THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse III Paul Sabatier (UT3 Paul Sabatier)

Discipline ou spécialité :
Mathématiques appliquées

Présentée et soutenue par
Cindie ANDRIEU
le : 24 septembre 2013

Titre :
Modélisation fonctionnelle de profils de vitesse en lien avec l'infrastructure et méthodologie de construction d'un profil agrégé

École doctorale :
Mathématiques Informatique Télécommunications (MITT)

Unités de recherche :
IFSTTAR/COSYS/LIVIC & IMT (UMR 5219)

Directeurs de thèse :
Xavier Bressaud, Université Toulouse III
Guillaume Saint Pierre, IFSTTAR

Rapporteurs :
Gérard Biau, Université Paris VI
Jérôme Saracco, Institut Polytechnique de Bordeaux

Autres membres du jury :
Christine Thomas-Agnan, Université Toulouse I
Patrice Aknin, IFSTTAR

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ TOULOUSE III PAUL SABATIER

Spécialité :

Mathématiques Appliquées

École doctorale :

Mathématiques, Informatique et Télécommunications de Toulouse (MITT)

Présentée par

Cindie ANDRIEU

le 24 septembre 2013

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ TOULOUSE III PAUL SABATIER

Modélisation fonctionnelle de profils de vitesse en lien avec l'infrastructure et méthodologie de construction d'un profil agrégé

Thèse présentée devant le jury composé de :

Xavier Bressaud	Professeur, Université Toulouse III	*(Directeur de thèse)*
Guillaume Saint Pierre	Chargé de recherche, IFSTTAR	*(Co-directeur de thèse)*
Gérard Biau	Professeur, Université Paris VI	*(Rapporteur)*
Jérôme Saracco	Professeur, Institut Polytechnique de Bordeaux	*(Rapporteur)*
Christine Thomas-Agnan	Professeur, Université Toulouse I	*(Examinateur)*
Patrice Aknin	Directeur de recherche, IFSTTAR	*(Examinateur)*

Remerciements

Mes premiers remerciements vont à mes deux directeurs de thèse Xavier Bressaud et Guillaume Saint Pierre. Un grand merci à Guillaume Saint Pierre pour m'avoir toujours fait confiance et guidé depuis mon stage de M2 en 2009, et pour m'avoir convaincu de me lancer dans cette difficile mais très enrichissante aventure qu'est la thèse ! Je le remercie pour ses nombreux conseils et son soutien tout au long de ces années au LIVIC, sans oublier les discussions philosophiques à la cantine ou à la salle café ! Je remercie Xavier Bressaud pour m'avoir mis sur le chemin des belles mathématiques. Ma 1ère année à Toulouse a été très enrichissante d'un point de vue scientifique et les précieux conseils (même par téléphone) de Xavier m'ont beaucoup aidé durant ces trois années de thèse.

Je remercie Gérard Biau et Jérôme Saracco d'avoir accepté de rapporter cette thèse. Je les remercie pour leur lecture rigoureuse de mon manuscrit et pour l'intérêt qu'ils ont manifesté pour mon travail.

Je remercie Patrice Aknin et Christine Thomas-Agnan d'avoir accepté de faire partie du jury de cette thèse. Merci à Patrice Aknin d'avoir contribuer à ce que ce sujet de thèse existe lors des candidatures pour l'obtention d'un financement de thèse à l'IFSTTAR. Un grand merci à Christine Thomas-Agnan pour sa gentillesse et son aide précieuse d'experte lors de ma dernière année de thèse.

Je tiens également à remercier tous les membres du LIVIC pour l'ambiance sympathique et accueillante qui y règne. Tout d'abord je remercie Jacques Ehrlich de m'avoir accueilli dans ce laboratoire. Puis je remercie Aurélien, Dominique, Sébastien, Olivier, Clément, Mickaël, Rachid, Jean-Marie, Jérôme (ex-livicien), Mariana, Alain, Redouane, Vincent, Isnié (pour son aide à surmonter les difficultés administratives) et je dois en oublier... Plus personnellement, un grand merci à Joëlle (et Cassis) pour nos longues discussions à propos de nos compagnons respectifs. Un grand merci également à Lydie pour sa gentillesse et ses délicieux macarons !

Je remercie également les membres de l'IMT et plus particulièrement Jean-Michel Loubes et Jérémie Bigot pour leurs précieux conseils scientifiques. Un grand merci également à Agnès Requis et Martine Labruyère de l'ED MITT pour m'avoir aidé dans les procédures administratives et dans le remplissage pas toujours simple de l'ADUM...

Un grand merci à mon groupe de statisticiens préférés : Tata Nacera, Mohammed, Rémi, Mory (Bouh !), Manu (et Suzanna et Julia), Assia, Toby. Merci également aux

non-statisticiennes mais amies Aude et Nathalie.

Je tiens également à remercier Jean-François Korngut (JFK !) pour m'avoir guidé sur le chemin des mathématiques alors que je n'étais encore qu'au lycée et que je m'orientais vers des étude musicales. Son soutien amical et nos discussions mathématiques et musicales ont toujours été très précieux pour moi.

Un grand merci à ma famille. Tout d'abord mes parents qui ont toujours été présents que ce soit avant ou pendant ma thèse. Leur soutien et leurs encouragements m'ont beaucoup aidé. Merci également à ma soeur Karine (et Philippe, Bastien et Lucie) pour les repas de famille à Mandres, toujours dans la bonne humeur, qui m'ont permis de recharger les batteries quand j'en avais besoin. Merci à mon parrain Gilbert, et à sa femme Marcelle pour les powerpoints distrayants (surtout ceux avec des chats !). Merci à ma famille Aveyronnaise et Cantalou (je pense en particulier à ma grand-mère, toujours au top de sa forme !), et particulièrement aux toulousains (ou ex-toulousains) Raymond, Marie-Rose, Carole, Daniel et MC pour leur soutien lors de mon année à Toulouse. Merci également à ma belle-famille pour son soutien et plus particulièrement à mon beau-père pour ses précieux conseils mathématiques.

Enfin un grand merci à Francis pour avoir supporté 24h/24 les moments de joie et de stress d'une thésarde, et les nombreux allers-retours Paris-Toulouse. Merci d'avoir toujours les mots pour me rassurer et d'avoir brillamment réussi tous mes défis LaTeX. Et merci à mon petit Filou pour les parties de jeu "rouleau-vole" le soir, et ses siestes sur mon ordinateur lors de la rédaction de ma thèse.

Table des matières

Introduction

Contexte

Aujourd'hui, l'exploitation du réseau routier nécessite la prise en compte de nombreux critères : rapidité, confort, sécurité, flexibilité, économie d'énergie... Il est alors important d'optimiser au mieux l'usage du réseau afin d'en augmenter l'efficacité, ce qui nécessite une connaissance détaillée de celui-ci et de son usage réel. Ainsi, il devient indispensable de disposer de moyens efficaces et fiables permettant de mesurer et d'évaluer l'usage réel du réseau, et tenant compte de la complexité des liens entre ces critères.

Les moyens utilisés pour observer et mesurer l'utilisation du réseau routier ont connu récemment des évolutions majeures : nous sommes passés de la collecte de données agrégées obtenues grâce à des équipements statiques (ex : boucles magnétiques, radars), à des mesures beaucoup plus fines obtenues à l'aide de boîtiers connectés aux ordinateurs de bord des véhicules. Cela a contribué à l'apparition du concept de véhicules traceurs, assimilé à celui de capteurs mobiles explorant le réseau routier en continu. Si jusqu'à présent, la plupart des véhicules traceurs étaient réduits à des flottes privées (ex : bus, taxis...), le développement des smartphones a permis d'accroître le nombre de "traces" numériques laissées par les véhicules, et ainsi d'étudier le comportement des usagers sur l'ensemble du réseau routier. Cependant, ces données très riches ne sont pas facilement accessibles, et demandent des équipements ou adaptateurs spéciaux : un boîtier enregistreur connecté au réseau électrique sur lequel transitent les données des capteurs (bus CAN), ou plus récemment un smartphone relié à ce même réseau par un adaptateur. Avant de s'intéresser à toutes ces données, deux mesures contiennent déjà beaucoup d'information sur le trafic et le comportement du conducteurs, et sont facilement accessibles dès lors que l'on dispose d'un système de positionnement satellitaire (GPS), à savoir la position et la vitesse du véhicule.

La connaissance des vitesses pratiquées est une caractéristique majeure de l'usage réel du réseau routier. En effet, plusieurs études (Ericsson [52], Laureshyn [93]) ont montré que le choix de la vitesse était l'une des composantes les plus importantes du comportement des usagers de la route. La connaissance des vitesses pratiquées

permet notamment de détecter les points noirs de circulation, d'améliorer la prédiction des temps de parcours, ou encore d'évaluer les effets d'une modification de l'infrastructure (ex : ajout de ralentisseurs ou de ronds-points). Cette information est rendue accessible grâce aux véhicules traceurs qui permettent l'enregistrement en continu des mesures de position et de vitesse des véhicules. La collecte de ces profils de vitesse constitués de mesures de position et de vitesse, et reflétant le comportement individuel des usagers sur la route, produit de grands volumes de données ("Big Data") et leur exploitation nécessite donc l'utilisation de méthodes appropriées.

En parallèle, la généralisation des systèmes embarqués d'aide à la conduite, et plus particulièrement les systèmes d'assistance ou de contrôle de la vitesse (ex : régulateur ou limiteur de vitesse), ainsi que développement des véhicules autonomes (ex : Google Car), conduit à s'interroger sur la consigne de vitesse la plus pertinente à adopter. En effet, ces systèmes nécessitent de fournir au véhicule une vitesse servant de consigne qui soit adaptée à la fois au style de conduite du conducteur (conduite nerveuse, éco-conduite), aux conditions de trafic (trafic libre, congestion), et aux spécificités de l'infrastructure (état des feux). Cependant, la notion de vitesse de référence est généralement associée à de nombreux concepts qui ne reflètent pas toujours l'usage réel du réseau. Ainsi, les limitations de vitesse constituent une borne supérieure à ne pas dépasser, et non une consigne, et ne sont pas toujours adaptées au contexte de conduite (trafic, visibilité...). De même, la vitesse de conception est difficile à calculer car elle nécessite une information très précise de la géométrie de la route (pente, devers...). Si, la V85, définie comme la vitesse en dessous de laquelle circulent 85% des véhicules en conditions nominales, est actuellement la vitesse de référence utilisée dans de nombreux domaines, celle-ci est généralement calculée de manière ponctuelle à partir de mesures de vitesse issues de boucles magnétiques ou de radars. Le développement des véhicules traceurs conduit donc à s'interroger sur une méthodologie adaptée à l'estimation d'un profil V85 ou d'un profil moyen reflétant le comportement des usagers sur une section de route.

Objectifs

L'objectif de cette thèse est de proposer une approche statistique fondée sur le comportement longitudinal adopté par les utilisateurs de la route. Nous proposons d'utiliser les mesures répétées de vitesse réellement pratiquées issues de véhicules traceurs, et de développer une méthodologie statistique permettant d'extraire divers profils de vitesse de référence, chacun étant adapté à une situation de conduite (ex : feu rouge ou vert). En effet, la récolte des traces numériques constituées des mesures de vitesse et de position des véhicules, conduit à l'obtention d'une multitude de profils spatiaux de vitesses reflétant le comportement individuel des usagers de la route, et permet ainsi d'évaluer l'usage réel du réseau routier. En pratique, ces profils spatiaux de vitesse correspondent à une succession de mesures horodatées de vitesse et de position, et sont représentés par des vecteurs de $\mathbb{R}^n$ où n est un entier généra-

lement très grand. En effet, le développement des technologies des capteurs permet de collecter des données avec des fréquences d'échantillonnage pouvant être très élevées, ce qui conduit à des vecteurs de grandes dimensions. Les méthodes usuelles de statistique multivariée ne sont alors plus adéquates (fléau de la dimension, corrélation entre les observations), et il devient nécessaire d'utiliser des méthodes plus appropriées au traitement des profils de vitesse.

L'originalité de notre approche consiste à représenter les profils spatiaux de vitesse par des fonctions et non par des vecteurs de $\mathbb{R}^n$. Cette approche est fondée sur l'Analyse des Données Fonctionnelles (en anglais, "Functional Data Analysis") qui s'est considérablement développée depuis une vingtaine d'années, notamment sous l'impulsion des travaux de J.O. Ramsay (Ramsay et Silverman [129]). Cette approche fonctionnelle est particulièrement appropriée au traitement des profils spatiaux de vitesse afin de préserver la cohérence physique entre vitesse et position (et implicitement temps). En effet, le fait de se placer dans un cadre fonctionnel permet de tenir compte de la structure sous-jacente continue de ce type de données, ainsi que de leurs caractéristiques fonctionnelles : calcul des dérivées, régularité, contraintes de forme...

Les mesures issues de capteurs étant bruitées, la conversion des données brutes de nature vectorielle en objet fonctionnel nécessite une étape de lissage afin d'estimer des profils spatiaux de vitesse valides, ce qui revient à la résolution d'un problème de régression non paramétrique. Nous proposons dans cette thèse, une méthodologie permettant de construire un estimateur d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse, et fondée sur l'utilisation des splines de lissage. Cette approche correspondant à la résolution d'un problème variationnel dans un espace de Hilbert, permet de contrôler le compromis entre la fidélité aux données et la régularité de la solution par un unique paramètre de lissage, contrairement aux méthodes à noyau ou par polynômes locaux généralement utilisées pour l'estimation de profils de vitesse, qui nécessitent la sélection de nombreux paramètres. De plus, cette approche permet d'obtenir une expression explicite de l'estimateur, celui-ci appartenant à un espace de dimension finie même si l'espace d'étude est de dimension infinie.

Enfin, nous proposons une méthodologie permettant de représenter l'ensemble des profils spatiaux de vitesse individuels par un profil de référence représentatif tel que le profil moyen ou le profil médian. En effet, lorsque l'on dispose de gros volume de données, il devient nécessaire de résumer l'ensemble des profils de vitesse par des profils agrégés afin d'en faciliter l'étude. Cependant, si l'utilisation d'un profil de vitesse de référence permet d'étudier la variabilité des vitesses sur une section de route, elle ne permet pas de conserver la variabilité entre les individus. Nous proposons donc de pallier ce problème par la construction d'enveloppes de vitesse reflétant la dispersion des vitesses sur une section de route donnée. L'outil proposé correspond à une extension de la notion de boîtes à moustaches (ou boxplots), classiquement

utilisée en statistique univariée, aux données fonctionnelles.

Organisation du manuscrit

Le premier chapitre est consacré à la notion de vitesse et à l'importance de la connaissance des vitesses pratiquées comme caractéristique de l'usage réel du réseau routier. Nous abordons le problème de la gestion de la vitesse et de l'impact des vitesses excessives ou inappropriées sur la sécurité et l'environnement. Nous évoquons également les principales catégories de systèmes d'adaptation intelligente de la vitesse permettant d'aider les conducteurs à adopter une vitesse adéquate. Puis, après une présentation des principaux indicateurs de vitesse utilisés comme référence dans le domaine de la sécurité routière et de la conception des routes, nous montrons les enjeux et les problématiques d'une connaissance des vitesses pratiquées. Nous montrons notamment que la généralisation des véhicules traceurs permet l'obtention de profils de vitesse permettant une analyse fine du comportement du conducteur. Enfin, après avoir énoncé de nombreux exemples d'application fondés sur l'étude des profils de vitesse, nous passons en revue les différentes méthodes d'estimation de ces profils et notons leurs limites et les problèmes à résoudre.

Dans le deuxième chapitre, nous proposons une modélisation fonctionnelle des profils spatiaux de vitesse. Après une présentation de l'analyse des données fonctionnelles et des différentes méthodes relatives à ce type de données, nous donnons une définition de l'espace fonctionnel des profils spatiaux de vitesse et étudions certaines propriétés de ces fonctions. Nous montrons notamment que les profils spatiaux de vitesse ne sont pas dérivables aux points où la vitesse s'annule, ce qui pose le problème de l'estimation des profils lorsque le véhicule s'arrête, et complique également le calcul du profil moyen.

Le troisième chapitre est consacré aux différentes méthodes de lissage fondées sur les splines. Après un rappel sur la régression non paramétrique et les principales méthodes de lissage actuellement utilisées pour l'estimation des profils spatiaux de vitesse, à savoir la méthode du noyau ou par polynômes locaux, nous introduisons la notion de splines et passons en revue les différentes méthodes de lissage fondées sur cette approche. Nous évoquons notamment en détails les splines de régression, les splines de lissage et les splines pénalisées, et montrons sur un jeu de données réelles l'efficacité des splines de lissage pour l'estimation des profils temporels de vitesse et d'accélération. Enfin, nous présentons le principe de régularisation dans le cas particulier des espaces de Hilbert à noyau reproduisant (RKHS), et montrons que cette généralisation des splines de lissage permet de traiter de manière unifiée de nombreux modèles de régression.

Dans le quatrième chapitre, nous proposons une méthodologie permettant de construire un estimateur d'un profil spatial de vitesse à partir de mesures bruitées

de position et de vitesse, et fondée sur l'utilisation des splines de lissage. Dans un premier temps, nous proposons d'améliorer la qualité des mesures de position en construisant divers estimateurs ponctuels par fenêtre glissante de la position du véhicule en chaque instant d'échantillonnage, et fusionnant les informations issues des deux principaux capteurs de localisation, à savoir l'odomètre et le GPS. Après cette étape optionnelle de pré-traitement des données de position, nous nous intéressons à l'étape de lissage de profils spatiaux de vitesse à partir d'observations bruitées de la position et de la vitesse du véhicule. Nous montrons que suite à différentes contraintes difficiles à prendre en compte dans l'estimation directe d'un profil spatial de vitesse, il est préférable de changer d'espace d'étude, et de commencer par construire un estimateur de la fonction de régression représentant la distance parcourue en fonction du temps. Nous nous ramenons ainsi à un problème de régression non paramétrique sous les contraintes suivantes : estimer une fonction de régression à partir d'observations bruitées sur celle-ci et sur sa dérivée, ainsi qu'une contrainte de monotonie. Dans un premier temps, nous proposons un estimateur par spline de lissage de la fonction représentant la distance parcourue au cours du temps, et utilisant à la fois les mesures de position et de vitesse. Puis dans un second temps, nous proposons de "monotoniser" cet estimateur afin d'obtenir un estimateur strictement croissant. Nous montrons les performances et les limites de notre estimateur sur données simulées et réelles.

Le cinquième chapitre est consacré à la construction de divers profils de vitesse de référence représentatifs d'un ensemble de profils spatiaux de vitesse individuels. Après avoir appliqué la méthode de lissage développée au chapitre précédent sur un ensemble de profils de vitesse issu d'un jeu de données réelles relativement conséquent (78 courbes), nous montrons l'importance de recaler les profils de vitesse, notamment au niveau des arrêts, afin d'obtenir un profil moyen représentatif. La section étudiée comprenant un stop et un feu de signalisation, nous proposons d'utiliser un alignement par landmarks et de distinguer les cas feu rouge/feu vert afin de construire un profil moyen pour chacune de ces situations de conduite. Nous proposons également la construction d'enveloppes de vitesse reflétant la dispersion des vitesses, fondée sur une extension de la notion de boîtes à moustaches (ou boxplots), classiquement utilisée en statistique univariée, aux données fonctionnelles. Enfin, nous proposons un exemple d'application à l'étude de l'éco-conduite.

Finalement, trois annexes concluent ce manuscrit. La première annexe décrit les moyens embarqués pour acquérir la position et la vitesse du véhicule et définit les différents termes techniques en usage dans le contexte automobile. La seconde est consacrée à l'estimation par fenêtre glissante de la position du véhicule à partir de l'odomètre et du GPS, et contient l'original d'un article sur ce sujet, actuellement en attente d'acceptation. Enfin, la troisième est consacrée à l'éco-conduite, thématique sur laquelle j'ai travaillé avant et en parallèle de cette thèse, et présente les travaux et les publications scientifiques réalisées sur ce sujet.

Chapitre 1

La connaissance des vitesses pratiquées : vers une approche statistique basée sur l'étude des profils de vitesses

Ce premier chapitre est consacré à la notion de vitesse et aux différents indicateurs associés. En effet, si la vitesse est un paramètre essentiel du domaine routier, ses définitions sont nombreuses et varient énormément selon les différents intervenants du système routier. De plus, les moyens de mesure de la vitesse sont nombreux et permettent, selon l'usage, d'obtenir des mesures ponctuelles ou continues des vitesses pratiquées.

Dans une première partie, nous abordons le problème de la gestion de la vitesse. En effet, si la vitesse présente de nombreux avantages en terme de mobilité et de réduction des temps de parcours, elle a également des conséquences néfastes sur la sécurité et sur l'environnement. Nous évoquons les liens entre les vitesses excessives et le risque d'accidents ainsi que leur gravité. Nous abordons également l'impact de la vitesse sur la consommation de carburant et sur les émissions de gaz à effet de serre. Puis, nous présentons les différents systèmes d'aide à la conduite qui permettent d'inciter ou d'obliger les conducteurs à respecter les limitations de vitesse et à rouler aux vitesses appropriées.

Dans une seconde partie, nous explicitons les principaux indicateurs de vitesse utilisés comme référence dans le domaine de la sécurité routière et de la conception des routes. Nous définissons la V85 qui est actuellement la vitesse dite de référence, puis nous présentons les différentes règles de fixation des limitations de vitesse ainsi que les différents modèles utilisés pour définir les vitesses de conception.

