimation de la trajectoire pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes	54
5.9	Erreur de prévision pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes	54
5.10	Estimation de la trajectoire pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes	55
5.11	Erreur de prévision pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes	55
5.12	Estimation de la trajectoire pour un « pas » de 10s et à horizon 10, pour trois valeurs de h différentes	56
5.13	Erreur de prévision pour un « pas » de 10s et à horizon 10, pour trois valeurs de h différentes	56
5.14	Estimation de la trajectoire pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes	58
5.15	Erreur de prévision pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes	58
5.16	Estimation de la trajectoire pour un « pas » de 10s et à horizon 20, pour trois valeurs de h différentes	59

5.17	Erreur de prévision pour un « pas » de 10s et à horizon 20, pour trois valeurs de h différentes	59
5.18	Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes	60
5.19	Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes	61
5.20	Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes	62

Chapitre 1

Problématique

« *Tout ce qui bouge bouge parce qu'autre chose bouge* »
Saint Thomas d'Aquin

1.1 Introduction

Le problème de la prévision de trajectoires est un problème très simple dans sa formulation :

« Etant donné l'observation de la trajectoire passée d'un avion, comment prédire au mieux sa trajectoire future? »

Ce problème en appelle d'autres qui seront l'essence même de ce mémoire: Quelles méthodes de prévision peut on utiliser? Quelles sont les critères d'une bonne prévision? Comment prévoir en pratique?

Au delà des problèmes liés à la prévision elle-même nous pouvons nous intéresser, dans un premier temps, au support de cette prévision : le passé. Le concept « d'observations passées » n'est, en effet, pas dépourvu d'ambiguïté: il y a le passé « récent », le passé « lointain », et enfin « tout » le passé. Si nous prenons comme exemple la fonction f représentée dans la figure 1.1, pour laquelle nous observons des réalisations jusqu'à la date T_0 , nous pouvons donner trois prévisions simples basées sur trois « passés » différents :

- En utilisant le passé *récent*, et en traçant la droite la plus proche des trois derniers points observés, avant T_0 , nous obtenons une première prévision : P_0 ;
- en utilisant un passé *moins récent* (de T_1 à T_0), et toujours par la même méthode de la droite la plus proche de ce nouvel ensemble de points observés, nous obtenons une deuxième prévision : P_1 ;
- enfin, en utilisant *tout* le passé, et donc toutes les observations, nous obtenons une autre prévision : P_2 .

Un examen plus approfondi montre que la prévision P_0 est peut-être la meilleure à court terme, mais qu'il s'avère plus prudent, pour prévoir à long terme, de prendre les prévisions P_2 ou P_1 . Ces trois prévisions utilisent pourtant la même méthode et ne diffèrent que par le *passé* pris en compte pour la prévision. Plus exactement, la différence fondamentale entre ces trois prévisions réside dans l'importance donnée aux différentes observations du passé, leur « poids ». Pour P_0 on accorde une grande importance aux dernières observations et on néglige les autres alors que pour P_2 toutes les observations sont prises en compte avec la même importance. Nous avons donc déjà un premier problème concernant la pondération et l'utilisation des données observées. Il sera beaucoup question de cette notion dans ce mémoire, et de la question du « bon choix » des pondérations pour la prévision de trajectoires.

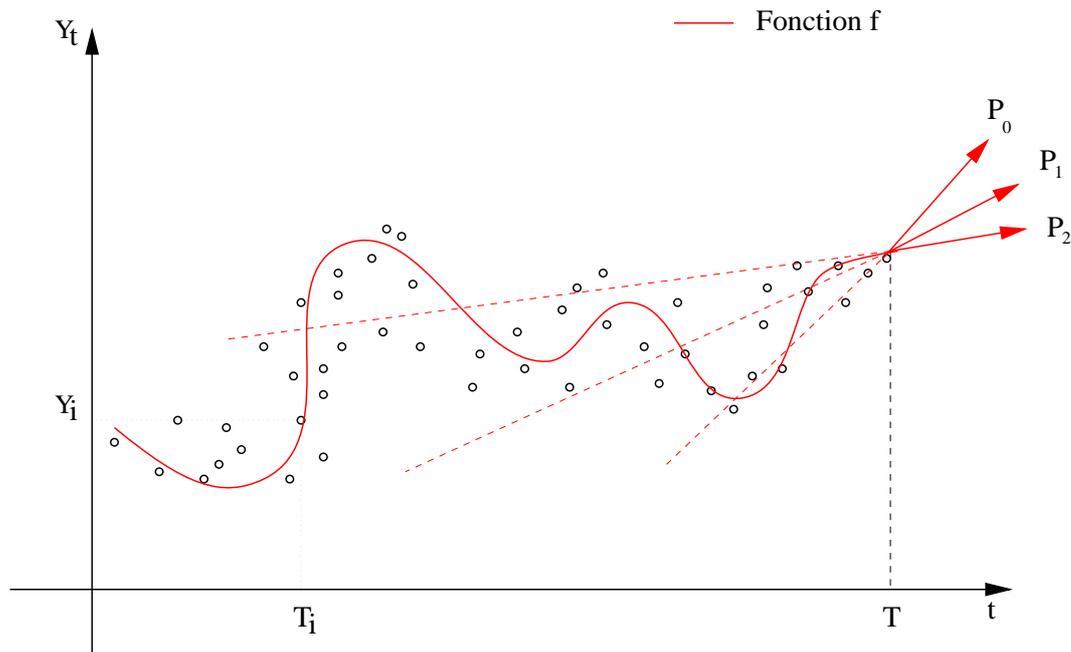


FIG. 1.1 - Les prévisions dépendent du poids accordé au passé.

Notre objectif (et les méthodes que nous présenterons ici devront nous y aider) est de prédire « au mieux » la trajectoire d'un appareil. Nous devons pour cela nous donner quelques repères concernant la qualité d'une prévision. Au delà de l'indispensable *justesse* de la prévision, il nous semble que la *fiabilité* et la *robustesse* sont deux points importants, trop souvent négligés. La fiabilité apporte la confiance et la robustesse une souplesse d'utilisation à la prévision, quelle que soit la méthode utilisée.

Nous avons choisi pour cela une approche statistique prenant en compte les éléments que nous venons de décrire. Notre approche sera donc axée sur *l'inférence* statistique : de l'observation de la trajectoire passée nous essayerons de tirer une estimation du processus¹ ayant engendré cette trajectoire. La connaissance de ce processus n'étant évidemment qu'utopie, nous appliquerons le schéma général de la statistique inférentielle que nous pouvons détailler comme suit :

- *modélisation* du processus inconnu, ayant engendré les données. Dans cette phase, des hypothèses simplificatrices sur la nature supposée de la trajectoire seront nécessaires.
- *estimation* du processus, ou de son modèle simplifié. C'est la partie calculatoire proprement dite synthétisant les informations apprises dans le passé. Les choix concernant les poids à attribuer aux points du passé interviendront à ce stade.
- *prévision* d'une valeur future à l'aide de la modélisation effectuée.
- *test de fiabilité*. L'utilisation de structures stochastiques, sur la nature des erreurs entre modèle approché et réalisations, permet le calcul d'*intervalles de confiance* pour la prévision, donnant la fiabilité de la prévision.

Nous présenterons diverses méthodes stochastiques pour réaliser les prévisions de trajectoires, toutes suivront ce schéma général. En l'absence d'information extérieure, nous supposerons donc ici que l'avion *bouge* parce que le temps *bouge*. D'autres approches sont cependant possibles, nous mentionnerons l'étude de C. Leclaire du CENA [18], qui utilise une toute autre méthode. Le modèle de prévision repose en effet sur l'exploitation des données du plan vol et sur l'estimation des

1. que nous supposerons stochastique

différentes variables liées à la trajectoire d'un appareil. Chacun des modèles permet la prévision des variables pertinentes (consommation, taux de montée, météo, etc), les équations (déterministes) de la mécanique du vol donnent ensuite la trajectoire. Il s'agit donc là d'une réelle approche *aéronautique* basée sur un « modèle avion » et sur la mécanique du vol.

Une autre démarche, très récente, se base sur l'apprentissage de trajectoires par réseaux de neurones. Cette approche prometteuse développée au sein du LOG par Yann Le Fablec, n'est pas très éloignée, de certaines méthodes décrites ici, comme nous le verrons dans notre dernier chapitre.

Ce mémoire s'articule autour de quatre chapitres principaux ; deux chapitres sont consacrés aux méthodes paramétriques classiques, et deux chapitres proposent une approche nouvelle de ce problème. Dans le chapitre 3, diverses modélisations paramétrées de la fonction trajectoire seront éprouvées sur la base d'une trajectoire type. Les résultats seront présentés avec leurs intervalles de confiance, pour différentes familles de modèles linéaires et non-linéaires. Les méthodes d'estimation de ces modèles, ainsi que le calcul des critères d'ajustement et des intervalles de confiance, auront été rappelés au préalable dans le chapitre 2.

Nous présenterons également dans le chapitre 4 de nouvelles méthodes d'estimation dites méthodes « non-paramétriques ». Contrairement aux méthodes « classiques » du chapitre 2, ces méthodes reposent sur l'estimation directe de la fonction (inconnue) ayant engendrée les données, sans imposer de modèle paramétré prédéfini. Le lecteur pressé pourra toutefois se reporter rapidement à la mise en œuvre et aux résultats de ces procédures qui seront présentés et discutés dans le chapitre 5.

Nous nous proposons de discuter au préalable des hypothèses de travail, contraintes, et autres partis pris sur le travail effectué dans les section suivantes.

1.2 Contraintes

Le travail qui nous a été demandé consiste en la réalisation d'un outil de prévision de trajectoires, pour différents types d'appareils, à horizon « lointain », sur la base de trajectoires observées. Cet outil devra, par la suite, être incorporé à la procédure de résolution de conflits en vol en cours de développement au LOG. En fait, les prévisions ne concernent la trajectoire « que » dans sa composante verticale, c'est à dire qu'il faudra être capable de déterminer le plus précisément possible l'altitude d'un avion.

Pour cela, nous disposons de la trajectoire passée² de l'avion. Cette information unique est plus ou moins riche suivant la période et suivant la fréquence des observations. Ici, nous observerons les trajectoires toutes les 5, 10 ou 15 secondes (voir section 5.2), et proposerons des prévisions jusqu'à 200 secondes. Bien entendu, la qualité des observations, et l'horizon de prévision auront une forte influence sur les résultats (voir section 5.2.1).

Il est à noter que nous n'étudierons que les trajectoires en montée. En effet, c'est la partie de la trajectoire qui connaît la plus forte non-linéarité. Les phases en niveau de vol et en descente étant, elles, quasi-linéaires, ne présentent que peu d'intérêt pour notre étude. La section 1.3 ci-après, nous permettra de présenter plus en détail les trajectoires sur lesquelles nous nous proposons de travailler.

1.3 Les trajectoires

Par souci de clarté et d'homogénéité, nous illustrerons notre propos sur la base d'un unique exemple présenté dans la figure 5.1. Il s'agit de la trajectoire simulée d'un Airbus A320, en montée avec stabilisation à 31000 pieds. La simulation est issue du logiciel PERFO en développement au CENA, en partenariat avec l'Aérospatiale. Le travail de prévision concerne les phases de montées

2. C'est à dire du relevé régulier des altitudes de l'avion considéré

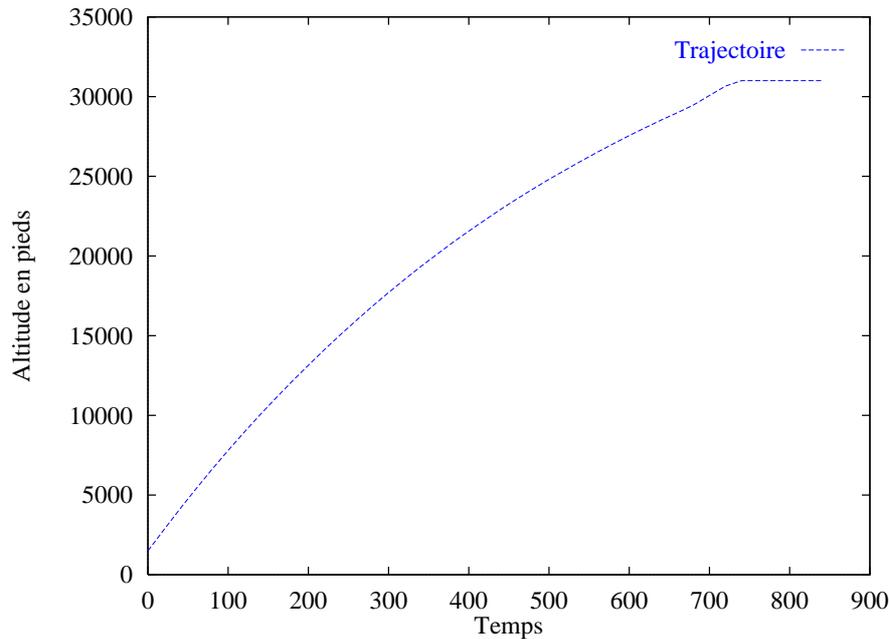


FIG. 1.2 - Exemple de trajectoire d'un avion en montée vers son niveau de vol .

(ou de descente³). Nous pouvons découper artificiellement cette trajectoire en trois grandes phases, présentant du point de vue de la prévision, des difficultés différentes :

- la trajectoire « initiale » (de 0 à 300 secondes environ). La difficulté réside dans l'absence de passé, et le manque de points pour estimer la trajectoire. L'estimation et donc la prévision peuvent y être difficiles, surtout pour un horizon de prévision lointain.
- la « montée » (de 300 à 650 secondes). Nous pouvons nous servir du passé de la trajectoire pour l'estimation. Nous sommes dans une zone dans laquelle nous pouvons à la fois prévoir et comparer les prévisions à la réalité. La difficulté est ici double : il s'agira de tenir compte du passé sans en propager les erreurs, et de prévoir le plus justement l'avenir sans quitter la réalité des données. C'est dans cette zone que nous espérons avoir la plus grande précision.
- la « stabilisation » (de 650 à 750 secondes). La trajectoire change rapidement, nous pouvons toujours effectuer les prévisions sur la base du passé mais celles-ci seront grossièrement fausses tant que quelques points du palier (à 31000 pieds) n'auront pas été observés. Il s'agira de prendre en compte rapidement la variation de la trajectoire, pour l'incorporer dans la prévision et rectifier le tir.

1.4 Fonction réelle ou série temporelle ?

D'un point de vue statistique, nous pouvons voir une trajectoire telle que celle présentée en 1.2, sous deux aspects très différents, menant à des procédures de prévision de nature différente :

- La trajectoire est une fonction du *temps*.
En T_i on observe la fonction inconnue f telle que $Y_i = f(T_i)$. Dans ce cas on aura donc à estimer une *fonction* pour pouvoir prévoir Y à une date future. Les procédures d'estimation de fonctions seront alors nécessaires pour la prévision.

³. En niveau de vol, la trajectoire est connue

- La trajectoire est une fonction de son *passé*.
On observe la série chronologique $Y_i = f(Y_{i-1}, Y_{i-2}, \dots, Y_{i-p})$. Dans ce cas, il conviendra de modéliser le *processus temporel* par une série chronologique. On aura alors à déterminer le nombre de points du passé (c'est à dire le réel p) intervenant dans la réalisation de Y_i ainsi que la forme du processus.

Il ne s'agit donc là que de deux approches pour une même courbe. Nous avons choisi de privilégier la première, et donc l'estimation de la « *fonction trajectoire* », pour deux raisons. D'une part, les outils de modélisation d'une série chronologiques sont plus complexes et plus long à mettre en œuvre, d'autre part, la modélisation d'une série chronologique ne s'applique que pour des séries stationnaires⁴ et la série des Y_i étant croissante, n'est pas stationnaire⁵.

1.5 Estimation paramétrique ou non-paramétrique ?

Au vu de l'abondante littérature⁶ traitant de l'estimation « non-paramétrique », on peut être amené à s'interroger sur l'intérêt de ces nouveaux outils. Ces procédures seront présentées et développées dans le chapitre 4, mais nous pouvons d'ores et déjà en décrire les intérêts principaux :

- L'absence d'hypothèses sur la forme de la fonction à prévoir :

Les méthodes développées dans le chapitre 4 reposent sur l'estimation d'une fonction en tout point de l'espace sur lequel elle est définie, sans hypothèse sur sa forme permettant ainsi d'estimer la valeur au delà de la dernière observation. Cette méthode contrairement aux méthodes classiques (voir chapitre 2), ne repose donc pas sur un « *modèle* » (linéaire, polynômial, trigonométrique, etc), et est donc exempte de tout *a priori* sur la forme de la trajectoire.

- L'intégration de toute l'information passée :

L'estimation de la fonction ayant engendrée les données observées, qui servira de base à notre prévision, repose sur l'ensemble des observations passées. Nous verrons par la suite (section 4.2) que le passé récent aura un poids supérieur au passé lointain dans la prévision. Toutefois celui-ci interviendra dans l'estimation de la trajectoire et donc dans la prévision. La prévision non-paramétrique incorpore, en quelque sorte, les observations de manière plus globale.

- L'automatisation et la robustesse :

Nous nous attacherons à développer une méthode « guidée » par les données, pour l'estimation comme pour la prévision, en outre de part la souplesse d'utilisation de ces méthodes, des critères d'ajustement de la courbe aux données peuvent être développés. Cet aspect est particulièrement important dans le contexte d'évitement développé au sein du LOG. D'autre part, les trajectoires d'avions connaissent de brusques et imprévisibles⁷ variations. Les méthodes « classiques » ne peuvent que très rarement s'ajuster à une rupture de la trajectoire avec le modèle proposé. L'absence de modèle est donc, une nouvelle fois, un « plus » appréciable.

Nous développerons cependant des procédures paramétriques classiques dans les chapitres 3 et comparerons sur la base de nos prévisions les deux approches, paramétriques et non-paramétriques.

4. Une série $(Y_i)_{i \in \mathbb{N}}$ est stationnaire ssi : $\forall n, \forall p$ la loi de $(Y_1, Y_2, \dots, Y_n) =$ la loi de $(Y_{p+1}, Y_{p+2}, \dots, Y_{p+n})$. En particulier on doit avoir que $\mathbf{E}[Y_n] = \text{Constante}, \forall n$.

5. Nous verrons toutefois dans la section 5.1.1 que nous pouvons « filtrer » la série des Y_i , et obtenir une nouvelle série stationnaire.

6. Voir par exemple Bosq et Lecoutre [2], Collomb [5], Eubank [8], Härdle [15], Vieu [31] ou Linton [16] pour une revue, ainsi que Fan et Muller [9] ou Chen et Härdle [4] pour des travaux sur les séries temporelles.

7. Hors information supplémentaire

Chapitre 2

Méthodes de régression « classiques »

« Mais de tous ces principes, celui des moindres carrés est le plus simple : avec les autres nous serions conduits aux calculs les plus complexes. »

K. F. Gauss, (1809)

2.1 Un peu de théorie

Considérons un couple de v.a.r. (X, Y) ; lorsque l'on pense que le phénomène aléatoire représenté par X peut servir à expliquer, ou prédire celui représenté par Y (causalité, concomitance, etc.), on est conduit à rechercher une formule de liaison entre une variable à *expliquer* Y par une (ou des) variable(s) *explicative(s)* X , du type :

$$Y = f(X) + \varepsilon$$

On cherche bien évidemment à estimer cette fonction f par \hat{f} et on cherche $\hat{Y} = \hat{f}(X)$, le meilleur possible c'est à dire :

- tel que l'on ait pas de « biais », soit : $E[Y - \hat{Y}] = 0$
- minimisant l'erreur d'estimation, (erreur quantifiée par la variance des résidus ε), c'est à dire réalisant :

$$\hat{f} = \text{Arg min}_f \sum_{i=1}^n E[(Y - f(X))^2] \quad (2.1)$$

Par définition l'espérance conditionnelle, (notée $f(x) = E[Y | X]$) est la projection orthogonale de Y sur l'espace W_X engendré par X et les constantes et réalise de fait le minimum de cette quantité.

La qualité de l'approximation de Y par $E[Y | X]$ est alors mesurée par le critère du R^2 défini comme le rapport :

$$R^2 = \frac{\text{Var}(E[Y | X])}{\text{Var}(Y)} = \frac{\text{Variance Expliquée}}{\text{Variance Totale}} = (\cos(\Theta))^2$$

Dans cette expression Θ est défini conformément à la figure 2.1 comme l'angle entre Y et sa projection.

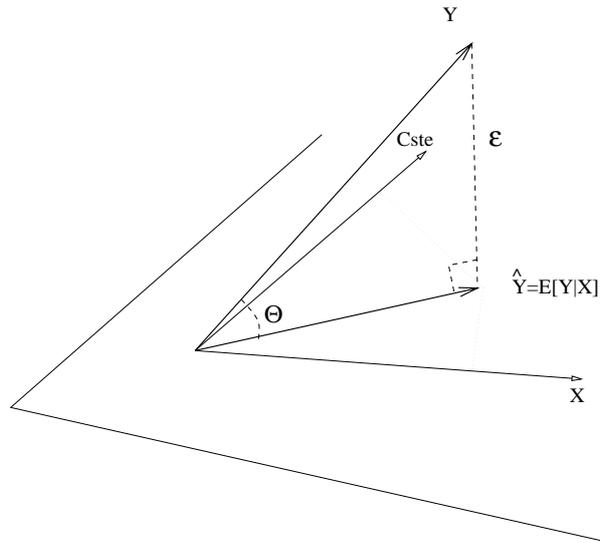


FIG. 2.1 - $\hat{Y} = E[Y | X]$ est la projection de Y sur W_X

La fonction $f(x) = E[Y | X = x]$ est la fonction de régression de Y sur X . On pose alors :

$$Y = E[Y | X] + \varepsilon$$

avec :

- $E[\varepsilon] = 0$ car $E[Y] = E[E[Y | X]]$
- ε non corrélé linéairement avec X et avec $E[Y | X]$ car il est orthogonal au plan engendré par X et les constantes

Remarque :

$\varepsilon = Y - E[Y | X]$ est orthogonal à l'espace engendré par X , on peut donc appliquer le théorème de Pythagore¹ (voir figure 2.1) :

$$Var(Y) = Var(E[Y | X]) + Var(\varepsilon) \tag{2.2}$$

i.e.

$$variance\ totale = variance\ expliquée + variance\ résiduelle$$

2.2 Régression linéaire simple

2.2.1 Méthode

D'un point de vue pratique, l'estimation de la régression linéaire par "les moindres carrés" repose sur des idées simples dues à Gauss lui-même.

On observe $(X_i, Y_i)_{i=1, \dots, n}$ n réalisations d'un couple de variables aléatoires (X, Y) . On cherche la fonction linéaire

$$f(x) = \alpha + \beta \cdot x$$

1. On peut interpréter la variance comme la norme euclidienne des variables aléatoires dans l'espace \mathbf{R}^p , de même la covariance entre deux variable s'écrit alors comme le produit scalaire.

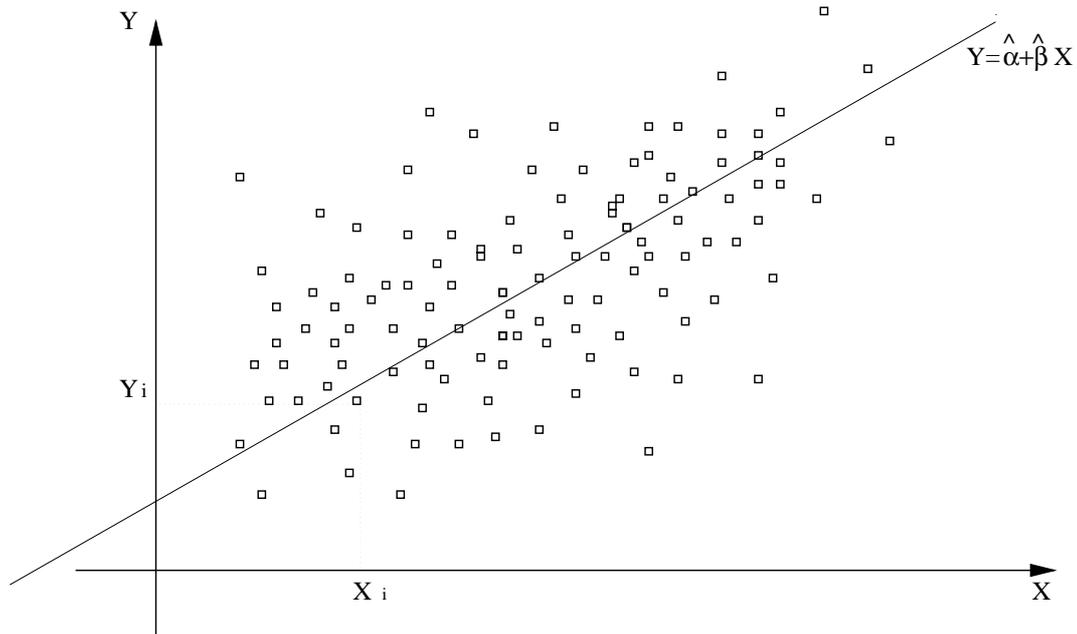


FIG. 2.2 - Nuage de points et droite de régression

“la plus proche possible” de ce nuage de points (X_i, Y_i) au sens quadratique, c’est à dire que l’on cherche $(\hat{\alpha}, \hat{\beta})$ tels que :

$$(\hat{\alpha}, \hat{\beta}) = \text{Arg min}_{(\alpha, \beta)} \sum_{i=1}^n [Y_i - (\alpha + \beta \cdot X_i)]^2$$

de telle sorte que la somme des écarts des (X_i, Y_i) a la droite $\hat{\alpha} + \hat{\beta} \cdot x$ soient “moindres” .