Enfin, dans une dernière partie, nous abordons l'importance de la connaissance des vitesses pratiquées, et présentons les différents moyens de mesure des vitesses

pratiquées ainsi que les différents indicateurs permettant de les analyser. En effet, le choix de la vitesse est l'une des composantes les plus importantes du comportement des usagers de la route, et le développement des moyens de mesures a permis de collecter des mesures continues de vitesse tout au long d'un trajet. Ces mesures continues conduisent à l'étude des profils de vitesse, permettant ainsi d'analyser finement le comportement des conducteurs. Nous passons en revue différents types d'applications fondées sur l'analyse des profils de vitesse, puis nous présentons les différentes méthodes d'estimation de ces profils de vitesse.

1.1 La gestion de la vitesse : problématique et enjeux

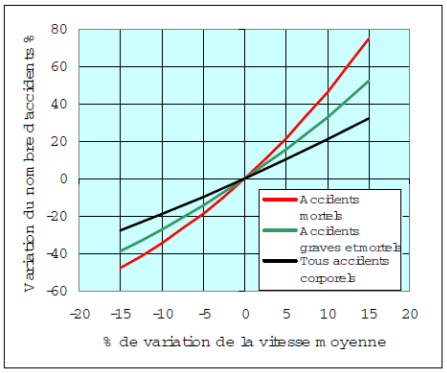
1.1.1 Le rôle de la vitesse

La vitesse est un paramètre déterminant du risque routier et de nombreuses études continuent à porter leur attention sur le thème de la vitesse. En effet, le choix de la vitesse est une caractéristique importante du comportement du conducteur et est influencé par de nombreux facteurs : facteurs relatifs au conducteur (âge, sexe, alcoolémie, fatigue...), au véhicule (catégorie, puissance...), à la route (géométrie, adhérence...), et à l'environnement en général (conditions de trafic, conditions météorologiques, limitations de vitesse...). L'amélioration des réseaux routiers et des performances des véhicules a permis une augmentation des vitesses de déplacement. Cependant, si cette augmentation des vitesses a contribué à l'amélioration de la mobilité et à la réduction des temps de parcours, elle a également eu des conséquences néfastes sur la sécurité routière (accidentologie) et sur l'environnement (émissions de gaz d'échappement, nuisances sonores). La plupart des pays ont admis la nécessité de faire face à ce dilemme de la vitesse et à agir contre les vitesses dites excessives, i.e. les vitesses supérieures aux vitesses réglementaires mais également les vitesses trop élevées par rapport aux conditions locales (route, trafic, infrastructure, environnement). Ainsi, de nombreuses stratégies de gestion de la vitesse ont été mises en place afin de faire circuler les véhicules à des vitesses appropriées sur l'ensemble du réseau routier. A ce propos, le rapport "Gestion de la vitesse" de l'OCDE [117] contient une bonne revue des différentes mesures de gestion de la vitesse (aménagement de l'infrastructure, signalisation, contrôle-sanction...).

1.1.2 Vitesse et sécurité

La vitesse joue un rôle important dans les accidents de la route. En effet, si la vitesse n'est pas toujours la cause des accidents, elle est cependant l'un des principaux facteurs de risque et de gravité des accidents. Ainsi, selon les chiffres de la sécurité routière, le bilan provisoire pour l'année 2012 en France fait état de 60 556 accidents corporels et de 3 645 personnes tués sur les routes, et il semblerait qu'environ 26 % des accidents mortels aient pour cause identifiée la vitesse (source : http://www.securite-routiere.gouv.fr). Les principaux facteurs de risque liés à une vitesse élevée sont une diminution du temps de réaction, une augmentation de la distance

a) Modèle de Nilsson.



b) Courbe en U de Solomon.

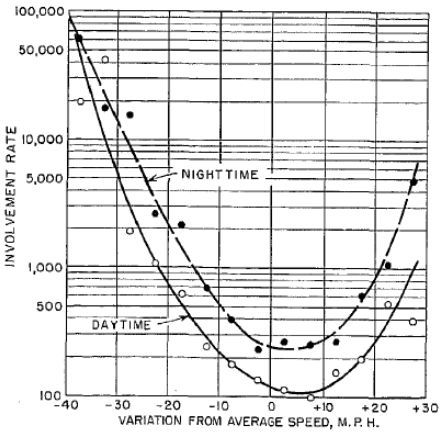


Figure 1.1 - Relation entre vitesse et accidents : a) relation entre vitesse moyenne et accidents (source : Nilsson [115] ; extrait de : OCDE [117]) ; b) relation entre écart de vitesse à la moyenne et nombre d'accidents (source : Solomon [158]).

d'arrêt et une réduction du champ de vision. De nombreux chercheurs ont étudié le lien entre vitesse et accidents, dont l'article de Aarts et Van Schagen [1] propose une revue assez complète sur le sujet. L'un des modèles les plus répandus est le modèle puissance proposé par Nilsson (Nilsson [114] et Nilsson [115]) qui modélise les effets d'un changement de la vitesse moyenne sur le nombre d'accidents comme l'illustre la figure 1.1.a.

D'après ce modèle, une augmentation de 5 % de la vitesse moyenne entraîne une hausse d'environ 10 % du nombre total d'accidents corporels et de 20 % du nombre d'accidents mortels, même si d'autres facteurs tels que l'environnement, l'infrastructure, ainsi que le facteur humain doivent également être pris en compte. D'autres études ont montré un effet de l'hétérogénéité des vitesses sur le taux d'accidents. Solomon [158] a notamment montré que la relation entre l'écart à la vitesse moyenne et le risque d'accident était caractérisée par une courbe en U (figure 1.1.b), même si des études plus récentes (Kloeden *et al.* [89]) ont un regard plus critique sur les effets d'une vitesse inférieure à la vitesse moyenne sur le risque d'accident.

1.1.3 Vitesse et environnement

Si, comme nous venons de le voir à la section précédente, les vitesses excessives ont des conséquences néfastes sur la sécurité, elles ont aussi un impact sur l'environnement. En effet, les émissions des véhicules contiennent divers polluants tels que le monoxyde de carbone (CO), les hydrocarbures (HC) et les oxydes d'azote (NO_x), dont la quantité varie selon la vitesse. De manière générale, les émissions sont optimisées pour une vitesse constante comprise entre 50 et 70 km/h comme l'illustre la figure 1.2.

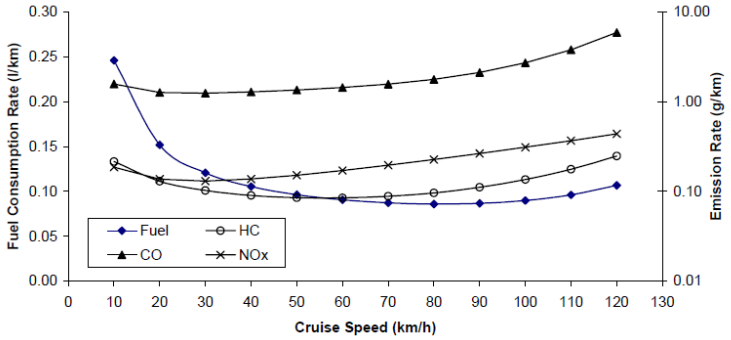


Figure 1.2 - Variations de la consommation et des émissions de polluants selon la vitesse (source : Rakha et Ding [124]).

La vitesse et plus généralement le mode de conduite ont également un impact sur la consommation de carburant. Ainsi, une conduite agressive avec de fortes accélérations et/ou décélérations peut entraîner une hausse allant jusqu'à 30 % de carburant. Afin de réduire la consommation et les émissions de polluant, un nouveau style de conduite appelé eco-conduite est apparu depuis quelques années et consiste en une sollicitation modérée du véhicule basée sur un ensemble de règles telles que le maintien d'une vitesse constante, l'anticipation du trafic ou l'utilisation du frein moteur (`http://www.ecodrive.org`). L'éco-conduite sera abordée plus en détails dans l'annexe C de la thèse.

La modélisation de la consommation et des émissions des véhicules est également un sujet étudié par de nombreux chercheurs, et le lien avec la vitesse est évidemment la base de ces modèles. Ainsi, la plupart des modèles macroscopiques sont basés essentiellement sur la vitesse moyenne du trafic, et les modèles microscopiques sont basés sur la vitesse et l'accélération instantanées du véhicule (Rakha et Ding [124], Luu [101] section 2.2).

1.1.4 Systèmes d'adaptation intelligente de la vitesse (ISA)

Nous venons de voir aux sections précédentes que les vitesses excessives avaient des conséquences néfastes sur l'environnement et la sécurité. A l'inverse, si le cas extrême d'une vitesse de 0 km/h permet de réduire à néant les inconvénients de la vitesse, ce n'est évidemment pas la solution adéquate. La difficulté consiste donc à trouver une vitesse appropriée résultant d'un compromis entre sécurité et mobilité. Si les différentes politiques de contrôle-sanction (ex : radars fixes et mobiles) ont permis de réduire les excès de vitesse et d'améliorer la sécurité routière, de nombreuses recherches sont faites sur les systèmes dits "d'adaptation intelligente de la vitesse" (ISA, "*Intelligent Speed Adaptation system*") afin d'aider les conducteurs à respecter les limitations de vitesse et à rouler aux vitesses appropriées. Les systèmes ISA peuvent être divisées en deux catégories : les systèmes d'assistance à la vitesse, dits

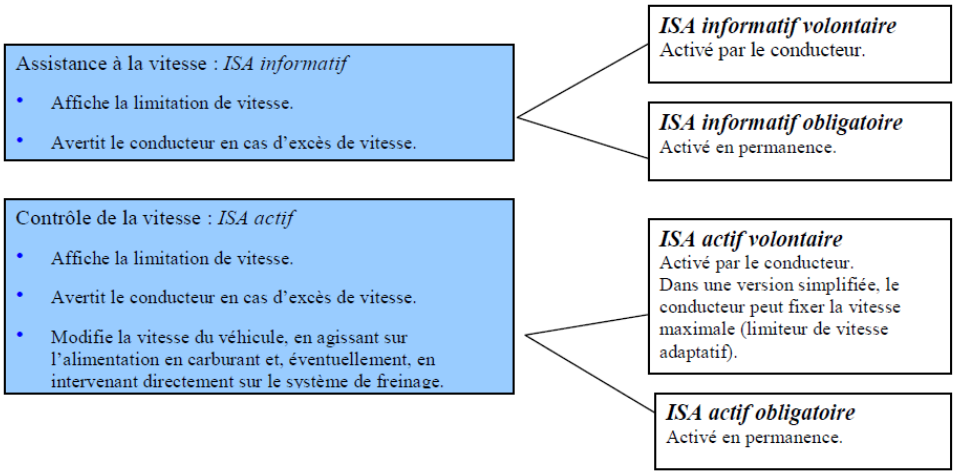


Figure 1.3 - Types de systèmes ISA (source : OCDE [117]).

informatifs, qui consistent à informer le conducteur par une alerte visuelle et/ou sonore de la limitation de vitesse ou de la vitesse appropriée, et les systèmes de contrôle de la vitesse, dits actifs, qui agissent directement sur le système d'injection, le système de freinage ou la pédale d'accélérateur afin de réduire la vitesse du véhicule (voir figure 1.3).

Si certains systèmes consistent à informer ou imposer au conducteur la vitesse limite autorisée (ex : projet SpeedAlert, ou projet LAVIA), d'autres systèmes prennent en compte la géométrie de la route ou les conditions météorologiques (Gallen *et al.* [66]). Un bon aperçu des différents systèmes européens est donné dans Ehrlich [49] et OCDE [117] (chap. 10). Cependant, le déploiement de ces systèmes reste encore limité car leur implantation pose des problèmes de mise en oeuvre (acquisition et maintien de la base de données des limitations de vitesse) et des problèmes juridiques (responsabilité en cas d'accidents). De plus, si les avantages de ces systèmes du point de vue de la sécurité semblent acquis, les effets sur la consommation de carburant ne sont pas toujours positifs (Saint Pierre et Ehrlich [141]).

Enfin, notons également le développement de systèmes d'aide à l'éco-conduite (EDAS, "*Ecological Driving Assistance System*") dont l'objectif est d'informer ou d'assister le conducteur afin de limiter sa consommation de carburant. Dans la catégorie des systèmes informatifs, on peut citer les travaux de Luu [101] dont l'objectif est de suggérer au conducteur une vitesse approprié à la route et à son environnement tout en tenant compte de la consommation de carburant. Dans la catégorie des systèmes actifs, on peut citer les travaux d'Hellstrom [78] dont l'objectif est la conception d'un régulateur de vitesse pour poids lourds tenant compte de la consommation de carburant. Enfin, citons également le projet européen "ecoDriver" (`http://www.ecodriver-project.eu`), qui a débuté en octobre 2011, et dont l'un

des objectifs est de tester l'acceptabilité de différents systèmes d'aide à l'éco-conduite (systèmes embarqués et nomades).

1.2 Les principaux indicateurs de vitesse

Afin de limiter les vitesses excessives, il est nécessaire de disposer d'une vitesse de référence caractérisant la vitesse appropriée à adopter. Par exemple, les systèmes ISA nécessitent de fournir au conducteur une vitesse de référence afin de l'aider à rouler aux vitesses appropriées. Une question se pose alors : quel indicateur de vitesse choisir comme vitesse de référence ? En effet, la notion de vitesse de référence est généralement associée à de nombreux concepts tels que la vitesse moyenne, la V85, la vitesse réglementaire, la vitesse de conception... La difficulté est que la notion de vitesse de référence n'est pas la même pour tous les acteurs de la route. Par exemple, pour le gestionnaire routier chargé de l'entretien des routes, la vitesse appropriée sera liée à la notion d'adhérence du véhicule sur la route, alors que pour le concessionnaire autoroutier la notion de fluidité du trafic devra également être prise en compte. De plus, un même indicateur de vitesse peut être défini de plusieurs manières. Ainsi, en 1998, le programme de recherche américain "National Cooperative Highway Research Program" (NCHRP) finança un projet de recherche afin d'étudier les différents concepts liés à la vitesse de conception ("*design speed*"), la vitesse pratiquée ("*operating speed*") et la vitesse réglementaire ("*posted speed*"). Le rapport résultant dénommé "NCHRP Report 504" (Fitzpatrick *et al.* [59]) passe en revue les différentes définitions associées à ces concepts. Le tableau 1.4 extrait de Ogle [118] d'après Fitzpatrick *et al.* [59], contient les définitions publiées dans le manuel MUTCD ("Manual on Uniform Traffic Control Devices") (2000) qui définit les normes utilisées par les gestionnaires des routes nationales américaines, et dans le célèbre "Livre vert" ("Policy on Geometric Design of Highways and Streets", AASHTO Green Book) (2001) qui porte sur la conception des routes et autoroutes. Nous explicitons dans les sections suivantes les principaux indicateurs de vitesse utilisés comme référence dans le domaine de la sécurité routière et de la conception des routes.

1.2.1 La V85, la vitesse de référence ?

La V85 est actuellement la vitesse de référence utilisée dans de nombreux domaines allant de la conception des routes à la définition des vitesses réglementaires. Elle est définie de manière générale comme le $85^{ème}$ centile des vitesses pratiquées. Cependant, la V85 peut prendre des valeurs très différentes selon le type de véhicule considéré (tous types ou uniquement les véhicules légers (VL)) et selon les conditions de trafic (trafic libre ou contraint). Ainsi, le SETRA [152] définit la V85 comme la vitesse en dessous de laquelle circulent 85 % des véhicules légers libres (i.e. dont le temps inter-véhiculaire est supérieur à 4 s). Généralement, la V85 est une mesure ponctuelle calculée à partir de mesures de vitesses issues de boucles magnétiques

Measure	Reference	Definition
Operating Speed	MUTCD 2000	**Operating Speed** – a speed at which a typical vehicle or the overall traffic operates. Operating speed may be defined with speed values such as the average, pace, or 85th percentile speeds.
	AASHTO Green Book 2001	**Operating Speed** is the speed at which drivers are observed operating their vehicles during free-flow conditions. The 85th percentile of the distribution of observed speeds is the most frequently used measure of the operating speed associated with a particular location or geometric feature.
85th Percentile Speed	MUTCD 2000	**85th Percentile Speed** – The speed at or below which 85 percent of the motorized vehicles travel.
Average Speed	MUTCD 2000	**Average Speed** – The summation of the instantaneous or spot-measured speeds at a specific location of vehicles divided by the number of vehicles observed.
Pace Speed	MUTCD 2000	**Pace Speed** – The highest speed within a specific range of speeds that represents more vehicles than in any other like range of speed. The range of speeds typically used is 10 km/hr or 10 mph.
Design Speed	MUTCD 2000	**The Design Speed** is a selected speed used to determine the various geometric design features of the roadway.
	AASHTO Green Book 2001	**The Design Speed** is a selected speed used to determine the various geometric design features of the roadway. Provisions: • The assumed design speed should be logical for the topography, adjacent land use, and highway functional classification (paraphrased). • "All of the pertinent features of the highway should be related to the design speed to obtain a balanced design." • "Above-minimum design values should be used where feasible…" • "The design speed chosen should be consistent with the speed a driver is likely to expect." • "The speed selected for design should fit the travel desires and habits of nearly all driers…The design speed chosen should be a high-percentile value…i.e., nearly all inclusive…whenever feasible."

Figure 1.4 - Tableau des définitions des différentes notions de vitesse (source : Fitzpatrick *et al.* [59] ; extrait de : Ogle [118]).

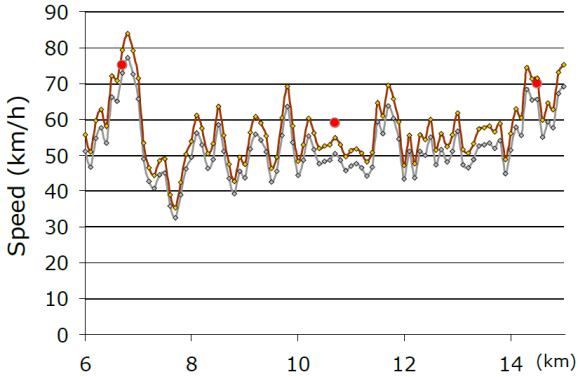


Figure 1.5 - Exemple de profil de vitesse avant (en gris) et après recalage (en rouge) à partir de mesures ponctuelles de V85 (disques rouges) (source : Edgar [48]).

ou de radars (voir section 1.3.2 pour plus de détails sur ces capteurs). Reste alors le problème de l'estimation d'un profil de vitesse V85 afin d'étendre la notion de V85 au niveau d'un itinéraire. Si le développement des véhicules traceurs et des smartphones facilite l'accès à des données de vitesse et de position tout au long d'un itinéraire (voir section 1.3.2), il existe peu d'études sur l'estimation d'un profil V85. Une solution simple proposée par Louah [98], puis reprise dans Edgar [48] et Gallen [65], consiste à fusionner des mesures issues de dispositifs de bord de voie avec des mesures obtenues en continu avec un véhicule instrumenté. Le principe général de cette méthode consiste à recaler chaque profil de vitesse individuel à partir d'une mesure ponctuelle de V85 comme l'illustre la figure 1.5.

Cependant cette méthode repose sur deux hypothèses fondamentales qui limitent son utilisation : d'une part, le comportement du conducteur doit être uniforme tout au long du trajet, et d'autre part, les zones de gêne (dépassement, feu stop...) ne doivent pas être prises en compte. Nous proposerons au chapitre 5 d'étendre l'estimation d'un profil V85 à la construction d'enveloppes de vitesse reflétant la dispersion des vitesses pratiquées sur une section de route donnée.

Enfin, notons que si la V85 a été définie ici à partir des vitesses pratiquées, elle peut également être estimée à partir d'un modèle basé sur les caractéristiques géométriques de la route (voir section 1.2.3).

1.2.2 Vitesses réglementaires

Les limitations de vitesse sont utilisées afin d'informer les usagers des vitesses appropriées d'un point de vue sécuritaire. Généralement, la vitesse réglementaire correspond à la vitesse maximale de sécurité pour les véhicules légers dans des conditions "normales" de circulation. Les limitations générales de vitesse varient selon le type de route et doivent tenir compte de la sécurité, de la mobilité, et de plus en plus de l'environnement (voir figure 1.6).

	Vitesses appropriées répondant aux objectifs spécifiés			
Fonction et catégorie de routes	**Sécurité**	**Environnement**	**Économie et mobilité**	**Qualité de vie des riverains**
Autoroutes et grandes routes interurbaines	**90-130 km/h**[1]	**70-90 km/h**	**Haut de la fourchette des vitesses**	**Bas de la fourchette des vitesses**
Réseau de grande qualité conçu pour le transport de personnes, de marchandises et de services à longue distance et à vitesse élevée.	Une réduction de la vitesse peut être appropriée en conditions météorologiques difficiles.	L'augmentation de la vitesse entraîne l'accroissement des émissions et du bruit. La réduction de la vitesse est nécessaire lorsque les aspects liés à la pollution atmosphérique et aux nuisances sonores sont importants.	Cette question est très importante pour les déplacements professionnels et privés.	S'il existe des aménagements adjacents, ce qui est relativement rare, les vitesses doivent en tenir compte, afin de réduire les nuisances sonores, la pollution atmosphérique et l'effet de coupure.
Routes et grands axes urbains	**50-60-70 km/h**	**30-60 km/h**	**Haut de la fourchette des vitesses**	**Bas de la fourchette des vitesses**
Réseau urbain de grande qualité, conçu pour supporter le trafic de transit.	Réduction à 30 km/h en présence de nombreux usagers vulnérables.	Vitesses optimales en ce qui concerne les émissions.	Trafic local et trafic de transit. Dans de nombreux cas, zones commerciales et résidentielles. Nécessité d'équilibrer les besoins de sécurité et de mobilité.	Importante lorsque la zone adjacente est à usage résidentiel. Nécessité de gérer les vitesses en fonction de la pollution atmosphérique, des nuisances sonores et de l'effet de coupure.
Rues résidentielles	**30 km/h**	?[2]		
Réseau conçu pour les riverains et uniquement accessible au trafic local.	Modération du trafic lorsque nécessaire, pour réduire la vitesse.	Au-dessous des vitesses optimales en ce qui concerne les émissions. Certains éléments verticaux de modération du trafic peuvent augmenter les nuisances sonores.	En deuxième position, après la sécurité et la qualité de vie.	Très importante sur toute la voirie résidentielle.
Routes principales de rase campagne	**70-90 km/h**	**60-90 km/h**		
Hors autoroutes interurbaines Conçues pour le trafic de transit local.	Selon la qualité[3] Réduction de la vitesse dans les virages et les intersections.		Importantes.	
Petites routes de rase campagne	**40-60 km/h**			
Conçues pour le trafic local avec des accès, en présence d'usagers vulnérables.	Selon la nature et le nombre d'usagers vulnérables.	Vitesses optimales.	En deuxième position après la qualité de vie.	

1. Certaines autorités fixent des limitations de vitesse plus élevées sur une partie de ce réseau.
2. Il n'existe pas de résultats de recherche suffisants pour définir une limitation de vitesse.
3. Plus la qualité de la route est mauvaise, en termes de virages et d'intersections, plus la vitesse doit être faible.

Figure 1.6 - Vitesses appropriées pour différents types de routes (source : OCDE [117]).

De plus, selon les pays, elles peuvent également varier selon le type de véhicule (ex : vitesses spécifiques pour les poids lourds), le type de conducteurs (ex : vitesses spécifiques pour les jeunes conducteurs) ou les conditions météorologiques (ex : vitesses spécifiques en cas de pluie ou de brouillard). Enfin, certaines limitations de vitesse peuvent également être fixés temporairement comme par exemple en France dans le cas de pics de pollution ou en cas de fort trafic. Il existe également des limitations de vitesse locales fixées dans les zones à risque (nombreux accidents). En effet, les limitations générales de vitesse ne reflètent pas toujours certaines caractéristiques de la route, ce qui nécessite une information plus spécifique dans certaines zones. Si ces limitations locales de vitesse ont souvent été déterminés à partir de la vitesse V85 ou à partir de vitesses moyennes comme en Grande-Bretagne, il est recommandé de tenir compte du risque et de fixer une vitesse inférieure à la vitesse enregistrant un risque d'accident moyen (OCDE [117]). Cependant, pour être réellement efficaces, ces limitations de vitesse doivent rester crédibles pour les usagers. Plusieurs études ont comparées les vitesses pratiquées par rapport aux vitesse réglementaires. Ces études montrent que la vitesse V85 est généralement supérieure à la vitesse réglementaire. Par exemple, Fitzpatrick *et al.* [59] établit le modèle de régression linéaire suivant pour la vitesse V85 (exprimée en mph, 1 mph=1.61 km/h) :

$$V85 = 7.675 + 0.98 \times \text{Posted Speed Limit.} \tag{1.1}$$

Cette équation montre que, en moyenne, la vitesse V85 est supérieure de 7.675 mph (i.e. 12.35 km/h) à la vitesse réglementaire. Les résultats de cette étude montrent également que pour tous types de routes confondus, le meilleur modèle reliant un percentile des vitesses à la vitesse réglementaire est celui associé à la vitesse V50, bien que le choix du meilleur modèle varie énormément lorsque l'on distingue les différents types de routes (Fitzpatrick *et al.* [59] Table 24).

1.2.3 Vitesses de conception

A l'origine, la vitesse de conception permet de fixer les caractéristiques géométriques minimales des routes afin d'assurer aux usagers des conditions satisfaisantes d'adhérence, de stabilité, de visibilité... (SETRA [151]). Cependant, depuis les années 80, les normes de conception géométrique des projets routiers ont évolué dans de nombreux pays, et la prise en compte des vitesses pratiquées se généralise afin d'assurer une cohérence entre les comportement supposées et les comportements réels (SETRA [153]). En effet, pour qu'elle soit lisible par l'usager, une route de doit pas présenter de trop grandes différences entre la vitesse de conception, la vitesse limite et la vitesse réellement pratiquée par les usagers. De plus, la vitesse de conception ne doit jamais être inférieure à la vitesse réglementaire. Ces critères nécessitent donc une bonne connaissance des vitesses pratiquées, et plus particulièrement de la vitesse V85. Cependant, si la V85 peut être déterminée par des mesures sur les itinéraires existants, elle ne peut être qu'estimée pour les projets neufs, à partir de formules établies pour chaque type de route en fonction des principales caractéristiques géo-

Type de route	V_{85} théorique (km/h)	Vitesse réglementaire (km/h)
Autoroute	150	130
2x2 voies en interurbain	120	110
2 voies (6 à 7 m) ou 3 voies	102	90
2 voies (5 m)	92	90

Figure 1.7 - Vitesses V85 estimées et vitesses réglementaires pour différents types de routes (source : SETRA [151]).

métriques (rayon du virage, degré de courbure...). Une revue des différents modèles établis est donnée dans l'étude bibliographique de Louah [99] réalisé par le CETE de l'Ouest pour le compte du SETRA, on citera également le rapport de 2008 du SETRA [152] dont l'objectif était une vérification des formules établies dans le rapport de 1986 (SETRA [153]) pour les virages sur les 2 voies (6 et 7 m) et les 3 voies. La figure 1.7 donnent les principales valeurs de V85 estimées pour différents types de routes, mais il est dorénavant recommandé d'appliquer le principe d'écrêtement à la vitesse réglementaire (excepté pour la visibilité en carrefour plan où c'est la V85 qui est prise en compte pour des impératifs de sécurité).

1.3 La connaissance des vitesse pratiquées : enjeux et problématiques

La connaissance et la maîtrise des vitesses pratiquées est une composante fondamentale des objectifs de sécurité routière, de mobilité et de protection de l'environnement. En effet, du point de vue de la sécurité, connaître les vitesses réelles des usagers permet de détecter les zones critiques du réseau routier et d'y remédier par une modification de l'infrastructure ou par une signalisation appropriée. La connaissance des vitesses pratiquées permet également d'améliorer la connaissance des temps de parcours qui est un indicateur important permettant d'évaluer les conditions d'exploitation du réseau et d'informer les usagers de l'état du trafic (Maza [108], Allain [4]). La vitesse est également un bon indicateur du comportement du conducteur et de ses interactions avec son environnement et avec les autres usagers. Nous abordons plus en détails ce sujet dans les sections suivantes, et nous explicitons également les différents moyens de collecte de mesures des vitesses pratiquées et les différentes méthodes permettant d'analyser ces mesures de vitesses.

1.3.1 La vitesse comme indicateur du comportement du conducteur

Le comportement des conducteurs sur la route est un phénomène complexe qui est influencé par de nombreux facteurs : facteurs relatifs au conducteur (âge, sexe, alcoolémie, fatigue...), au véhicule (catégorie, puissance...), à la route (géométrie,

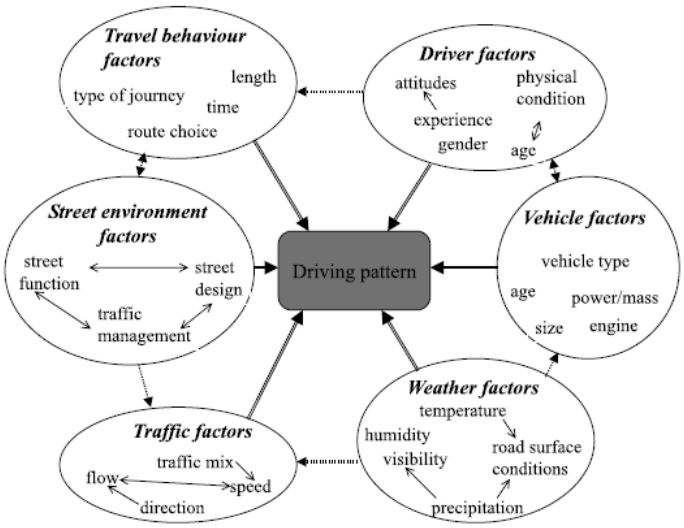


Figure 1.8 - Facteurs influençant le comportement du conducteur (source : Ericsson [52]).

adhérence...), et à l'environnement en général (conditions de trafic, conditions météorologiques, limitations de vitesse...) (figure 1.8). Ces nombreux facteurs impliquent une grande variabilité des comportements de conduite comme l'illustre l'étude réalisée en Suède par Ericsson [52].

Plusieurs paramètres relatifs à l'accélération ou la décélération du véhicule ainsi qu'au régime moteur peuvent être analysés pour caractériser le comportement du conducteur, mais les indicateurs de vitesse comme la vitesse moyenne sont les plus couramment utilisés (Ericsson [52], Ericsson [51]). En effet, le choix de la vitesse est l'une des composantes les plus importantes du comportement des usagers de la route. L'article de Laureshyn [93] contient une bonne revue des différentes études utilisant un indicateur basé sur la vitesse, ces études portant principalement sur la sécurité mais également sur l'environnement. Nous verrons notamment dans les sections suivantes que les profils de vitesse sont des indicateurs très riches en information et permettent une analyse fine du comportement du conducteur.

1.3.2 Développement des moyens de mesures : d'une mesure ponctuelle à une mesure continue

Les moyens utilisés pour observer et mesurer l'utilisation du réseau routier, et en particulier les vitesses pratiquées, ont connu récemment des évolutions majeures : nous sommes passés de la collecte de données agrégées obtenues grâce à des équipements statiques, à des mesures beaucoup plus fines enregistrées à l'aide de boîtiers connectés aux ordinateurs de bord des véhicules (données CAN). En effet, jusqu'à présent, la plupart des capteurs utilisés permettent d'obtenir une mesure unique-

ment ponctuelle des vitesses pratiquées. Parmi les capteurs les plus fréquemment utilisés, on peut citer les boucles magnétiques qui sont les capteurs les plus utilisés notamment sur le réseau autoroutier (stations SIREDO), les tubes pneumatiques utilisés pour des mesures temporaires sur des voies à faible trafic, et les radars à effet Doppler-Fizeau qui permettent d'obtenir la vitesse instantanée des véhicules en un point précis. Les capteurs vidéos permettent d'étendre les mesures à l'étude d'une courte section (ex : carrefour), cependant leurs performances sont très dépendantes des conditions météorologiques et de la luminosité, et l'extraction des mesures de vitesse nécessite un lourd travail de traitement d'images (Laureshyn et Ardö [94]). Pour plus de détails sur ces capteurs, le lecteur est invité à consulter le livre de Klein [88] ou le site du ministère français des transports consacré aux transports intelligents (http://www.transport-intelligent.net).

L'étude des vitesses pratiquées sur de longues sections devient possible avec la généralisation des véhicules traceurs. Ces véhicules instrumentés peuvent être considérés comme des capteurs mobiles explorant le réseau en continu. L'instrumentation de ces véhicules peut être légère (GPS, smartphone) mais peut également être plus lourde (boîtier enregistreur connecté au bus CAN, radar, caméras...). Une description des différents types d'instrumentation et des principaux capteurs est donnée dans l'annexe A. Par exemple, l'analyse de l'état du trafic nécessite de recueillir la position, la vitesse et le sens de déplacement des véhicules et de les transférer en temps réel à un opérateur de données de trafic. L'étude de ces données issues de véhicules traceurs, et appelées "Floating Car Data" (FCD), correspond généralement à une analyse en temps réel de traces de véhicules composées de la position GPS de chacun des véhicules au cours du temps (Allain [4] chap. 4). Une nouvelle source de données issues des réseaux des opérateurs de téléphonie mobile, et appelées "Floating Mobile Data" (FMD), est également utilisée pour l'analyse du trafic. Le principe consiste à exploiter de manière anonyme les données issues des téléphones portables, chaque mobile en communication devenant source d'information, et à localiser par triangulation le mobile dans une cellule du réseau GSM. Le fournisseur d'informations trafic Médiamobile s'est notamment associé à l'opérateur de téléphonie mobile Orange, afin d'enrichir les services V-Trafic édités par Médiamobile grâce à la technologie FMD développée par Orange. Une expérimentation nommée TraficZen a également eu lieu à Toulouse en 2010 associant la technologie FMD d'Orange avec les données de boucles magnétiques d'ASF (Autoroutes du Sud de la France). On peut ainsi noter que si jusqu'à présent, la plupart des véhicules traceurs étaient réduit à des flottes privées (bus, taxis...) et n'étaient donc pas représentatif de l'ensemble du réseau routier, le développement des smartphones permet de faciliter l'accès aux vitesses pratiquées par les usagers sur l'ensemble du réseau routier.

a) Agrégation sur le temps/espace. b) Agrégation sur les individus.

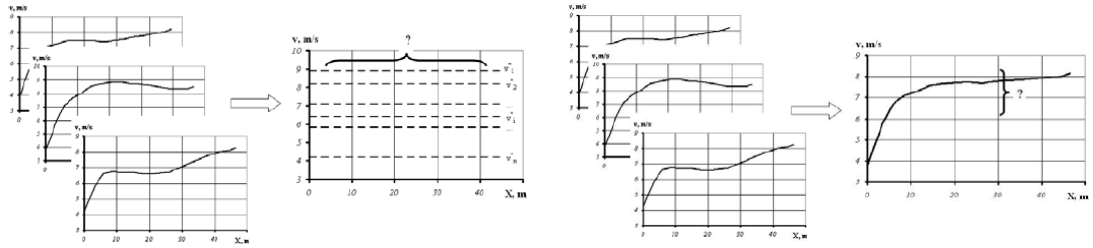


Figure 1.9 - Deux types d'agrégation des profils de vitesse (source : Laureshyn [93]).

1.3.3 Vers une analyse fine du comportement du conducteur : l'étude des profils de vitesse

Le développement des moyens de mesures des vitesses pratiquées a permis une évolution des indicateurs statistiques utilisés dans les différentes études. En effet, si les capteurs classiquement utilisés comme les boucles magnétiques et les radars permettent d'obtenir uniquement une mesure ponctuelle de la vitesse, l'utilisation des capteurs vidéos et plus particulièrement des véhicules traceurs permet d'obtenir une mesure continue de la vitesse. Ces mesures continues de vitesse permettent d'obtenir, selon l'usage, des profils temporels de vitesses (vitesse en fonction du temps), ou des profils spatiaux de vitesse (vitesse en fonction de la position). Ces profils de vitesse sont riches d'information car ils permettent d'étudier le comportement individuel des usagers sur la route, contrairement à une vitesse moyenne obtenue en un point qui ne reflète pas les variations de vitesse entre les individus. Dans le cas de gros volume de données, il est cependant nécessaire de résumer l'ensemble des profils de vitesse par des indicateurs agrégés afin d'en faciliter l'étude. On distingue généralement deux types d'agrégation illustrés à la figure 1.9.

La première méthode consiste à représenter chaque profil individuel par un indicateur agrégé (moyenne, médiane...). Cependant, cette méthode ne permet pas de garder le caractère continu des données de vitesse même si la variabilité entre les individus est conservée. Une autre méthode plus utilisée consiste à représenter l'ensemble des profils de vitesse par un profil de vitesse représentatif (profil moyen, profil V85...). Dans ce cas, les variations de vitesses au cours du trajet (ou au cours du temps) sont conservées mais on perd la variabilité entre les individus. Nous verrons au chapitre 6, une façon de combiner ces deux approches en utilisant des "boxplots" fonctionnels (boîtes à moustaches étendues aux données fonctionnelles).

Afin de montrer la richesse de l'étude des profils de vitesse, nous proposons d'énoncer quelques exemples d'études utilisant ce type de données.

Une première application de l'étude des profils de vitesse est l'analyse et la prévision du trafic routier. Sur ce sujet, on peut citer les thèses de Maza [108] puis de

Allain [4] issues d'une convention entre l'université Paul Sabatier de Toulouse et la société Médiamobile, qui ont permis l'élaboration de méthodes de prévision à court terme à partir de l'étude de profils de vitesse temporels. On peut également citer l'article de Quiroga et Bullock [122] décrivant une méthode de prévision de temps de parcours à partir de profils spatiaux de vitesse obtenus à partir de véhicules traceurs équipés de GPS. Enfin, les travaux de Ko *et al.* [91] utilisent les profils de vitesses pour caractériser les temps d'arrêts en distinguant la phase de décélération, la phase d'arrêt du véhicule, et la phase d'accélération jusqu'à la vitesse de croisière.

Une autre application est l'étude de l'impact d'une modification de l'infrastructure. Par exemple, les études de Barbosa *et al.* [14] et de Moreno et García [112] montrent l'impact de l'ajout de ralentisseurs ou de chicanes sur les vitesses pratiquées obtenues à partir de tubes pneumatiques pour la première étude, et à partir de véhicules traceurs équipés de GPS pour la seconde. Hyden et Varhelyi [80] utilisent des capteurs vidéos pour obtenir des profils spatiaux de vitesses et étudier les effets de l'ajout de rond-points sur la sécurité et sur les émissions de gaz. Laureshyn *et al.* [95] utilisent également la vidéo, mais pour étudier le comportement des conducteurs à une intersection lorsque le véhicule tourne à gauche. Dans cette étude, les profils spatiaux de vitesse sont considérés comme des vecteurs et différentes méthodes de classification sont comparés.

Dans le domaine de la sécurité routière, l'observation de la conduite en situation naturelle tend à remplacer les expérimentations sur simulateurs ou sur circuits fermés. Ces études appelées couramment "Naturalistic Driving Studies" (NDS) consistent à recueillir les "traces" numériques d'un ensemble de conducteurs dans leurs déplacements quotidiens effectués avec leur propre véhicule, sans qu'aucune consigne particulière ne leur soit donnée. Cette collecte de données réalisée sur une grande échelle et sur une période de temps longue (1 an par exemple) permet notamment de mettre en évidence des situations de conflits ou de presque-accident afin de détecter les points noirs du réseau et d'anticiper le risque d'accident. Le rapport de master du finlandais Nygard [116] utilise les profils de vitesse et les profils dérivées, i.e. les profils d'accélération et les profils de jerks (dérivée de l'accélération) pour caractériser les situations de conflits correspondant à un freinage brusque (pics dans le profil de jerk). En effet, des études plus récentes ont montré que le jerk était un bon indicateur des comportements à risque (Bagdadi et Várhelyi [13]). Les rapport de thèse de Ogle [118] et Boonsiripant [22] étudient également les situations de presque-accidents à partir de l'étude des vitesses pratiquées. Les données sont issues d'une étude de type NDS, "the Commute Atlanta program", avec plus de 460 véhicules traceurs observés pendant 1 an (en 2004) à Atlanta. L'étude de Ogle [118] est consacrée aux excès de vitesse, alors que l'étude de Boonsiripant [22] s'intéresse aux variations de vitesses comme mesure du risque.

Toujours dans le domaine de la sécurité, l'étude des profils de vitesse permet d'évaluer les effets de l'utilisation d'un système ISA sur le comportement du conducteur. Par exemple, le sujet de l'étude de Comte [30] porte sur l'évaluation d'un limiteur de vitesse (projet européen MASTER) à partir de profils de vitesse obtenus

à partir d'un simulateur de conduite. On peut également citer les travaux de Várhelyi *et al.* [168] qui calculent un profil de vitesse moyen afin d'étudier les effets d'un système ISA actif agissant sur la pédale d'accélérateur ("Active Accelerator Pedal") sur la sécurité et sur le comportement du conducteur.

L'étude des profils de vitesse est également utilisée dans un objectif environnemental. L'étude de Mandava *et al.* [105] porte sur la conception d'un système ISA informatif donnant une vitesse de consigne calculée en fonction de l'état des feux afin de réduire la consommation de carburant. On peut citer également les travaux de Kerper *et al.* [84] du groupe Volkswagen basés sur l'analyse des traces des véhicules (profils de vitesse) afin de prédire l'état des feux et de donner une consigne de vitesse adaptée au conducteur dans le but de réduire sa consommation de carburant.

Enfin, le calcul d'un profil V85 à partir de plusieurs profils de vitesse individuels est un outil important pour les gestionnaires routiers. La connaissance du profil V85 permet notamment de calculer les distances de visibilité que l'infrastructure doit offrir aux usagers, de vérifier la crédibilité des limitations de vitesse et d'analyser la sécurité d'un itinéraire (exemple d'application dans Violette *et al.* [170]). Le profil V85 peut également être utilisé dans le cadre de la réalisation d'un système ISA. Ainsi, dans le cadre du projet DIVAS, Gallen *et al.* [66] proposent d'utiliser le profil V85 obtenu en conditions idéales de circulation (route sèche, bonne visibilité atmosphérique) et calculé avec la méthode de Louah [98] présentée à la section 1.2.1, puis de le modifier en cas de conditions dégradés (pluie, brouillard) en tenant compte du risque (probabilité d'accident et gravité).