Il ne s’agit donc ici que d’une autre formulation de la projection de la variable Y sur l’espace engendré par la variable X et les constantes. Autrement dit, on impose une forme particulière (linéaire) pour l’espérance conditionnelle. On suppose en fait que $E[Y | X] = \alpha + \beta \cdot X$, α et β étant bien sûr inconnus, et l’on cherche à estimer ces paramètres pour estimer la régression. Le problème de minimisation exprimé de manière générale en 2.1 sur un espace fonctionnel se ramène donc à un problème de minimisation sur un espace réel via l’estimation de deux paramètres α et β .

Pour estimer ces paramètres, posons $F(\alpha, \beta) = \sum_{i=1}^n [Y_i - (\alpha + \beta \cdot X_i)]^2$. On a alors à résoudre le système :

$$\begin{aligned} \frac{\partial F(\alpha, \beta)}{\partial \alpha} = 0 &\Leftrightarrow \hat{\alpha} = \bar{Y} - \hat{\beta} \cdot \bar{X} \\ \frac{\partial F(\alpha, \beta)}{\partial \beta} = 0 &\Leftrightarrow \hat{\beta} = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{Cov}(X, Y)}{S_x^2} \end{aligned}$$

La droite de régression (voir figure 2.2) a pour équation :

$$Y = \hat{\alpha} + \hat{\beta} \cdot x$$

Supposons que les variables X et Y soient centrées $((\bar{X}, \bar{Y}) = (0, 0))$, on a alors :

$$\begin{cases} \hat{\alpha} = 0 \\ \hat{\beta} = \frac{\sum_{i=1}^n Y_i X_i}{\sum_{i=1}^n X_i^2} \end{cases}$$

Remarques :

- Elle passe par le centre de gravité du nuage (\bar{X}, \bar{Y})
- Dans ce cas, $\hat{\beta}$ correspond au coefficient de corrélation de X et Y :

$$\hat{\beta} = \frac{\frac{1}{n} \sum_{i=1}^n Y_i X_i}{\frac{1}{n} \sum_{i=1}^n X_i^2}$$

- Puisque les (Y_i, X_i) sont des réalisations de variables aléatoires, $\hat{\alpha}$ et $\hat{\beta}$ (de même que \bar{X}, \bar{Y} , etc..) sont des réalisations de variables aléatoires.

2.2.2 Qualité de l'ajustement

Posons $\hat{Y}_i = \hat{\alpha} + \hat{\beta} \cdot X_i$, on peut effectuer la décomposition :

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

On voit facilement que $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})(\hat{Y}_i - Y_i) = 0$, et donc :

$$\begin{array}{lcl} \sum_{i=1}^n (Y_i - \bar{Y})^2 & = & \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\ \text{Somme des carrés} & = & \text{Somme des carrés} + \text{Somme des carrés} \\ \text{totale} & & \text{résiduelle} \quad \text{expliqués} \end{array}$$

Le rapport défini par

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

détermine la part d'information expliquée par X .

Plus ce terme est proche de 1, meilleure est l'approximation linéaire.

2.2.3 Tests

Si l'on suppose que les résidus suivent une loi gaussienne, de variance σ^2 inconnue, alors étant donné que les estimateurs $\hat{\alpha}$ et $\hat{\beta}$ sont des combinaisons linéaires de ε_i , ils suivent également des lois gaussiennes et on a :

$$\hat{\alpha} \sim N \left(\alpha, \sigma^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n X_i^2} \right) \right) \quad (2.3)$$

et

$$\hat{\beta} \sim N \left(\beta, \sigma^2 \left(\frac{\sigma^2}{\sum_{i=1}^n X_i^2} \right) \right) \quad (2.4)$$

Le problème est que σ^2 est inconnu ! On sait cependant que :

$$\varepsilon_i = Y_i - (\hat{\alpha} + \hat{\beta} \cdot X_i) \text{ et que } \varepsilon_i \sim N(0, \sigma^2) \quad (2.5)$$

On peut alors « ruser » et essayer d'éliminer σ^2 en effectuant le rapport de deux expressions utilisant similairement ce terme. En utilisant les expressions 2.5 on a :

$$\frac{\sum_{i=1}^n \varepsilon_i^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sigma^2} \sim \chi_{n-2}^2$$

D'autre part, si l'hypothèse $H_O : \beta = 0$ est vraie, et dans ce cas seulement, on a :

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sigma^2} \sim \chi_1^2$$

et donc on peut tester la significativité de la régression, i.e. $H_O : \beta = 0$, sans connaître σ^2 , en examinant si le rapport

$$\frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

dont on connaît la loi² correspond à une valeur admissible à 95% pour cette loi. C'est le « test de Fisher », qui permet de décider si l'hypothèse $H_O : \beta = 0$ peut être statistiquement rejetée, c'est à dire si la régression est significative³.

2.3 Prévisions et intervalles de confiance

Une fois la forme de la fonction choisie, ici linéaire, et les paramètres estimés, la prévision consiste simplement à se déplacer sur le modèle estimé, c'est à dire ici le long de la droite de régression. On a ainsi une estimation \hat{Y}_0 de la valeur donnée par le modèle pour toute valeur de x_0 de la variable X , y compris hors du dernier point observé :

$$\hat{Y}_0 = \hat{f}(x_0) = \hat{\alpha} + \hat{\beta} \cdot x_0$$

Bien entendu, la qualité de cette prévision dépendra de la qualité de l'ajustement (le R^2) et de la fiabilité de l'estimation des paramètres $\hat{\alpha}$ et $\hat{\beta}$, fiabilité que nous pouvons tester (test de Fisher).

Il est clair cependant que la prévision « brute » ne suffit généralement pas au praticien, l'incertitude liée à cette prévision est un élément qui peut avoir une importance capitale. Nous nous proposons donc de détailler succinctement le calcul des intervalles de confiance liés à l'estimation et à la prévision⁴.

Si l'on suppose que les résidus suivent une loi gaussienne, de variance σ^2 inconnue, alors $\hat{\alpha}$ et $\hat{\beta}$ suivent également des lois gaussiennes conformément à 2.3 et 2.4. On peut alors calculer la variance, c'est à dire l'incertitude attachée à \hat{Y}_0 par :

2. Il s'agit d'une loi $F(1, n - 2)$ définie comme le rapport de lois χ^2 , c'est à dire $\frac{\chi_1^2}{\chi_{n-2}^2}$, qui correspond bien aux rapports des lois des deux termes exprimés.

3. Voir également le test de Student décrit dans les ouvrages statistiques cités en référence.

4. Nous laissons le lecteur intéressé par de plus amples informations, se reporter à la lecture d'ouvrages statistiques cités en bibliographie (voir par exemple Saporta [26] ou Wonnacott et Wonnacott [33])

$$\begin{aligned} \text{Var}[\widehat{Y}_0] &= \text{Var}[\widehat{\alpha} + \widehat{\beta} \cdot x_0] \\ &= \text{Var}[\widehat{\alpha}] + x_0^2 \cdot \text{var}[\widehat{\beta}] \end{aligned}$$

En utilisant les variances de $\widehat{\alpha}$ et $\widehat{\beta}$, données par les équations 2.3 et 2.4, on obtient finalement :

$$\begin{aligned} \text{Var}[\widehat{Y}_0] &= \frac{\sigma^2}{n} + x_0^2 \cdot \frac{\sigma^2}{\sum_{i=1}^n X_i^2} \\ &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^n X_i^2} \right) \end{aligned}$$

Puisque $\widehat{\alpha}$ et $\widehat{\beta}$ suivent des lois normales, il en est de même pour \widehat{Y}_0 :

$$\widehat{Y}_0 \sim N \left(\alpha + \beta \cdot x_0, \sigma^2 \left(\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^n X_i^2} \right) \right)$$

Grâce à cette équation, nous pouvons déduire les éléments pour construire un intervalle de confiance pour la « vraie » valeur inconnue $Y_0 = \alpha + \beta \cdot x_0$. Cette vraie valeur a été estimée et l'on a $Y_0 = \widehat{Y}_0 + \varepsilon_0$, donc Y_0 suit également une loi normale. L'utilisation de la formule 2.2 (théorème de Pythagore) nous permet de calculer sa variance :

$$\begin{aligned} \text{Var}(Y_0) &= \text{Var}(\widehat{Y}_0) + \text{Var}(\varepsilon_0) \\ &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^n X_i^2} \right) + \sigma^2 \\ &= \sigma^2 \cdot \left(\frac{1}{n} + \frac{x_0^2}{\sum_{i=1}^n X_i^2} + 1 \right) \\ &= \sigma_0^2 \end{aligned}$$

D'où :

$$Y_0 \sim N \left(\widehat{Y}_0, \sigma_0^2 \right)$$

La connaissance de la loi de Y_0 , nous permet ainsi de déterminer quelles sont les valeurs « probables » de sa réalisation, autour de sa valeur moyenne \widehat{Y}_0 . Ainsi notre estimation de \widehat{Y}_0 et de la variance σ_0^2 nous permet de déterminer avec un seuil de confiance η , que la vraie valeur Y_0 se situe dans l'intervalle de confiance⁵ :

$$Y_0 \in \left[\widehat{\alpha} + \widehat{\beta} \cdot x_0 \pm t_\eta \cdot \sigma_0 \right]$$

Dans cette expression le nombre t_η représente la valeur au delà de laquelle une variable suivant une loi normale⁶ $N(0, 1)$ a une probabilité de $1 - \eta$ de se réaliser (en valeur absolue).

Remarques :

- La variance σ^2 est inconnue mais on peut lui substituer un estimateur, S_n et obtenir le même type de résultat, la loi de Y_0 et donc le réel t_η étant légèrement modifiés lors de cette substitution.

5. En général $\eta = 95\%$.

6. En effet si $Y_0 \sim N(\widehat{Y}_0, \sigma_0^2)$ alors la variable $\frac{Y_0 - \widehat{Y}_0}{\sigma_0} \sim N(0, 1)$.

- Il est clair que l'intervalle de confiance associé à Y_0 sera d'autant plus petit que l'incertitude sur l'estimation, σ_0 , sera, elle aussi, petite. En outre, toutes choses égales par ailleurs, si n augmente σ_0 diminue et cet intervalle se réduit.
- Les intervalles de confiance sont basés sur l'hypothèse de normalité des résidus, c'est à dire sur la « bonne » répartition des erreurs ε_i entre réalité et modèle. Cette hypothèse peut sembler arbitraire et ne doit pas être oubliée lors de l'utilisation de cet intervalle de confiance. Toutefois, si l'on observe un grand nombre de points, le *théorème central limite* nous indique que cette hypothèse est réalisée.

Les intervalles de confiance qui accompagnent les prévisions présentées dans le chapitre 3 sont basés sur ces principes, et sur les formules que nous venons de donner.

Chapitre 3

Prévisions « classiques »

Au vu d'une trajectoire « type » comme celle de la figure 1.2, il semble évident que seule une modélisation non-linéaire peut nous permettre d'obtenir une prévision correcte¹. Nous avons donc testé des modèles non-linéaires « classiques », basés sur des familles paramétrées de fonctions de la forme $f(t, \theta)$, où θ est un paramètre, éventuellement vectoriel, à estimer. La méthode d'estimation est celle des « moindres carrés » que nous décrirons succinctement et qui repose sur les mêmes méthodes que celles présentés dans le chapitre précédent.

3.1 Estimation paramétrique

Sur la base des $(Y_i)_{i=1,\dots,n}$ observés aux temps $(T_i)_{i=1,\dots,n}$, on peut estimer le paramètre θ tel que la famille $f(t, \theta)$ « colle » le mieux aux données observées. L'estimateur $\hat{\theta}_n$ doit pour cela réaliser le minimum de la distance entre le modèle $f(T_i, \theta)$ et les Y_i observés, soit :

$$\hat{\theta}_n = \text{Arg} \min_{\theta} \sum_{i=1}^n (Y_i - f(T_i, \theta))^2 \quad (3.1)$$

C'est le principe général des moindres carrés, que la fonction $f(t, \theta)$ soit linéaire ou non.

Nous proposons ici quelques modèles non-linéaires, estimés sur la base d'une trajectoire donnée par le simulateur développé au CENA, et représentés avec leurs intervalles de confiance dans les figures 3.2, 3.3 et 3.4 pour $T_n = 530$. Nous nous plaçons ainsi dans un contexte assez favorable puisqu'une grande partie de la trajectoire est observée, permettant un bon ajustement. Dans ces figures la trajectoire réelle est représentée par des points bleus, la trajectoire prévue, ainsi que les intervalles de confiance à 95%, sont des lignes roses. Ces estimations ont été réalisées à l'aide du logiciel statgraphics.

Une comparaison de ces différents modèles sur la base de leurs prévisions est donnée par la figure 3.5 et permet d'évaluer ces modélisations en fonction de notre critère principal : l'erreur de prévision.

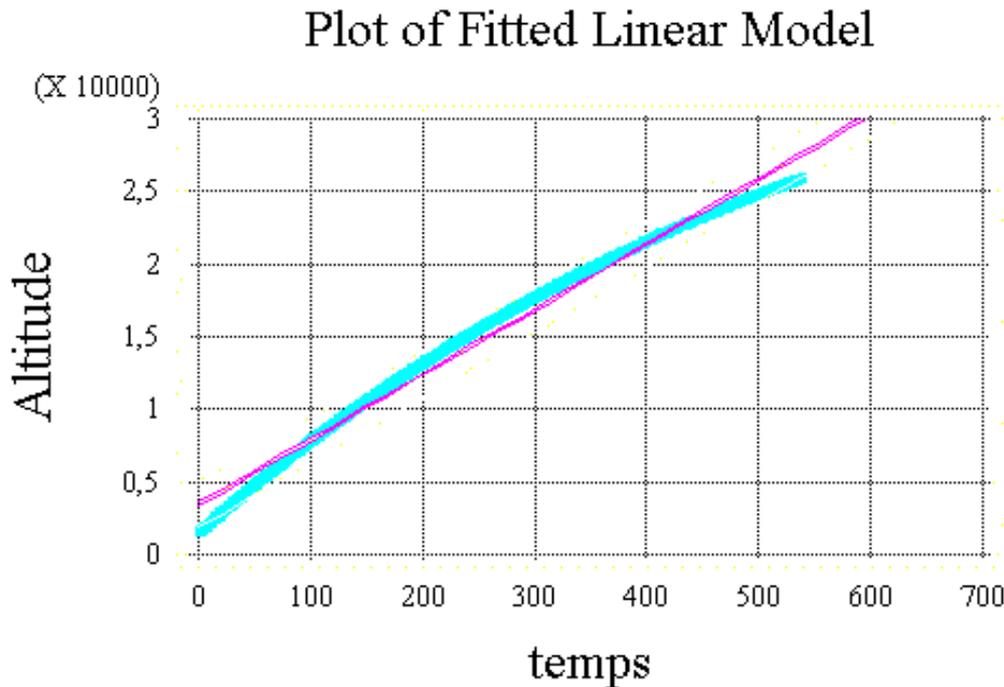
3.1.1 Modèle linéaire

Par acquis de conscience et pour mieux visualiser les apports d'un modèle non-linéaire nous donnons ici (figure 3.1), l'estimation linéaire de la trajectoire.

Le modèle retenu ici est donc simplement :

$$\text{Altitude} = \alpha + \beta \cdot \text{temps}$$

1. L'estimation d'un modèle linéaire viendra d'ailleurs confirmer cette intuition.

FIG. 3.1 - *Modèle de prévision linéaire*

L'estimation des paramètres inconnus α et β par la minimisation des moindres carrés, c'est à dire en utilisant 3.1, nous donne les valeurs : $\hat{\alpha} = 3438,77$ et $\hat{\beta} = 44,78$ et un ajustement calculé² par le R^2 de 98,69 %.

Ce modèle ne sera donc pas retenu pour notre problème et nous ne l'évoquerons que comme référence dans la suite de ce travail.

Remarque :

Ce modèle est trop simpliste pour pouvoir répondre à notre besoin de prévision. Cependant, une approche « locale » consistant à estimer un modèle sur la base des p dernières valeurs observées peut s'avérer plus efficace. L'idée sous-jacente consiste en fait à linéariser localement la trajectoire, c'est à dire calculer une sorte de « tangente » et l'extrapoler. Le problème de ce type de méthode est quelle est par essence locale³, de plus le choix du nombre de points pertinents pour le calcul de cette « tangente » est délicat. Nous retrouvons là le problème évoqué en introduction, à savoir du choix du « poids » du passé (voir section 1.1, ou B. Coutrot [6] pour une analyse plus précise).

Cette remarque peut se généraliser à l'ensemble des modèles non-linéaires proposés ci-dessous, pour lesquels une version locale pourrait être envisagée.

3.1.2 Modèle « racine »

En raison de la forme de la trajectoire observée nous avons testé le modèle :

$$\text{Altitude} = \alpha + \beta \cdot \sqrt{\text{temps}}$$

donnant comme paramètres optimaux : $\hat{\alpha} = -4282,84$ et $\hat{\beta} = 1278,32$ et un R^2 de 98,9 %.

Ce modèle fournit le meilleur ajustement sur la base des données observées. La courbe prévue, en rose sur la figure 3.2, au delà du dernier point observé, semble s'infléchir un peu trop pour

2. Le R^2 mesure le rapport entre la variance expliquée par le modèle sur la variance totale, il est l'indicateur de la qualité de l'ajustement. Ce critère ne détermine cependant pas la qualité de la prévision. On trouvera de plus amples informations sur le bon usage du R^2 dans Saporta [26], ou dans n'importe quel ouvrage de statistique.

3. C'est une des raisons pour laquelle nous n'avons pas testé cette méthode « d'approximation locale ».

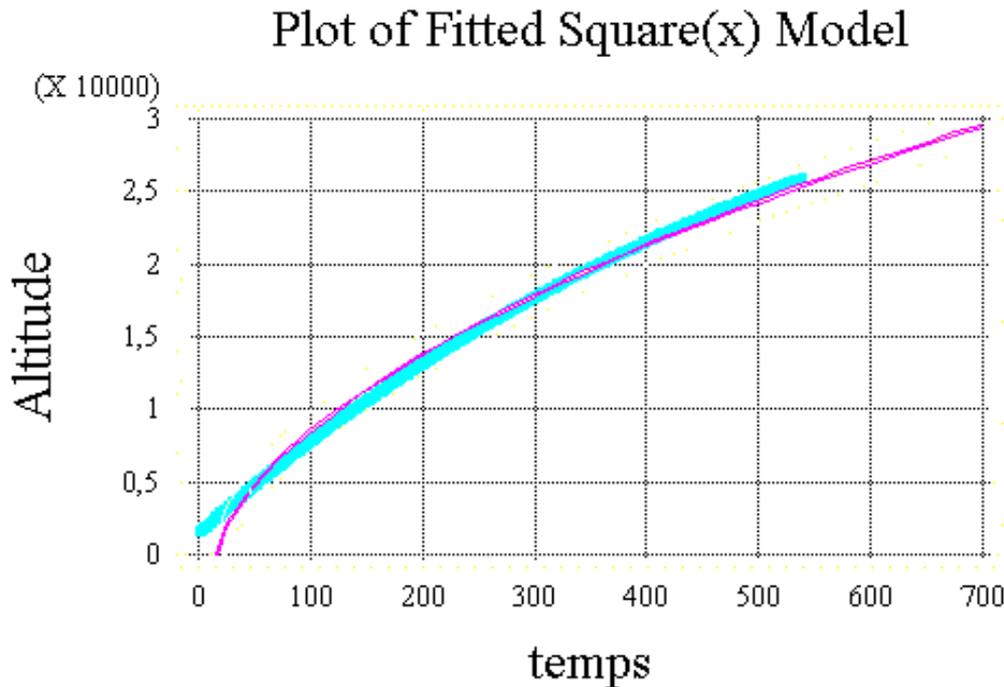


FIG. 3.2 - Modèle de prévision « racine »

pouvoir donner une bonne prévision. Toutefois l'examen de la figure 3.5, montre que les erreurs de prévision restent en dessous de 500 pieds, quel que soit l'horizon de prévision. Il semble donc qu'un modèle de type « puissance » du temps soit approprié pour l'approximation de la trajectoire.

3.1.3 Modèle « multiplicatif »

Une autre façon de modéliser la trajectoire en s'affranchissant de l'additivité consiste en un modèle multiplicatif, ce modèle généralise également le modèle « racine » précédent.

$$\text{Altitude} = \alpha \cdot \text{temps}^\beta$$

L'estimation donne : $\hat{\alpha} = 438,72$, $\hat{\beta} = 0,643$, et un R^2 de 97,69 %, l'ajustement représenté en 3.3 est correct.

Ici, les prévisions semblent bonnes, meilleures en tout cas que pour les autres modèles, si l'on en juge par la figure 3.5.

Le coefficient β a une valeur proche, mais supérieure à $\frac{1}{2}$, et le modèle a un meilleur comportement à long terme que le modèle « racine ». Pour t grand on a en effet $\sqrt{t} \leq t^{0,64}$, la courbe s'infléchit donc moins pour les prévisions à long terme. Ce résultat confirme donc les bons résultats du modèle « racine », en tout état de cause une puissance du temps n'est pas forcément une mauvaise approche pour la modélisation de trajectoires.

Ce résultat est cependant à prendre avec prudence puisque l'on dispose ici d'un nombre d'observations élevé⁴. Nous verrons dans la section suivante 3.2 que ce modèle est trop rigide pour pouvoir s'adapter aux contraintes de prévision avec peu de points observés.

Remarque:

Il est à noter que si l'on prend le logarithme de part et d'autre de l'équation définissant le modèle multiplicatif, on tombe sur un modèle linéaire dans lequel les variables sont transformées logarithmiquement.

4. Une observation toutes les secondes pendant 530 secondes.

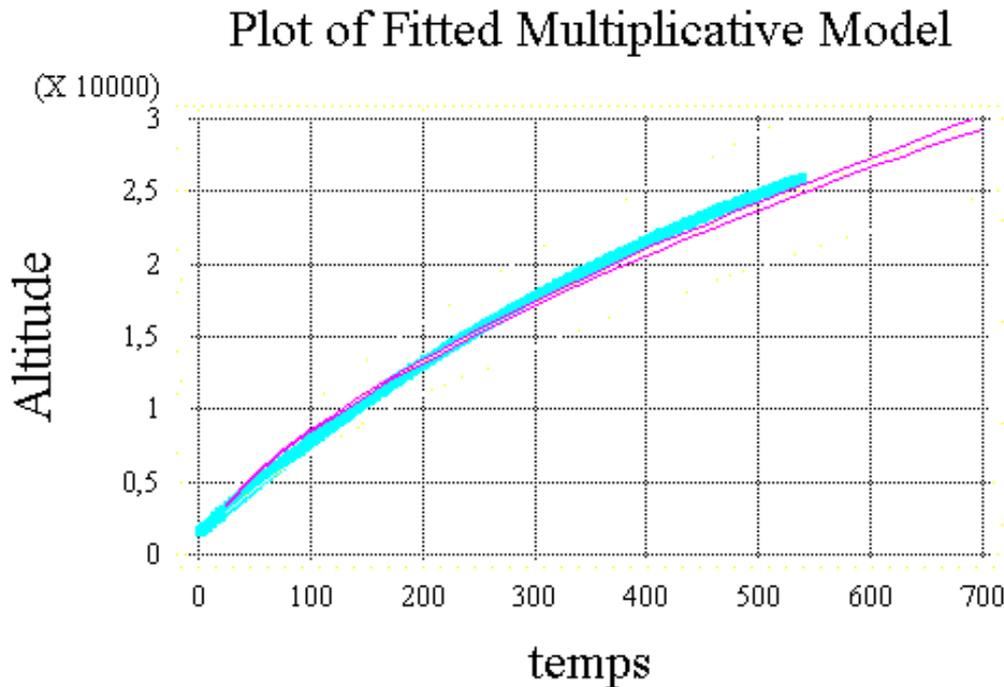


FIG. 3.3 - Modèle de prévision multiplicatif

$$\log(\text{Altitude}) = \alpha' + \beta \cdot \log(\text{temps})$$

Il en est de même pour tous les modèles proposés ici, qui ne sont que des transformations du modèle linéaire.

3.1.4 Modèle « Logarithmique »

Une nouvelle fois l'aspect visuel, sans élément théorique, de la courbe nous inspire une modélisation « ad hoc » logarithmique :

$$\text{Altitude} = \alpha + \beta \cdot \log(\text{temps})$$

La minimisation par moindres carrés nous donne : $\hat{\alpha} = -19298,2$ et $\hat{\beta} = 6574,66$ et un R^2 de 83,76 . Soit le moins bon de tous les ajustements proposés par les différents modèles.

Ce modèle « sous-estime » énormément la trajectoire. L'observation de la courbe prévue 3.4 et de l'erreur de prévision 3.5 renforce le sentiment de modélisation inadéquate.

3.2 Prévisions opérationnelles

De manière quelque peu surprenante, les modèles non-linéaires « multiplicatif » et « racine » estiment relativement bien la trajectoire, l'étude des erreurs de prévision données par la figure 3.5 nous permet d'espérer des prévisions fiables. Les résultats obtenus pour un nombre important de données, et pour une longue série sont assez bons puisque l'erreur de prévision pour ces deux modèles ne dépasse pas 500 pieds. De par leur simplicité⁵, ces modèles permettent également le calcul d'intervalles de confiance autour de la valeur prédite, comme nous pouvons le voir sur les figures 3.2 et 3.3.

5. Ils ne sont l'un et l'autre que des transformations du modèle linéaire!

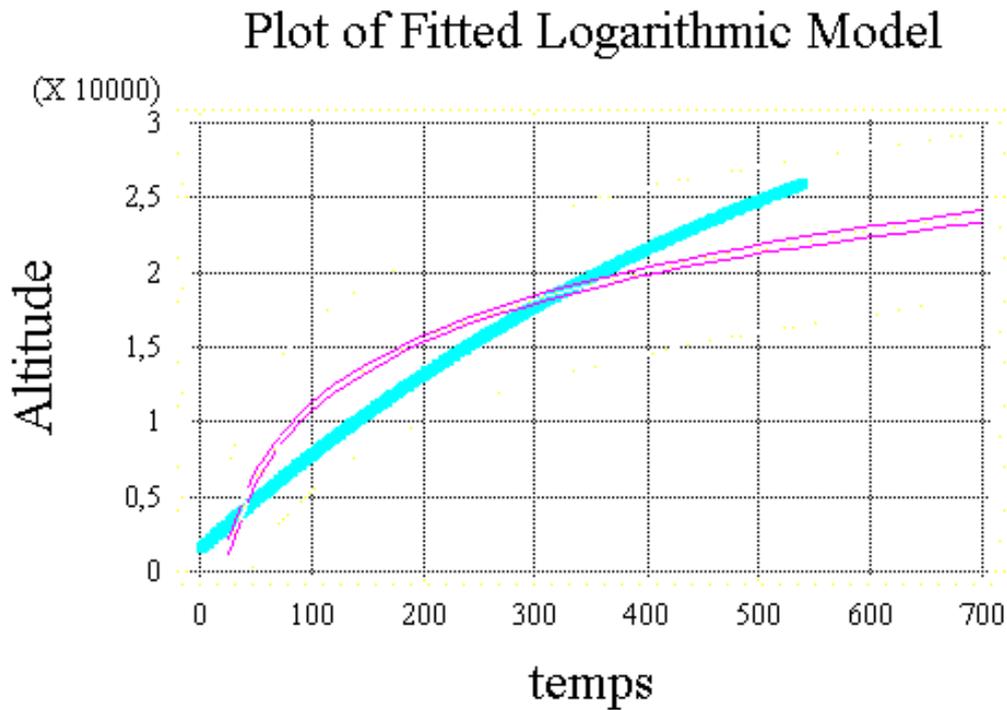


FIG. 3.4 - *Modèle de prévision logarithmique*

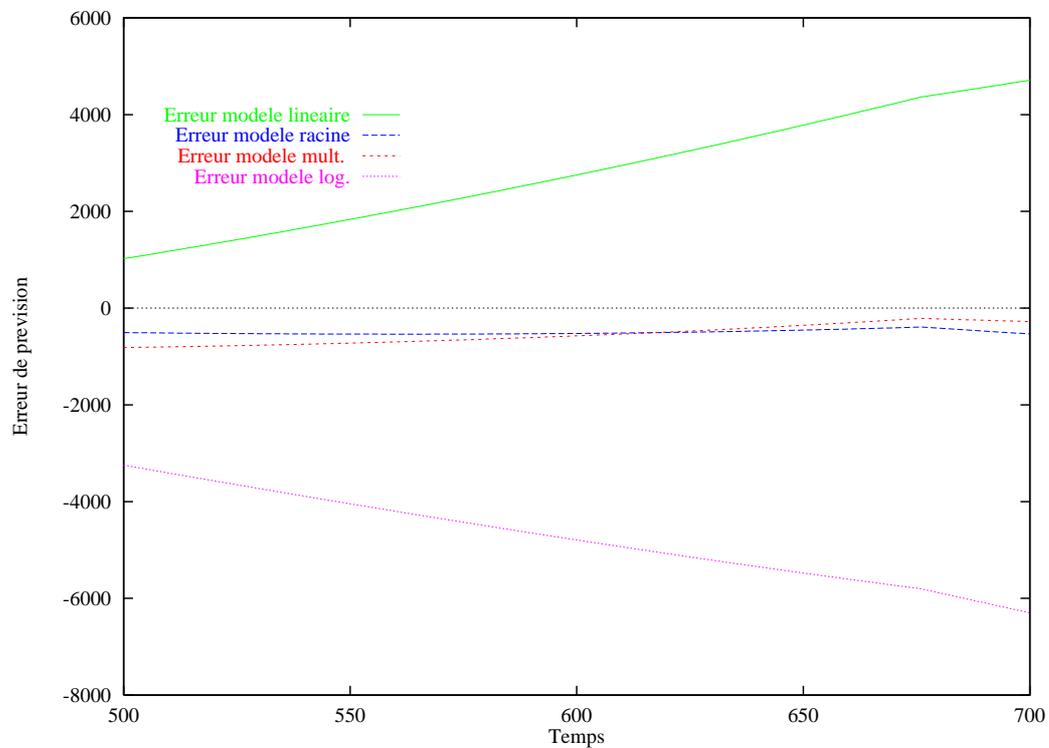


FIG. 3.5 - *Erreurs de prévision pour les différents modèles.*

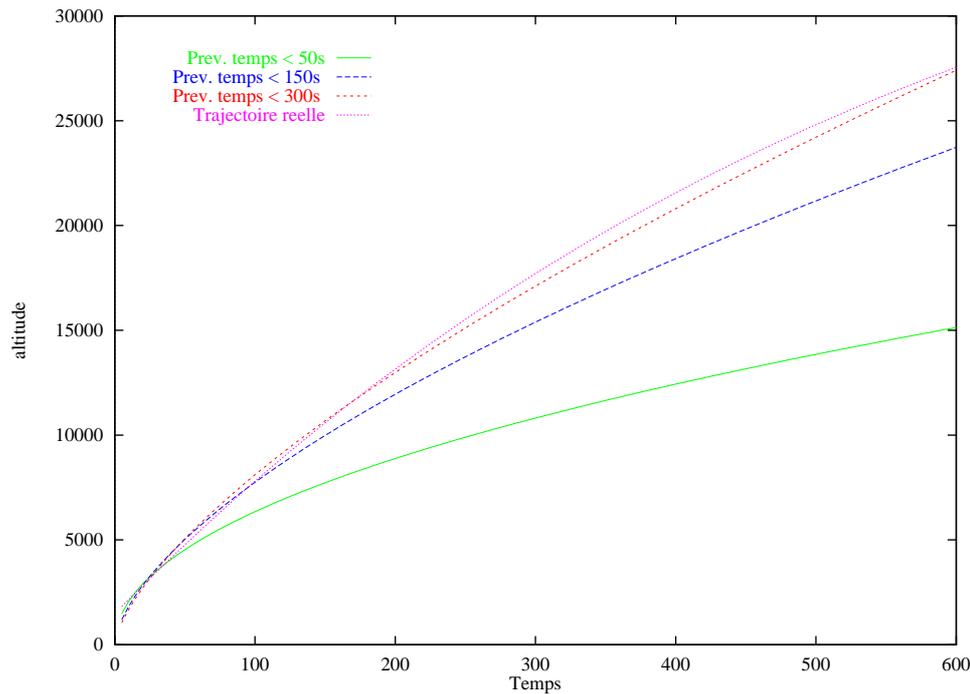


FIG. 3.6 - Prévisions pour différentes durées d'observation (Modèle « multiplicatif »).

Nous avons donc utilisé ces modèles en situation réelle, c'est à dire en nous plaçant à $T = 50s$, $T = 150s$ et $T = 300s$ et en estimant les paramètres α et β à partir des points observés avant ces dates. Les trajectoires estimées et prévues ont été représentées jusqu'à $T = 600$ de manière à nous rendre compte des erreurs de prévision à court et à long terme. Par souci de cohérence, nous avons effectué ces prévisions sur la base des trajectoires issues du générateur du CENA. Dans ce cadre nous nous sommes restreint à n'observer qu'un point toutes les 5 secondes. Les trajectoires prévues sont représentées dans les figures 3.6 et 3.7 pour les modèles « multiplicatif » et « racine » respectivement.

Les résultats sont assez mitigés :

- On remarque que ces modèles sous-estiment la trajectoire réelle quelle que soit la configuration⁶. Le modèle « multiplicatif » se comporte toutefois mieux que « racine » avec des écarts se réduisant peu à peu au cours du temps (et lorsque le nombre d'observations augmente).
- La prévision peu de temps après le début de l'observation ($T = 50s$), est assez catastrophique dans les deux modélisations, y compris à court terme ou l'on observe une divergence très importante de la courbe prévue par rapport à la trajectoire réelle.
- Par contre les prévisions au temps $T = 300s$, du modèle « multiplicatif », donnent de bons résultats, surtout à long terme. Les erreurs de prévision restent cependant de l'ordre de 1000 pieds. En fin d'observation ($T = 600s$), alors que l'avion est à 100 secondes et 3000 pieds de sa stabilisation en croisière, la trajectoire réelle s'infléchit, ce qui explique que le modèle « recolle » à la réalité, ce qui semble inespéré dans d'autres configurations.
- Les intervalles de confiance, pour la prévision, (non représentés sur les graphiques 3.6 et 3.7), ne contiennent que rarement la trajectoire réelle, ce qui rend la prévision peu fiable.

6. Ceci est certainement dû à l'estimation des moindres carrés : l'estimateur de β agit en effectuant une somme pondérée des observations, une sorte de « barycentre ». La trajectoire étant croissante, la courbe prévue est donc « tirée » vers le bas par les poids des observations passées, de valeurs plus faibles.

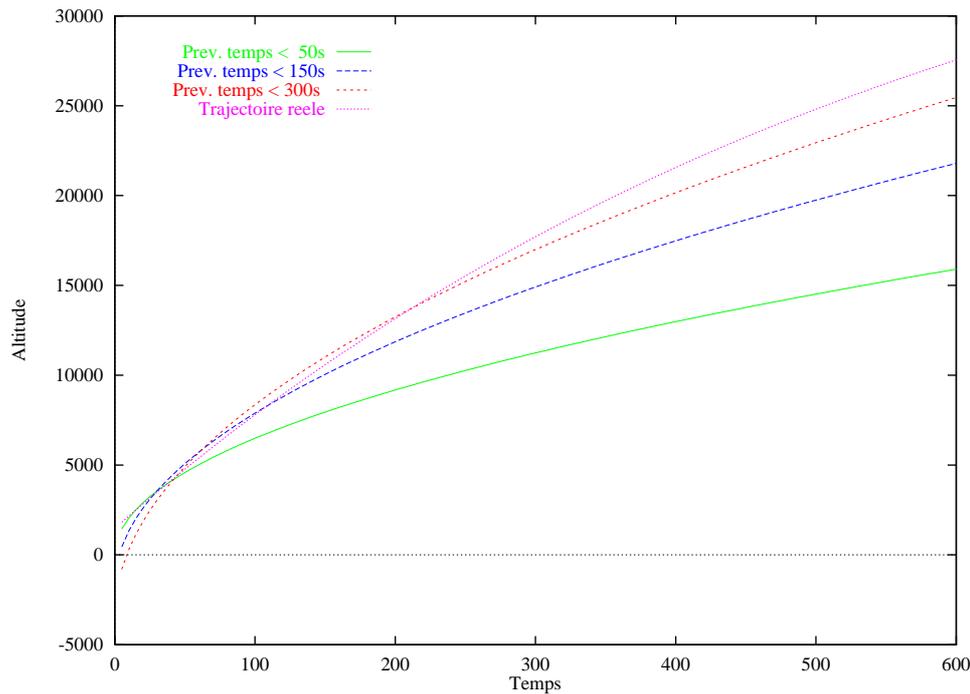


FIG. 3.7 - Prévisions pour différentes durées d'observation (Modèle « racine »).

3.3 Bilan

Il existe de nombreuses autres modélisations pour des courbes de la forme de celles étudiées. On pourrait par exemple envisager d'autres modèles non-linéaires plus sophistiqués comme des modèles polynômiaux :

$$\text{Altitude} = \alpha + \beta \cdot \sqrt{\text{temps}} + \gamma \cdot \text{temps} + \delta \cdot \text{temps}^2 + \dots$$

ou des modèles mixtes du type :

$$\text{Altitude} = \alpha + \beta \cdot \sqrt[4]{\text{temps}} + \gamma \cdot \log(\text{temps}) + \delta \cdot \text{temps}^n$$

Les variations sur ce thème sont, hélas, infinies. Quelle que soit la modélisation choisie, nous sommes limités à l'estimation, plus ou moins complexe, de paramètres prédéfinissant la forme de la fonction à estimer, en l'occurrence la trajectoire.

Outre les problèmes d'estimation des paramètres⁷ dans des équations non-linéaires comme celles-ci, le problème du choix « ad hoc » de la modélisation s'avère crucial. En effet, nous ne savons pas répondre à la question :

Comment trouver une bonne modélisation ?

L'approche proposée ici est très réductrice puisque l'on cherche à approcher une courbe originale, la trajectoire - c'est à dire un élément d'un espace fonctionnel- par une fonction connue, le modèle - c'est à dire un élément d'une famille paramétrée. La dimension de l'espace de recherche est donc trop faible pour approcher convenablement la fonction inconnue. Les erreurs de prévision sont alors beaucoup trop importantes pour être acceptables, même si les intervalles de confiance présentent un attrait considérable.

Nous nous proposons donc de contourner ce problème dans le chapitre suivant par l'utilisation de méthodes non-paramétriques ne reposant sur aucun modèle préétabli.

⁷ Les variables temps , $\log(\text{temps})$ et temps^n étant très fortement corrélées, le problème d'identification des paramètres peut être important, et l'estimation faussée.

Chapitre 4

Méthodes non-paramétriques

Ce chapitre se veut une introduction, théorique et pratique, aux estimateurs non-paramétriques. Après avoir présenté succinctement différents estimateurs fonctionnels nous détaillerons plus particulièrement les propriétés relatives à l'estimateur utilisant la méthode du noyau, et discuterons de son utilisation pour les problèmes de prévision. Nous nous préoccupons également du problème crucial de sélection du paramètre de fenêtre, et de son impact sur la « *douceur* » de l'estimateur fonctionnel. Enfin nous étudierons l'influence de ce paramètre sur la prévision et proposerons des méthodes pour le sélectionner. Le lecteur intéressé pourra trouver en annexe les fondements théoriques ainsi que les principaux résultats de convergence de ces méthodes.

4.1 Introduction

« *On dit qu'un problème d'estimation est non-paramétrique lorsqu'il ne peut pas se ramener au problème de l'estimation d'un élément d'un espace vectoriel de dimension finie* »

Gérard Collomb (1976)

D'après G. Collomb [5], l'estimation non-paramétrique se présente comme une « *non-définition* », rejetant l'estimation d'un paramètre sans que ne soit explicitement exposé l'objet à estimer. Dans cette approche non-paramétrique, l'objet d'intérêt est une fonction tout-à-fait générale, appartenant à un espace fonctionnel¹. Dans le cadre de ce travail, et afin de clarifier notre propos, nous entendrons par « non-paramétrique » l'estimation de la fonction f dans le modèle :

$$Y = f(X) + u$$

où (X, Y) est un couple de variables aléatoires définies sur l'espace mesuré $(\mathbb{R}^p \times \mathbb{R}, \mathcal{B}_{\mathbb{R}^p+1}, \lambda)$

et où $E[u | X] = 0$, *presque sûrement*

dans laquelle, ni la forme de la fonction f (fonction de régression), ni la distribution des résidus ne seront spécifiés.

Ceci est la double négation d'un modèle paramétrique où, par exemple, la forme linéaire est imposée et où la distribution des résidus est spécifiée (voir chapitre 3).

Une classification précise entre estimation paramétrique, non-paramétrique, semi-paramétrique et semi-non-paramétrique a été détaillée par M. Delecroix [7]. Cette classification se base sur l'objet d'intérêt de l'estimation, et nous ne la détaillerons pas davantage, laissant le lecteur intéressé se reporter à cette référence.

1. Nous verrons que cette procédure n'exclut cependant pas tout paramètre de l'estimation.

L'accent sera mis dans ce chapitre sur l'importance des choix arbitraires intervenant dans l'estimation non-paramétrique, et en premier lieu, sur le choix du paramètre déterminant le degré de « *douceur* » de l'estimateur non-paramétrique. En effet, « *non-paramétrique* » ne signifie pas absence de paramètre, bien au contraire et un paramètre de lissage interviendra de manière cruciale dans l'estimation. La sélection de ce paramètre dans le cadre de la méthode du noyau de convolution sera étudiée et illustrée afin de mieux cerner l'impact de ce choix sur l'estimateur et donc sur la prévision. Pour la suite de notre travail nous tenterons de déterminer un critère *objectif* face à l'*arbitraire* des choix *ad hoc*.

Comme dans tout travail d'estimation, nous supposons l'observation d'un échantillon i.i.d., noté $(X_i, Y_i)_{i=1, \dots, n}$ d'un couple de variable aléatoires (X, Y) définies sur l'espace réel mesuré $(\mathfrak{R}^p \times \mathfrak{R}, \mathcal{B}_{\mathfrak{R}^{p+1}}, \lambda)$ et de densités jointe et marginales notées respectivement $\varphi(x, y)$ et $\varphi(x), \varphi(y)$.

Afin d'assurer l'existence de $f(x) = E[Y | X = x]$, nous supposons que $E[|Y|] < \infty$. La fonction f n'étant définie sur \mathfrak{R}^p qu'à une équivalence près, nous supposons également qu'il en existe une version continue f . Par convention, on posera $f(x) = 0$ si la densité de X est nulle ($\varphi(x) = 0$). Par soucis de clarté nous nous restreindrons au cas particulier univarié ($p=1$).

4.2 Définition des estimateurs

L'estimation non-paramétrique d'une fonction repose sur l'idée intuitive que l'estimateur $\hat{f}(\cdot)$, qui est donc également une fonction, en un point x doit être « *proche* » de Y_i si x est « *proche* » de X_i . La même propriété se répétant sur l'ensemble des observations, les estimateurs non-paramétriques de la régression s'écriront donc comme des moyennes pondérées des Y_i , la pondération prenant en compte l'éloignement de X_i au point considéré.

La forme générale d'un estimateur non-paramétrique d'une fonction, tel que nous venons de le présenter, sera donc :

$$\hat{f}(x) = \sum_{i=1}^n Y_i \cdot W_m(X_i, x) \quad (4.1)$$

De manière à obtenir une pondération de somme totale unitaire, on posera :

$$W_m(X_i, x) = \frac{w_m(X_i, x)}{\sum_{i=1}^n w_m(X_i, x)}$$

Suivant le type de pondération utilisé, nous obtiendrons différents types d'estimateurs, chacun dépendant d'un paramètre dont le choix permet de déterminer la « *douceur* » de l'estimateur. Parmi les méthodes les plus utilisées nous pouvons citer :

- L'estimateur des k points les plus proches :

$$w_k(X_i, x) = \mathbf{1}_{[\text{les } k \text{ } X_j \text{ les plus proches de } x]}(X_i)$$

Cette méthode propose d'approcher localement $f(x)$ par une moyenne des k points observés « *les plus proches* » du point considéré. Le paramètre k intervenant dans cette définition détermine le nombre de points considérés comme pertinents pour estimer f en x , les autres points ne participant pas au calcul de $\hat{f}(x)$.

Si k est « petit », par exemple pour $k = 3$, l'estimateur connaîtra de fréquentes (et possible-ment importantes) variations, au contraire, pour k « grand » (de l'ordre du nombre total de points), l'estimateur sera « plat » voire constant puisqu'effectuant une moyenne de presque tous les points.

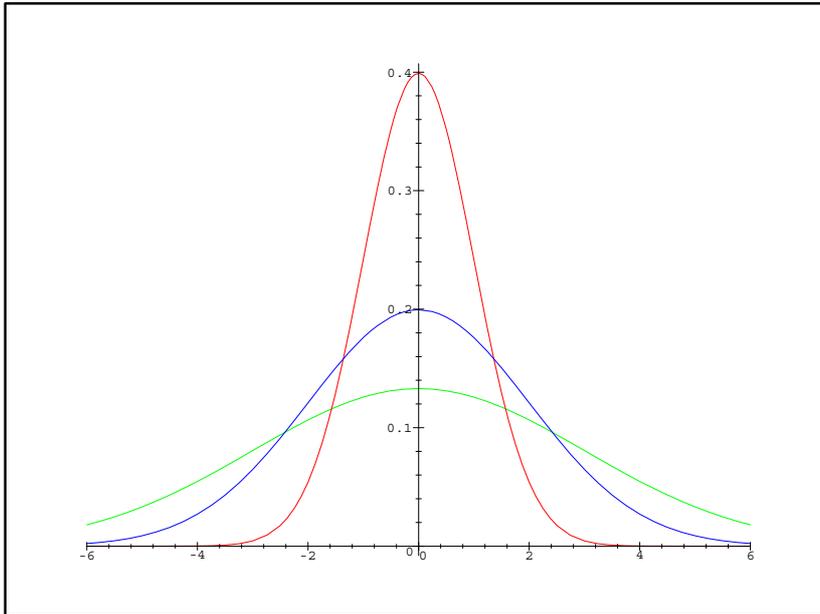


FIG. 4.1 - Exemple de Noyau $K\left(\frac{u}{h}\right)$ pour 3 valeurs de h .

- L'estimateur de la fenêtre mobile (ou régressogramme) est obtenu en posant² :

$$w_h(X_i, x) = \mathbf{1}_{[x-h, x+h]}(X_i)$$

Ici, le paramètre est la largeur $2h$ de la fenêtre³ construite autour de x dans laquelle sont retenues les observations sur lesquelles porte la moyenne des Y_i . Là encore, suivant la valeur du paramètre de lissage h , la fonction estimée prendra des allures plus ou moins « lisses ».

Ces deux estimateurs, s'ils permettent l'estimation d'une fonction en tous points, ne sont pas continus et présentent des sauts dus à la fonction indicatrice intervenant dans leur définition, en outre ils attribuent le même poids à chacun des points retenus, indépendamment de leur proximité au point x .

- L'estimateur du noyau de convolution

$$w_h(X_i, x) = K\left(\frac{X_i - x}{h}\right)$$

où K est une fonction continue prenant en compte l'ensemble des points de l'échantillon, dont la valeur diminue avec l'éloignement entre x et X_i , et où h est un réel positif permettant de *relativiser* cet éloignement. Typiquement, la fonction K choisie donnera plus de poids à des individus proches de x , et moins à ceux qui s'en éloignent. Un exemple de fonction K est la densité gaussienne représenté dans la figure 4.1, pour différentes valeurs de la « fenêtre » h .

Cet estimateur, continu et régulier par construction, est le plus utilisé et le mieux cerné par la littérature. Nous l'utiliserons pour la suite de cette étude et en préciserons les avantages et propriétés dans la section 4.3.

2. la fonction $\mathbf{1}_I(x)$ est l'indicatrice de l'intervalle I : $\mathbf{1}_I(x) = \begin{cases} 1 & \text{si } x \in I \\ 0 & \text{Sinon} \end{cases}$

3. C'est à cette technique que le paramètre de lissage doit le terme usuel de fenêtre.

- L'estimateur du noyau de convolution récursif

$$w_h(X_i, x) = K\left(\frac{X_i - x}{h_i}\right)$$

est une variante de cet estimateur introduit par Devroye et Wagner (voir Härdle [15]), pour lequel h est remplacé par la suite $(h_i)_{i=1, \dots, n}$, h_i variant avec le point X_i considéré.

- L'estimateur des fonctions orthogonales

$$w_M(X_i, x) = \sum_{j=1}^M e_j(X_i) \cdot e_j(x)$$

où la suite de fonctions $(e_j(\cdot))_{j \in \mathbb{Z}}$ constitue une base orthogonale de l'espace Hilbertien $L^2(\mathfrak{R})$ supposé contenir la fonction de régression f .

Il s'agit là d'un estimateur basé sur la projection de f sur une base de fonctions orthogonale, dans laquelle les coefficients directeurs de la base sont estimés. L'entier M représente alors le nombre d'éléments de la base intervenant dans cet estimateur, et donc un paramètre arbitrant la « fidélité aux données ».

- L'estimateur à ondelettes orthogonales

$$w_p(X_i, x) = \sum_{j \in \mathbb{Z}} 2^p \Phi(2^p X_i - j) \cdot \Phi(2^p x - j)$$

où $\Phi(\cdot) \in L^2(\mathfrak{R})$ est issue d'une analyse multirésolution, qui est une décomposition de l'espace $L^2(\mathfrak{R})$ en une suite croissante d'espaces vectoriels fermés V_j . On établit que pour tout entier relatif j , la suite de fonctions $\left(2^{\frac{j}{2}} \Phi(2^j \cdot -k)\right)_{k \in \mathbb{Z}}$ constitue une base orthonormée de V_j . Le paramètre p (ou plutôt 2^p) détermine la « finesse » de cette décomposition (voir Gasquet et Witomsky [10]).