1.3.4 Estimation des profils de vitesse : revue et limites des méthodes actuelles

La précision des données de vitesse et de position est très importante dans l'étude des profils de vitesse. En effet, une erreur de positionnement ou de vitesse peut impliquer de graves conséquences dans les résultats d'une étude sur la sécurité. Les données issues de capteurs n'étant pas toujours très précises, il est nécessaire de chercher à minimiser les erreurs de mesures en calculant un estimateur du "vrai" profil de vitesse. Cette étape de pré-traitement des données est généralement appelée étape de lissage. Plusieurs études ont été consacrées à l'estimation de profils de vitesse ainsi qu'à l'estimation de profils d'accélération ou de jerks. Nous proposons dans la suite de cette section, de passer en revue les différentes méthodes de lissage utilisées en pratique et de noter leurs limites.

L'article de Bratt et Ericsson [24] (1999) propose d'utiliser la méthode des polynômes locaux pour estimer les profils temporels de vitesse et d'accélération. Cette méthode est proposée comme une alternative à l'estimation par moyenne mobile couramment utilisée en pratique pour estimer la vitesse mais inappropriée au calcul de la dérivée (estimateur du profil d'accélération instable). Les données de vitesses sont mesurées sur un véhicule équipé d'un enregistreur de données qui récupèrent les mesures de vitesse dérivées de l'odomètre (vitesse CAN (voir annexe A)) avec

une fréquence d'échantillonnage de 10 Hz (i.e. 10 mes/s). Bratt et Ericsson [24] proposent d'utiliser un estimateur par polynômes locaux avec les paramètres suivants : noyau d'Epanechnikov, degré p=2 pour l'estimation du profil de vitesse et p=3 pour le profil d'accélération, et selection de la fenêtre de lissage en utilisant une procédure développée par Ruppert *et al.* [139] basée sur une minimisation de l'erreur quadratique moyenne (cf. section 3.1.3 pour plus de détails sur l'estimation par polynômes locaux). La fonction de variance ainsi que le biais des estimateurs de la vitesse et de l'accélération sont représentés sur deux jeux de données. Cette étude montre que le profil de vitesse estimé n'est pas très lisse lorsque les erreurs de mesures de vitesse sont importantes (figure 2(a) dans Bratt et Ericsson [24]) et que les pics d'accélération ont tendance à être sur-lissés. Les auteurs soulignent également le problème du temps de calcul qui peut être long pour de gros jeux de données avec cette méthode, et proposent en conclusion de tester d'autres méthodes de lissage comme les splines de régression ou les ondelettes.

Rakha *et al.* [125] proposent de comparer différentes méthodes de lissage afin d'estimer un profil temporel de vitesse et un profil d'accélération à partir de mesures de vitesses issues d'un GPS (avec une fréquence de 1 Hz), dans l'objectif d'estimer la consommation de carburant et les émissions de gaz. En effet, les auteurs soulignent que si la précision des mesures de vitesses des GPS actuels (l'article datant de 2001) est sensé être de l'ordre de 1 m/s, les données sont régulièrement entachées d'erreurs plus importantes. Dans un premier temps, les auteurs comparent différentes méthodes d'estimation de l'accélération obtenue par des techniques de différentiation numérique ("backward-difference", "forward-difference" et "central-difference") (Burden et Faires [26] chap. 4). Puis différentes méthodes de lissage sont comparées : "data trimming", lissage exponentiel simple, estimateur à noyau en utilisant le noyau d'Epanechnikov, et une version robuste du lissage exponentiel et de l'estimateur à noyau. Les méthodes robustes qui consistent à mettre un poids nul sur les valeurs aberrantes (valeurs de vitesse et d'accélération irréalisables) donnent les meilleurs résultats mais ces méthodes nécessitent de prédéfinir une zone de valeurs acceptables.

Les travaux de Jun *et al.* [82] consistent également en une comparaison de différentes méthodes de lissage afin d'estimer des profils temporels de vitesse et d'accélération à partir de mesures GPS. Les mesures de vitesses sont issues d'un GPS avec une fréquence de 1 Hz et sont obtenues par effet Doppler, donc indépendantes des mesures de position du véhicule (voir annexe A pour plus de détails). Les auteurs comparent trois méthodes de lissage afin d'estimer le profil de vitesse : la méthode des polynômes locaux avec une fonction de poids constante, un degré égal à 2 pour les polynômes et une fenêtre de lissage de 3 s ; l'estimateur de Nadaraya-Watson avec un noyau gaussien et une fenêtre de lissage de 3 s (cf. section 3.1.2 pour plus de détails) ; et un filtre de Kalman modifié. Les valeurs estimées sont comparées aux valeurs de vitesses dérivées de l'odomètre ("Vehicle Speed Sensor", VSS). Les résultats montrent que l'estimateur à noyau n'est pas un bon estimateur en général. L'estimateur par polynômes locaux corrige bien les valeurs aberrantes, mais à tendance à affecter les

valeurs fiables situées à proximité de ces valeurs et à sur-estimer ou sous-estimer les vitesses. Enfin, le filtre de Kalman modifié est un bon estimateur des vitesses mais tend à être affecté par les valeurs aberrantes. L'étude s'intéresse également à l'estimation du profil d'accélération et de la distance parcourue et conclut que le filtre de Kalman modifié présente les meilleurs résultats et requiert le moins de temps de calcul.

Dans des études plus récentes, certains auteurs confirment le choix de Bratt et Ericsson [24] d'utiliser la méthode des polynômes locaux pour estimer les vitesses (Ko *et al.* [91], Laureshyn et Ardö [94]) même si le choix des différents paramètres (choix du noyau et de la fenêtre de lissage essentiellement) n'est pas toujours facile, et soulignent le problème du calcul des dérivées premières et secondes correspondant aux valeurs d'accélération et de jerk (voir Bagdadi et Várhelyi [13] pour le calcul des valeurs de jerk). En effet, les erreurs d'estimation sont amplifiés lors du passage à la dérivée et le calcul par différentiation numérique rajoute des erreurs dans les valeurs des dérivées. Andersson [6] propose d'estimer les valeurs de vitesse après un lissage de la distance parcourue en fonction du temps puis un calcul de la dérivée par "backward-difference", et met en évidence l'importance du choix de la fréquence d'échantillonnage dans le calcul des vitesses : une fréquence trop élevée entraîne un profil de vitesse très bruité, alors qu'une fréquence trop faible entraîne un profil de vitesse trop lissé.

Après l'étape de lissage, des difficultés apparaissent également dans le calcul d'un profil de vitesse de référence (profil moyen ou profil V85). Les travaux du CETE Normandie-Centre (Violette *et al.* [170]) et du groupe Volkswagen (Kerper *et al.* [84]) mettent en évidence le problème du décalage des profils spatiaux de vitesse, notamment au niveau des arrêts (tous les véhicules ne s'arrêtent pas exactement au même endroit). Le calcul d'un profil de référence nécessite donc de recaler au préalable l'ensemble des profils de vitesse. Si Violette *et al.* [170] proposent une méthode basique basée sur une simple translation des profils, Kerper *et al.* [84] proposent une méthode plus élaborée. En effet, leur méthode consiste dans un premier temps, à discriminer les conducteurs qui s'arrêtent au feu (feu rouge) et ceux qui ne s'arrêtent pas (feu vert) par classification des profils de vitesse en utilisant la distance "Dynamic Time Warping" (DTW), puis dans la classe des conducteurs qui se sont arrêtés, les profils de vitesse sont recalés par un algorithme développé par les auteurs et basé sur une minimisation de la distance euclidienne. Ces travaux montrent que le problème du recalage de profils de vitesse est un problème qui nécessite d'être traité afin de pouvoir calculer des profils agrégés crédibles à partir d'un ensemble de profils de vitesse individuels (figure 1.9 b.). Ce problème sera abordé au chapitre 5.

1.4 Bilan et perspectives

Nous avons montré que la connaissance des vitesses pratiquées était essentielle à la fois d'un point de vue sécuritaire et environnemental, mais également pour

l'analyse et la prévision du trafic. En effet, le choix de la vitesse est une caractéristique importante du comportement du conducteur et est influencé par de nombreux facteurs relatifs au conducteur, au véhicule, à la route et à son environnement. L'information des vitesses réellement pratiquées devient accessible avec la généralisation des "véhicules traceurs". Ces véhicules équipés d'un GPS peuvent être vus comme des capteurs mobiles explorant le réseau routier en continu et sont capables de transmettre leur position et leur vitesse. Les informations recueillies par ces véhicules sont riches d'enseignement pour connaître l'usage réel du réseau et peuvent à ce titre intéresser les gestionnaires d'infrastructures comme les fournisseurs d'outils de navigation (Navteq et TeleAtlas). Si jusqu'à présent, la majorité des véhicules traceurs étaient des flottes privées (taxis, bus) et n'étaient pas représentatif de l'ensemble du réseau, le développement des études de conduite en situation naturelle (NDS) permet l'observation à grande échelle et sur une période de temps longue d'un ensemble de conducteurs lors de ses trajets quotidiens. En outre, la généralisation des smartphones équipés de GPS contribue à faire croître le nombre de "traceurs" susceptibles de renvoyer leur position et leur vitesse en temps réel, et permet ainsi de recueillir de l'information sur le réseau secondaire dont l'usage est mal connu.

Nous avons vu que les profils de vitesse étaient des outils très performants pour l'analyse des vitesses pratiquées et étaient riches d'information pour l'étude du comportement des conducteurs. En effet, contrairement à un indicateur agrégé ponctuel tel que la moyenne, les profils de vitesse permettent une analyse fine du comportement individuel des conducteurs sur de longues sections. Cependant, nous avons vu que les mesures de vitesse et de position étaient généralement entachées d'erreurs et qu'une étape de lissage était nécessaire afin d'obtenir une bonne estimation du "vrai" profil de vitesse. La revue des différentes méthodes de lissage utilisées jusqu'à maintenant a montré l'enjeu de proposer d'autres méthodes plus appropriées à l'estimation des profils de vitesse et des profils dérivés (accélération et jerk). L'estimation de la vitesse par moyenne mobile ou plus généralement par la méthode du noyau, ou par l'utilisation d'un filtre de Kalman, ne présente généralement pas de bons résultats en présence de valeurs aberrantes et pose le problème du calcul des dérivées. En effet, l'estimation de l'accélération ou du jerk se fait par différentiation numérique et se pose alors le problème du choix de la méthode ("backward", "forward" ou "central") et de la fréquence d'échantillonnage afin de minimiser les erreurs d'estimation qui sont amplifiées lors du passage à la dérivée. La plupart des chercheurs utilisent actuellement la méthode des polynômes locaux qui présente l'avantage d'être simple et de faciliter le calcul des dérivées. Cependant, si cette méthode corrige bien les valeurs aberrantes, elle a tendance à affecter les valeurs fiables situées à proximité de ces valeurs aberrantes et à sur-estimer ou sous-estimer les vitesses. Des problèmes similaires apparaissent dans l'estimation des valeurs dérivées (accélération) : les pics ont tendance à être aplatis, ce qui est un grave problème pour la reconnaissance des situations à risque (presque-accidents). En outre, cette méthode exige la détermination de nombreux paramètres qu'il n'est pas toujours facile de choisir : choix du noyau et de la fenêtre de lissage. Nous avons également soulevé le problème du calcul

d'un profil de vitesse de référence (profil V85 par exemple) à partir d'un ensemble de profils de vitesse. En effet, les profils de vitesse sont décalés au niveau des arrêts (tous les véhicules ne s'arrêtent pas exactement au même endroit) et il est donc nécessaire de les recaler avant de calculer un profil moyen ou un profil V85.

Dans les chapitres suivants, nous nous intéressons plus particulièrement à l'estimation des profils spatiaux de vitesse (i.e. vitesse en fonction de la position du véhicule) et nous proposons une méthodologie de construction d'un profil de vitesse de référence basée sur une approche fonctionnelle, permettant de palier les difficultés évoquées dans ce chapitre. Nous présentons dans un premier temps, l'intérêt d'utiliser une approche fonctionnelle et présentons une modélisation fonctionnelle des profils spatiaux de vitesse. Puis nous présentons une méthode d'estimation des profils spatiaux de vitesse basée sur les splines de lissage. Nous présentons également une méthodologie de construction de divers profils de référence, chacun étant adapté à une situation de conduite (ex : feu rouge/feu vert). Enfin, nous proposons d'étendre la notion de profil V85 à la notion de boxplots fonctionnels afin de représenter à la fois les variations de vitesse au cours du trajet, mais également la variabilité entre les individus.

Chapitre 2

Modélisation fonctionnelle des profils spatiaux de vitesse

L'objet de ce chapitre est la modélisation fonctionnelle des profils spatiaux de vitesse. En effet, nous avons vu au chapitre précédent que contrairement aux boucles magnétiques qui ne fournissent que des mesures ponctuelles de la vitesse, l'utilisation de boîtiers connectés aux ordinateurs de bord des véhicules ("data-logger") permet d'obtenir des mesures de vitesse d'un véhicule tout au long de son trajet. Ainsi l'utilisation de véhicules traceurs permet d'obtenir une mesure continue de la vitesse. En pratique, les mesures étant observées en des instants discrétisés, ces mesures sont en fait représentées comme des vecteurs de $\mathbb{R}^n$ où n représente le nombre de mesures. Cependant, le développement des technologies de capteurs permet d'obtenir des mesures avec des fréquences d'échantillonnage de plus en plus élevées. Ainsi, les données qui circulent sur le bus CAN sont en général disponibles à une fréquence de 20 Hz, soit 20 mesures par seconde. On est alors ramené à l'étude de vecteurs de grande dimension (n est très grand), ce qui pose des difficultés d'utilisation des méthodes classiques de statistique multivariée (fléau de la dimension, corrélation entre les observations proches...).

Une solution consiste alors à prendre en compte la structure sous-jacente continue de ces données et à traiter les profils de vitesse comme des fonctions et non comme des vecteurs de $\mathbb{R}^n$. L'étude des données fonctionnelles appelée "Analyse des Données Fonctionnelles" (AFD) s'est considérablement développée depuis une vingtaine d'années notamment sous l'impulsion des travaux de J.O. Ramsay.

Nous présentons donc dans la première section de ce chapitre les avantages de l'analyse des données fonctionnelles, ainsi que les méthodes permettant de convertir les données vectorielles en courbes.

Puis, dans une deuxième section, nous explicitons d'un point de vue mathématique l'espace fonctionnel des profils spatiaux de vitesse, i.e. les fonctions représentant la vitesse d'un véhicule en fonction de sa position. Nous proposons une définition de ces fonctions et nous étudions certaines de leurs propriétés (continuité, dérivabilité).

2.1 L'analyse des données fonctionnelles

2.1.1 Qu'est-ce que l'analyse des données fonctionnelles ?

De nombreuses méthodes ont été développées dans le cas où les observations sont des vecteurs de $\mathbb{R}^n$. Cependant, le développement des technologies des capteurs permet de collecter des données sur des grilles temporelles ou spatiales de plus en plus fines, ce qui conduit à traiter ces données comme des courbes ou plus généralement comme des fonctions de variables continues. Ces observations appelées données fonctionnelles sont l'objet d'étude d'un nouveau domaine de la Statistique appelée Analyse des Données Fonctionnelles (AFD ou FDA pour "*Functional Data Analysis*" en anglais) qui s'est considérablement développé depuis une vingtaine d'années. Si historiquement les premiers travaux sur les données fonctionnelles concernent la décomposition de Karhunen-Loève qui est une extension fonctionnelle de l'analyse en composante principale (ACP) aux processus stochastiques (Deville [44], Dauxois et Pousse [36]), les prémices de l'approche actuelle de l'analyse des données fonctionnelles sont attribués notamment aux travaux de Besse [17], Grenander [70] et Ramsay [126]. Le terme "*functional data analysis*" est apparu pour la première fois dans Ramsay et Dalzell [127] et a été popularisé avec le premier ouvrage de Ramsay et Silverman [129], puis avec les ouvrages de Ramsay et Silverman [130], Ramsay et Silverman [128], Bosq [23] et Ferraty et Vieu [57].

Les données de nature fonctionnelle apparaissent dans de nombreux domaines comme la climatologie, la chimiométrie, la biologie, le traitement du signal... Les figures 2.1.a et 2.1.b représentent des exemples de jeux de données issus de la chimiométrie et de la linguistique dont les données sont disponibles sur le site associé à l'ouvrage de Ferraty et Vieu [57] (`http://www.math.univ-toulouse.fr/staph/npfda/`). La figure 2.1.a représente 215 courbes spectrométriques permettant de déterminer le taux de graisse contenu dans chaque morceau de viande. Ces données sont obtenues en mesurant pour chaque morceau de viande l'absorbance de lumière pour 100 valeurs différentes de longueur d'onde. La figure 2.1.b concerne le problème de la reconnaissance vocale dont l'objectif est de pouvoir retranscrire phonétiquement des mots et des phrases prononcés par un individu. La figure 2.1.b représente un échantillon de 10 log-périodogrammes correspondant à 10 enregistrements sonores pour le phonème "sh".

En pratique, ces données continues sont observées en un ensemble fini de points de discrétisation et peuvent donc être traités comme des vecteurs de $\mathbb{R}^n$. Cependant, si les méthodes classiques de statistique multivariée sont bien adaptées au cas où les points de discrétisation sont peu nombreux et relativement distants des uns des autres, elles ne sont pas appropriées lorsque la fréquence d'échantillonnage devient élevée. En effet, on est alors confronté à des problèmes de grande dimension liés au nombre important de points de discrétisation par rapport à la taille de l'échantillon (fléau de la dimension,) ainsi qu'à des problèmes de corrélation entre des observations issues d'un même processus de nature continue. L'analyse des données fonctionnelles permet de résoudre ces problèmes en tenant compte de la structure sous-jacente

a) Courbes spectrométriques. b) Enregistrements sonores.

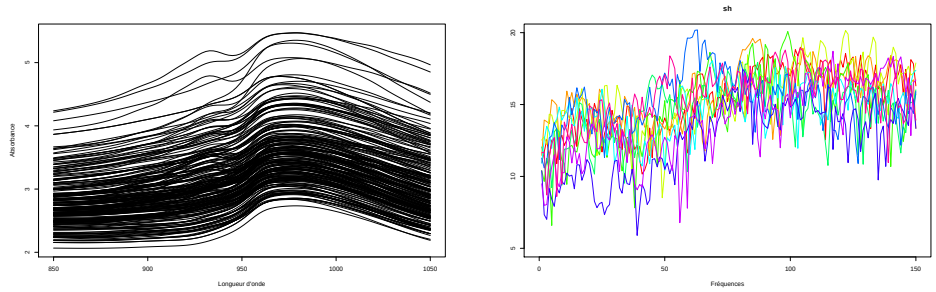


Figure 2.1 - Exemples de données fonctionnelles : la figure a) représente un échantillon de 215 courbes spectrométriques et la figure b) représente un échantillon de 10 enregistrements vocaux (log-périodogrammes) pour le phonème "sh".

a) Courbes de croissance. b) Courbes d'accélération.

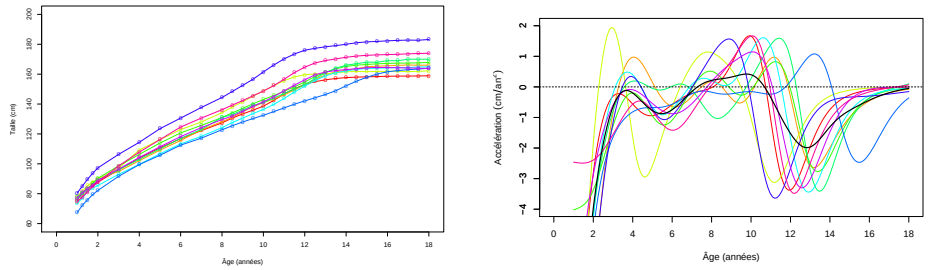


Figure 2.2 - Exemples de courbes de croissance : la figure a) représente les courbes de taille de 10 filles entre 1 et 18 ans (les cercles représentant les mesures), et la figure b) représente les courbes d'accélération associées (la courbe noire représentant la courbe moyenne).

continue de ce type de données et permet ainsi de prendre en compte certaines caractéristiques fonctionnelles des données : dérivées, périodicité, régularité... Par exemple, la figure 2.2.a représente un échantillon de courbes de taille de 10 filles pour lesquelles 31 mesures (cercles sur la figure) ont été effectuées entre 1 et 18 ans. Ces données sont extraites d'une étude californienne intitulée "Berkeley Growth Study" et sont disponibles sur le site très complet associé à l'ouvrage de Ramsay *et al.* [131] (`http://www.functionaldata.org`). Certaines propriétés remarquables de ces courbes apparaissent plus nettement sur les dérivées secondes comme le montre la figure 2.2.b. On peut également noter sur cette figure que la courbe d'accélération moyenne (en noir) n'est pas très représentative des courbes observées et qu'il serait intéressant de distinguer les variations d'amplitude et les variations temporelles (voir Ramsay et Silverman [128] section 6.5).

Enfin, l'analyse des données fonctionnelles permet également de traiter le cas

où les instants d'observation varient d'un individu à l'autre, ce qui revient à un problème de données manquantes dans le cas d'une approche multivariée classique. De plus, le fait de prendre en compte le caractère fonctionnel des données permet d'ajouter certaines contraintes de forme (positivité, monotonie, convexité...) dans la représentation fonctionnelle des données (voir chapitre 4).