D'autres estimateurs peuvent s'écrire sous cette forme avec une pondération non unitaire et sont proposés par Ullah et Vinod [29].

- L'estimateur de Mack et Müller [19]

$$W_h(X_i, x) = \frac{1}{nh} \frac{1}{\hat{\varphi}(X_i)} \cdot K\left(\frac{X_i - x}{h}\right)$$

où $\hat{\varphi}(X_i) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{X_i - X_j}{h}\right)$ est l'estimateur du noyau de la densité φ au point X_i .

Cet estimateur est l'estimateur du noyau pour lequel le dénominateur est $\hat{\varphi}(X_i)$ au lieu de $\hat{\varphi}(x)$. L'avantage de cet estimateur réside dans le calcul de l'estimateur de la dérivée de $f(x)$ puisque seul le numérateur est alors à calculer.

- L'estimateur de Gasser et Müller [11]

$$W_h(X_i, x) = \frac{1}{h} \int_{A_i} K\left(\frac{u - x}{h}\right) du$$

où $A_i = (S_{i-1}, S_{i+1})$, et $S_i = (X_i + X_{i+1})/2$ est le milieu des points X_i et X_{i+1} . Cet estimateur est développé dans le cadre des régresseurs fixes.

Tous ces estimateurs sont construits et ont été pensés comme des sommes pondérées des Y_i , la pondération étant « régulée » à chaque fois par un paramètre réel (h, k, h_i, M, p , etc) D'autres techniques d'estimation non-paramétrique existent néanmoins et ne sont pas issues de la même logique. Certains estimateurs sont ainsi définis non comme une somme pondérée, mais comme minimisant un critère sur un ensemble de fonctions.

- On peut ainsi citer les « *fonctions splines cubiques* » déterminées comme les fonctions deux fois différentiables minimisant :

$$S_\lambda = \sum_{i=1}^n (Y_i - f(X_i))^2 + \lambda \int (f''(x))^2 dx$$

sur un intervalle compact. La solution unique de cette minimisation est \hat{f}_λ .

On montre que \hat{f}_λ est un polynôme cubique entre deux observations successives (voir Eubank [8]).

Le paramètre λ sert en fait d'arbitrage entre la « *fidélité aux données* », $(\sum_i (Y_i - f(X_i))^2)$ et la « *douceur de la courbe* », $(\int (f''(x))^2 dx)$.

Silverman (voir Härdle[15]), remarque toutefois que l'on peut encore approximer \hat{f}_λ par une somme pondérée de Y_i avec la pondération suivante :

$$W_\lambda(x, X_i) = \frac{1}{\varphi(x)} \frac{1}{h_\lambda(x)} \cdot K\left(\frac{x - X_i}{h_\lambda(x)}\right)$$

où $h_\lambda(x)$ est une fenêtre locale dépendant de λ et de la densité φ des X_i et où K est une fonction telle que $\lim_{|u| \rightarrow \infty} K(u) \rightarrow 0$

- Une autre classe d'estimateurs définis dans la même logique de minimisation est celle des M -estimateurs à noyau solutions d'un problème d'optimisation du type :

$$\hat{f}_n(x) = Arg \min_f \sum_{i=1}^n \frac{1}{h} \cdot K\left(\frac{S(W_i, \alpha_i) - x}{h}\right) \cdot \psi(W_i, \alpha_n, x, f) \quad (4.2)$$

où S et ψ sont des fonctions connues définissant la variable conditionnante $S(W_i, \alpha_i)$, et la fonction d'objectif paramétrique respectivement. Les $(W_i)_{i=1, \dots, n}$ sont les observations et α_i tend vers une limite α_∞ lorsque n tend vers l'infini. Cette formulation générale permet également de retrouver certains des estimateurs que nous venons de citer dans une optique de type « *condition de moment* »⁴ (voir Härdle[15]).

Remarques :

- Quelle que soit la méthode d'estimation proposée, le problème de « *douceur de l'estimateur* » évoqué ci-dessus, se posera par l'intermédiaire du paramètre de lissage. Son choix permettra d'arbitrer entre « *variabilité* » et « *lissage* » ou, d'une manière plus formelle, entre « *variance* » et « *biais* » comme nous le verrons section 4.4. Nous pouvons nous représenter l'impact de la fenêtre sur l'estimation d'une fonction par la méthode du noyau en examinant la figure 4.2. Il est à noter que cette représentation est valable pour n'importe quelle méthode présentées ici.

4. L'estimateur à noyau de la régression est ainsi solution de (4.2) lorsque :

$$\psi(W_i, \alpha_n, x, f) = (Y_i - f(x))^2$$

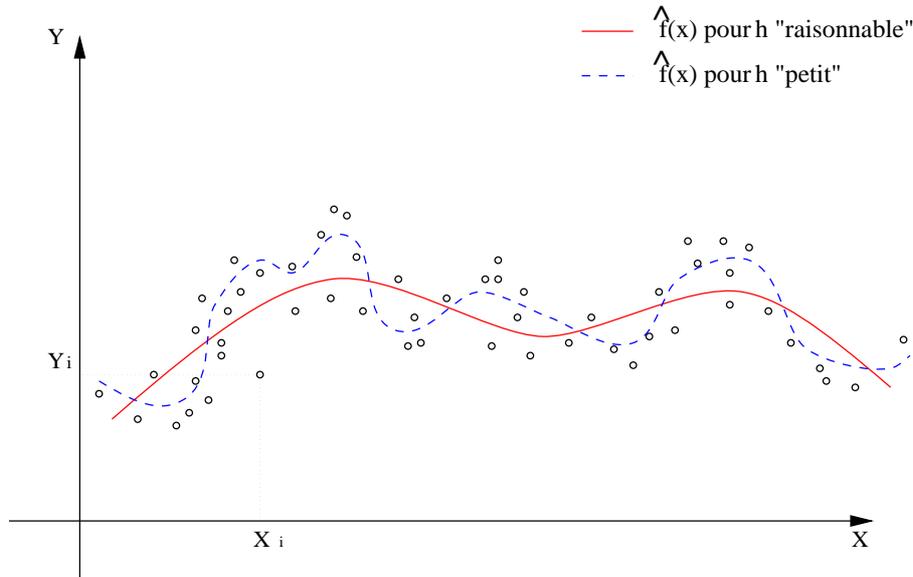


FIG. 4.2 - Impact de la fenêtre (h) sur l'estimateur \hat{f} d'une fonction f .

- Pour chacune des méthodes envisagées ci-dessus, ce paramètre est, en réalité, dépendant de la taille de l'échantillon, voire de l'échantillon lui-même si l'on souhaite le sélectionner convenablement. Les estimateurs proposés ci-dessus seront donc retrouvés dans la littérature sous la forme d'une suite de paramètres, $h(n)$, $M(n)$, etc. Les propriétés de convergence de ces estimateurs seront conditionnées par des hypothèses sur la vitesse de convergence (ou de divergence) de ces suites.

4.3 Estimateur du noyau de convolution

Définition 1 (Noyau de Parzen-Rosenblatt)

Un noyau K est une application de \mathbb{R}^p dans \mathbb{R} , bornée, intégrable pour la mesure de Lebesgue, d'intégrale unitaire. Un noyau de Parzen-Rosenblatt vérifie de plus :

$$\lim_{\|x\| \rightarrow \infty} \|x\|^p K(x) = 0$$

où $\|\cdot\|$ désigne la norme de \mathbb{R}^p .

Un exemple de noyau de Parzen-Rosenblatt est la densité normale standard (voir figure 4.1) qui vérifie cette condition. Nous utiliserons ces noyaux dans la suite de ce travail.

On définit également des classes de noyaux correspondant à des propriétés de régularité particulières.

Définition 2 (Noyau d'ordre m)

Le noyau K appartient à la classe $\mathcal{K}_m(\mathbb{R}^p)$ des noyaux d'ordre m si :

$$\int_{\mathbb{R}^p} \prod_{i=1}^p x_i^{a_i} K(x_1, x_2, \dots, x_p) dx_1 \cdots dx_p = \begin{cases} 1 & \text{si } a_i = 0, \forall i = 1, \dots, p \\ 0 & \text{si } 0 < \sum_{i=1}^p a_i < m \end{cases}$$

$$\text{et } \int_{\mathbb{R}^p} |x_i|^m |K(x_1, x_2, \dots, x_p)| dx_1 \cdots dx_p < \infty, \forall x \in \mathbb{R}^p$$

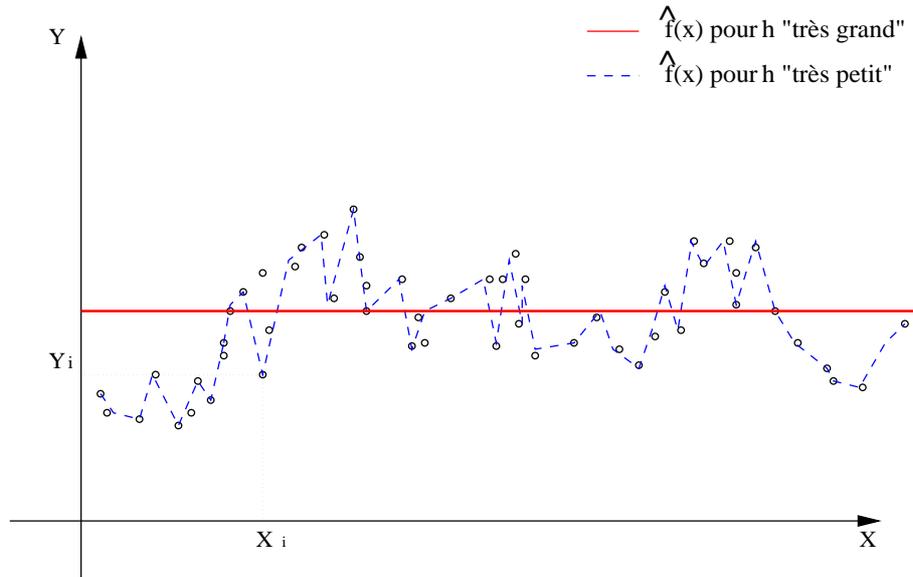


FIG. 4.3 - Cas limites pour le choix de la fenêtre h dans \hat{f} .

Cette propriété est standard en statistique non-paramétrique et est couramment utilisée comme hypothèse de régularité pour les noyaux dans les théorèmes de convergence asymptotique.

Il est à noter que pour $m \geq 3$, les noyaux de $\mathcal{K}_m(\mathbb{R}^p)$ ne sont plus des densités, et pourront prendre des valeurs négatives sur certains intervalles.

Définition 3 (Estimateur du noyau de la régression)

Soit $(X_i, Y_i)_{i=1, \dots, n}$ n observations d'un couple (X, Y) de variables aléatoires définies sur l'espace réel mesuré $(\mathbb{R}^p \times \mathbb{R}, \mathcal{B}_{\mathbb{R}^{p+1}}, \lambda)$. L'estimateur du noyau de convolution de la régression $f(x) = E[Y | X = x]$ associé au noyau K et à la fenêtre h_n , un nombre réel dépendant de n , est défini par :

$$\hat{f}_n(x) = \frac{\frac{1}{nh_n^p} \sum_{i=1}^n Y_i K\left(\frac{X_i - x}{h_n}\right)}{\frac{1}{nh_n^p} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)} \quad \forall x \in \mathbb{R}^p \tag{4.3}$$

avec la convention $\hat{f}_n(x) = 0$ si le dénominateur $\frac{1}{nh_n^p} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right) = 0$.

Cette formulation ayant été introduite simultanément par Nadaraya [22] et Watson [32] en 1964, cet estimateur est également appelé estimateur de Nadaraya - Watson.

Le rôle de la fenêtre h_n dans cet estimateur peut être visualisé en se représentant deux cas limites :

- si $h_n \rightarrow \infty$ alors $K\left(\frac{X_i - x}{h_n}\right) \rightarrow Cste = K(0)$, $\forall x$ et $\hat{f}_n(x) \rightarrow Cste \cdot \frac{1}{n} \sum_i Y_i = Cste \cdot \bar{Y}$

L'estimateur dégénère en une fonction d'une douceur extrême puisque constante, mais sans estimer réellement $f(x)$.

- si $h_n \rightarrow 0$ sans restriction, $K\left(\frac{X_i - x}{h_n}\right) \rightarrow \begin{cases} 1 & \text{si } x = X_i \\ 0 & \text{Sinon} \end{cases}$

L'estimateur devient alors extrêmement sensible à l'éloignement de x à X_i et tend vers une fonction discontinue passant par tous les points (X_i, Y_i) . Ces cas sont représenté dans la figure 4.3.

Remarque: Le dénominateur de l'expression (4.3) est un estimateur de la densité marginale $\varphi(x)$, tandis que le numérateur constitue un estimateur de $\Phi(x) = \int y\varphi(x, y)dy$. Nous pouvons donc écrire $\widehat{f}_n(x)$ sous la forme :

$$\widehat{f}_n(x) = \frac{\widehat{\Phi}_n(x)}{\widehat{\varphi}_n(x)}$$

Si, en particulier, K est une densité de probabilité alors l'estimateur $\widehat{\varphi}_n(x)$ de $\varphi(x)$ est donné par la densité de la somme de deux variables aléatoires :

- l'une suivant la densité empirique des X_i , $\mu_n = \frac{1}{n} \sum_i \delta_{X_i}$
- l'autre suivant la distribution de densité $K_n(\cdot) = \frac{1}{h_n^p} K\left(\frac{\cdot}{h_n}\right)$

La convolution ainsi réalisée suit une distribution de densité :

$$\widehat{\varphi}_n(x) = (K_n * \mu_n)(x) = \frac{1}{nh_n^p} \sum_{i=1}^n K\left(\frac{X_i - x}{h_n}\right)$$

d'où l'estimateur tire son nom « d'estimateur du noyau de convolution ».

4.3.1 Propriétés générales

Le premier résultat de convergence est dû à Bochner [1], dont le lemme est à la base des principaux théorèmes de convergence. Nous tirons son énoncé de l'ouvrage de Bosq et Lecoutre [2] (p. 61).

Lemme (de Bochner) :

- i) Soit K un noyau de Parzen-Rosenblatt et g une fonction de \mathbf{L}^1 .
Alors, en tout point x , où g est continue :

$$\lim_{h_n \rightarrow 0} (K_n * g)(x) = g(x)$$

- ii) Soit K un noyau quelconque et g une fonction de \mathbf{L}^1 uniformément continue, alors

$$\lim_{h_n \rightarrow 0} \sup_x |(K_n * g)(x) - g(x)| = 0$$

L'interprétation de ce lemme est que lorsque la fenêtre h est « petite », la convolution d'une fonction de \mathbf{L}^1 avec K_n perturbe peu cette fonction.

4.3.2 Propriétés de convergence ponctuelle de l'estimateur \widehat{f}_n :

Quelques hypothèses standard permettant l'application du lemme de Bochner sont présentées ici.

Hypothèses H_0

- i) Les observations $(X_i, Y_i)_{i=1, \dots, n}$ sont des observations indépendantes du couple de variables aléatoires (X, Y) de $\mathfrak{R}^p \times \mathfrak{R}$.

- ii) Le noyau $K(\cdot)$ est de Parzen-Rosenblatt

iii) La fenêtre h_n vérifie $\lim_{n \rightarrow \infty} h_n = 0$ et $\lim_{n \rightarrow \infty} nh_n^p = \infty$

Cette dernière hypothèse (iii) sur les fenêtres est la traduction d'un arbitrage entre variabilité et douceur de l'estimateur déjà évoqué au travers des figures 4.2 et 4.3. Elle nous impose un « juste milieu », nécessaire à la convergence de l'estimateur \widehat{f}_n .

Dans la suite de ce travail nous supposons ces hypothèses vérifiées, et nous ne mentionnerons que les hypothèses supplémentaires.

Nous rappelons, tout d'abord un résultat concernant la convergence des estimateurs $\widehat{\varphi}_n(x)$ de la densité $\varphi(x)$ et de $\widehat{\Phi}_n(x)$, estimateur de $\Phi(x) = \int y\varphi(x, y)dy$, c'est à dire des estimateurs du numérateur et du dénominateur de la fonction de régression f .

Théorème 1 (Convergence en moyenne quadratique de $\widehat{\varphi}_n(x)$)

Supposons $E[Y^2] < \infty$ et posons $v(\cdot) = \text{Var}[Y | X = \cdot]$,

Si

- $\varphi(\cdot)$ est continue au point x

Alors $\widehat{\varphi}_n(x)$ converge en moyenne quadratique vers $\varphi(x)$.

Si de plus :

- $f(\cdot)$ et $v(\cdot)$ sont continues au point x

Alors $\widehat{\Phi}_n(x)$ converge en moyenne quadratique vers $\Phi(x)$.

La démonstration de ce théorème découle de la définition de l'erreur quadratique de $\widehat{\varphi}_n(x)$ et du lemme de Bochner (voir Bosq et Lecoutre [2]). Cette erreur a été étudiée notamment par Collomb [5] pour l'étude des moments conditionnels $E[Y^a | X]$ lorsque $a \in \mathbf{N}$.

Ce théorème permet de vérifier que l'estimateur $\widehat{f}_n(x)$ est un estimateur convergent de l'espérance conditionnelle $E[Y | X = x]$, comme l'indique le corollaire suivant.

Théorème 2 (Convergence simple en probabilité)

Sous les hypothèses du théorème précédent et :

Si $\varphi(x) \neq 0$, alors :

$$\widehat{f}_n(x) \xrightarrow{p} f(x).$$

4.4 Du bon choix de la fenêtre

« However there is a price to be paid for the great flexibility of nonparametric methods, which is that the smoothing parameter must be chosen »

J. S. Marron (1988)

Nous avons relevé dans la section précédente l'importance de la fenêtre intervenant dans l'estimation non-paramétrique, et nous nous proposons d'apporter des éléments de réponse, fournis par la littérature, à la question :

« Comment choisir la fenêtre? »

En l'absence de techniques de sélection, l'utilisateur de la statistique non-paramétrique sélectionne généralement la fenêtre au vu de la courbe, par essais successifs, cette technique visuelle ne pouvant d'ailleurs s'appliquer que pour des dimensions de régresseurs faibles ($p < 3$). L'aspect *arbitraire* de ce choix étant peu souhaitable, (en particulier dans un cadre comme le nôtre), il nous semble nécessaire de déterminer des critères de sélection *objectifs* de ce paramètre.

Une manière de s'assurer « *objectivement* » du comportement de l'estimateur en fonction de la fenêtre, est d'examiner un critère d'erreur entre la fonction de régression et son estimée. Certains de ces critères seront globaux ou locaux, ils peuvent être basés sur la prévision ou sur l'écart au sens d'une norme fonctionnelle, qui variera suivant l'utilisation ou les propriétés souhaitées de l'estimateur.

L'un des critères le plus utilisé repose sur l'Erreur Quadratique Intégrée (*EQI*) définie par⁵ :

$$EQI(h) = \int_{\mathbb{R}^p} \left(\widehat{f}_n(x) - f(x) \right)^2 \varphi(x) \varpi(x) \lambda(dx)$$

où $\varpi(\cdot)$ est une fonction de poids⁶.

Ce critère étant théorique, on peut en trouver une version empirique, l'Erreur Quadratique Empirique (*EQE*), en remplaçant la mesure de Lebesgue par la loi empirique des X_i :

$$EQE(h) = \sum_{i=1}^n \left(\widehat{f}_n(X_i) - f(X_i) \right)^2 \varpi(X_i)$$

ou une version globale, l'Erreur Quadratique Intégrée Moyenne (*EQIM*) obtenue en prenant l'espérance de l'*EQI* :

$$EQIM(h) = \mathbf{E}[EQI(h)] = \mathbf{E} \left[\int_{\mathbb{R}^p} \left(\widehat{f}_n(x) - f(x) \right)^2 \varphi(x) \varpi(x) \lambda(dx) \right]$$

Härdle et Marron [21] montrent que ces trois mesures quadratiques sont asymptotiquement équivalentes pour une grande variété d'estimateurs.

La fenêtre idéale, h_{opt} doit alors réaliser le minimum de l'un de ces critères qui dépendent des fonctions inconnues f et φ . On la définit par :

$$h_{opt} = Arg \min_{h \in H_n} EQI(h)$$

Le rôle de la fenêtre h dans ces critères d'erreur peut être vu à travers une formulation de Vieu [31] donnant une évaluation asymptotique de l'*EQI* pour des fonctions f et φ d -fois continûment différentiables et pour des noyaux d'ordre d :

$$EQI(h) = B \cdot h^{2d} + \frac{V}{n \cdot h^p} + o_p \left(h^{2d} + \frac{1}{n \cdot h^p} \right) \quad (4.4)$$

où : B et V sont deux nombres réels finis

et p est la dimension du vecteur aléatoire X .

Nous retrouvons ici l'aspect d'arbitrage entre *Biais* et *Variance* joué par la fenêtre, puisque les termes B et V correspondent respectivement à des termes de biais et de variance approchés (voir également Rice [24], ou Härdle et Marron [21]). A propos de cette expression Härdle [15] écrit d'ailleurs :

« (\dots) one gets a feeling of what the smoothing problem is about : Balance the variance versus the biais. »

5. Il existe bien d'autres critères et nous ne citons ici que les plus « populaires », voir Härdle [15] ou Marron [21].

6. Cette fonction de poids est généralement introduite dans ces définitions pour compenser les problèmes d'estimation lorsque la densité des régresseurs devient faible (« *Effets de bord* »). Les conditions sur cette fonction sont donc liées à la densité inconnue $\varphi(x)$, nécessitant une information supplémentaire.

Cette expression permet de dégager deux optiques menant à deux procédures de choix de la fenêtre.

- Soit on estime la fenêtre minimisant le critère retenu (ici l' $EQI(h)$) en estimant B et V ,
- Soit on estime le critère (l' $EQI(h)$) et on sélectionne la fenêtre minimisant ce nouveau critère.

La première solution mène aux techniques de « *Plug-in* », la seconde aux méthodes de « *validation croisée* ».

4.4.1 Le « *plug-in* »

La méthode du « *Plug-in* » repose sur la sélection de la fenêtre h_d minimisant l' $EQI(h)$ donné par (4.4). Cette fenêtre peut être approchée par \widehat{h}_d obtenue en estimant les termes B et V qui dépendent des fonctions inconnues f , φ (et de leurs dérivées) ainsi que du noyau K .

Cette technique est très satisfaisante théoriquement puisque l'expression de h_d minimisant l' $EQI(h)$ dans l'équation (4.4) est :

$$h_d = \left(\frac{p \cdot V}{2d \cdot B} \right)^{\frac{1}{2d+p}} \cdot n^{-\frac{1}{2d+p}}$$

et que $EQI(h_d)$ est alors de l'ordre

$$EQI(h_d) = O_p \left(n^{-\frac{2d}{2d+p}} \right)$$

Cette vitesse de convergence est donnée par Stone [28] comme étant la vitesse de convergence optimale dans la classe des fonctions de régression d -fois continûment différentiables. De plus la vitesse de convergence de h_d et donc de \widehat{h}_d est explicitement $n^{-\frac{1}{2d+p}}$.

De plus cette technique nous confine à l'étude de fonctions de régression suffisamment régulières (d -fois continûment différentiables). Enfin, des difficultés importantes se posent en pratique : pour calculer \widehat{h}_d il faut, en effet, estimer les constantes B et V et donc les dérivées des fonctions f et φ ce qui s'avère techniquement délicat (voir Vieu [31]).

4.4.2 La « *Validation Croisée* »

L'idée de base consiste à trouver une fonction de score $CV(h)$ ayant la même structure que l' $EQI(h)$ et dont le calcul soit plus simple. On sélectionne alors la fenêtre \widehat{h}_{cv} minimisant ce critère dont on attend le même comportement asymptotique que h_{opt} .

Le critère $CV(h)$ est obtenu à partir de l'Erreur Quadratique Empirique ($EQE(h)$) dans laquelle l'estimateur $\widehat{f}_n(X_i)$ est remplacé par l'estimateur de « *leave-one-out* » $\widehat{f}_n^{-i}(X_i)$ et $f(X_i)$ est estimé naïvement par Y_i (voir Härdle[15] pp.152-153).

On choisit alors $\widehat{h}_{cv} = Arg \min_{h \in H_n} CV(h)$ où :

$$CV(h) = \frac{1}{n} \sum_{i=1}^n \left(Y_i - \widehat{f}_n^{-i}(X_i) \right)^2 \varpi(X_i)$$

et

$$\widehat{f}_n^{-i}(X_i) = \frac{\frac{1}{nh_n^p} \sum_{j \neq i} Y_j K \left(\frac{X_j - X_i}{h_n} \right)}{\frac{1}{nh_n^p} \sum_{j \neq i} K \left(\frac{X_j - X_i}{h_n} \right)}$$

Härdle et Marron [21] démontrent que la fenêtre ainsi obtenue vérifie la propriété d'optimalité asymptotique suivante :

$$\frac{EQI(\widehat{h}_{cv})}{EQI(h_{opt})} \longrightarrow 1 \quad p.s.$$

sous les hypothèses suivantes.

Hypothèses F_0 (Optimalité asymptotique) :

- $H_n = [\underline{h}, \overline{h}] = [a_n \cdot n^{-\frac{1}{p}}, a_n^{-1}]$
où $a_n = C \cdot n^\delta$ et C, δ sont des constantes positives, $\delta \in \left[0, \frac{1}{p}\right]$
- Les fonctions f, φ et K sont Hölder continues⁷
- Les moments conditionnels de $Y | X$ sont bornés, c'est-à-dire :
 $\forall q > 0, \exists A_q > 0$ tel que $E[|Y|^q | X = x] < A_q, \forall x$
- La fonction de poids $\varpi(\cdot)$ est à support compact S
- La densité marginale $\varphi(x)$ est bornée inférieurement sur l'intérieur de ce support S

L'inconvénient principal est que la fenêtre \widehat{h}_{cv} , qui est ici un estimateur, présente une grande variabilité, c'est-à-dire que pour deux échantillons distincts issus de la même distribution, les fenêtres obtenues seront très différentes. Ce problème a été étudié par Härdle et Marron [15] où il est montré que \widehat{h}_{cv} converge « très lentement » vers h_{opt} . Une technique pour pallier à cet inconvénient consiste peut-être à utiliser le « double smoothing » proposé par Härdle, Hall et Marron [21] et évoqué plus haut.

Un autre inconvénient réside dans la complexité du calcul de la fenêtre « optimale ». Pour chaque échantillon de taille n , et avant l'estimation de \widehat{f} , on doit calculer l'estimateur *leave-one-out*, ce qui nécessite n opérations, l'évaluation du critère $CV(h)$ demande également n calculs, soit une complexité en $O(n^2)$ et cela pour une seule fenêtre. Reste ensuite à minimiser ce critère!

Dans le contexte de notre travail, cette approche est donc très couteuse compte tenu du fait que notre échantillon de points observés évolue d'une période sur l'autre. Nous verrons dans le chapitre 6 que cette approche pourrait être utilisée ou modifiée dans un contexte plus restreint. Notamment dans le cas où les prévisions ne concernent qu'un seul type d'avion, l'estimation de la fenêtre pourrait être calibrée par validation croisée, préalablement à la prévision de la trajectoire, par l'utilisation de trajectoires types.

4.4.3 L'approche « naïve » de Bosq et Lecoutre

Dans leur ouvrage, Bosq et Lecoutre [2], préconisent une méthode particulièrement simple, à partir de l'estimation (cf 4.4) de l'Erreur Quadratique Intégrée (EQI), pour une fonction supposée 2-fois continûment différentiable ($d = 2$), et pour x réel ($p = 1$), on peut estimer la fenêtre minimisant ce critère.

En effet l'EQI s'écrit alors :

$$EQI(h) = B \cdot h^4 + \frac{V}{n \cdot h} + o_p \left(h^4 + \frac{1}{n \cdot h} \right) \quad (4.5)$$

d'où l'on déduit l'expression de la fenêtre minimisant ce critère. L'idée de Bosq et Lecoutre est

7. Une fonction g est Hölder continue s'il existe des constantes positives M et ξ , telles que :

$$|g(x) - g(t)| \leq M \cdot \|x - t\|^\xi$$

d'estimer naïvement les termes de variance (V) et de biais (B) inconnus, tout comme dans la méthode du « *plug-in* », mais de manière très simple.

Au lieu de la fenêtre minimisant 4.5, soit :

$$\hat{h} = \left(\frac{V}{B} \right)^{\frac{1}{5}} \cdot n^{-\frac{1}{5}}$$

ces auteurs nous proposent donc (p.85) :

$$\hat{h} = \alpha \cdot S_n \cdot n^{-\frac{1}{5}}$$

où S_n est l'écart type empirique des X_i (et donc un estimateur de $V^{\frac{1}{2}}$) et où α ne dépend que du noyau choisi⁸.

Cette méthode présente de nombreux avantages : outre le fait qu'elle ne demande pour être applicable, que des hypothèses faibles sur le degré de différentiabilité de f , c'est une méthode automatique entièrement guidée par les données, au travers de l'écart type et de la taille de l'échantillon. Ce point est particulièrement satisfaisant dans notre contexte, où l'échantillon observé est amené à varier entre deux prévisions. De plus, cette procédure s'adaptera automatiquement si l'intervalle de temps entre deux observations est amené à varier. Nous verrons dans les sections suivantes comment nous pourrions adapter cette procédure à notre problème de prévision.

4.4.4 Autres méthodes

Il existe de nombreuses autres méthodes et notamment des raffinements de la validation croisée. Vieu [30] propose de sélectionner la fenêtre localement en utilisant un critère de validation croisé local, tenant compte de la densité autour de chaque observation. Ce critère, particulièrement appréciable dans le cas d'observations « mal réparties », est malheureusement encore assez coûteux en temps de calcul pour être utilisé en pratique.

Une autre variante consiste à détruire plusieurs points en utilisant un estimateur de « *leave-several-out* » dans la définition du critère CV .

L'introduction d'une fonction pénalisante Ξ dans l'estimateur de l' EQE permet d'obtenir un critère de Score sur la base duquel est estimée la fenêtre. Härdle [15] (pp.155-167) nous donne une étude comparative sur un échantillon, de différentes fonctions pénalisantes.

D'autres méthodes sont exposées dans la revue sur ce sujet réalisée par Vieu [31], parmi lesquelles les méthodes de Bootstrap semblent également particulièrement prometteuses (voir également Härdle [15]).

4.5 Conclusion

Nous avons exposé dans ce chapitre quelques unes des méthodes d'estimation fonctionnelle. Ces méthodes permettent l'étude et l'estimation des modèles de régression en l'absence de forme fonctionnelle prédéfinie, et en l'absence de spécification de la loi des résidus. Cette liberté dans la spécification (ou plutôt dans l'absence de spécification) des modèles de régression n'est cependant pas exempte de règles. La sélection du paramètre de lissage, dans chacune de ces méthodes est soumise à des contraintes et les règles de sélection pratiques de ce paramètre sont encore à l'étude pour certaines méthodes.

Une autre contrainte nous est donnée par l'inexistence d'un estimateur sans biais de la régression montré par Collomb [5], il en résulte une approche asymptotique nécessitant un nombre

8. Dans le cas d'un noyau gaussien, α vaut 1,059.

important d'observations. Ce point est aggravé par la perte d'une vitesse de convergence « *paramétrique* » (en \sqrt{n}), montré par Stone [28] ; la convergence non-paramétrique étant plus lente, ces méthodes exigent un plus grand nombre de données.

Nous avons présenté ces méthodes sous une forme unifiée de somme pondérée des observations. Dans cette optique, l'on comprend mieux le rôle joué par la pondération fonctionnelle dans chacune des méthodes. Ce rôle est arbitré pour chaque méthode par un « paramètre de lissage », que ce soit la fenêtre, le nombre de fonctions de base dans les méthodes de projection orthogonale, ou le paramètre de pénalisation dans les méthodes « splines ».

Nous avons choisi de développer plus particulièrement la méthode du noyau pour de simples raisons : cette méthode est la plus développée à ce jour, et les propriétés des estimateurs sont maintenant bien connues. En outre elle bénéficie d'une abondante littérature sur des problèmes théoriques et appliqués. Enfin, des procédures d'estimation de la fenêtre sont possibles dans le cadre de cette méthode. Certaines sont entièrement guidées par les données et présentent donc un caractère « *objectif* » particulièrement appréciable dans notre contexte.

Nous utiliserons principalement l'estimation « naïve » due à Bosq et Lecoutre comme point de départ pour le « calibrer » nos fenêtres et jugerons de la pertinence de cette fenêtre au vu des prévisions obtenues, dans le chapitre suivant.

Une question importante reste toutefois sans éléments théoriques de réponse, il semblerait naturel de voir notre fenêtre varier avec l'horizon de prévision. En effet, plus on veut prévoir loin, plus l'on s'intéresse à la « tendance générale » de la courbe, et donc plus l'estimateur de la trajectoire doit être lisse. On s'attend donc à ce que la fenêtre s'allonge avec l'horizon de prévision. Nous verrons que cela n'est pas toujours le cas dans nos prévisions.

4.6 Annexe au chapitre 3

Notations

Les « petits- o » et les « grands- O », que l'on trouve couramment dans la littérature sont rappelés de manière précise ici. Ces symboles ont été introduits par Landau pour simplifier les relations entre quantités (stochastiques ou non) de même Ordre de grandeur, ou d'un ordre de grandeur inférieur asymptotiquement. Nous nous servirons de ces notations dans les propriétés à venir.

Définition 4 Si f et g sont deux fonctions réelles de la variable entière n , alors la notation $f(n) = o(g(n))$ signifie que :

$$\lim_{n \rightarrow \infty} \left(\frac{f(n)}{g(n)} \right) = 0$$

Il est important de noter que $g(n)$ peut avoir n'importe quel comportement lorsque $n \rightarrow \infty$, en particulier la notation $f(n) = o(1)$ signifie simplement que la suite $f(n) \rightarrow 0$ lorsque $n \rightarrow \infty$.

Définition 5 Si f et g sont deux fonctions réelles de la variable entière n , alors la notation $f(n) = O(g(n))$ signifie qu'il existe une constante $K > 0$, indépendante de n , et un entier N tels que :

$$\left| \frac{f(n)}{g(n)} \right| < K \quad , \quad \forall n > N$$

Ceci signifie donc que f et g ont le « même ordre de grandeur »⁹

De même, des relations liant les ordres de grandeur de quantités stochastiques sont exprimées par les célèbres « petits- o_p » et « grands- O_p » définis comme suit.

Définition 6 Si a_n est une suite de variables aléatoires et g est une fonction réelle de la variable entière n , alors la notation $a_n = o_p(g(n))$ signifie que :

$$p \lim_{n \rightarrow \infty} \left(\frac{a_n}{g(n)} \right) = 0$$

De manière similaire, la notation $a_n = O_p(g(n))$ signifie que il existe une constante $K > 0$, telle que $\forall \varepsilon > 0, \exists$ un entier N_ε tel que :

$$P_r \left(\left| \frac{a_n}{g(n)} \right| > K \right) < \varepsilon \quad , \quad \forall n > N_\varepsilon$$

Propriétés de convergence uniforme de \widehat{f}_n

La formulation du théorème de convergence uniforme que nous reproduisons ici est tirée de l'ouvrage de Györfi, Härdle, Sarda et Vieu ([13] pp. 24-30), choisie pour la simplicité des hypothèses. Cette formulation nous donne explicitement la vitesse de convergence de l'estimateur.

Théorème 3 (Convergence uniforme de \widehat{f}_n):

Soit G un compact de \mathbb{R}^p et \widehat{G} un voisinage de ce compact ($G \subset \widehat{G}$),

Supposons $E[Y^2] < \infty$ et posons $\sigma^2(\cdot) = \text{Var}[Y | X = \cdot]$, sous les hypothèses suivantes :

- La densité $\varphi(x) > 0, \forall x \in G$

9. Cette définition n'exclut pas la possible nullité de ce rapport, l'expression « de même ordre que » peut être trompeuse.

- $\forall x \in \widehat{G}, \sigma^2(x) < \infty$
- $\varphi(\cdot)$ et $f(\cdot)$ sont d -fois continûment différentiables, et ont des dérivées bornées,
- le noyau K est de Parzen-Rosenblatt d'ordre d

et si la fenêtre h est telle que V_n

$$V_n = h^d + \sqrt{\frac{\log(n)}{n \cdot h^p}}$$

vérifie $V_n \xrightarrow{n \rightarrow \infty} 0$, alors

$$\sup_{x \in G} |\widehat{f}_n(x) - f(x)| = O_p(V_n)$$

La preuve de ce résultat est donné par Györfi et alii dans le cadre de processus φ -mélangeant et n'est pas reproduite ici, laissant le lecteur intéressé se reporter à la difficile lecture de ces articles. Bosq et Lecoutre [2] nous donnent d'autres résultats de convergence de l'estimateur \widehat{f}_n suivant le type de norme considérée pour mesurer l'écart de \widehat{f}_n à f . Des résultats plus complets sur la convergence uniforme sont donnés également par Sarda et Vieu [31].

Distribution limite

Nous donnons ici le résultat principal concernant la distribution asymptotique de \widehat{f}_n . Ce résultat a été obtenu par Schuster pour le cas univarié et à Collomb [5] dans le cas multidimensionnel.

Des hypothèses supplémentaires sont nécessaires à ce résultat et sont similaires à celles rencontrées usuellement. De plus celles-ci sont explicitement utilisées dans la démonstration, ce qui rend leur interprétation plus facile.

Théorème 4 (Normalité asymptotique de \widehat{f}_n)

Sous les hypothèses suivantes :

- $\exists \delta$ tel que la fonction $\sigma^{2+\delta}(x) \cdot \varphi(x)$ est uniformément bornée
- Les fonctions $f(x)^2 \cdot \varphi(x)$ et $\sigma^2(x) \cdot \varphi(x)$ sont continues et uniformément bornées
- Les fonctions $\varphi(x)$ et $f(x) \cdot \varphi(x)$, ainsi que leurs dérivées et dérivées secondes sont continues et uniformément bornées.
- Le noyau K est de Parzen Rosenblatt d'ordre 2
- $\varphi(x) > 0$

On a,

$$\text{Si } h_n^2 \sqrt{n \cdot h_n^p} \xrightarrow{n \rightarrow \infty} l \text{ avec } 0 \leq l < \infty$$

Alors,

$$\sqrt{n \cdot h_n^p} (\widehat{f}_n(x) - f(x)) \xrightarrow{d} \mathcal{LN} \left(\frac{l \cdot b(x)}{\varphi(x)}, \frac{\sigma^2(x)}{\varphi(x)} \int_{\mathbb{R}^p} K^2(z) dz \right) \quad (4.6)$$

De plus,

$$\text{Si } h_n^2 \sqrt{n \cdot h_n^p} \xrightarrow{n \rightarrow \infty} \infty$$

Alors,

$$p \lim_{n \rightarrow \infty} h_n^{-2} \left(\widehat{f}_n(x) - f(x) \right) = \frac{b(x)}{\varphi(x)} \quad (4.7)$$

avec:

$$b(x) = \frac{1}{2} \text{Tr} \left(\Xi \frac{\partial}{\partial x'} \frac{\partial}{\partial x'} [f(x)\varphi(x)] \right) - \frac{1}{2} f(x) \text{Tr} \left(\Xi \frac{\partial}{\partial x'} \frac{\partial}{\partial x'} [\varphi(x)] \right)$$

$$\text{où : } \Xi = \int x' x K(x) dx.$$

Ce dernier résultat pouvant également être interprété comme la convergence en distribution vers une loi dégénérée.

La démonstration de ce résultat par Bierens est fort instructive et permet une décomposition intéressante entre termes « *asymptotiquement normaux* » et termes « *générant du biais* ». Cette dichotomie est très caractéristique des estimateurs non-paramétriques : ces estimateurs étant biaisés en échantillon fini, tout l'art consiste à « tuer » ce biais asymptotiquement.

Pour d'autres propriétés de convergence nous renvoyons à Bosq et Lecoutre [2], Collomb [5], ainsi qu'à Härdle [15].

Chapitre 5

Prévisions « non-paramétriques »

5.1 Mise en œuvre

Nous avons envisagé plusieurs approches de mise en œuvre des méthodes non-paramétriques à notre problème de prévision de trajectoires. Nous discuterons tout d'abord de la manière d'envisager une prévision, puis des différents problèmes d'implantation de ces méthodes, enfin nous proposerons des méthodes originales et détaillerons la méthode utilisée, finalement nous présenterons quelques résultats.

Il est important de noter que notre problème n'est pas un problème de prévision standard. La nature même de la fonction à prévoir est simple en apparence (monotone par morceaux) mais très difficile à cerner dans le détail. En outre les observations ne sont pas très nombreuses et les horizons de prévision souhaités importants, etc. Le réglage du paramètre de fenêtre constituera donc l'un des problèmes majeurs de la mise en œuvre de nos méthodes pour la prévision.

5.1.1 Prévision et prévisions

Il y a plusieurs façons de concevoir une prévision, à l'horizon p , de la valeur de l'altitude Y_{n+p} , à partir des observations jusqu'à la date T_n : On peut en effet

- soit estimer directement le point Y_{n+p} à partir des observations passées, $(Y_i)_{i=1,\dots,n}$;
- soit utiliser une procédure donnant la prévision à « horizon 1 » de manière récursive.

Dans ce deuxième cas, l'idée est de prévoir Y_{n+p} sur la base des $(Y_i)_{i=1,\dots,n}$ observés aux temps $(T_i)_{i=1,\dots,n}$ et des prévisions successives $\hat{Y}_{n+1}, \hat{Y}_{n+2}, \dots, \hat{Y}_{n+p-1}$ obtenues.

Nous nous proposons de discuter ici des deux approches afin d'en déterminer les avantages et inconvénients pour notre étude.

Méthode directe

L'estimateur à noyau de f défini section 4.3, sur la base de n observations $(T_i, Y_i)_{i=1,\dots,n}$ a pour expression, pour tout t :

$$\hat{f}(t) = \frac{\sum_{i=1}^n Y_i \cdot K\left(\frac{T_i-t}{h}\right)}{\sum_{i=1}^n K\left(\frac{T_i-t}{h}\right)} \quad (5.1)$$

où $K(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{u^2}{2}}$ est la densité gaussienne standard.

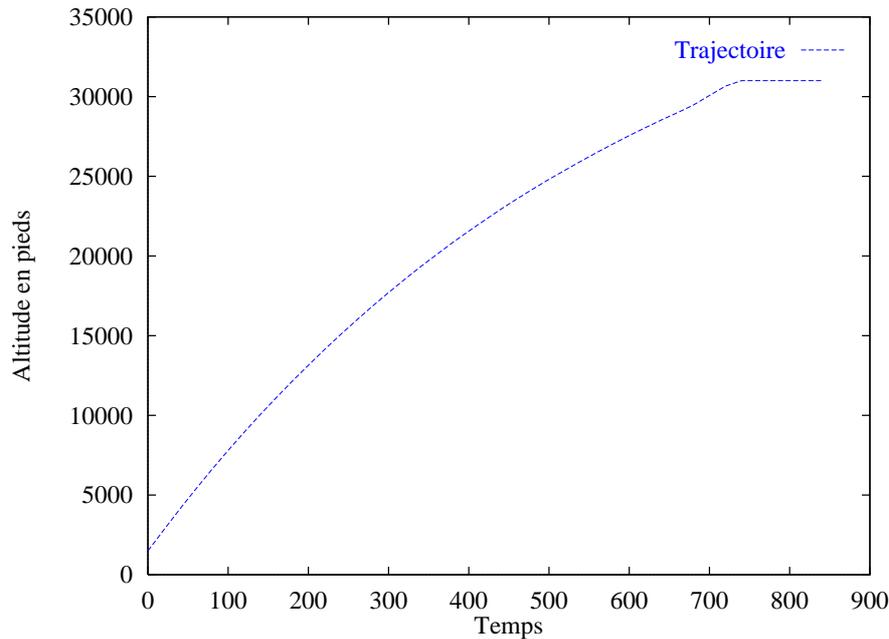


FIG. 5.1 - Trajectoire d'un A320 en montée .

Cet estimateur fonctionnel est donc défini **pour tout** t , ce qui signifie que l'on connaît la trajectoire, y compris au delà de la dernière observation. Une méthode « directe » d'estimation de Y_{n+p} consiste donc à remplacer t par T_{n+p} dans la formule 5.1.

Méthode réursive

Une autre manière de procéder pour prévoir Y_{n+p} consiste à appliquer séquentiellement, à court terme, l'estimateur \hat{f} , tout d'abord aux observations $(T_i, Y_i)_{i=1, \dots, n}$, d'en déduire la valeur prédite \hat{Y}_{n+1} de Y_{n+1} , laquelle valeur est introduite comme une nouvelle observation pour l'estimation de Y_{n+2} et ainsi de suite, pas à pas, jusqu'à l'horizon souhaité.

L'algorithme de prévision suivant cette optique sera récursif, et l'estimateur ne servira qu'à prédire à très court terme. Il est important de noter que l'erreur de prévision initiale peut se propager de manière catastrophique au cours de cette procédure, surtout si p est « grand », c'est à dire l'horizon « lointain ».

De la nécessité de « filtrer »

Si nous examinons une trajectoire d'avion en montée comme celle représentée par la figure 5.1, nous voyons que, en tout point, la fonction f à estimer (c'est à dire la trajectoire), est monotone. Or, l'estimateur \hat{f} de f est construit comme une somme pondérée des observations passées. La pondération étant unitaire, on ne peut que « sous estimer » la trajectoire en utilisant un estimateur non-paramétrique.

En effet, l'expression de \hat{f} reproduite en 5.1 peut également s'écrire :

$$\hat{f}(t) = \sum_{i=1}^n Y_i \cdot W_h(T_i, t)$$

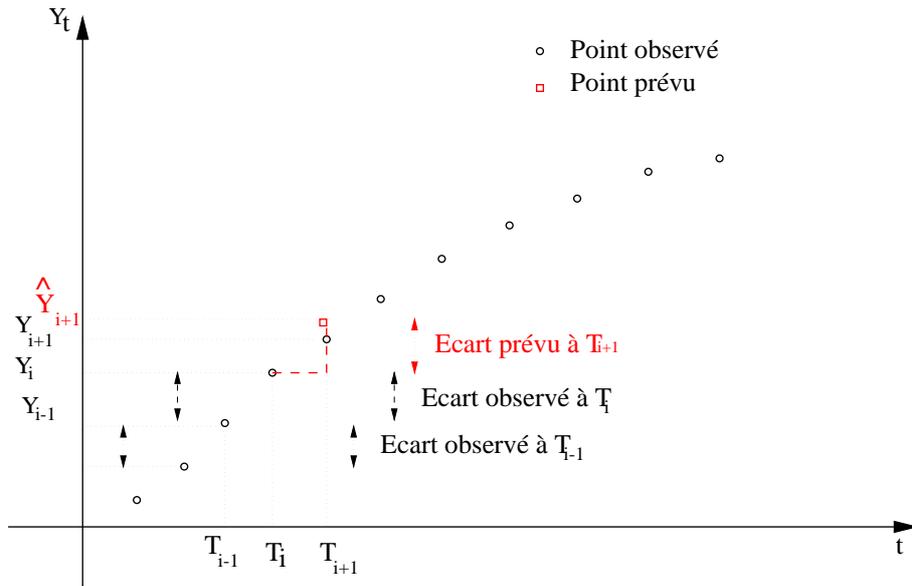


FIG. 5.2 - Estimation par filtrage.

avec la pondération unitaire¹ $W_h(T_i, t)$ vérifiant :

$$W_h(T_i, t) = \frac{w_h(T_i, t)}{\sum_{i=1}^n w_h(T_i, t)} \text{ et donc } \sum_{i=1}^n W_h(T_i, t) = 1$$

En un point t , l'estimateur, qui se comporte comme un « barycentre pondéré » des observations, ne peut dépasser, en valeur, la plus forte valeur observée, c'est à dire ici, la dernière. Ce phénomène est bien connu en statistique² et peut être contourné simplement par la technique du filtrage. Pour cela on peut remarquer que :

$$Y_{i+1} = Y_i + (Y_{i+1} - Y_i) = Y_i + e_{i+1}$$

avec $(e_{i+1}) = Y_{i+1} - Y_i$.

L'idée du filtrage consiste à éliminer la tendance commune aux Y_i par différenciation. Cette technique élimine la tendance commune aux Y_i , pour ne laisser que « l'innovation », représentée par la série (e_i) des écarts ainsi créés. D'où l'idée d'estimer cette « innovation » c'est à dire e_i pour l'estimation de la fonction f . Ainsi la prévision de Y_{j+1} connaissant $(Y_i)_{i=1, \dots, j}$ sera basée sur la prévision $\widehat{e_{j+1}}$ de e_{j+1} connaissant $(e_i)_{i=1, \dots, j}$ en utilisant l'expression :

$$\widehat{Y_{j+1}} = Y_j + \widehat{e_{j+1}}$$

Il suffit donc d'estimer et de prévoir la série des innovations $(e_i)_{i=1, \dots, j}$ pour prévoir la trajectoire. Ce principe est illustré par la figure 5.2.

Remarque:

1. Dans notre cas, l'estimateur à noyau est construit avec la pondération :

$$w_h(T_i, t) = K\left(\frac{T_i - t}{h}\right)$$

2. On connaît le même type de problème pour l'estimation de séries temporelles non-stationnaires (voir par exemple Hamilton [14]), ou en utilisant des méthodes d'estimation classiques (voir Bourbonnais [3]).