2.1.2 Quand les données deviennent des courbes

La première étape d'une analyse de données fonctionnelle consiste à convertir les données brutes de nature vectorielle en objet fonctionnel. Lorsque les observations sont faites sans erreur (ou avec des erreurs négligeables), on fait une interpolation et dans ce cas les points d'observation appartiennent à la courbe. Lorsque les observations sont bruitées, on fait du lissage ("*smoothing*" en anglais) et dans ce cas les points d'observation n'appartiennent pas à la courbe. Le lissage et l'interpolation correspondent en fait à la résolution d'un problème de régression non paramétrique dont les principales méthodes seront détaillées au chapitre 3.

Une procédure classique permettant pour un individu donné de représenter les observations discrètes comme une fonction consiste à décomposer la fonction $f(t)$ que l'on cherche à estimer dans une base de fonctions ("*basis expansion*"), i.e. à approximer f par une combinaison linéaire finie de fonctions de base :

$$f(t) \approx \sum_{k=1}^{K} c_k \phi_k(t), \tag{2.1}$$

où les $\phi_k,\ k = 1, \ldots, K$, sont des fonctions de base initialement choisies, et où les coefficients $c_k,\ k = 1, \ldots, K$, sont estimés à partir des observations et représentent le poids de chaque fonction de base dans la construction de la fonction f. L'intérêt de cette méthode est de se ramener à un cadre d'étude de dimension finie même si f appartient à un espace de dimension infinie. De plus, le nombre de fonctions de base K est généralement très inférieur au nombre n d'observations, ce qui permet de réduire la dimension du problème et donc de pouvoir utiliser des méthodes classiques de statistique multidimensionnelle sur le vecteur des coefficients estimées $(c_1, \ldots, c_K)$. Il existe plusieurs choix de systèmes de fonctions de base. Les principales bases sont les suivantes :

- La base des monômes : $1,\ t,\ t^2, \ldots$, est utilisée pour construire des fonctions polynomiales. Ces fonctions de base ne sont pas très flexibles et ne sont utilisées que dans des cas simples.
- La base de Fourier : $1,\ sin(\omega t),\ cos(\omega t),\ sin(2\omega t),\ cos(2\omega t), \ldots$, est utilisée pour des fonctions périodiques.
- Les bases de splines permettent de construire des fonctions splines, i.e. des polynômes par morceaux avec conditions de continuité sur f et ses dérivées au niveau des points de jonction appelés noeuds. Les bases de splines sont très flexibles et donc bien adaptées au cas de fonctions complexes non périodiques. Les systèmes de base de splines sont définis par l'ordre de la spline, et par le

nombre et la position des noeuds. La base de splines la plus utilisée est la base des B-splines qui présente de bonnes propriétés numériques. La définition des splines polynomiales et l'étude des bases de spline sont détaillées respectivement aux sections 3.2.1.1 et 3.2.1.2 du chapitre 3.

– Les bases d'ondelettes combinent les avantages de la base de Fourier et des bases de splines. Elles fournissent une analyse en fréquence et permettent également la localisation temporelle de changement abrupts. De plus, en raison de leurs bonnes propriétés numériques, elles sont souvent utilisées en compression d'images.

Une fois la base choisie, la principale difficulté est le choix du nombre K de fonctions dans la base. En effet, le nombre de fonctions de base permet de contrôler le degré de lissage de la fonction f et résulte d'un compromis biais-variance : un nombre de noeuds important conduira à une fonction interpolant les données mais peu lisse (sur-lissage), alors qu'un faible nombre de noeuds conduira à une fonction trop lisse ne reflétant pas assez les caractéristiques des données (sous-lissage). Le choix du nombre de noeuds dans le cas de l'utilisation d'une base de splines est discuté au chapitre 3.

Une autre méthode permettant de construire une fonction à partir de données discrètes est la méthode de régularisation. Cette méthode consiste à trouver une fonction f qui ajuste bien les données, tout en pénalisant (voire interdisant) les solutions indésirables par l'ajout d'un terme de pénalité afin d'éviter un sur-lissage des données. La méthode de régularisation sera étudiée en détails au chapitre 3.

2.1.3 Méthodes d'analyse de données fonctionnelles

Contrairement à l'analyse statistique multivariée classique où l'on manipulait des vecteurs de $\mathbb{R}^n$, la difficulté de l'analyse des données fonctionnelles est de manipuler des fonctions. Ainsi, plutôt que de considérer chaque observation comme la réalisation d'une variable aléatoire multivariée de grande dimension, celle-ci sera considérée comme une discrétisation de la réalisation d'une variable aléatoire dite fonctionnelle, i.e. à valeurs dans un espace de dimension infinie. L'analyse des données fonctionnelles requièrent donc la construction de nouveaux outils théoriques et appliqués adaptés à l'étude de variables aléatoires fonctionnelles.

La plupart des méthodes utilisées en statistique multivariée ont été adaptée au cadre fonctionnel : le modèle linéaire, l'analyse en composantes principales, l'analyse discriminante, l'analyse canonique... De plus, le fait de prendre en compte le caractère fonctionnel des données a permis l'apparition de nouvelles problématiques : recalage de courbes, modèles fonctionnels définis par des équations différentielles ou des systèmes dynamiques... La richesse et la complexité des données fonctionnelles ont conduit à un véritable engouement de la communauté scientifique envers ce type de données et à la publication de nombreux travaux sur le sujet (Ferraty et Vieu [58]). Nous ne détaillerons pas ici ces nombreuses méthodes et nous renvoyons le lecteur aux ouvrages généraux de Ramsay et Silverman [130], Ramsay et Silverman [128] et Ferraty et Vieu [57] pour un aperçu des différentes méthodes, ainsi qu'à l'ar-

ticle de Levitin *et al.* [96] pour un aperçu rapide des principales méthodes. Un état de l'art assez complet de ces méthodes est également disponible dans les thèses en français de Conan-Guez [31] (chapitre 2) et Delsol [40] (chapitres 2 et 3). Les notions d'alignement de courbes et de boxplots fonctionnels seront abordées au chapitre 5.

2.1.4 Conclusion

La section précédente a montré la richesse des données fonctionnelles et l'importance de prendre en compte le caractère fonctionnel de données issues d'un processus continu. De plus, nous avons vu que la plupart des méthodes classiques de statistique multivariée avaient été adaptées aux données fonctionnelles, et que ces données avaient également permis l'apparition de nouvelles méthodes tenant compte de leurs caractéristiques fonctionnelles. L'analyse des données fonctionnelles est particulièrement bien adaptée aux profils spatiaux de vitesse et permet de tenir compte de la richesse des données issues de véhicules traceurs. Dans la suite de cette thèse, nous nous placerons donc dans un cadre fonctionnel, et nous traiterons les profils spatiaux de vitesse comme des fonctions. Ceci nécessite de proposer une modélisation fonctionnelle des profils spatiaux de vitesse et de répondre aux questions suivantes : Quelles sont les fonctions pouvant représenter un profil spatial de vitesse ? Quelles sont les propriétés de ces fonctions ? Les réponses à ces questions sont l'objet de la section suivante.

2.2 Définition et propriétés des profils spatiaux de vitesse

Afin de se placer dans le cadre de l'analyse des données fonctionnelles, nous devons définir l'espace fonctionnel des profils spatiaux de vitesse. En effet, toute fonction $f : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ n'est pas un profil spatial de vitesse. Par exemple, la fonction représentée à la figure 2.3.a n'est pas un profil spatial de vitesse puisqu'elle est nulle sur l'intervalle $[x_1, x_2]$ avec $x_1 \neq x_2$, or un véhicule ne peut pas se déplacer si sa vitesse est nulle. Plus surprenant, la figure 2.3.b n'est pas non plus la représentation d'un profil spatial de vitesse. En effet, nous allons voir à la section suivante qu'une telle fonction affine par morceaux n'est pas un profil spatial de vitesse.

On propose donc dans cette section de définir l'ensemble des fonctions pouvant être la représentation d'un profil spatial de vitesse et d'étudier les propriétés de telles fonctions.

2.2.1 Definition de l'espace fonctionnel des profils spatiaux de vitesse

En pratique, un profil spatial de vitesse est une succession de mesures horodatées de position et de vitesse, et peut donc être manipulé dans les 3 espaces suivants : `distance×temps`, `vitesse×temps` ou `vitesse×distance`. La figure 2.4 illustre le lien entre ces 3 espaces : la distance parcourue en fonction du temps est représentée dans l'espace `distance×temps`, le profil temporel de vitesse ("*time-speed profile*") est

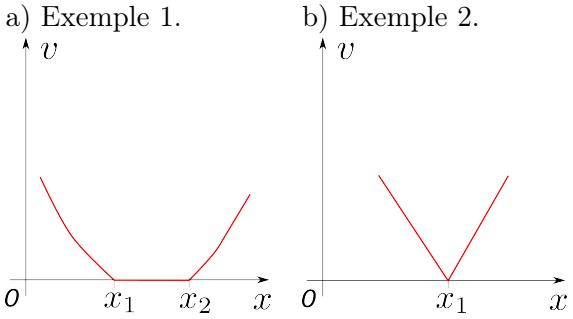


FIGURE 2.3 - Exemples de fonctions qui ne sont pas des profils spatiaux de vitesse.

représenté dans l'espace `vitesse×temps`, et le profil spatial de vitesse ("*space-speed profile*") est représenté dans l'espace `vitesse×distance`.

Nous proposons de définir l'espace fonctionnel des profils spatiaux de vitesse de la manière suivante :

Définition 2.1. *Soit* $x_f \in \mathbb{R}^+$. *Alors l'espace des profils spatiaux de vitesse sur* $[0, x_f]$, *noté* $\mathcal{E}_{SSP}$, *est défini de la manière suivante :*
$\mathcal{E}_{SSP} = \{v_S : [0, x_f] \longrightarrow \mathbb{R}^+$ *tel qu'il existe un réel positif* T *et une fonction* $\mathcal{C}^2$ *croissante* $F : [0, T] \longrightarrow [0, x_f]$ *avec* $F(0) = 0$ *tels que*
$v_S(x) = F' \circ F^{-1}(x)$, $x \in [0, x_f]\}$,
où F' *est la dérivée de* F *par rapport à* t, *et* F^{-1} *est l'inverse généralisée de* F *définie par* $F^{-1}(x) = inf\{t \in [0, T],\ F(t) = x\}$ *pour tout* $x \in [0, x_f]$.

Dans cette définition, les réels positifs x_f et T représentent respectivement la longueur et le temps de parcours de la section étudiée. La fonction croissante $F : [0, T] \longrightarrow [0, x_f]$ représente la distance parcourue en fonction du temps (espace `distance×temps`), et sa dérivée $F'(t)$ représente le profil temporel de vitesse (espace `vitesse×temps`). Le profil spatial de vitesse v_S (espace `vitesse×distance`) se déduit des deux fonctions précédentes par la transformation suivante : $v_S(x) = (F'(F^{-1})(x))$, où F^{-1} est l'inverse généralisée de F (voir section 2.2.2.1 pour les propriétés de l'inverse généralisée). Le diagramme de la figure 2.5 illustre le lien entre ces 3 espaces en utilisant les notations de la définition 2.1.

Remarque 2.1.
La condition que $F : [0, T] \longrightarrow [0, x_f]$ soit $\mathcal{C}^2$ est ici arbitraire et permet d'obtenir un profil temporel d'accélération, correspondant à la fonction $F''(t)$, qui soit continu. Dans ce manuscrit, nous ne nous intéresserons pas aux dérivées de F d'ordre supérieur à 2. Cependant, il est possible d'imposer que la fonction F soit $\mathcal{C}^m$ avec $m > 2$, voire même $\mathcal{C}^\infty$.

Exemple 2.1.
La fonction définie par $v_S(x) = 2\sqrt{x}$ pour $x \in [0, x_f]$ est un profil spatial de vitesse. En effet, il existe un réel $T \in \mathbb{R}^+$ tel que la fonction $F : [0, T] \longrightarrow [0, x_f]$ définie

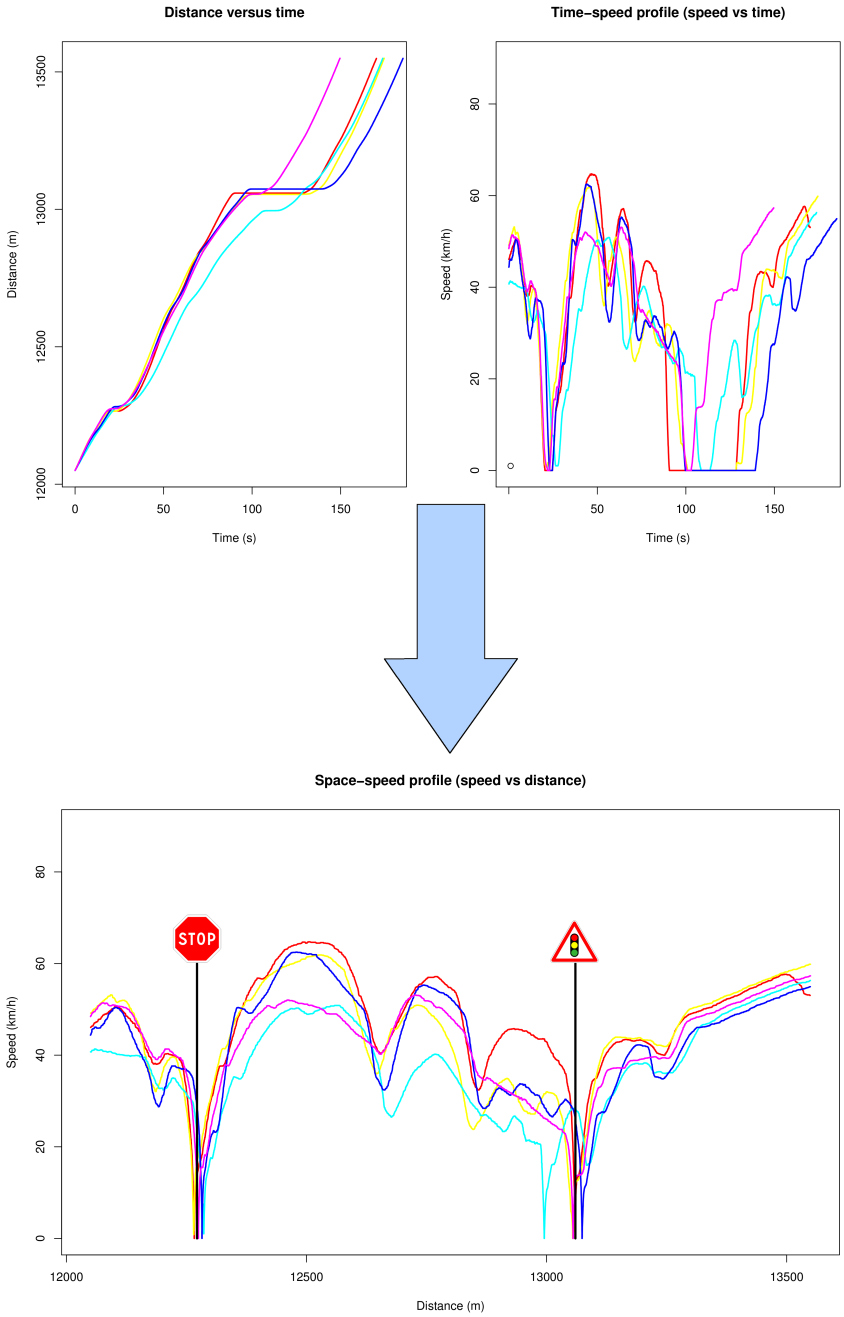


FIGURE 2.4 - Lien entre les trois espaces : [`distance`×`temps`, `vitesse`×`temps`] et [`vitesse`×`distance`].

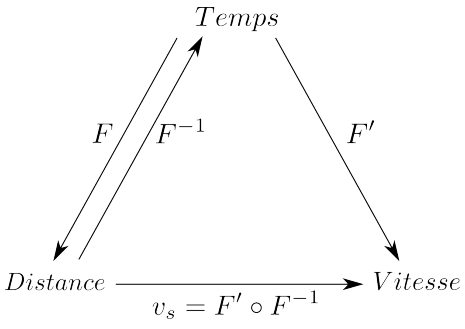


Figure 2.5 - Diagramme fonctionnel illustrant la définition des profils spatiaux de vitesse.

par $F(t) = t^2$ vérifie les conditions de la définition 2.1 puisque pour tout $t \in [0, T]$, $F'(t) = 2t$, pour tout $x \in [0, x_f]$, $F^{-1}(x) = \sqrt{x}$, $F(0) = 0$, et pour tout $x \in [0, x_f]$, $v_S(x) = (F' \circ F^{-1})(x)$.

Exemple 2.2.
On pose $v_S(x) = ax + b$, avec $a \neq 0$, $b \neq 0$, et $x \in [0, x_f]$. On considère la fonction $F : [0, T] \longrightarrow [0, x_f]$ définie par $F(t) = exp(at) - b/a$. On a pour tout $t \in [0, T]$, $F'(t) = a\, exp(at)$, et pour tout $x \in [0, x_f]$, $F^{-1}(x) = (1/a)\, log(x+b/a)$. La condition $F(0) = 0$ étant vérifiée pour $a = b$, on en déduit que $v_S(s) = ax + b$ est un profil spatial de vitesse si $a = b$.

Remarque 2.2.
De la relation $\frac{d}{dt}F(F^{-1}(x)) = v_S(x)$, on déduit en posant $z = F^{-1}(x)$, l'équation suivante :

$$\frac{d}{dt}F(z) = v_S(F(z)). \tag{2.2}$$

Si v_S est continue, on se ramène alors à une équation différentielle autonome du 1er ordre de la forme $y' = f(y)$ (voir par exemple dans Demailly [41]). Ainsi, connaissant le profil spatial de vitesse v_S, trouver la fonction F associée représentant la distance parcourue en fonction du temps revient à résoudre l'équation différentielle (2.2). Sous certaines hypothèses de régularité sur v_S, l'équation différentielle (2.2) avec la condition initiale $F(0) = 0$ admet une unique solution (Théorème de Cauchy-Lipschitz).

Exemple 2.3 (Contre-exemple)**.**
Supposons que pour $x \in [0, x_f]$, $v_S(x) = \lambda x$ avec $\lambda \neq 0$. On en déduit alors que

$$a(t) = \frac{dv}{dt} = \frac{dv}{dx}\frac{dx}{dt} = \lambda v(t).$$

Or, à partir de l'équation $a(t) = \lambda v(t)$, si on pose $\dot{x} = dx/dt$, on se ramène à l'équation différentielle du 2^{nd} ordre suivante : $\ddot{x} - \lambda\dot{x} = 0$. En posant $y = \dot{x}$, on

se ramène en fait à une équation différentielle du 1er ordre $\dot{y} - \lambda y = 0$, dont la solution est de la forme $y = Kexp(\lambda t)$ où K est une constante. On en déduit donc que $x(t) = (K/\lambda)exp(\lambda t)$. Si on pose comme condition initiale $x(0) = 0$, on en déduit que $K = 0$ et donc que l'unique solution est $x(t) = 0$ pour tout $t \in [0, T]$. On en déduit donc que les fonctions linéaires de la forme $v_S(x) = \lambda x$ avec $\lambda \neq 0$ ne sont pas des profils spatiaux de vitesse.

On déduit des exemples 2.2 et 2.3 que les fonctions affines par morceaux ne sont pas des profils spatiaux de vitesse. En effet, d'après l'exemple 2.2, les seules fonctions affines qui sont des profils spatiaux de vitesse appartenant à l'espace $\mathcal{E}_{SSP}$ (i.e. avec la condition initiale $F(0) = 0$) sont de la forme $v_S(x) = ax + a$ avec $a \neq 0$. Cependant ces fonctions ne s'annulent pas sur l'intervalle $[0, x_f]$, avec $x_f > 0$ (ces fonctions s'annulent uniquement en $x = -1$), excluant tout arrêt du véhicule sur $[0, x_f]$ (cas où la vitesse s'annule). De plus, l'exemple 2.3 montre que dans le cas où v_S est une fonction linéaire, la seule solution de l'équation différentielle (2.2) avec la condition initiale $F(0) = 0$ est la fonction nulle $F(t) = 0$. Ainsi, une interpolation linéaire des données de vitesse dans l'espace `vitesse`×`distance` est à exclure, bien que cette méthode soit couramment utilisée en pratique.

2.2.2 Propriétés des profils spatiaux de vitesse

Nous avons défini à la définition 2.1 les fonctions correspondant à des profils spatiaux de vitesse. On propose alors d'étudier certaines propriétés de ces fonctions. Une première propriété évidente est la positivité des profils spatiaux de vitesse.

Théorème 2.1. *Toute fonction $v_S \in \mathcal{E}_{SSP}$ est positive.*

Démonstration. Toute fonction $v_S \in \mathcal{E}_{SSP}$ étant définie par $v_S(x) = F'(F^{-1}(x))$ où F est une fonction croissante, on en déduit que v_S est une fonction positive, i.e. $\forall x \in [0, x_f], \quad v_S(x) \geq 0$. □

Avant d'étudier les propriétés de continuité et de dérivabilité des profils spatiaux de vitesse, nous allons commencer par énoncer certaines propriétés de l'inverse généralisée.

2.2.2.1 Préliminaires : propriétés de l'inverse généralisée

Commençons par énoncer quelques rappels sur la fonction réciproque d'une fonction.