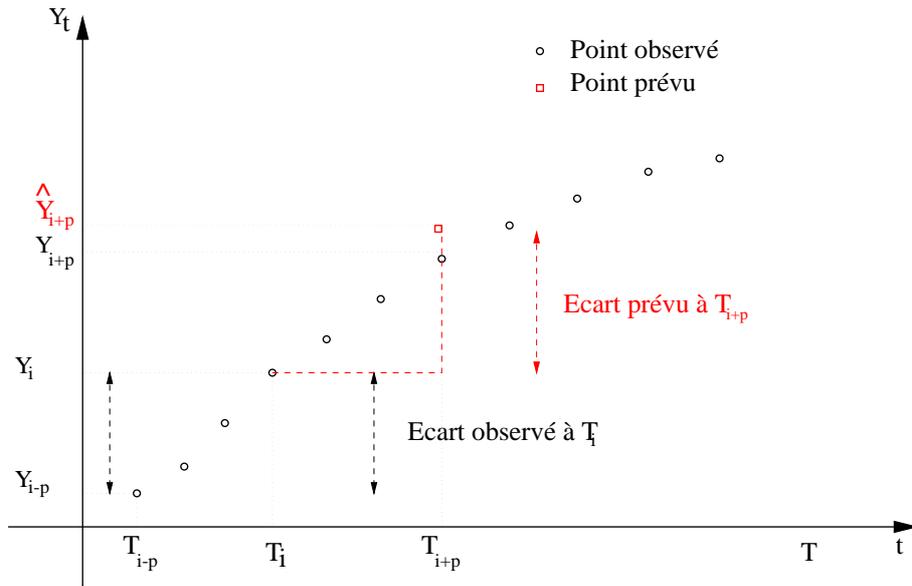


FIG. 5.3 - Généralisation du filtrage.

Nous pouvons généraliser ce principe pour l'estimation à long terme de la trajectoire. Si nous cherchons à estimer Y_{n+p} à partir des observations passées : $(Y_j)_{j=1,\dots,n}$, la même méthode s'applique en observant les écarts de longueur p .

Pour cela, on construit $u_j = Y_j - Y_{j-p}$ pour $j = p, \dots, n$. Les écarts observés servent alors d'éléments pour la prévision de $u_j + 1$ comme l'illustre la figure 5.3.

Une méthode de prévision à long terme basée sur les $(u_j)_{j=1,\dots,n}$ a été testée sur nos trajectoires³ en utilisant ce principe. Les résultats (non reproduits) ne sont pas satisfaisants en raison de la réduction de la taille de l'échantillon. En effet, l'échantillon des écarts observés (e_i) comporte un élément de moins que l'échantillon original des Y_i , de même l'échantillon des (u_i) comporte p observations de moins.

Nous utiliserons, en revanche, la méthode de filtrage simple ($p = 1$) pour éliminer la tendance de nos trajectoires, sans souffrir de ce phénomène.

La méthode « D »

Une méthode suggérée par N. Durand nous apparaît intéressante à développer, puisqu'elle ne nécessite pas de filtrage. Il s'agit de supposer que l'on observe le point que l'on cherche à prévoir Y_{n+1} et d'écrire l'équation permettant d'estimer le dernier point observé Y_n par une méthode non-paramétrique.

Supposons, en effet, que l'on observe $(T_i, Y_i)_i$ pour $i = 1, \dots, n+1$ et « omettons » volontairement d'utiliser la n -ième observation. On peut estimer f en tout point t par la formule classique 4.1 de somme pondérées, légèrement modifiée puisque l'on n'utilise pas le point (T_n, Y_n) et la somme fait donc intervenir $i = 1, \dots, n-1, n+1$.

$$\hat{f}(t) = \sum_{i=1, i \neq n}^{n+1} Y_i \cdot W_h(T_i, t)$$

Comme on le remarque aisément, cette expression fait intervenir explicitement Y_{n+1} .

$$\hat{f}(t) = \sum_{i=1}^{n-1} Y_i \cdot W_h(T_i, t) + Y_{n+1} \cdot W_h(T_{n+1}, t)$$

3. La procédure Ecartloin en annexe est basée sur cette méthode.

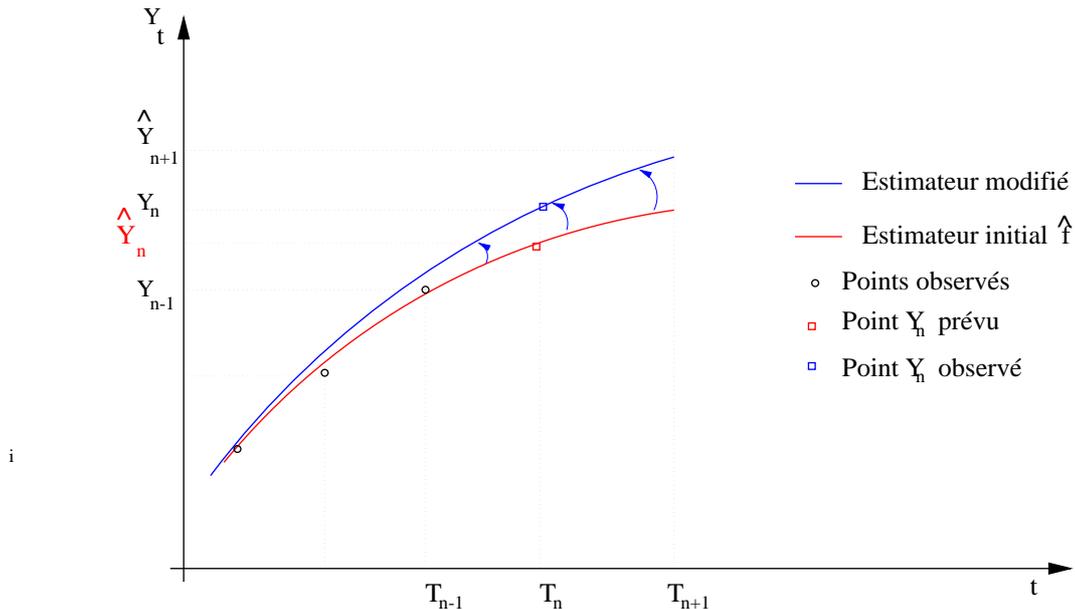


FIG. 5.4 - Principe de la méthode « D ».

En « retournant » cette expression on peut ainsi extraire une formule d'estimation de Y_{n+1} .

$$Y_{n+1} \cdot W_h(T_{n+1}, t) = \hat{f}(t) - \sum_{i=1}^{n-1} Y_i \cdot W_h(T_i, t) \quad (5.2)$$

Nous effectuons donc un barycentre entre Y_{n+1} , d'une part, et les $(Y_i)_{i=1, \dots, n-1}$, d'autre part. L'idée est alors de déterminer les pondérations nous donnant un ajustement parfait de Y_n avec son estimée $\hat{Y}_n = \hat{f}(T_n)$, et d'en déduire un estimateur de Y_{n+1} . Pour cela, il nous suffit de remplacer dans (5.2) le temps t par T_n , $\hat{f}(T_n)$ par sa valeur observée Y_n . Ce qui nous donne :

$$Y_{n+1} \cdot W_h(T_{n+1}, T_n) = Y_n - \sum_{i=1}^{n-1} Y_i \cdot W_h(T_i, T_n)$$

Un estimateur de Y_{n+1} est donc :

$$\widehat{Y}_{n+1} = \frac{Y_n - \sum_{i=1}^{n-1} Y_i \cdot W_h(T_i, T_n)}{W_h(T_{n+1}, T_n)} \quad (5.3)$$

Inconvénients

L'idée de départ est difficile à appliquer pour deux simples raisons

- Pour que le barycentre \widehat{Y}_n entre $(Y_i)_{i=1, \dots, n-1}$, d'une part, et Y_{n+1} d'autre part, tombe exactement en Y_n , il faut que la pondération associée à Y_{n+1} (i.e. $W_h(T_{n+1}, T_n)$), soit supérieure au total des poids associés aux $(Y_i)_{i=1, \dots, n-1}$. En prenant une pondération conforme à celles développées dans le chapitre précédant, nous donnons le même poids à deux points situés à la même distance du point de prévision. Ici le point à prévoir est (T_n, Y_n) , situé à égale distance de (T_{n-1}, Y_{n-1}) et de (T_{n+1}, Y_{n+1}) . Il faut donc prévoir une pondération adéquate pour Y_{n+1} , nettement supérieure à un.
- Le réglage des paramètres intervenant dans cette méthode, la fenêtre et la pondération associée à Y_{n+1} , s'avère délicat et les éléments théoriques d'aide au choix de ces facteurs

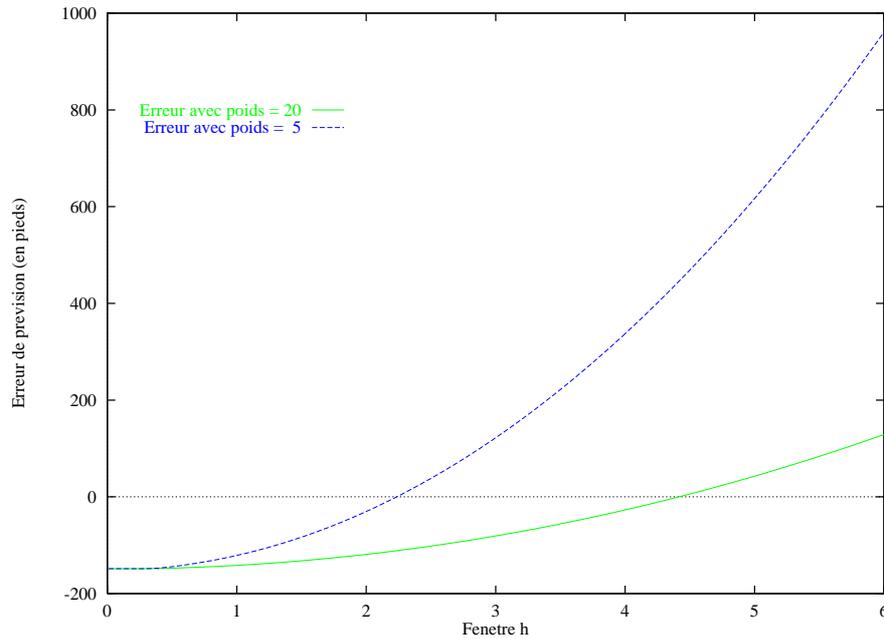


FIG. 5.5 - Impact de la fenêtre h et des « poids » sur l'erreur de prévision (Méthode « D »).

déterminants pour la prévision manquent cruellement. De plus, ces paramètres ne sont évidemment pas indépendants l'un de l'autre. On peut visualiser leurs effets sur l'erreur de prévision à horizon un effectuée à la date $T_n = 100s$ et représentée par la figure 5.5. La prévision est basée sur des observations toutes les 5 secondes. Les poids choisis pour Y_{n+1} sont fixés arbitrairement aux valeurs 5. et 20.

Cette procédure a cependant été testée avec des choix arbitraires de fenêtre et de *poids*, sans amélioration sensible de nos prévisions.

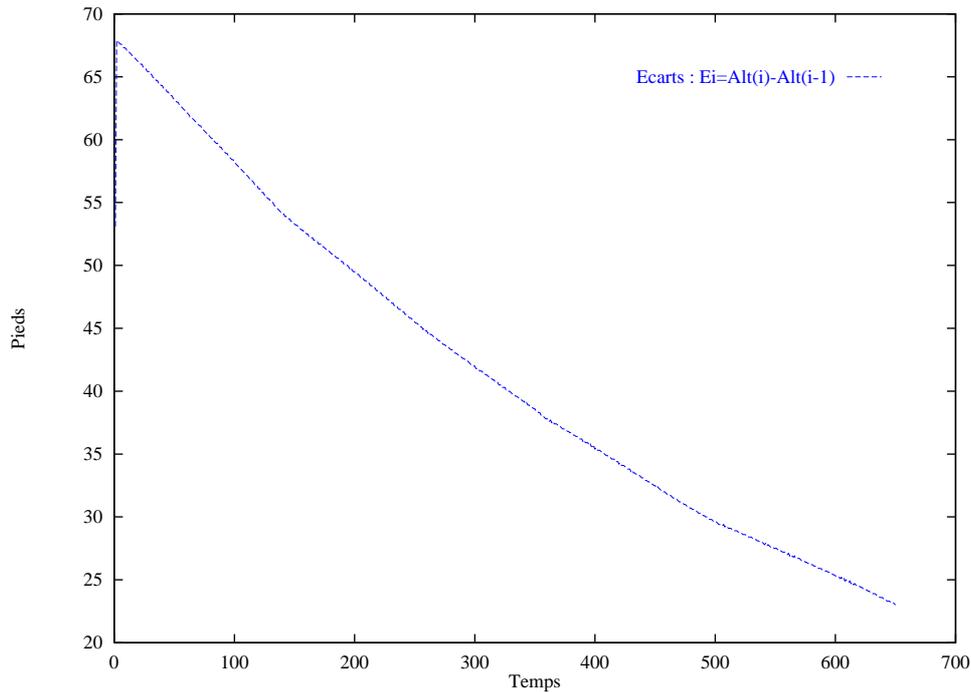
5.1.2 Méthode utilisée

Les inconvénients s'accumulent donc pour utiliser une méthode non-réursive :

- Pour la méthode directe : outre le nombre important de données nécessaires à une bonne estimation de f^4 , nous avons vu dans la section 5.1.1 que cette méthode ne peut être appliquée dans le cas d'observations strictement croissantes (ou décroissantes), c'est à dire pour des trajectoires monotones.
- Pour la méthode « D » : le nombre de paramètres à régler en vue d'une bonne prévision passe de un à deux, comme nous l'avons vu, de plus ces paramètres sont directement liés et peuvent varier suivant la date de la prévision, et l'horizon retenu.

Nous utiliserons donc une méthode réursive dont le danger principal est le risque d'amplification de l'erreur de prévision initiale, mais dont les résultats semblent satisfaisants (voir section 5.2). Cette méthode est également préféré par les théoriciens (voir A. C. Rosa [25]). Bien entendu nous filtrerons les données originales conformément à (5.2). Ainsi notre estimateur non-paramétrique sera appliqué à la série des « innovations » pour une prévision à court terme répétée. Nous nous attacherons donc à bien estimer cette série des écarts représentée en 5.6. En outre, le choix de la fenêtre est « facilité » par l'aspect local de la prévision opérée par \hat{f} à horizon un uniquement (voir section 5.2.2), sa détermination restera cependant un élément crucial pour la prévision.

4. Le fait d'estimer f hors de son support (au delà du dernier point observé notamment) aggrave ce phénomène.

FIG. 5.6 - Série des innovations: $e_i = Y_i - Y_{i-1}$

5.1.3 Et les séries chronologiques ?

Si nous détaillons encore un peu notre méthode, nous nous apercevons que nous ne sommes pas loin d'une modélisation de type série chronologique. En effet, notre optique est désormais d'estimer les écarts $(e_i)_{i=1,\dots,n}$, de prévoir non-paramétriquement e_{i+1} et d'en déduire une prévision de Y_{n+1} . Mais la série des écarts présente elle aussi une tendance comme on peut le constater en (5.6). Nous pouvons donc réitérer notre démarche et travailler sur les écarts « d'ordre 2 », notés $(\varepsilon_i)_{i=1,\dots,n}$ et définis comme suit :

Comme :

$$Y_{i+1} = Y_i + (Y_{i+1} - Y_i) = Y_i + e_{i+1}$$

avec $(e_{i+1}) = Y_{i+1} - Y_i$; De même :

$$e_{i+1} = e_i + (e_{i+1} - e_i) = e_i + \varepsilon_{i+1}$$

avec $(\varepsilon_{i+1}) = e_{i+1} - e_i$.

Pour estimer Y_{i+1} il nous suffit donc d'estimer e_{i+1} ce qui peut se faire à l'aide des $(e_i)_{i=1,\dots,n}$ ou bien à l'aide des écarts d'ordre 2, les $(\varepsilon_i)_{i=1,\dots,n}$.

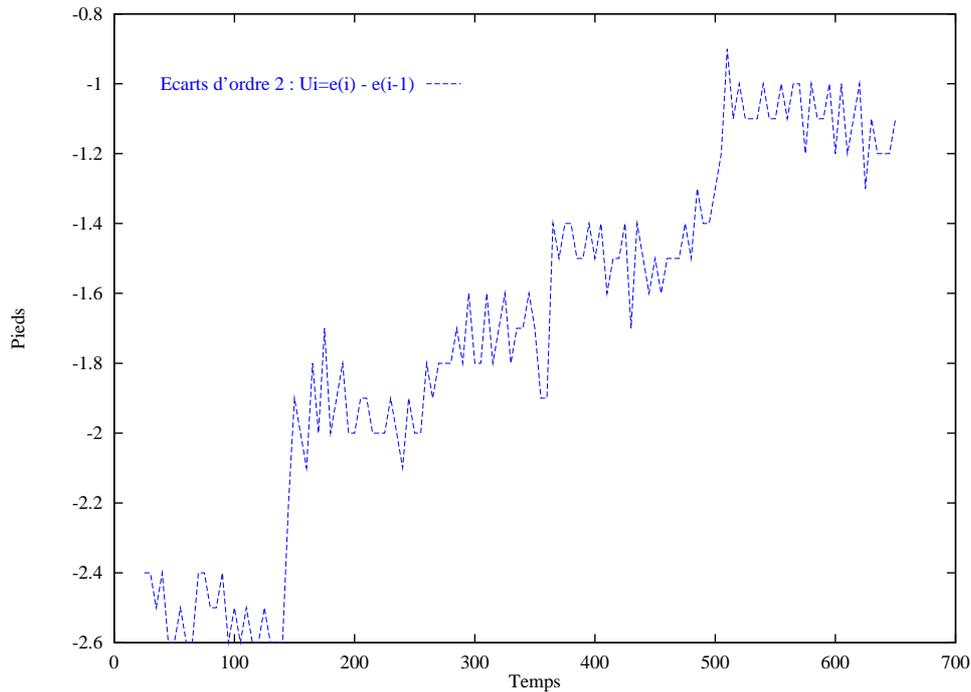
On peut alors écrire :

$$Y_{i+1} = Y_i + e_{i+1} = Y_i + e_i + \varepsilon_{i+1}$$

La première équation correspond à une modélisation de série chronologique d'ordre un ($Y_{i+1} = Y_i +$ une innovation), la seconde à une série chronologique d'ordre deux puisqu'alors, en remplaçant e_i par $Y_i - Y_{i-1}$, on obtient :

$$Y_{i+1} = 2 \cdot Y_i - Y_{i-1} + \varepsilon_{i+1}$$

Le type de modélisation que nous avons choisi (avec filtrage), s'apparente donc à une série chronologique dans laquelle nous estimons les innovations. Dans le cas d'un filtrage simple il s'agit d'un

FIG. 5.7 - Série des innovations « d'ordre 2 » : $u_i = e_i - e_{i-1}$

processus « autoregressif d'ordre un », dans le cas d'un double filtrage d'un processus « autoregressif d'ordre deux⁵ ». On espère ainsi obtenir finalement une série « stationnaire », ou au minimum, sans tendance. La représentation 5.7 des écarts d'ordre deux $(\varepsilon_i)_{i=1,\dots,n}$ n'est malheureusement pas conforme à nos attentes : non stationnaire, et extrêmement perturbée. Nous n'entrerons donc pas plus dans la voie de l'estimation de ε_{i+1} pour la prévision de Y_{i+1} , et nous contenterons d'un filtrage simple.

5.2 Prévisions à « court » terme

Nous avons donc utilisé la méthode du noyau de convolution sur les écarts (avec filtrage), de manière itérative afin de prévoir une trajectoire de 800 secondes issue du simulateur du CENA. Nous présentons les résultats dans les graphiques 5.8 et suivants.

Pour bien comprendre les résultats exposés nous nous devons de préciser certains termes utilisés :

- Le « *pas* » désigne le temps écoulé entre deux observations. Typiquement, 5s, 10s et 15s seront des pas rencontrés.
- L'« *horizon* » désigne l'horizon de prévision dans le sens « *nombre de points prévus* », ou si l'on préfère, le nombre d'itérations de la procédure récursive. Pour obtenir l'horizon de prévision en secondes il suffit de multiplier le « *pas* » par l'« *horizon* ».

5. Un processus autoregressif est défini par l'équation générale suivante:

$$Y_i = \sum_{i=1}^p \alpha_i \cdot Y_{i-p} + \varepsilon_i$$

où ε_i est un « bruit blanc » et donc $E[\varepsilon_i] = 0$ et $Var[\varepsilon_i] = \sigma^2$ constante. L'entier p est l'ordre du processus.

- Les «fenêtres» correspondent aux paramètres utilisés dans l'estimateur non-paramétrique. La fenêtre «*hopt*» est une fenêtre estimée naïvement et sert de référence, d'échelle, pour la recherche de fenêtres adaptées. Son calcul est détaillé dans la section 5.2.2.
- Les «trajectoires prévues» sont représentées de la manière suivante. A l'instant T_n on prévoit la trajectoire pour l'instant $T_{n+h_{or}}$, c'est à dire $\hat{Y}_{n+h_{or}}$. La courbe des «trajectoires prévues» est celle des couples $(T_{n+h_{or}}, \hat{Y}_{n+h_{or}})$.
- Les «erreurs de prévision» sont calculées par la différence «prévu - réalisé» en chaque point d'observation. C'est à dire que l'on reproduit le couple (T_i, \hat{d}_i) où \hat{d}_i est calculé comme $\hat{d}_i = \hat{Y}_i - Y_i$.

Dans cette section nous présenterons les résultats de prévision pour des «pas» de 5 secondes, et pour des horizons allant de 10 à 40 soit des prévisions maximales de 200 secondes. En ce sens certaines prévisions dépassent le «court» terme, que nous pouvons limiter arbitrairement à environ 100 secondes.

5.2.1 Du rôle du «pas»

Le «pas» entre deux observations joue, nous le verrons un grand rôle dans la prévision. Le nombre de points observés est évidemment très restreint lorsque l'on observe qu'un point toutes les 10 ou 15 secondes. En outre, l'espacement entre les points nuit à la précision de la prévision.

La fenêtre devra également s'adapter à ce fait. Plus le «pas» sera grand, plus la fenêtre devra être grande pour incorporer un nombre de points suffisants pour l'estimation. Il ne s'agit pas là que d'un problème d'échelle, mais également d'une prise en compte de la densité des points d'observation.

Un autre phénomène vient s'ajouter à cette description, nous ne disposons que peu d'observations pour prévoir puisque la trajectoire se stabilise au bout de 700 secondes. Avec un «pas» de 15 secondes cela nous laisse à peine 47 points pour la trajectoire complète. L'estimation s'effectue donc le plus souvent avec quelques dizaines de points seulement, et quelques dizaines d'observations de l'erreur de prévision. Si nous voulions raffiner le choix de la fenêtre par une procédure se basant sur l'observation de l'erreur de prévision sur le début de la trajectoire, celle-ci ne serait sûrement pas opérationnelle, faute de points, pour un «pas» important.

5.2.2 Du choix pratique de la fenêtre

La figure 5.8 qui donne la prévision pour un pas de 5 secondes et à horizon 10, nous servira d'exemple pour discuter des choix effectués pour sélectionner la fenêtre.

Comme nous pouvons le voir sur les graphiques 5.8, 5.9 et suivants, et comme nous l'avons évoqué précédemment, selon le choix de la fenêtre, la prévision donne des valeurs très différentes. Nous présenterons ici toutes nos prévisions pour trois valeurs de h différentes. Nous pourrions ainsi visualiser comment la fenêtre, en arbitrant sur le degré de variabilité et de douceur, modifie l'estimateur non-paramétrique aux différentes contraintes de chacune des grandes phases de la trajectoire décrites section 1.3). Nous voyons également que les erreurs de prévision peuvent varier du simple au double.

Quelques remarques s'imposent au vu de ces premiers résultats, concernant le choix du paramètre de fenêtre.

- La fenêtre h_{opt} sert ici de «calibre» pour le calcul de fenêtres mieux adaptées⁶. Il s'agit de la fenêtre naïve de Bosq et Lecoutre [2] décrite dans la section 4.4.3. Elle est calculée sur la

6. Il est notoire que cette fenêtre est généralement trop grande et que l'estimation effectuée avec h_{opt} est trop «lisse», trop régulière, pour donner une prévision correcte (voir Härdle [15], ou Bosq et Lecoutre [2]). Nous le vérifierons dans nos prévisions.