Définition 2.2. *Soit I un intervalle de $\mathbb{R}$ et soit f une fonction bijective de I sur J où J est un intervalle de $\mathbb{R}$. On appelle fonction réciproque de f l'application, notée f^{-1}, définie sur J par $f^{-1}(y) = x$ où x est l'unique élément de I tel que $f(x) = y$.*

Proposition 2.1. *Soit I un intervalle de $\mathbb{R}$ et soit f une fonction définie sur I, continue et strictement monotone. Alors f réalise une bijection de I sur $J = f(I)$ qui est un intervalle de $\mathbb{R}$ et sa fonction réciproque f^{-1} vérifie les propriétés suivantes :*

1. *f^{-1} est strictement monotone de J sur I, de même sens de variation que f.*
2. *f^{-1} est continue sur J.*
3. *Si f est dérivable sur I et ne s'annule pas sur I, alors f^{-1} est dérivable sur J et $(f^{-1})' = \frac{1}{f' \circ f^{-1}}$.*

Lorsqu'une fonction f définie de I sur J est continue et monotone (non strictement), sa fonction réciproque n'est pas définie (puisque f n'est pas bijective). Cependant, on peut définir son inverse généralisée f^{-1}, définie pour tout $y \in J$ par $f^{-1}(x) = inf\{x \in I,\ f(x) = y\}$. Cette notion d'inverse généralisée (ou pseudo-inverse) généralise la notion de réciproque d'une fonction continue, et on a alors les propriétés suivantes :

Théorème 2.2. *Soit $F : [0,T] \longrightarrow [0, x_f]$ une fonction continue croissante d'inverse généralisée F^{-1} définie par :*

$$\forall x \in [0, x_f], F^{-1}(x) = inf\{t \in [0,T],\ F(t) = x\}.$$

Alors on a les propriétés suivantes :

1. *$F \circ F^{-1} = id_{[0,x_f]}$.*
2. *$\forall t \in [0,T], F^{-1} \circ F(t) \leq t$.*
3. *F^{-1} est strictement croissante.*
4. *Pour tout $t \in [0,T]$ et tout $x \in [0, x_f]$, on a l'équivalence suivante :*

$$F(t) \geq x \Leftrightarrow t \geq F^{-1}(x).$$

5. *F^{-1} est continue à gauche.*

Démonstration.

1. Soit $x \in [0, x_f]$. On pose $t = F^{-1}(x)$. Par définition de $F^{-1}(x)$ comme borne inférieure, il existe une suite (t_n) de $[0,T]$ de limite t avec $F(t_n) = x$. Par continuité de F et passage à la limite quand $n \longrightarrow +\infty$, on en déduit que $F(t) = x$. On obtient donc que : $\forall x \in [0, x_f], F \circ F^{-1}(x) = x$.

2. Soit $t \in [0,T]$ et $x = F(t)$. Par définition de $F^{-1}(x)$, on a $F^{-1}(x) \leq t$, soit $F^{-1} \circ F(t) \leq t$.

3. Soient x_1 et x_2 appartenant à $[0, x_f]$ avec $x_1 < x_2$. Supposons $F^{-1}(x_1) \geq F^{-1}(x_2)$. Alors par la croissance de F, on a $F(F^{-1}(x_1)) \geq F(F^{-1}(x_2))$, soit $x_1 \geq x_2$ ce qui est impossible. Donc $F^{-1}(x_1) < F^{-1}(x_2)$. On en déduit donc que F^{-1} est strictement croissante.

4. Soient $t \in [0,T]$ et $x \in [0, x_f]$. Si $F(t) \geq x$, alors comme F^{-1} est croissante d'après 3, on a $F^{-1} \circ F(t) \geq F^{-1}(x)$. Or d'après 2, $t \geq F^{-1} \circ F(t)$, soit $t \geq F^{-1}(x)$. Réciproquement, si $t \geq F^{-1}(x)$, alors comme F est croissante, on a $F(t) \geq F \circ F^{-1}(x)$, soit $F(t) \geq x$ d'après 1.

5. Soit $x \in [0, x_f]$. On distingue 2 cas :

1er cas : $F^{-1}(x)$ appartient à un intervalle ouvert sur lequel F n'est pas constante. Dans ce cas, F coïncide localement avec une fonction ϕ strictement croissante continue. Or selon les théorèmes usuels, ϕ^{-1} est continue en x, donc continue à gauche en x. Donc F^{-1} est également continue à gauche.

2ème cas : $F^{-1}(x)$ est l'extrémité gauche d'un intervalle fermé sur lequel F est constante.

Alors il existe $t_1 < F^{-1}(x)$ tel que F soit continue et strictement croissante sur $[t_1, F^{-1}(x)]$. Sur $[F^{-1}(x), T]$, on peut remplacer F par une fonction ϕ continue et strictement croissante (figure 2.6). La fonction ψ définie sur $[t_1, T]$ coïncidant avec F sur $[t_1, F^{-1}(x)]$ et avec ϕ sur $[F^{-1}(x), T]$ est continue et strictement croissante sur $[t_1, T]$.

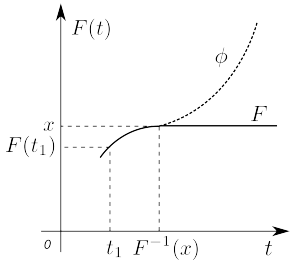


FIGURE 2.6 - Cas où $F^{-1}(x)$ est l'extrémité gauche d'un intervalle fermé sur lequel F est constante.

D'après les théorèmes usuels, sa réciproque est continue en x, et donc continue à gauche en x. La fonction F coïncidant avec ψ sur $[t_1, F^{-1}(x)]$, on en déduit également que sa réciproque F^{-1} est continue à gauche en x.

□

Remarque 2.3.

1. $F^{-1} \circ F$ n'est pas toujours égal à $id_{[0,T]}$. En effet, dans l'exemple illustré à la figure 2.7, on a $F^{-1} \circ F(t_2) = t_1 \neq t_2$. Ainsi, on a seulement la propriété 2 du théorème précédent : $\forall t \in [0, T], F^{-1} \circ F(t) \leq t$.

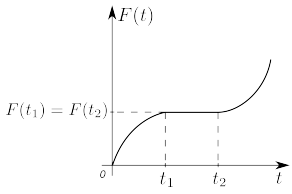


FIGURE 2.7 - Cas où $F^{-1} \circ F = id_{[0,T]}$ n'est pas vrai.

2. Soit $x \in [0, x_f]$. F^{-1} est discontinue à droite en x si $\exists t_g \neq t_d$ tel que $F(t_g) = F(t_d) = x$ (i.e. il existe un "plateau" en x).
En effet, soit $A = \{t \in [0, T], F(t) = x\}$. On note $t_g = inf(A)$ et $t_d = sup(A)$ (figure

2.8).

Montrons dans un premier temps que A est un intervalle fermé.
On a $F(t_g) = F(t_d) = F(x)$ par continuité de F en t_g et en t_d. Donc F est constante sur $[t_g, t_d]$. D'autre part, $\forall t < t_g$, on a $F(t) < F(t_g)$ (car si $F(t) = F(t_g)$, t_g n'est plus un minorant de A). De même, $\forall t > t_d$, on a $F(t) > F(t_d)$. On en déduit donc que A est un intervalle fermé (non singleton).

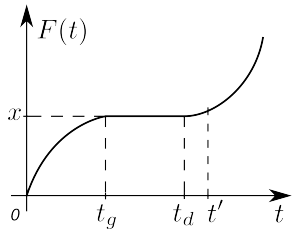


Figure 2.8 - Démonstration de la non continuité à droite de F^{-1}.

Ainsi $A = [t_g, t_d]$ avec $t_g = F^{-1}(x)$ par définition de F^{-1}. De plus, F coïncide sur un certain intervalle $[t_d, t']$ $(t' > t_d)$ avec une fonction continue strictement croissante. On en déduit que $\lim\limits_{\substack{x' \to x \\ x' > x}} F^{-1}(x') = t_d$. Or $t_d \neq F^{-1}(x)$. Donc F^{-1} n'est pas continue à droite en x.

2.2.2.2 Continuité des profils spatiaux de vitesse

Lemme 2.1. *Soit $x_0 \in [0, x_f]$ un point de discontinuité de F^{-1}. Alors la vitesse est nulle en ce point, i.e. $v_S(x_0) = F' \circ F^{-1}(x_0) = 0$.*

Démonstration. Soit $x_0 \in [0, x_f]$ un point de discontinuité de F^{-1}. On se ramène au cas 2 de la remarque 2.3. Alors $F^{-1}(x_0) = t_g$ où $[t_g, t_d]$ est l'intervalle sur lequel F est constante et égale à x_0. On en déduit que $F'_d(t_g) = 0$, et comme F est supposée dérivable, on a également $F'_g(t_g) = 0$. Finalement, $F'(t_g) = 0$. Donc $F' \circ F^{-1}(x_0) = 0$ (ou $v_S(x_0) = 0$). □

Remarque 2.4. La réciproque est fausse (exemple d'un point d'inflexion en x_0).

Théorème 2.3. *Toute fonction $v_S : [0, x_f] \longrightarrow \mathbb{R}^+$ appartenant à l'espace des profils spatiaux de vitesse $\mathcal{E}_{SSP}$ défini en 2.1 est continue sur $[0, x_f]$.*

Démonstration. Soit $x_0 \in [0, x_f]$. On distingue deux cas :
1er cas : x_0 est un point de continuité de F^{-1}.
Alors par composition de deux fonctions continues, on en déduit que $F' \circ F^{-1}$ est continue en x_0.
2ème cas : x_0 est un point de discontinuité de F^{-1}.
$(v_S)_g(x_0) = \lim\limits_{\substack{x \to x_0 \\ x < x_0}} v_S(x) = \lim\limits_{\substack{x \to x_0 \\ x < x_0}} F' \circ F^{-1}(x)$. On reprend les notations du lemme 2.1, i.e. $F^{-1}(x_0) = t_g$ où $[t_g, t_d]$ est l'intervalle sur lequel F est constante et égale à x_0. Lorsque $x \to x_0$ par valeurs inférieures, $t \to t_g$ par valeurs inférieures. Or

$\lim_{\substack{t\to t_g \\ t<t_g}} F'(t) = 0$ car $F' = 0$ sur $[t_g, t_d]$ et F' est continue en t_g. Donc $(v_S)_g(x_0) = 0$. De même, $(v_S)_d(x_0) = \lim_{\substack{x\to x_0 \\ x>x_0}} v_S(x) = \lim_{\substack{x\to x_0 \\ x>x_0}} F' \circ F^{-1}(x)$. Or lorsque $x \to x_0$ par valeurs supérieures, $t \to t_g$ par valeurs supérieures, et $\lim_{\substack{t\to t_g \\ t>t_g}} F'(t) = 0$ car $F' = 0$ sur $[t_g, t_d]$. Donc $(v_S)_d(x_0) = 0$. Et comme, d'après le lemme 2.1, on a $v_S(x_0) = 0$, on en déduit que v_S est continue en x_0. □

2.2.2.3 Dérivabilité des profils spatiaux de vitesse

Commençons par étudier le cas où la vitesse n'est jamais nulle, ce qui correspond au cas où le véhicule ne s'arrête pas.

Théorème 2.4. *Soit $v_S : [0, x_f] \longrightarrow \mathbb{R}^+$ appartenant à l'espace des profils spatiaux de vitesse $\mathcal{E}_{SSP}$ défini en 2.1. Soit $H_+ = \{x \in [0, x_f], v_S(x) > 0\}$. Alors v_S est dérivable sur H_+.*

Démonstration. Soit $x_0 \in H_+$, i.e. $v_S(x_0) = F' \circ F^{-1}(x_0) > 0$. On note $t_0 = F^{-1}(x_0)$. Alors $F'(t_0) \neq 0$, et comme F est dérivable en t_0 (par définition de $\mathcal{E}_{SSP}$), on en déduit que F^{-1} est dérivable en $F(t_0) = x_0$. D'autre part, F étant de classe $\mathcal{C}^2$, F' est dérivable en $t_0 = F^{-1}(x_0)$. Donc d'après le théorème de dérivabilité d'une fonction composée, on en déduit que $v_S = F' \circ F^{-1}$ est dérivable en x_0. □

Théorème 2.5. *Soit $v_S : [0, x_f] \longrightarrow \mathbb{R}^+$ appartenant à l'espace des profils spatiaux de vitesse $\mathcal{E}_{SSP}$ défini en 2.1. Soit $H_0 = \{x \in [0, x_f], v_S(x) = 0\}$. On introduit les hypothèses suivantes :*

(H_1) *F est de classe $\mathcal{C}^2$ sur $[0, T]$, strictement croissante, $\exists t_0 \in]0, T[$ tel que $F'(t_0) = 0$, $F'''(t_0)$ existe avec $F'''(t_0) \neq 0$.*

(H_2) *F est de classe $\mathcal{C}^2$ sur $[0, T]$, croissante, $\exists t_0, t_1 \in]0, T[$, $t_0 \neq t_1$ tel que $F'(t) = 0$ sur $[t_0, t_1]$, et la fonction G définie sur $[0, T - (t_1 - t_0)]$ par :*

$$\begin{cases} \text{pour } t \leq t_0, & G(t) = F(t) \\ \text{pour } t \geq t_0, & G(t) = F(t + t_1 - t_0) \end{cases}$$

vérifie les hypothèses (H_1).

Si F vérifie les hypothèses (H_1) ou (H_2), alors $v_S = F' \circ F^{-1}$ n'est pas dérivable sur H_0.

Démonstration.
1er cas : On suppose que F vérifie les hypothèses (H_1).
On définit x_0 tel que $t_0 = F^{-1}(x_0)$. Comme $F'(t_0) = 0$, on a $v_S(x_0) = 0$, i.e. $x_0 \in H_0$. Sous les hypothèses (H_1), on peut appliquer la formule de Taylor-Young à F' : pour tout θ dans un voisinage de t_0, on a
$F'(t_0 + \theta) = F'(t_0) + \theta F''(t_0) + \frac{\theta^2}{2} F'''(t_0) + \theta^2 \varepsilon(\theta)$, où $\varepsilon(\theta) \to 0$ quand $\theta \to 0$.

Or $F'(t_0) = 0$. Ainsi, si on avait $F''(t_0) \neq 0$, F' changerait de signe en t_0, ce qui contredirait la stricte monotonie de F. Donc nécessairement $F''(t_0) = 0$. Alors $F'(t_0 + \theta) \underset{\theta \to 0}{\sim} \frac{\theta^2}{2} F'''(t_0)$ (car $F'''(t_0) \neq 0$ par hypothèse).
On pose $h = F(t_0 + \theta) - F(t_0)$. On applique la formule de Taylor-Young à F : $h = F(t_0+\theta) - F(t_0) = \theta F'(t_0) + \frac{\theta^2}{2} F''(t_0) + \frac{\theta^3}{6} F'''(t_0) + \theta^3 \varepsilon'(\theta)$ où $\varepsilon'(\theta) \to 0$ quand $\theta \to 0$. Afin d'étudier la dérivabilité de v_s en x_0, on forme alors le taux d'accroissement suivant :
$\frac{v_S(x_0+h)-v_S(x_0)}{h} = \frac{F'(t_0+\theta)-F'(t_0)}{F(t_0+\theta)-F(t_0)} \underset{\theta \to 0}{\sim} \frac{\frac{\theta^2}{2} F'''(t_0)}{\frac{\theta^3}{6} F'''(t_0)} = \frac{3}{\theta}$. Le taux d'accroissement est sans limite quand $\theta \to 0$. Cependant ceci ne prouve pas qu'il soit également sans limite quand $h \to 0$.
On raisonne alors par l'absurde. Supposons que $\frac{v_S(x_0+h)-v_S(x_0)}{h} \underset{h \to 0}{\to} \ell \in \mathbb{R}$. Alors, par définition,
$\forall \varepsilon > 0$, $\exists \alpha > 0$ tel que $|h| < \alpha \Rightarrow |\frac{v_S(x_0+h)-v_S(x_0)}{h} - \ell| < \varepsilon$.
Or F est continue en t_0, donc $\exists \beta > 0$ tel que $|t - t_0| < \beta \Rightarrow |F(t) - F(t_0)| < \alpha$, ou encore $|\theta| < \beta \Rightarrow |\underbrace{F(t_0 + \theta) - F(t_0)}_{h}| < \alpha$.
On en déduit que $\forall \varepsilon > 0$, $\exists \beta > 0$ tel que $|\theta| < \beta \Rightarrow |\frac{v_S(x_0+h)-v_S(x_0)}{h} - \ell| < \varepsilon$, ce qui signifie que le taux d'accroissement aurait une limite $\ell \in \mathbb{R}$ lorsque $\theta \to 0$, ce qui est contradictoire. On en déduit que sous les hypothèses (H_1), v_S n'est pas dérivable en x_0.

$2^{\text{ème}}$ cas : On suppose que F vérifie les hypothèses (H_2).

Comme dans le 1^{er} cas, on définit x_0 tel que $t_0 = F^{-1}(x_0)$. La courbe représentative de G :
- coïncide avec celle de F sur $[0, t_0]$,
- se déduit de celle de F par la translation de vecteur $(t_0 - t_1)\vec{i}$ sur $[t_0, T - (t_1 - t_0)]$.

La fonction G définit donc le même mouvement que la fonction F mais en faisant abstraction de la période de temps $[t_0, t_1]$ pendant laquelle il y a immobilité (figure 2.9).

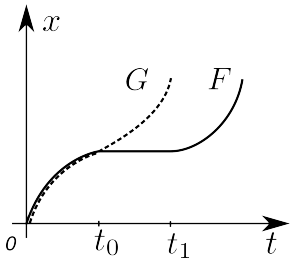


Figure 2.9 - Illustration des hypothèses 2.

Ainsi c'est le même taux d'accroissement qui intervient, et donc si v_S n'est pas dérivable en x_0 pour l'un des mouvements, elle ne l'est pas pour l'autre. Autrement dit, le résultat de (H_1) où $F' = 0$ en un seul point t_0 s'étend au cas plus général où F' est nulle sur un intervalle $[t_0, t_1]$ $(t_0 \neq t_1)$, sous réserve des hypothèses (H_1)

exprimées sur la fonction G. □

Remarque 2.5. Dans le cas où F vérifie les hypothèses (H_2), l'existence de $F'''(t_0)$ n'est pas nécessaire. En effet, nous verrons dans l'exemple 2.5, un cas où F est croissante, constante et égal à x_0 sur $[t_0, t_1]$ $(t_0 \neq t_1)$, et $F'''(t_0)$ n'est pas définie, et où v_S n'est pas dérivable en x_0. Ainsi l'existence de $F'''(t_0)$ n'est pas une hypothèse restrictive. Plus généralement, les hypothèses (H_1) et (H_2) ne sont pas restrictives et sont satisfaites dans la plupart des cas.

Ce théorème montre que les profils spatiaux de vitesse ne sont pas dérivables aux points où la vitesse est nulle (points correspondant aux arrêts du véhicule). D'un point de vue géométrique, on peut étudier la forme de la courbe représentative d'un profil spatial de vitesse au voisinage d'un tel point. Si on considère $x_0 \in H_0$ (i.e. $v_S(x_0) = 0$), on a :

$$\frac{v_S(x_0+h) - v_S(x_0)}{h} \to \begin{cases} +\infty & \text{si } h \to 0^+ \\ -\infty & \text{si } h \to 0^- \end{cases}.$$

Ainsi, on a une demi-tangente parallèle à l'axe des y en x_0. Autrement dit, la courbe représentative de v_S présente un point de rebroussement de première espèce au point $(x_0, 0)$ (figure 2.10).

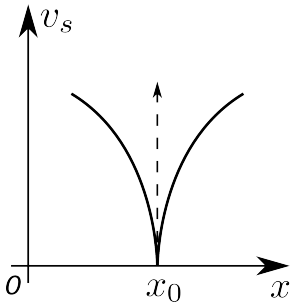


Figure 2.10 - Point de rebroussement.

Exemple 2.4.

$F(t) = (t-1)^3 + 1$ où $t \in [0, 2]$ (figure 2.11).
On prend ici $t_0 = 1$ et donc $x_0 = F(t_0) = 1$.
Les hypothèses (H_1) du théorème 2.5 sont satisfaites : F est de classe $\mathcal{C}^2$ (même $\mathcal{C}^\infty$), strictement croissante, $F'(1) = 0$, $F'''(1)$ existe $(F'''(t) = 6)$ avec $F'(1) \neq 0$.

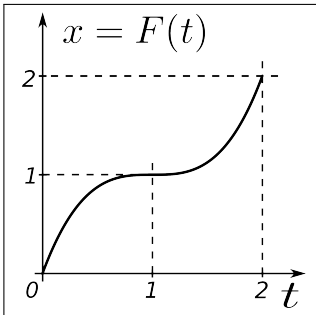


Figure 2.11 - Exemple de fonction F vérifiant les hypothèses $(H1)$.

On a $F'(t) = 3(t-1)^2$ et
$F^{-1}(x) = (x-1)^{\frac{1}{3}} + 1$, d'où
$v_S(x) = (F' \circ F^{-1})(x) = 3(x-1)^{\frac{2}{3}}$ non dérivable en 1 (figure 2.12). De plus, si on pose $x = 1 + h$, on a :
$\frac{v_S(1+h)-v_S(1)}{h} = 3\frac{h^{\frac{2}{3}}}{h}$
$= 3h^{-\frac{1}{3}} \to \begin{cases} +\infty & \text{si } h \to 0^+ \\ -\infty & \text{si } h \to 0^- \end{cases}$.
Donc on obtient bien un point de rebroussement de première espèce au point $(1, 0)$.

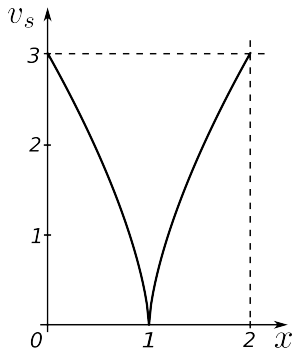


FIGURE 2.12 - Profil de vitesse associé.

Exemple 2.5.
Extension de l'exemple 2.4 dans le cas où F' est nulle sur un intervalle $[t_0, t_1]$ $(t_0 \neq t_1)$.

Soit $F : [0, 3] \longrightarrow \mathbb{R}$ illustrée à la figure 2.13 et définie par :
$F(t) = \begin{cases} (t-1)^3 + 1 & \text{si } t \leq 1 \\ 1 & \text{si } 1 \leq t \leq 2 \\ (t-2)^3 + 1 & \text{si } t \geq 2 \end{cases}$.
La fonction G définie sur $[0, 2]$ par :
$G(t) = \begin{cases} F(t) & \text{si } t \leq 1 \\ F(t+1) & \text{si } t \geq 1 \end{cases}$
est la fonction étudiée à l'exemple 2.4.

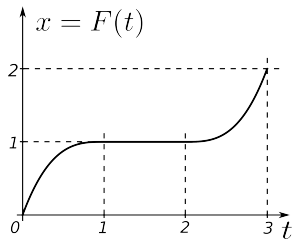


FIGURE 2.13 - Exemple de fonction F vérifiant les hypothèses $(H2)$.

Étudions la dérivabilité de $v_S = F' \circ F^{-1}$ sur $[0, 3]$.
On commence par montrer que F est dérivable sur $[0, 3]$:

$$F'(t) = \begin{cases} 3(t-1)^2 & \text{si } t < 1 \\ 0 & \text{si } 1 < t < 2 \\ 3(t-2)^2 & \text{si } t > 2 \end{cases} .$$

Pour $t = 1$, les dérivées à gauche et à droite sont $F'_g(1) = F'_d(1) = 0$. Donc F est dérivable en 1 avec $F'(1) = 0$. De même, on montre que F est dérivable en 2 avec $F'(2) = 0$. Ainsi, F est dérivable sur $[0, 3]$ avec

$$F'(t) = \begin{cases} 3(t-1)^2 & \text{si } t \leq 1 \\ 0 & \text{si } 1 \leq t \leq 2 \\ 3(t-2)^2 & \text{si } t \geq 2 \end{cases} .$$

On montre également que F'' existe sur $[0, 3]$ avec :

$$F''(t) = \begin{cases} 6(t-1) & \text{si } t \leq 1 \\ 0 & \text{si } 1 \leq t \leq 2 \\ 6(t-2) & \text{si } t \geq 2 \end{cases},$$

et que F'' est continue. Donc F est de classe $\mathcal{C}^2$. L'étude de F''' donne :

$$F'''(t) = \begin{cases} 6 & \text{si } t < 1 \\ 0 & \text{si } 1 < t < 2 \\ 6 & \text{si } t > 2 \end{cases},$$

avec $F'''_g(1) = 6 \neq F'''_d(1) = 0$ et $F'''_g(2) = 6 \neq F'''_d(2) = 0$. Donc en particulier, F''' n'est pas définie pour $t = 1$.
Enfin, on étudie la dérivabilité de $v_S = F' \circ F^{-1}$.
En utilisant l'expression de F', on en déduit l'expression de $v_S = F' \circ F^{-1}$:

$$v_S = \begin{cases} 3(x-1)^{\frac{2}{3}} & \text{si } x \leq 1 \\ 3(x-1)^{\frac{2}{3}} & \text{si } x > 1 \end{cases}.$$

On retrouve donc la fonction v_S obtenue à l'exemple 2.4. Donc v_S n'est pas dérivable en $x_0 = 1$.

2.2.3 Somme de deux profils spatiaux de vitesse

Proposition 2.2. *Soit v_1 et v_2 deux profils spatiaux de vitesse appartenant à $\mathcal{E}_{SSP}$ et définis sur $[0, x_f]$. Soit $F_1 : [0, T_1] \longrightarrow [0, x_f]$ la fonction associée à v_1 telle que $v_1(x) = F_1' \circ F_1^{-1}(x)$ et vérifiant les hypothèses (H_1) ou (H_2) du théorème 2.5. Soit $x_0 \in [0, x_f]$ tel que $v_1(x_0) = 0$ et $v_2(x_0) > 0$. Alors la somme $v_1 + v_2$ n'est pas un profil spatial de vitesse.*

Démonstration. Soit v_1 et v_2 appartenant à $\mathcal{E}_{SSP}$ et définis sur $[0, x_f]$. Par definition, il existe $T_1, T_2 \in \mathbb{R}^+$ et il existe deux fonctions $F_1 : [0, T_1] \longrightarrow [0, x_f]$ et $F_2 : [0, T_2] \longrightarrow [0, x_f]$ de classe $\mathcal{C}^2$, croissantes et nulles en zéro telles que $v_1(x) = F_1' \circ F_1^{-1}(x)$ et $v_2(x) = F_2' \circ F_2^{-1}(x)$. On suppose que F_1 vérifie les hypothèses (H_1) ou (H_2) du théorème 2.5, et on pose $x_0 \in [0, x_f]$ tel que $v_1(x_0) = 0$ et $v_2(x_0) > 0$. Alors, d'après les théorèmes 2.4 et 2.5, v_1 n'est pas dérivable en x_0 et v_2 est dérivable en x_0. Si on étudie les taux d'accroissement respectifs de v_1 et v_2 en x_0, on a :

$$\frac{v_1(x_0+h) - v_1(x_0)}{h} \to \begin{cases} +\infty & \text{si } h \to 0^+ \\ -\infty & \text{si } h \to 0^- \end{cases},$$

car la courbe représentative de v_1 présente un point de rebroussement de première espèce au point $(x_0, 0)$, et

$$\frac{v_2(x_0+h) - v_2(x_0)}{h} \to \ell \in \mathbb{R} \text{ quand } \ h \to 0$$

par définition de la dérivabilité. Ainsi, si on étudie le taux d'accroissement de la somme $v_1 + v_2$ en x_0, on obtient :

$$\begin{array}{rcl} \dfrac{(v_1+v_2)(x_0+h)-(v_1+v_2)(x_0)}{h} & = & \\ \dfrac{v_1(x_0+h)-v_1(x_0)}{h} + \dfrac{v_2(x_0+h)-v_2(x_0)}{h} & \to & \begin{cases} +\infty & \text{si } h \to 0^+ \\ -\infty & \text{si } h \to 0^- \end{cases} . \end{array}$$

Donc la courbe représentative de la fonction $v_1 + v_2$ présente un point de rebroussement de première espèce au point $(x_0, 0)$. Autrement dit, la fonction $v_1 + v_2$ n'est pas dérivable en x_0. Or $(v_1 + v_2)(x_0) > 0$ et on a montré que les profils spatiaux de vitesse étaient dérivables sur $H_+ = \{x \in [0, x_f], v_S(x) > 0\}$ (th. 2.4). On en déduit donc que la somme $v_1 + v_2$ n'est pas un profil spatial de vitesse. □

Cette proposition montre que la somme de deux profils spatiaux de vitesse n'est pas toujours un profil spatial de vitesse. Ainsi, faire la moyenne arithmétique de plusieurs profils spatiaux de vitesse n'a de sens que dans deux cas :

- soit lorsque tous les profils sont strictement positifs (pas d'arrêts),
- soit lorsque tous les profils passent par zéro aux mêmes points (i.e. tous les véhicules s'arrêtent exactement aux mêmes endroits).

La gestion des arrêts est donc un problème qu'il faudra prendre en compte lors de la construction du profil de référence. En effet, si on choisit le profil moyen (correspondant à la moyenne arithmétique de tous les profils) comme profil de référence, il faut soit exclure les arrêts, ce qui paraît être une solution irréaliste, soit recaler les profils afin qu'ils s'annulent tous aux mêmes points. Cette deuxième solution nécessite en fait deux étapes : dans un premier temps, classer les profils en groupes distinguant, pour chaque arrêt, ceux qui se sont arrêtés et ceux qui ne se sont pas arrêtés ; puis dans un deuxième temps, recaler les profils afin que les arrêts soient situés aux mêmes positions. Ce problème de recalage des profils de vitesse sera traité au chapitre 5.

2.3 Conclusion

Nous avons vu dans ce chapitre que l'analyse des données fonctionnelles avait connu un véritable essor depuis une vingtaine d'années et était utilisé dans des domaines très divers (chimiométrie, climatologie...). Cependant, à notre connaissance, l'analyse des données fonctionnelles est très peu utilisée dans le domaine routier. On citera essentiellement les travaux de Maza [108] et Allain [4], réalisés dans le cadre de thèses CIFRE entre l'Institut de Mathématiques de Toulouse et la société Médiamobile, qui représentent les profils temporels de vitesse comme des fonctions du temps dans un objectif d'analyse et de prévision du trafic routier à court terme. Allain [4] propose notamment un modèle linéaire fonctionnel des vitesse observées (page 37). On peut également citer les travaux de Besse et Cardot [19] consacrés à l'approximation spline avec contrainte de rang, sur un intervalle de temps, d'un processus

à temps continu, et appliqués à la prévision du trafic autoroutier. Enfin, on citera dans le domaine aérien, les travaux de Tastambekov *et al.* [162] qui s'intéressent à la prédiction de la trajectoire d'un avion et qui utilisent un modèle fonctionnel de régression linéaire avec une décomposition dans une base d'ondelettes.

Pourtant, nous avons montré dans ce chapitre l'importance de se placer dans un cadre fonctionnel pour l'étude des profils spatiaux de vitesse. En effet, nous avons vu les limites de la statistique multivariée dans le cas de données discrètes issues d'un processus sous-jacent continu et l'intérêt de prendre en compte le caractère fonctionnel de ces données (dérivées, régularité, contraintes de forme...). De plus, nous avons vu que, d'une part, la plupart des méthodes classiques de statistique multivariée ont été adaptées aux données fonctionnelles (analyses factorielles, modèles linéaires...), et que d'autre part, ces données ont permis l'apparition de nouvelles méthodes tenant compte de leurs caractéristiques fonctionnelles (recalage de courbes, modèles fonctionnels définis par des équations différentielles...). Notons que les problèmes de calcul des dérivées et de recalage de profils de vitesse ont été évoqués dans le chapitre précédent et nous avons vu à ce sujet, les limites de méthodes actuellement utilisées dans le domaine routier. Ainsi, l'approche fonctionnelle permet de résoudre ces problèmes et est donc bien adaptée à l'étude des profils de vitesse.

En particulier, l'analyse des données fonctionnelles est particulièrement bien adaptée aux profils spatiaux de vitesse puisqu'elle permet de tenir compte de la richesse des données issues de véhicules traceurs et de préserver la cohérence physique entre vitesse et position (et implicitement temps). Nous avons donc choisi dans cette thèse de nous placer dans un cadre fonctionnel et de traiter les profils spatiaux de vitesse comme des fonctions. Dans ce chapitre, nous avons proposé une modélisation fonctionnelle des profils spatiaux de vitesse. Nous avons défini les fonctions pouvant être la représentation d'un profil spatial de vitesse et nous avons étudié certaines propriétés de ces fonctions. En particulier, nous avons montré que les profils spatiaux de vitesse n'étaient pas dérivables aux points où la vitesse s'annule, ce qui pose le problème de la gestion des arrêts. Cette propriété devra être prise en compte dans l'étape de lissage permettant de convertir les données en courbes et dans la construction du profil de vitesse de référence que nous aborderons dans les chapitres suivants.

Chapitre 3

Méthodes de lissage : l'approche par splines

L'objet de ce chapitre est de présenter les principales méthodes de lissage et plus particulièrement les méthodes de lissage utilisant les splines. En effet, nous avons vu au chapitre 2 l'intérêt de traiter les profils spatiaux de vitesse comme des fonctions plutôt que des vecteurs de $\mathbb{R}^n$. Nous avons également vu que la première étape d'une analyse de données fonctionnelles consistait à convertir les données brutes de nature vectorielle en objet fonctionnel. Les mesures de vitesse et de position issues de capteurs étant bruitées, cette étape nécessite l'utilisation d'une méthode de lissage. On se ramène donc à un problème de régression non paramétrique où l'on cherche à estimer la fonction de régression.

Dans une première section, nous rappelons quelques généralités sur la régression non paramétrique et nous présentons les méthodes de lissage généralement utilisées pour le traitement des profils spatiaux de vitesse, à savoir la méthode du noyau et la méthode par polynômes locaux.

Puis nous présentons en détails les méthodes de lissage utilisant les fonctions splines en distinguant les splines de régression, les splines de lissage et les splines pénalisées. Nous montrons notamment sur un jeu de données réelles, l'efficacité des splines de lissage par rapport à la méthode du noyau et à la méthode des polynômes locaux, pour l'estimation des profils temporels de vitesse et d'accélération.

Enfin, nous terminons ce chapitre par une généralisation de la notion de splines dans le contexte plus large des méthodes de régularisation où les splines sont définies comme la solution d'un problème variationnel dans un espace de Hilbert à noyau reproduisant.

3.1 Méthodes classiques de lissage

3.1.1 Généralités sur la régression non paramétrique

La régression non paramétrique permet de décrire la relation entre une variable aléatoire dépendante Y et une variable aléatoire explicative X, sans supposer de forme particulière sur la relation entre ces deux variables. Considérons un échantillon aléatoire (x_i, y_i), $i = 1, \ldots, n$, du couple (X, Y). Alors le modèle de régression non paramétrique est le suivant :

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.1}$$

où les ε_i sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ^2, et où $f(x)$ est la fonction de régression que l'on cherche à estimer. Sous la condition que Y soit intégrable (i.e. $\mathbb{E}[|Y|] < \infty$), la fonction de régression $f(x)$ est définie comme la moyenne conditionnelle de Y sachant $X = x$, soit $f(x) = \mathbb{E}[Y|X = x]$. Le cas général où à la fois X et Y sont aléatoires est dit de dispositif expérimental à effets aléatoires (ou "*random design*"). Cependant dans certaines applications, X n'est pas toujours aléatoire et dans ce cas on parle de dispositif expérimental à effets fixes (ou "*fixed design*"). Un exemple classique de dispositif expérimental à effets fixes est le cas où les x_i sont une suite de points uniformément répartis $x_i = i/n,\ i = 0, \ldots, n$ sur $[0, 1]$.

Contrairement à la régression paramétrique qui suppose de connaître une forme explicite de la fonction de régression f (linéaire, logarithmique, ...), la régression non paramétrique est beaucoup plus flexible puisqu'elle ne nécessite aucune hypothèse a priori sur la forme de f. De plus, la régression non paramétrique permet de modéliser le lien "fonctionnel" entre X et Y avec certaines conditions de régularité (continuité, monotonie) sur la fonction f. Le problème consiste alors à estimer cette fonction f, a priori inconnue, et non plus uniquement les paramètres de cette fonction comme c'est le cas en régression paramétrique. Les estimateurs de f sont généralement appelés fonctions de lissage.

La flexibilité d'une fonction de lissage est généralement contrôlée par un (ou plusieurs) paramètre de lissage dépendant de la méthode de lissage choisie. Toute la difficulté de ces méthodes consiste à trouver le meilleur compromis entre lissage et flexibilité (allant de la régression paramétrique à l'interpolation simple) et couramment appelé *dualité biais-variance* (Hastie et Tibshirani [75]). La figure 3.1 illustre cette dualité correspondant à une minimisation de l'erreur quadratique moyenne (EQM, ou MSE pour "Mean Square Error" en anglais) définie par la relation suivante "$EQM = biais^2 + variance$".

En effet, augmenter la flexibilité permet de suivre plus fidèlement les données et donc de diminuer le biais mais augmente la variance (sous-lissage). Au contraire, choisir une courbe plus lisse augmente le biais mais diminue la variance (surlissage). La sélection du paramètre de lissage est donc un problème difficile que nous évoquerons pour chaque estimateur que nous étudierons dans la suite de ce chapitre.

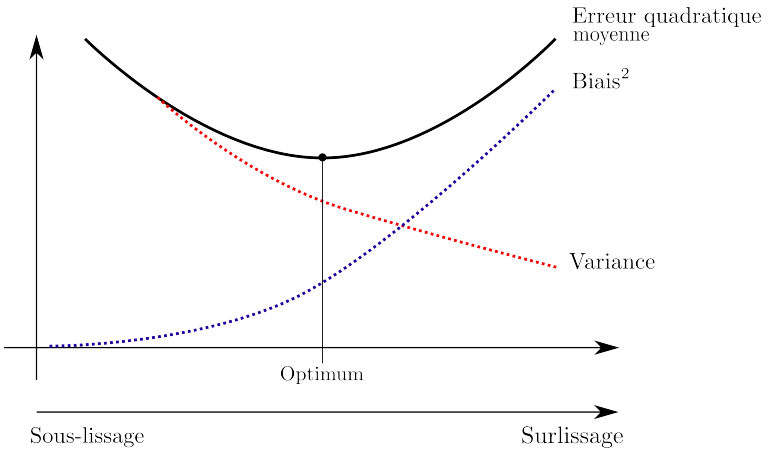


FIGURE 3.1 - Dualité biais-variance.

Une classe particulière des méthodes de lissage est celle des méthodes de lissage linéaires. Dans ce cas, la fonction de lissage s'écrit comme une combinaison linéaire des observations, i.e. pour tout x, il existe un vecteur $S(x) = (S_1(x), \ldots, S_n(x))^T$ tel que $\widehat{f}(x) = \sum_{k=1}^{n} S_k(x) y_k$ où $\widehat{f}$ désigne un estimateur de f. Si on note $\widehat{y} = (\widehat{f}(x_1), \ldots, \widehat{f}(x_n))^T$ le vecteur des valeurs estimées à partir du vecteur des observations $y = (y_1, \ldots, y_n)^T$, alors on peut écrire :

$$\widehat{y} = Sy, \tag{3.2}$$

où $S = \{S_{ik}\}_{i,k=1}^{n}$ avec $S_{ik} = S_k(x_i)$. La matrice S est appelée matrice de lissage, ou matrice chapeau ("*hat matrix*"), et le degré de liberté de la fonction de lissage est définie par $df = trace(S)$. La majorité des estimateurs et en particulier les estimateurs que nous étudierons par la suite, c'est-à-dire les estimateurs à noyau et les estimateurs splines, sont des estimateurs linéaires.

3.1.2 Estimateur à noyau

L'estimateur à noyau est une extension de l'estimateur par moyenne mobile. Le principe est d'estimer la fonction de régression $f(x)$ en chaque point x du domaine en calculant une moyenne des observations, pondérée par une fonction de distance appelée noyau. Ainsi plus les points x_i sont proches de x, plus le poids donné aux observations y_i associées sera important.

L'estimateur à noyau le plus connu est l'estimateur de Nadaraya-Watson défini de la manière suivante : pour tout x,

$$\widehat{f}(x) = \sum_{i=1}^{n} w_i(x) y_i \quad \text{avec} \quad w_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})}, \tag{3.3}$$

où $K(.)$ est une fonction noyau et $h > 0$ est un paramètre de lissage ("*bandwidth parameter*").

La fonction noyau K, qui correspond généralement à une densité de probabilité, est une fonction continue, bornée, non négative, symétrique telle que :

$$\int K(t)dt = 1, \quad \int tK(t)dt = 0, \quad \int t^2K(t)dt < \infty. \tag{3.4}$$

Les noyaux les plus connus sont :

- le noyau gaussien défini par $K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$,
- le noyau uniforme défini par $K(t) = \frac{1}{2}\mathbb{1}_{|t|\leq 1}$,
- le noyau d'Epanechnikov défini par $K(t) = \frac{3}{4}(1-t^2)\mathbb{1}_{|t|\leq 1}$.

D'autres noyaux sont définis notamment dans Wand et Jones [178], cependant le choix du noyau est peu important comparativement au choix du paramètre de lissage (Hastie et Tibshirani [75], Härdle [79]).

Le choix du paramètre de lissage h détermine le degré de lissage et résulte d'un compromis biais-variance : plus h est grand, plus la courbe est lisse (faible variance mais biais élevé) et vice-versa. Il existe différentes méthodes de sélection automatique du paramètre de lissage telles que la validation croisée ("*cross-validation*") ou la validation croisée généralisée ("*generalized cross-validation*") (Schimek [143] section 4.3, Wand et Jones [178] chapitre 3).

Si l'estimateur à noyau présente l'avantage d'être relativement simple et facile à comprendre, cet estimateur présente certains inconvénients, notamment des problèmes de biais aux extrémités de l'intervalle étudié (Simonoff [157], Härdle [79]). En effet, le biais est plus élevé aux extrémités où il est d'ordre $O(h)$ alors qu'il est d'ordre $O(h^2)$ à l'intérieur de l'intervalle. On note également un aplatissement des maxima et minima locaux. Enfin, le fait que la fenêtre de lissage soit fixe implique un manque de variation locale du lissage.