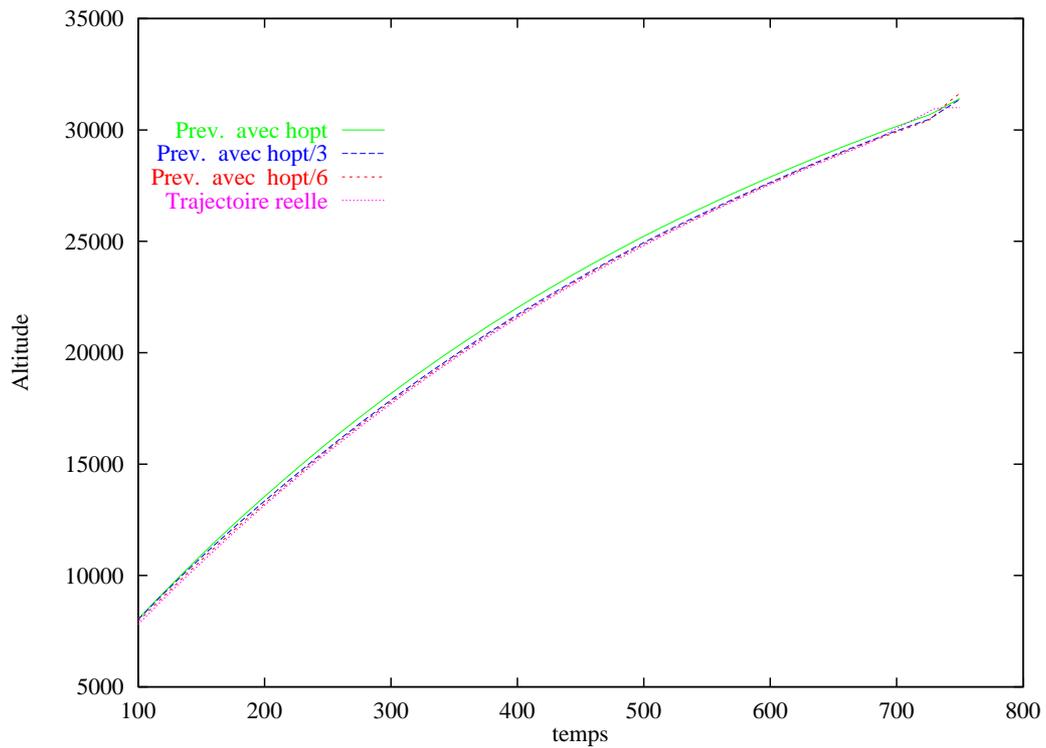


FIG. 5.8 - Estimation de la trajectoire pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes .

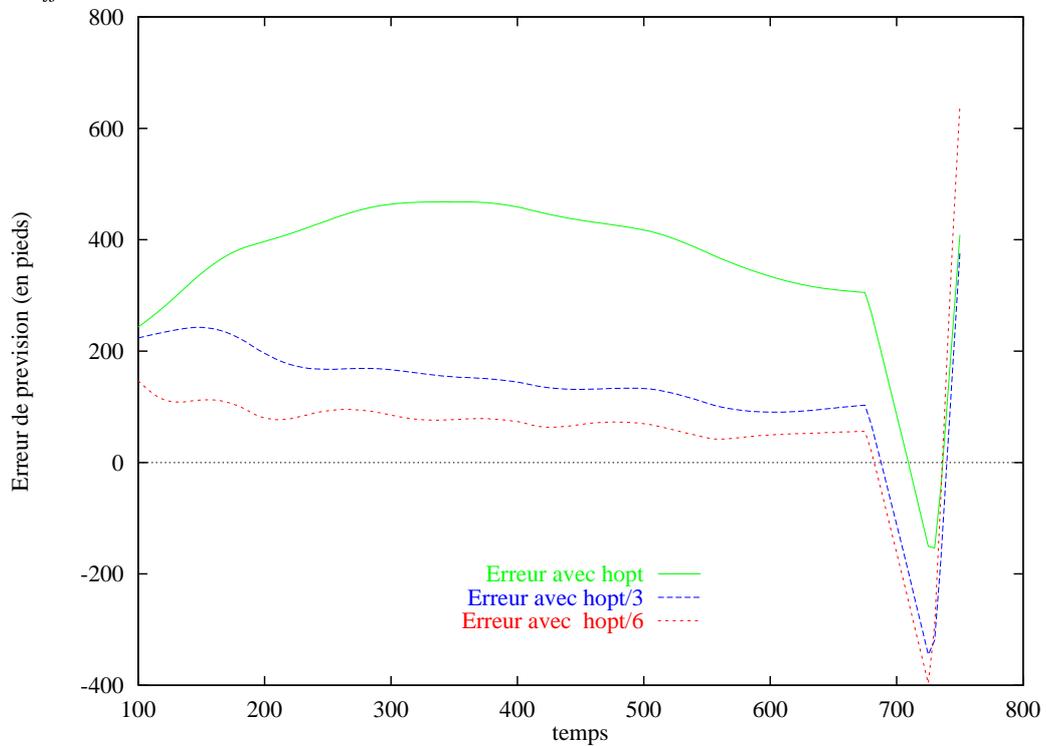


FIG. 5.9 - Erreur de prévision pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes .

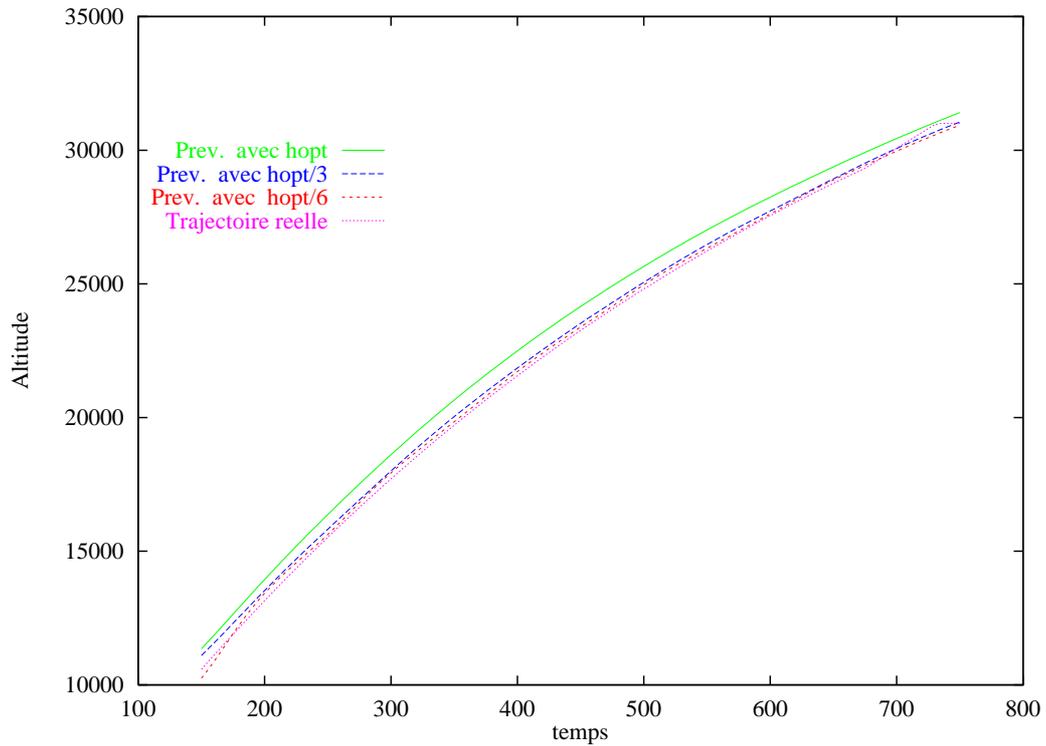


FIG. 5.10 - Estimation de la trajectoire pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes .

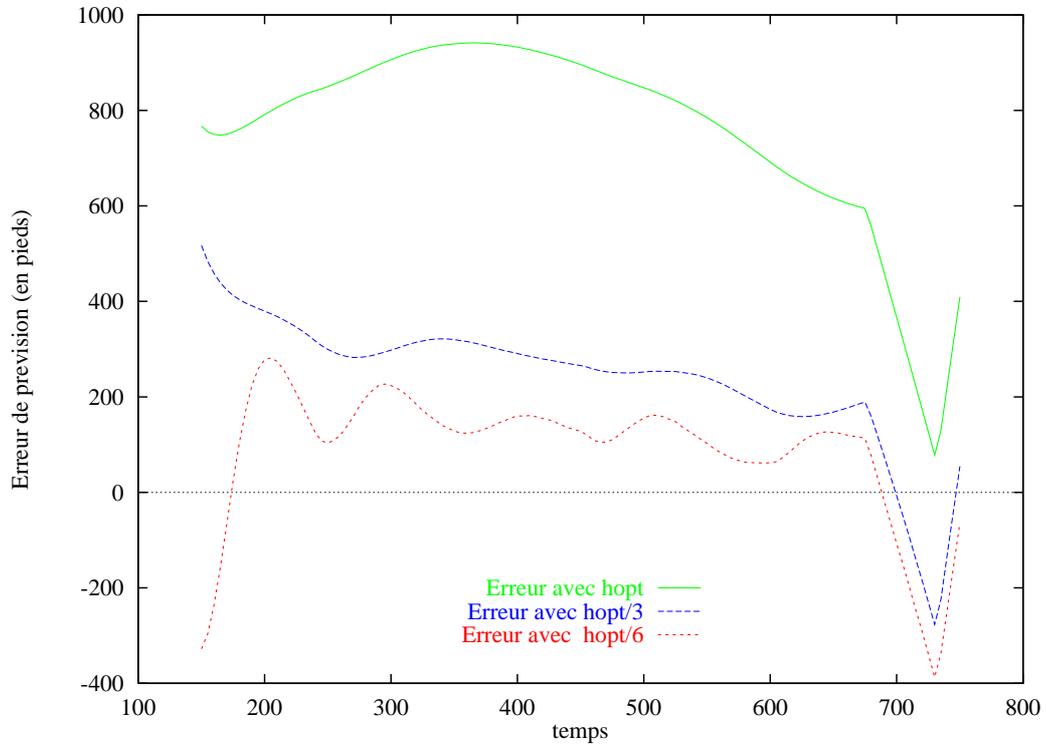


FIG. 5.11 - Erreur de prevision pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes .

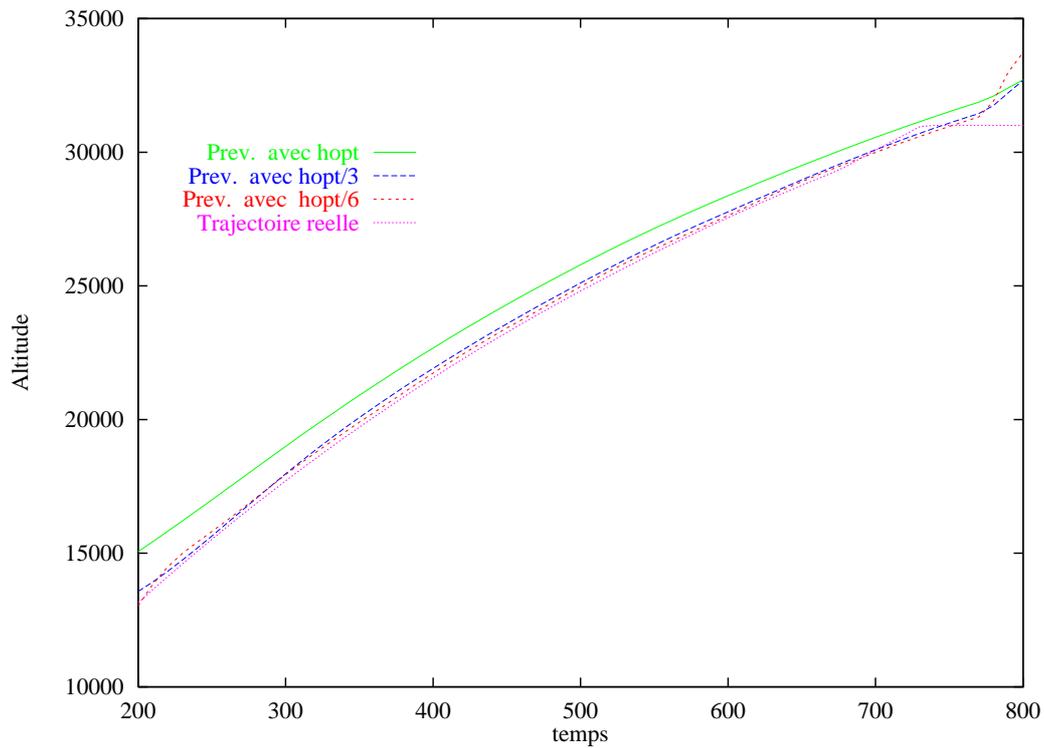


FIG. 5.12 - Estimation de la trajectoire pour un « pas » de 10s et à horizon 10, pour trois valeurs de h différentes .

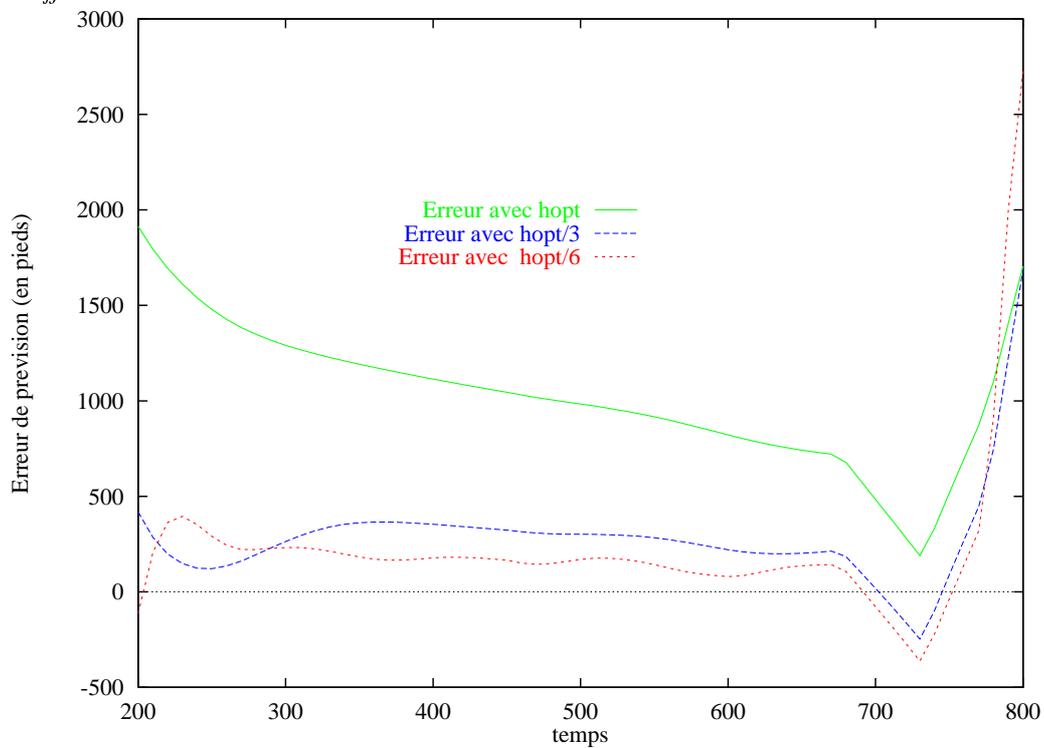


FIG. 5.13 - Erreur de prevision pour un « pas » de 10s et à horizon 10, pour trois valeurs de h différentes .

base de l'écart type des observations.

$$h_{opt} = \alpha \cdot S_n \cdot n^{-\frac{1}{5}}$$

où S_n est l'écart type empirique des T_i et où α ne dépend que du noyau choisi⁷. Cette fenêtre tient compte de l'espacement entre les points et de la dimension de l'échantillon observé (n ici). Plus le pas sera grand, plus l'écart type des T_i sera important et plus h_{opt} sera grande. D'autre part, plus n sera important, plus la fenêtre sera petite.

- Une fenêtre peut difficilement être préférée sur l'ensemble de la trajectoire. Nous avons systématiquement représenté les trajectoires pour deux fenêtres plus petites que h_{opt} . Souvent l'erreur de prévision est plus faible pour $\frac{h_{opt}}{6}$, notamment dans la deuxième phase de la trajectoire. Cependant, la variabilité de cette erreur augmente dangereusement avec une fenêtre petite (voir 5.10 et 5.11 par exemple).
- Réciproquement des fenêtres plus grandes, comme h_{opt} donnent une erreur plus régulière mais plus marquée.
- Plus la fenêtre est petite, plus la moindre variation de la trajectoire observée est incorporée voire amplifiée. Ce caractère s'observe notamment aux alentours de $t = 700s$ et pour des prévisions à long terme.

Nous nous proposons d'observer les résultats pour des pas de 5 et 10 secondes à horizons 10 et 20, obtenus avec différentes fenêtres, afin de pouvoir en tirer des éléments de discussion sur le choix des fenêtres.

5.2.3 Pour en finir avec les fenêtres...

Il semble désormais évident qu'une procédure de choix de la fenêtre doit faire intervenir 3 éléments principaux :

- le « pas » d'observation ;
- la « phase » de la trajectoire dans laquelle on opère la prévision ;
- l'« horizon » de prévision.

La fenêtre h_{opt} et ses sœurs ($\frac{h_{opt}}{3}$, $\frac{h_{opt}}{6}$, etc) ne fournissent qu'un guide pour la sélection. Une réelle estimation de « la » fenêtre \hat{h} répondant à notre objectif, c'est à dire minimisant l'erreur de prévision, doit être choisie empiriquement sur les données.

On pourrait par exemple choisir \hat{h} sur la base de l'observation de l'erreur de prévision sur une période passée. La fenêtre ainsi sélectionnée serait utilisée pour les prévisions suivantes pendant quelque temps tout en contrôlant l'erreur qu'elle engendre périodiquement. Si l'erreur devient trop importante, une nouvelle procédure de sélection pourra alors être effectuée. Ce type d'algorithme engendrerait donc une suite de fenêtres locales répondant bien aux différentes contraintes évoquées en début de chapitre. Cependant cette procédure s'avère gourmande en calculs.

5.3 Prévision à « Long » terme

Les mêmes procédures d'estimation ont été appliquées aux prévisions à long terme. Nous nous sommes heurtés à un problème nouveau quoique peu surprenant. La récursivité de la procédure de prévision engendre sur le long terme des erreurs gigantesques. Notre algorithme étant basé sur l'estimation des écarts à horizon un répété itérativement jusqu'à l'arrivée à l'horizon de prévision,

7. Ici α vaut 1,059.

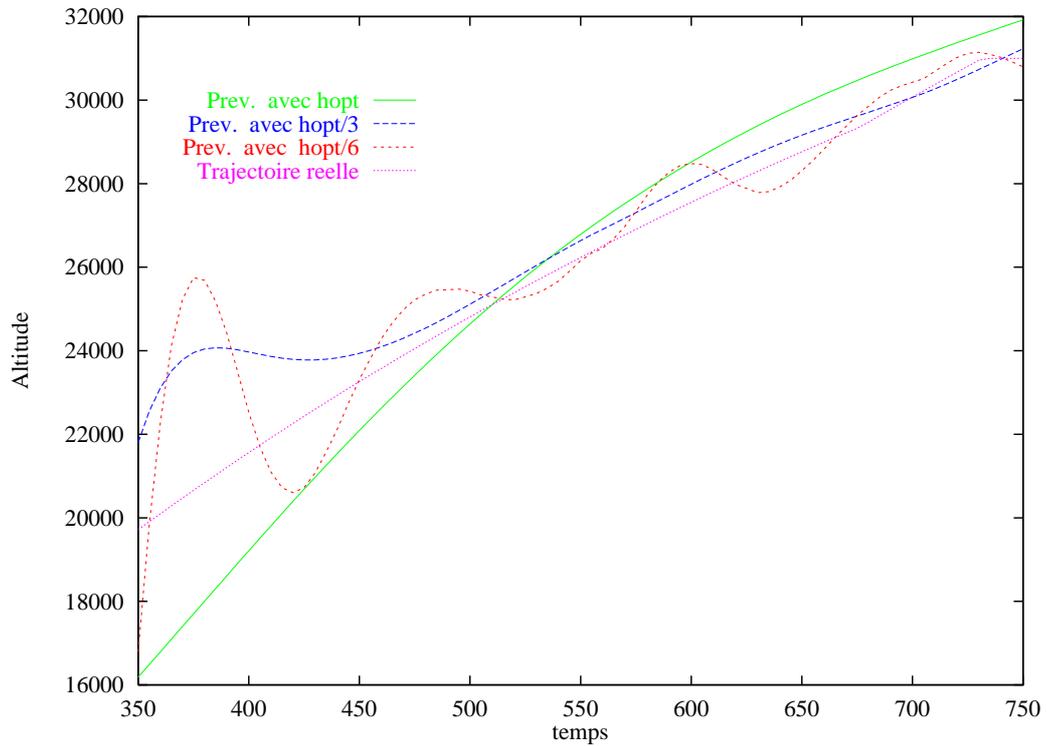


FIG. 5.14 - Estimation de la trajectoire pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes .

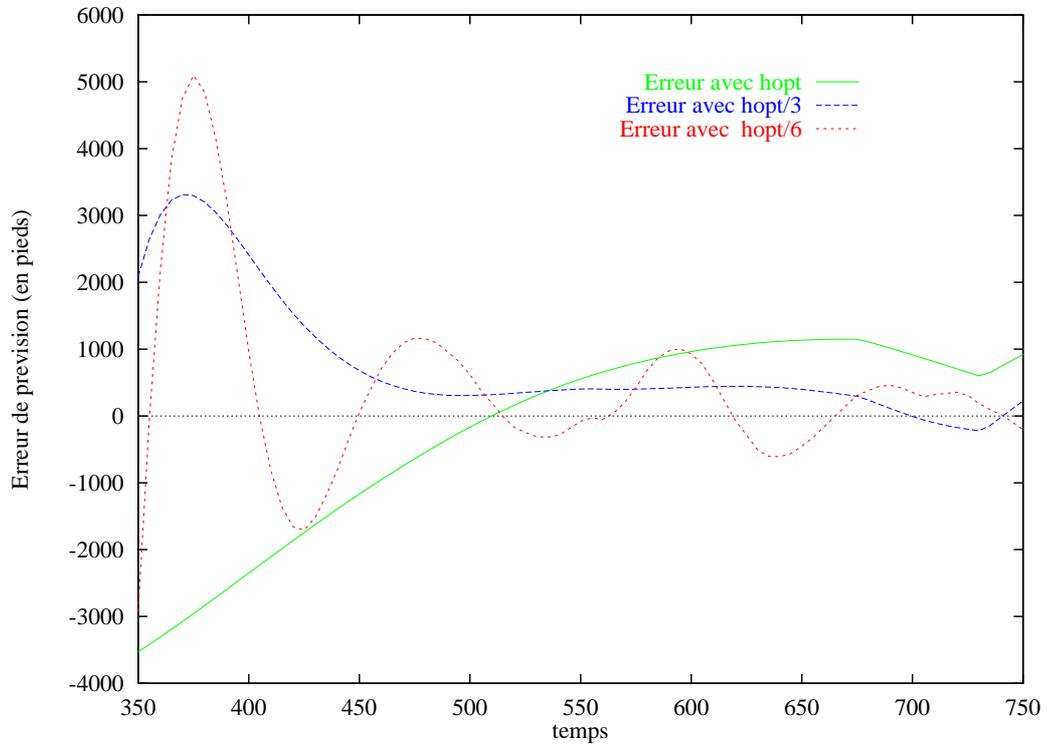


FIG. 5.15 - Erreur de prévision pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes .

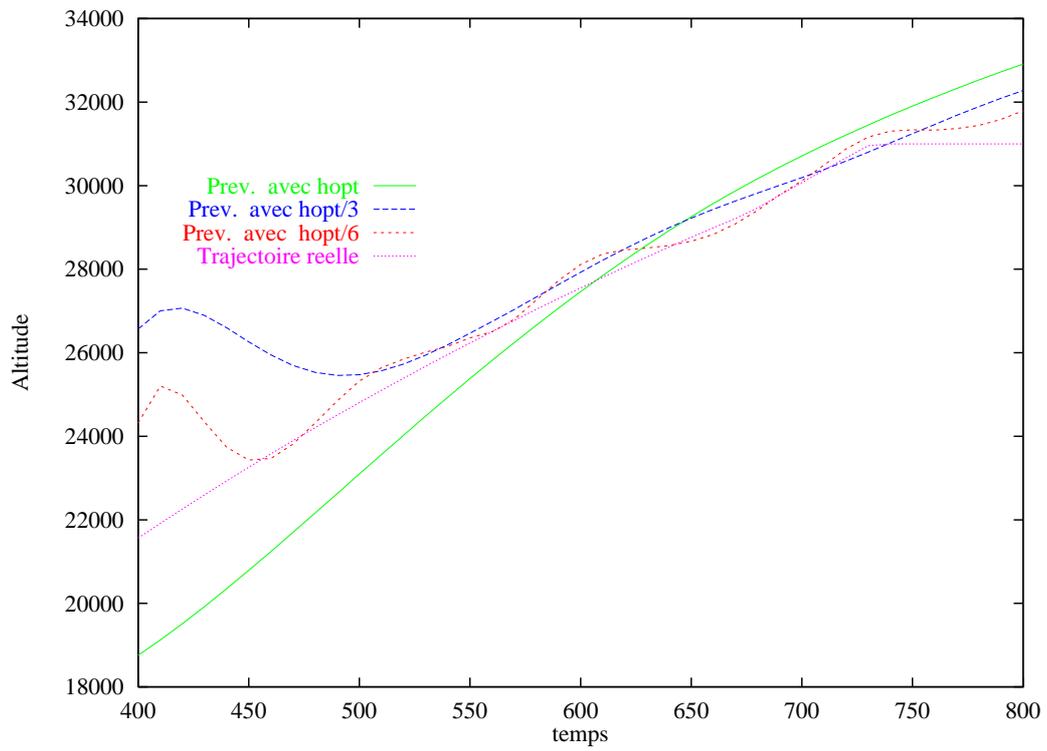


FIG. 5.16 - Estimation de la trajectoire pour un « pas » de 10s et à horizon 20, pour trois valeurs de h différentes .

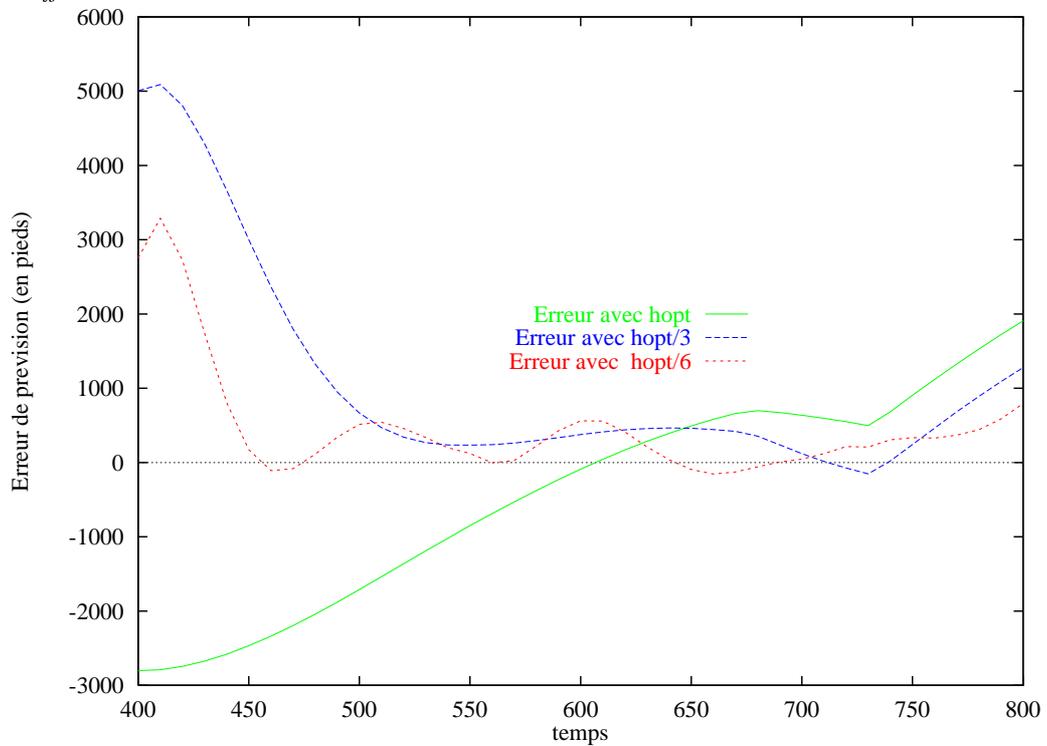


FIG. 5.17 - Erreur de prévision pour un « pas » de 10s et à horizon 20, pour trois valeurs de h différentes .

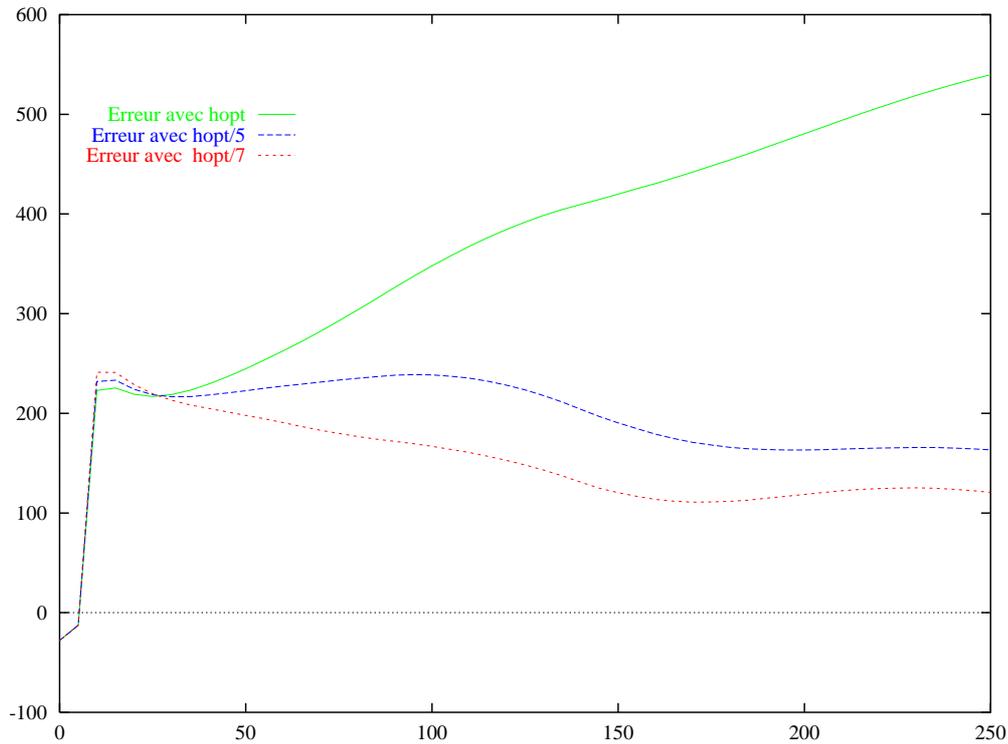


FIG. 5.18 - Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 10, pour trois valeurs de h différentes .

la moindre erreur initiale se trouve amplifiée. Nous pouvons d'ailleurs en juger sur les graphiques 5.18, 5.19 et 5.20 dans la section 5.3.1, représentant les erreurs en début de trajectoire (de 0 à 300 secondes environ).

Les résultats présentés concernent des prévisions effectuées :

- pour un pas de 5 secondes à horizon 40 soit une prévision de 200 secondes.
- pour un pas de 10 secondes à horizon 20 soit également une prévision de 200 secondes.

La comparaison est extrêmement instructive, on s'attend à obtenir des erreurs plus faibles pour un pas plus petit, puisqu'on observe plus de points. Cependant, la récursivité de notre procédure entraîne des « ondulations » beaucoup plus importantes dans le premier cas (pas 5s, horizon 40) et donc des variations des trajectoires assez importantes. Ce phénomène est d'autant plus important que la fenêtre est petite (voir figures 5.14 et 5.15). L'ordre de grandeur des erreurs reste sensiblement le même en phase de montée. La deuxième procédure utilisant moins de points (pas 10s, horizon 20) est certainement moins précise dans la prévision initiale, mais les erreurs se répercutent moins, d'où l'allure plus « lisse » des trajectoires, même pour de « petites » fenêtres(voir figures 5.16 et 5.17).

Au delà de l'horizon 200, les erreurs deviennent importantes sur la partie initiale de la trajectoire, les erreurs se propagent et la trajectoire prévue devient absolument farfelue.

5.3.1 Étude de l'erreur initiale

Les figures 5.18 5.19 et 5.20 donnent les écarts d'après notre prévision récursive, sur les premières observations.

Il est à noter qu'avec un pas de 5 secondes nous n'observons donc que 10 valeurs pendant les 50 premières secondes. On peut également observer que la trajectoire obtenue avec la plus

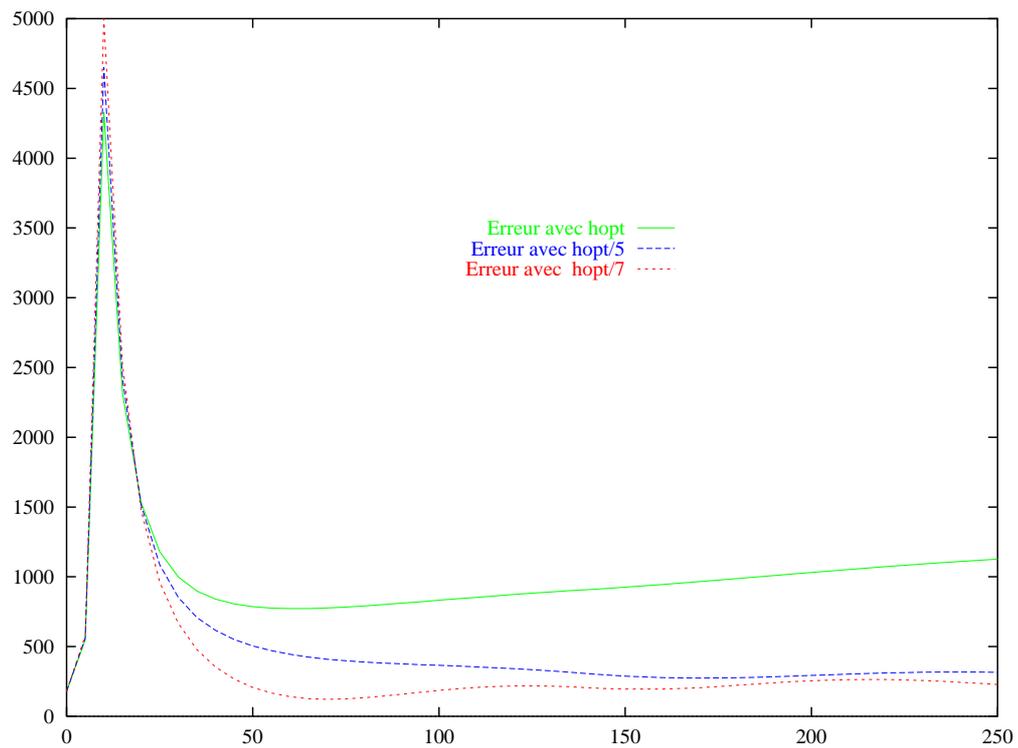


FIG. 5.19 - Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 20, pour trois valeurs de h différentes .

petite fenêtre (en rouge) est celle que recolle le plus aux données et qui rectifie les erreurs le plus rapidement. Toutefois, pour une prévision à horizon 40, les erreurs restent très importantes quelle que soit la fenêtre utilisée voir figure 5.20.

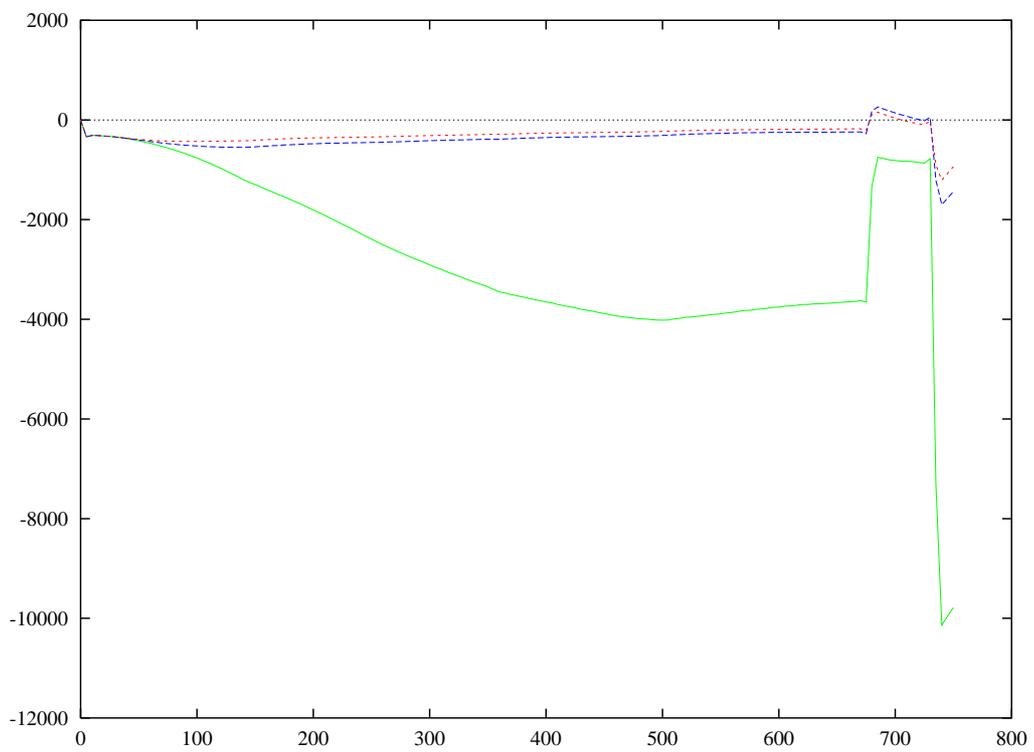


FIG. 5.20 - Erreur « initiale » de prévision pour un « pas » de 5s et à horizon 40, pour trois valeurs de h différentes .

Chapitre 6

Bilan et extensions

Les diverses méthodes de prévision de trajectoires que nous avons proposées, donnent des résultats mitigés sur nos applications. En l'absence de complément d'information sur les éléments déterminant la trajectoire d'un avion, il semble que l'approche stochastique soit toutefois une voie de recherche à poursuivre.

Les méthodes « classiques » que nous avons détaillées dans les chapitres 2 et 3 ont clairement montré leurs limites. Il est difficile, à vrai dire, d'approcher un élément d'un espace fonctionnel par un élément d'une famille paramétrée. La dimension de l'espace de recherche est en effet trop petite pour espérer approcher convenablement la fonction trajectoire inconnue, à moins d'explorer exhaustivement l'ensemble des familles paramétrées, ce qui ne semble pas raisonnable. Toutefois, si des informations concernant la nature de la trajectoire étaient connues, cette démarche serait facilitée et les résultats meilleurs. Par choix, nous n'avons pas cherché à caractériser les trajectoires par des éléments techniques faisant intervenir les contraintes physiques de mécanique du vol, laissant cette approche à ceux qui en ont la compétence (voir l'étude Leclair [18]).

L'approche fonctionnelle non-paramétrique évite, en théorie tout du moins, de restreindre l'espace de recherche et d'avoir à choisir *a priori* une spécification *ad hoc* pour la trajectoire. Toutefois non-paramétrique ne signifie pas absence de paramètres et nous l'avons vu, le choix de la fenêtre s'avère crucial. C'est d'ailleurs sur ce point que des efforts théoriques et pratiques restent à faire. Les critères de sélection proposés section 4.4 semblent efficaces à court terme, mais pas à long terme. Sur nos trajectoires, il est difficile de calibrer convenablement ce paramètre sur un petit nombre d'observations initiales sans s'exposer à des problèmes à long terme. Une approche similaire à la validation croisée (voir 6.2 ci-dessous) pourrait toutefois être développée en présence d'informations extérieures.

6.1 Raffinement des estimateurs et des procédures

De nombreuses extensions de ce travail sont possibles et nous nous proposons d'examiner les procédures qui nous semblent les plus adaptées à la prévision de trajectoire. Ces méthodes nécessitent un supplément de travail pour être développées ou/et un supplément d'information sur les caractéristiques *a priori* de la trajectoire. Nous mentionnerons ainsi le travail de A. C. Rosa [25] portant sur la prévision de séries chronologiques extrêmement générales, puisque seule l'hypothèse d'ergodicité est supposée. Les résultats théoriques peuvent sûrement être adaptés à notre problème, et notamment apporter des éléments de compréhension sur la structure autorégressive de la trajectoire. Toutefois, ces méthodes nécessitent un nombre important de données ; Pour être appliquées, ces méthodes nécessiteraient donc un supplément d'information.

L'utilisation d'une procédure non récursive pour la prévision à long terme est également l'un des points clés à développer. Les résultats à long terme souffrent d'un manque de précision lié à la répercussion des erreurs par une procédure récursive. Cette récursivité est liée à la nature de la trajectoire qui est strictement monotone, et qui nous force à travailler sur des données

filtrées (voir 5.1.1). Lever cet écueil contribuerait grandement à améliorer les prévisions. L'une des possibilités consisterait à ne travailler que sur les erreurs issues d'un autre modèle de prévision. Si ce modèle de base fournit des erreurs non monotones, c'est à dire s'il ne sous-estime ou ne sur-estime pas systématiquement la trajectoire, on peut appliquer les méthodes de prévision sur ces erreurs sans avoir à filtrer et donc sans avoir recours à une procédure récursive. Là encore, un supplément d'information (le modèle de prévision initial) ou de travail est nécessaire pour améliorer les prévisions.

6.2 Utilisation d'informations complémentaires

Deux types d'informations peuvent réalistement venir compléter et renforcer les données observables. Des informations sur le type d'avion (caractéristiques techniques, compagnies, ...) et des données sur la nature du vol (plan de vol, niveaux de vol, ...)

- Suivant les caractéristiques de l'avion (sa puissance, sa vitesse maximum, son poids, ...), sa nature (commerciale, privée, militaire, ...) l'évolution de la trajectoire diffère. Cette information peut être très utile pour la prévision. Dans le cas d'une estimation classique, le choix du modèle de trajectoire (la famille paramétrée), peut être conditionné par ces éléments. Dans le cas de procédures non-paramétriques, ces informations techniques permettraient d'envisager un pré-calibrage de la fenêtre d'estimation utilisée. La fenêtre pourrait alors être calibrée dès le début de la trajectoire par validation croisée, préalablement à la prévision de la trajectoire, par l'utilisation de trajectoires types simulées. Une observation périodique des erreurs de prévision permettrait d'ajuster au mieux la courbe en réduisant ou augmentant la valeur de cette fenêtre. Cette procédure se rapproche d'ailleurs des travaux de Vieu [30] sur l'estimation non-paramétrique avec fenêtres locales.
- On peut également tirer profit de données sur la nature du vol; à l'approche de la stabilisation (pour une trajectoire en montée) la forme de la trajectoire se modifie. Dans le cas d'une modélisation classique on peut envisager une rupture et séparer la trajectoire en deux morceaux régis par deux familles paramétrées différentes. Pour une procédure non-paramétrique la fenêtre doit être modifiée pour prendre en compte cette rupture dans la forme de la trajectoire. Dans les deux cas cette information s'incorpore avec profit avant toute estimation. Il est également probable qu'un avion se stabilisant en niveau de vol, ne dépasse pas de beaucoup ce niveau, cette information peut s'incorporer facilement.

6.3 Vers les réseaux de neurones ?

L'une des voies de recherche que nous n'avons pas abordé dans ce mémoire et qui semble adaptée à ces difficiles problèmes de prévision concerne les réseaux de neurones. Ces méthodes ont été utilisées dans différents domaines pour approcher des structures non-linéaires complexes. La structure de ces opérateurs rappelle (en plus simple) la structure des réseaux constituant le cerveau humain, d'où leur nom.

L'analyse par réseaux de neurones fait partie de l'intelligence artificielle, mais peut être reliée à la statistique non-paramétrique, comme le font remarquer Härdle et alii [17]. Nous pouvons illustrer cette similitude par un exemple traitant de séries temporelles (voir section 5.1.3). On cherche alors à estimer la fonction reliant Y_t à son passé $(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p})$ par l'intermédiaire d'une fonction inconnue f , soit :

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}) + \varepsilon_t \quad (6.1)$$

L'estimateur issu d'un réseau de neurones *feedforward* à une couche cachée¹ peut en effet s'écrire sous la forme d'un estimateur non-paramétrique.

1. Il s'agit du plus simple des réseaux de neurones.

L'estimateur² ainsi proposé est une somme pondérée, rappelant la forme générale des estimateurs à noyau³.

Les estimateurs de ce type peuvent approcher de vastes classes de fonctions pourvu que q soit suffisamment grand. Ces méthodes sont encore difficiles à mettre en oeuvre, mais semblent prometteuses, la thèse de Yann Le Fablec viendra sans doute le confirmer. En tout état des cause, elles sont suffisamment complètes pour pouvoir capter toute l'information disponible, afin d'estimer f et d'en tirer des prévisions plus précises.

2. En effet, l'estimateur $\hat{f}(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p})$ issu d'un réseau de neurones *feedforward* s'écrit :

$$\hat{f}(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}) = \beta_0 + \sum_{i=1}^q G(\gamma_{0i} + Y_t^T \gamma_i) \beta_i \quad (6.2)$$

où $Y_t^T = (Y_{t-1}, Y_{t-2}, \dots, Y_{t-p})^T$ où les vecteurs $\gamma_j = (\gamma_{1j}, \dots, \gamma_{pj})^T$ sont des vecteurs $(p \times 1)$ pour $j = 1, \dots, q$ et où β_1, \dots, β_q sont des coefficients scalaires.

Dans cette expression la fonction $G : \mathbb{R} \rightarrow [0, 1]$ est une fonction de répartition pré-spécifiée (typiquement on utilise la fonction logistique $G(x) = \frac{1}{1+e^{-x}}$ ou hyperbolique $G(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$).

3. Ici, la fonction G joue le rôle du noyau.

Bibliographie

- [1] Bochner S. (1955) : « Harmonic Analysis and the Theory of Probability », *University of California Press*.
- [2] Bosq D. et J.F. Lecoutre (1987) : « Théorie de l'estimation fonctionnelle » *Economica*.
- [3] Bourbonnais R.(1993) : « Econométrie », *Dunod, Paris*.
- [4] Chen R. and W. Härdle (1995) : « Nonparametric Time series Analysis, a selective review with examples », working paper No. 14, Sonderforschungsbereich 373, Humbolt Univerty, Berlin.
- [5] Collomb G. (1976) : « Estimation non-paramétrique de la régression par la méthode du noyau », *Thèse*, Université Paul Sabatier, Toulouse.
- [6] Coutrot B. et F. Drosbeke (1984) : « Les Méthodes de prévision » *Que sais-je? No. 2157, Presses Universitaires de France, Paris*.
- [7] Delecroix M. (1983) : « Histogrammes et Estimation de la Densité », *Que sais-je? No. 2055, Presses Universitaires de France*.
- [8] Eubank R. L. (1988) : « Smoothing Splines and Nonparametric Regression », *Marcel Dekker, New York*.
- [9] Fan J. and M. Muller (1995) : « Density and regression smoothing », working paper No. 1, Sonderforschungsbereich 373, Humbolt Univerty, Berlin.
- [10] Gasquet C. et P. Witomski (1989) : « Analyse de Fourier et Applications: Filtrage, Calcul Numérique, Ondelettes. » *Masson*.
- [11] Gasser T. and H. G. Müller (1984) : « Estimating Regression Function and their Derivatives by the Kernel Method », *Scandinavian Journal of Statistics*, Vol. 11, pp. 171-185.
- [12] Gourieroux C. et A. Monfort (1989) : « Statistique et modèles économétriques », Vol. 1 et 2, *Economica, Paris*.
- [13] Györfi L., W. Härdle, P. Sarda and P. Vieu (1989) : « Nonparametric Curve Estimation from Time Series », *Springer Verlag*.
- [14] Hamilton D. J. (1990) : « Time series analysis » *Princeton*.
- [15] Härdle W. (1990) : « Applied nonparametric regression ». *Econometric society monographs*, Cambridge University Press.
- [16] Härdle W and O. Linton (1995) : « Nonparametric regression », à paraître dans *Encyclopedia of statistical science*, vol. X, Wiley and Sons.
- [17] Härdle W, H. Lutkepohl and R. Chen (1996) : « A Review of Nonparametric Time Series Analysis », working paper, Sonderforschungsbereich 373, Humbolt Univerty, Berlin.

- [18] Leclaire C. (1997) : «Rapport d'études sur les fonctions de prévision de trajectoire et leurs évaluations », Rapport CENA/R97008.
- [19] Mack Y.P. and H.G. Müller (1989) : « Convolution type estimators for nonparametric regression », *Statistic and Probability Letters*, Vol. 7, pp. 229-239.
- [20] Métivier M. (1968) : « Notions fondamentales de la théorie des probabilités », Seconde édition, Dunod, Paris.
- [21] Marron J. S. (1988) : « Automatic Smoothing Parameter Selection : A Survey », *Empirical Economics*, Vol. 13, pp. 187-208.
- [22] Nadaraya E.A. (1964) : « On estimating regression », *Theory of Probability and its Applications*, 9, pp. 141-142.
- [23] Rao P. (1983) : « Nonparametric functional estimation », A Probability and mathematical statistics monographs and textbooks, *Academic press*.
- [24] Rice J. (1984) : « Bandwidth Choice for Nonparametric Regression » , *Annals of Statistics*, Vol. 12, no 4, pp. 1215-1230.
- [25] Rosa A. C. (1993) : « Prévision robuste sous une hypothèse ergodique », *Thèse de Mathématiques appliquées*, Université P. Sabatier, Toulouse.
- [26] Saporta G. (1990) : « Probabilités, analyse des données et statistique », *editions Technip, Paris*.
- [27] Spanos A. (1986) : « Statistical Foundations of Econometric Modelling », *Cambridge University Press*.
- [28] Stone, C. J. (1982) : « Optimal global rates of convergence for nonparametric regression », *Annals of Statistics*, Vol 10, no 4, pp. 1040-1053.
- [29] Ullah A. and H. D. Vinod (1993) : « General Nonparametric Regression Estimation and Testing in Econometrics », in *Handbooks of Statistics*, Vol.11, G. S. Maddala, C. Rao and H. D. Vinod editors, Elsevier Science Publishers.
- [30] Vieu P. (1991) : « Nonparametric Regression : Optimal Local Bandwidth Choice », *Journal of the Royal Statistical Society*, series B, no 53, pp. 453-464.
- [31] Vieu P. (1993) : « Bandwidth selection for kernel regression », *Computational Statistics and Data Analysis*, à paraître.
- [32] Watson, G.S. (1964) : « Smooth regression analysis », *Sankhya, Series A*, 26, pp. 359-372.
- [33] T. H. Wonnacott et R. J. Wonnacott (1984) : « Statistique », *Economica*.