Pour une étude complète des propriétés de l'estimateur à noyau, nous renvoyons le lecteur vers Wand et Jones [178] et Härdle [79].

3.1.3 Estimateur par polynômes locaux

La méthode par polynômes locaux généralise la méthode à noyau et consiste à approcher la fonction de régression f du modèle général (3.1), autour d'un point x, par un polynôme de degré p. Ainsi, si la fonction de régression f est p fois dérivable dans un voisinage de x, alors par le développement de Taylor, on obtient :

$$f(x') \approx \sum_{j=0}^{p} \frac{f^{(j)}(x)}{j!}(x'-x)^j = \sum_{j=0}^{p} \beta_j (x'-x)^j.$$

On se ramène alors à un problème de régression polynomiale locale dans un voisinage de x. La fonction de régression est estimée en chaque point en ajustant localement un polynôme de degré p par moindres carrés pondérés. Les coefficients $\beta = (\beta_0, \ldots, \beta_p)^T$ du polynôme local au point x sont estimés par minimisation du critère des moindres

carrés pondérés :

$$\min_{\beta} \sum_{i=1}^{n} K(\frac{x_i - x}{h})(y_i - \sum_{j=0}^{p} \beta_j (x_i - x)^j)^2,$$

où $K(.)$ est une fonction de poids (fonction noyau) et h est le paramètre définissant la largeur de la fenêtre de lissage (paramètre de lissage). Lorsque $p = 0$, on retrouve l'estimateur à noyau présenté à la section précédente. Généralement, on utilise des polynômes de degré 1 (régression localement linéaire) ou 2 (régression localement quadratique) car l'utilisation de polynômes de haut degré a tendance à surestimer les données et à rendre les calculs instables. Un cas particulier de régression localement linéaire ($p = 1$) consiste à utiliser la fonction tricube comme fonction de poids, et à choisir h de manière à obtenir exactement k observations (les k plus proches voisins) pour l'estimation au point x. Cette méthode proposée initialement par Cleveland [28] puis développée par Cleveland et Devlin [29] est appelée méthode loess ou lowess ("LOcally WEighted Scatterplot Smoothing").

En plus de sa simplicité et de sa flexibilité, l'estimateur par polynômes locaux présente un biais plus faible que l'estimateur à noyau, notamment aux extrémités. En effet, le biais de l'estimateur par polynômes locaux aux bornes est d'ordre $O(h^{p+1})$ alors qu'il est d'ordre $O(h)$ pour l'estimateur à noyau (Simonoff [157]). Cependant, l'estimateur par polynômes locaux peut être fortement influencé par la présence de valeurs aberrantes, même si Cleveland [28] présente une version robuste de l'estimateur loess permettant de réduire ce problème. De plus, la méthode des polynômes locaux ne permet pas d'obtenir une expression explicite de la fonction de régression. Plus de détails sur cette méthode sont donnés dans Fan et Gijbels [55] et Simonoff [157] (section 5.2).

3.2 Estimateurs splines

Si les méthodes locales de lissage tels que la méthode du noyau et les polynômes locaux sont simples à comprendre et présentent de bonnes caractéristiques numériques, nous avons vu que ces méthodes étaient instables aux bornes de l'intervalle, même si la méthode des polynômes locaux est meilleure que la méthode du noyau sur ce point. De plus, même si le choix de la largeur de la fenêtre de lissage permet de s'adapter aux variations locales des données, on observe généralement une surestimation ou sous-estimation des données dans les zones où le nombre de données est faible, ou lorsque la répartition des données est asymétrique autour du point que l'on cherche à estimer.

Les méthodes de lissage utilisant les fonctions splines permettent de pallier ces difficultés, et contrairement à la méthode des polynômes locaux où l'on effectue l'estimation en un point x donné, ces méthodes présentent l'avantage de garder le caractère fonctionnel des données et de fournir une expression explicite de la fonction de lissage.

Historiquement, le terme "fonctions spline" a été introduit dans la littérature mathématique par Schoenberg [147] (1946) même si l'idée originale est attribuée à Whittaker [185]. A l'origine, le mot anglais "spline" désigne une latte flexible utilisée par les dessinateurs pour matérialiser des lignes à courbure variable et passant par des points fixés a priori, le tracé ainsi réalisé minimisant l'énergie de déformation de la latte (Besse et Thomas-Agnan [18]). La théorie des splines s'est essentiellement développée au début des années 1960 avec notamment les travaux de Schoenberg [148] et Reinsch [135], qui introduisent les splines de lissage comme classe d'estimateurs. L'utilisation des splines en statistiques comme méthode de régression non paramétrique est attribuée aux travaux de Wahba (Wahba [172], Kimeldorf et Wahba [86], Craven et Wahba [33]). L'utilisation des splines s'est énormément développée dans de nombreux domaines des mathématiques tels que la théorie de l'approximation (Laurent [92]), l'analyse numérique, les équations intégrales et différentielles partielles, et les statistiques. La bibliographie sur le sujet est d'ailleurs considérable. Nous nous référons entre autres à l'article de Wegman et Wright [184] pour une revue des applications des splines en statistiques, à l'article en français de Besse et Thomas-Agnan [18] pour une utilisation des splines en régression non paramétrique, et aux ouvrages complets de De Boor [37], Schumaker [149], Eubank [53] et Wahba [173].

3.2.1 Splines de régression

Les splines de régression représentent un compromis entre la régression polynomiale globale et les méthodes de lissage locales présentées à la section 3.1. Le principe consiste à représenter la fonction de lissage f du modèle (3.1) par un polynôme par morceaux se raccordant de façon lisse en des points appelés noeuds.

En effet, une méthode simple de régression est la régression polynomiale qui consiste à représenter la fonction de régression par un polynôme de degré p. Cependant, si les polynômes sont des fonctions lisses et faciles à calculer, cette méthode manque de flexibilité et la fonction de régression à tendance à beaucoup osciller dès que p est grand. Une solution alternative consiste à représenter la fonction de régression par un polynôme par morceaux, mais si cette méthode permet de gagner en flexibilité, on perd en régularité puisque des points de discontinuité peuvent apparaître aux points de jonction (Fox [60], Schumaker [149]).

3.2.1.1 Splines polynomiales

La méthode des splines de régression consiste à représenter la fonction de lissage f du modèle (3.1) par un polynôme par morceaux d'ordre m (i.e. de degré $m-1$) et à imposer l'égalité des dérivées jusqu'à l'ordre $m-2$ en chaque point de jonction afin d'obtenir un certain niveau de régularité. Autrement dit, on considère que la fonction de régression f définie sur l'intervalle $[a,b]$ appartient à l'espace suivant :

$$\mathcal{S}_m(\xi) = \mathcal{PP}_m(\xi) \cap \mathcal{C}^{m-2}[a,b], \tag{3.5}$$

où $\xi = (\xi_1, \ldots, \xi_K)$ avec $\xi_1 < \ldots < \xi_K$ est une suite de points ordonnés correspondant aux points de jonction et appelés noeuds ("*knots*"), $\mathcal{PP}_m(\xi)$ est l'espace des polynômes par morceaux d'ordre m ayant pour séquence de noeuds ξ, et $\mathcal{C}^{m-2}[a,b] = \{f : f^{(j)}$ existe et est continue sur $[a,b],\ j = 1, \ldots, m-2\}$ (Schumaker [149]). L'espace $\mathcal{S}_m(\xi)$ dénote l'espace des fonctions splines polynomiales d'ordre m avec noeuds simples $\xi = (\xi_1, \ldots, \xi_K)$. La dimension de $\mathcal{S}_m(\xi)$ est égale à $m + K$.

Notons qu'il est possible de contrôler la régularité de la fonction de lissage en faisant coïncider plusieurs noeuds. Ces noeuds sont dits multiples et l'ordre de continuité en un noeud diminue avec la multiplicité du noeud (Ramsay et Silverman [130] section 3.5.1, Schumaker [149] def. 4.1). Par exemple, si $m = 3$, la fonction de régression est dérivable dans le cas d'un noeud simple, si le noeud est double on obtient un point anguleux, et si le noeud est triple on obtient un point de discontinuité. Cependant dans la plupart des cas, les noeuds sont distincts et on considérera par la suite uniquement des noeuds de multiplicité 1 (noeuds simples).

Une spline polynomiale est alors définie de la manière suivante :

Définition 3.1. *Pour un intervalle $[a,b]$ et une suite de K points $\xi_1 < \ldots < \xi_K$ dans $[a,b]$, on appelle spline polynomiale d'ordre m (m étant un entier ≥ 1) ayant pour noeuds simples les points $\xi_1, \ldots, \xi_K$, toute fonction f de $[a,b]$ dans $\mathbb{R}$ telle que :*

(i) f est continûment dérivable jusqu'à l'ordre $m-2$ (si $m \geq 2$),

(ii) sur chaque sous-intervalle $[a, \xi_1], \ldots, [\xi_i, \xi_{i+1}], \ldots, [\xi_K, b]$, f coïncide avec un polynôme de degré $m-1$.

Les splines les plus fréquemment utilisées sont les splines d'ordre 4 dites splines cubiques.

Enfin, citons également un cas particulier important des splines polynomiales : les splines naturelles. Si $m = 2r$ dans la définition 3.1, alors on appelle spline polynomiale naturelle d'ordre $2r$, une fonction f qui, en plus de satisfaire les conditions (i) et (ii), satisfait également la condition :

$$(iii)\ f^{(j)}(a) = f^{(j)}(b) = 0, \quad j = r, \ldots, 2r-1.$$

La condition (iii) implique que f coïncide avec un polynôme de degré $r-1$ en dehors de l'intervalle $[\xi_1, \xi_K]$. Par exemple, une spline cubique naturelle est une spline cubique linéaire sur $[a, \xi_1]$ et sur $[\xi_K, b]$. L'espace des splines polynomiales naturelles d'ordre $2r$ ayant pour noeuds $\xi_1, \ldots, \xi_K$ est noté $\mathcal{NS}_{2r}(\xi_1, \ldots, \xi_K)$ et sa dimension est égale à K.

3.2.1.2 Représentation dans une base appropriée

L'espace des splines polynomiales $\mathcal{S}_m(\xi_1, \ldots, \xi_K)$ est un sous-espace vectoriel de l'espace $\mathcal{C}^{m-2}[a,b]$, de dimension finie $m+K$. On peut alors déterminer une base de

fonctions appropriée pour représenter les fonctions splines. L'une des bases classiquement utilisée pour représenter une spline d'ordre m de noeuds simples $(\xi_1, \ldots, \xi_K)$ est la base des puissances tronquées ("*truncated power basis*") définie par :

$$\begin{cases} N_j(x) = x^{j-1}, & j = 1, \ldots, m \\ N_{m+j}(x) = (x - \xi_j)_+^{m-1}, & j = 1, \ldots, K \end{cases},$$

où $(x)_+ = max\{x; 0\}$ est la fonction partie positive. Une fonction spline $S \in \mathcal{S}_m(\xi_1, \ldots, \xi_K)$ est alors représentée dans la base des puissances tronquées par :

$$S(x) = \sum_{j=1}^{m+K} \beta_j N_j(x).$$

Cependant, si cette base présente certains avantages d'un point de vue théorique, elle n'est pas adaptée au calcul (problème d'instabilité numérique) (Eubank [53]).

Une base plus adaptée est celle des B-splines en raison de leurs propriétés numériques et de leur facilité de mise en oeuvre algorithmique. Nous commençons par ajouter des noeuds à la séquence initiale $(\xi_1, \ldots, \xi_K)$. Soient ξ_0 et ξ_{K+1} les noeuds correspondants aux bornes de l'intervalle $[a, b]$ sur lequel est définie la spline que l'on cherche à estimer, i.e. $\xi_0 = a < \xi_1$ et $\xi_K < \xi_{K+1} = b$. On définit la suite augmentée τ de noeuds telle que :

$$\begin{cases} \tau_1 \leq \tau_2 \leq \cdots \leq \tau_m \leq \xi_0 \\ \tau_{j+m} = \xi_j, & j = 1, \ldots, K \\ \xi_{K+1} \leq \tau_{K+m+1} \leq \tau_{K+m+2} \leq \cdots \leq \tau_{K+2m} \end{cases}.$$

Généralement, on choisit $\tau_1 = \cdots = \tau_m = \xi_0$ et $\xi_{K+1} = \tau_{K+m+1} = \cdots = \tau_{K+2m}$ (noeuds de multiplicité m), ce qui signifie que la spline est discontinue aux bornes de l'intervalle, et donc non définie en dehors des bornes. Les B-splines d'ordre l de noeuds τ sont alors définies récursivement par :

$$B_{i,l}(x) = \frac{x - \tau_i}{\tau_{i+l-1} - \tau_i} B_{i,l-1}(x) + \frac{\tau_{i+l} - x}{\tau_{i+l} - \tau_{i+1}} B_{i+1,l-1}(x), \quad i = 1, \ldots, K + 2m - l,$$

avec pour initialisation

$$B_{i,1}(x) = \begin{cases} 1 \text{ si } \tau_i \leq x < \tau_{i+1} \\ 0 \text{ sinon} \end{cases}.$$

Par convention, lorsque les noeuds sont confondus (noeuds multiples), on pose $0/0 = 0$. Par exemple, pour $m = 4$, $B_{i,4}$, $i = 1, \ldots, K + 4$ sont les $K + 4$ fonctions B-splines d'ordre 4 pour la séquence de noeuds $(\xi_1, \ldots, \xi_K)$. La figure 3.2 représente 13 fonctions B-splines d'ordre 4 avec 9 noeuds intérieurs uniformément répartis sur l'intervalle $[0, 1]$. Les B-splines sont des fonctions positives adaptées au calcul en raison de leur support compact. En effet, la B-spline $B_{i,l}$ d'ordre l est nulle en dehors de

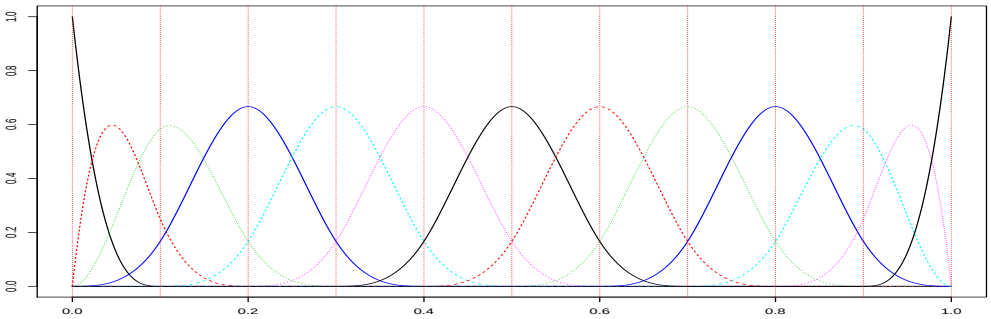


FIGURE 3.2 - Séquence de B-splines d'ordre 4 avec 9 noeuds intérieurs uniformément répartis sur l'intervalle $[0, 1]$.

l'intervalle $[\tau_i, \tau_{i+l}]$, ce qui implique que la matrice correspondante est une matrice bande. Par conséquent, les temps de calcul sont beaucoup plus faibles qu'avec la base des puissances tronquées surtout lorsque K est grand. En utilisant la base des B-splines, la représentation d'une fonction spline $S \in \mathcal{S}_m(\xi_1, \ldots, \xi_K)$ est

$$S(x) = \sum_{j=1}^{m+K} \beta_j B_{j,m}(x).$$

Pour plus de détails sur les propriétés des B-splines et leur implémentation, le lecteur pourra consulter le livre de De Boor [37] qui est une référence sur ce sujet, ainsi que les ouvrages de Eubank [53] (section 6.4), Green et Silverman [69] (section 3.6), et Hastie *et al.* [76] (appendice du chap. 5).

Enfin, il existe d'autre bases telles que la base des splines naturelles ou la base des fonctions radiales (Ruppert *et al.* [140]).

3.2.1.3 Estimation par moindres carrés

La méthode des splines de régression consiste à faire l'hypothèse que la fonction de lissage est une spline. Cette méthode tend ainsi vers les méthodes de régression paramétrique puisque la forme de la fonction de lissage est connue à l'exception des valeurs de ces paramètres (ordre de la spline, nombre et positions des noeuds). Dans, un premier temps, on considère que ces paramètres sont connus et on suppose que la fonction de lissage f du modèle (3.1) est une spline d'ordre m de noeuds $(\xi_1, \ldots, \xi_K)$, i.e. $f \in \mathcal{S}_m(\xi_1, \ldots, \xi_K)$. Considérons une base $(B_1, \ldots, B_{m+K})$ de l'ensemble $\mathcal{S}_m(\xi_1, \ldots, \xi_K)$. Alors le modèle (3.1) s'écrit :

$$y_i = \sum_{j=1}^{m+K} \beta_j B_j(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.6}$$

où le vecteur $\beta = (\beta_1, \ldots, \beta_{m+K})$ des coefficients est à estimer. L'estimateur de β est obtenu par minimisation des moindres carrés :

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{m+K} \beta_j B_j(x_i))^2. \tag{3.7}$$

Si on note B la matrice $n \times (m+K)$ des $B_j(x_i)$, alors l'estimateur spline de régression est linéaire et est donné, comme pour la régression linéaire, par

$$\widehat{\beta} = (B^T B)^{-1} B^T y. \tag{3.8}$$

Notons que si l'on utilise une base de B-splines, la matrice $B^T B$ est une matrice bande ayant $2m-1$ diagonales non nulles, ce qui facilite les calculs.

Les splines de régression, également appelées splines des moindres carrés, sont donc attractives en raison de leur facilité d'estimation lorsque les paramètres de la spline sont donnés. Cependant, si le choix de l'ordre de la spline ne pose généralement pas de difficultés (en général, on choisit $m = 4$ afin d'obtenir une spline cubique), le choix du nombre de noeuds et de leurs positions est un problème plus délicat.

3.2.1.4 Choix du nombre de noeuds et de leurs positions

Le choix du nombre de noeuds et de leurs positions est généralement un problème difficile à résoudre. Une méthode simple consiste à déterminer l'emplacement des noeuds en observant les données (Eubank [53] section 6.3). Le principe est de placer beaucoup de noeuds dans les zones où la courbe varie beaucoup afin que l'estimateur soit flexible dans ces zones. Cependant cette méthode heuristique n'est pas toujours simple à mettre en oeuvre. Une autre méthode généralement utilisée consiste à fixer le nombre de noeuds et à les placer de façon uniforme ou aux quantiles empirique de la variable X (Hastie et Tibshirani [75]). Il ne reste alors que le nombre de noeuds K comme paramètre de lissage, l'entier K variant dans l'intervalle $[0, n-m]$. Cependant, le choix du nombre de noeuds est difficile puisque, comme le choix de la largeur de la fenêtre de lissage pour les méthodes à noyau, celui-ci résulte d'un compromis biais-variance : un faible nombre de noeuds peut entraîner un phénomène de sous-lissage, alors qu'un grand nombre de noeuds peut entraîner un phénomène de sur-lissage. Enfin, certaines méthodes dites adaptatives consistent à choisir le nombre de noeuds et leur emplacement par minimisation d'un critère d'ajustement de modèle tel que le critère AIC, le critère de validation croisée généralisée (GCV), etc... (Ruppert *et al.* [140] section 3.4, Hastie et Tibshirani [75] chap.9). De nombreuses procédures automatiques de sélection des noeuds basées sur des procédures pas à pas ont été développées (voir par exemple Friedman et Silverman [62], Stone *et al.* [159], ou Wand [177] pour une comparaison de différentes approches), mais si leurs performances sont bonnes, elles sont généralement complexes et ont des temps de calcul relativement longs.

La méthode des splines de régression est donc attractive en raison de sa facilité d'estimation lorsque les noeuds sont donnés. Cependant le principal inconvénient de cette méthode est la difficulté du choix du nombre de noeuds et de leurs positions. De plus, cette méthode ne permet pas de contrôler le niveau de lissage de la courbe estimée à l'aide d'un unique paramètre de lissage.

3.2.2 Splines de lissage

Les splines de lissage sont une autre méthode de lissage utilisant les fonctions splines, mais contrairement aux splines de régression où l'on impose que la fonction de lissage soit une spline, l'estimateur par splines de lissage résulte d'un problème d'optimisation.

3.2.2.1 Principe de régularisation

On considère toujours le modèle de régression :

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{3.9}$$

où f est la fonction de lissage que l'on cherche à estimer. On suppose que les x_i appartiennent à l'intervalle borné $[a, b] \subset \mathbb{R}$.
Nous avons vu précédemment qu'un critère classique pour estimer f est la minimisation de la somme des carrés des résidus, appelé critère des moindres carrés, défini par

$$\min_{f \in \mathcal{E}} \sum_{i=1}^{n} (y_i - f(x_i))^2, \tag{3.10}$$

où $\mathcal{E}$ est un espace de fonctions. Si l'on n'impose aucune forme spécifique à f (ex : polynôme, spline), alors toute fonction f interpolant les données est solution de ce problème de minimisation. Une telle solution n'est pas satisfaisante car d'une part, elle n'est pas unique, et d'autre part elle ne permet pas d'obtenir une courbe lisse. D'après la définition d'un problème "bien posé" ("*well-posed problem*") au sens de Hadamard caractérisée par l'existence, l'unicité et la stabilité de la solution, on dit que le problème (3.10) est "mal posé" ("*ill-posed problem*"). Une solution permettant de se ramener à un problème bien posé consiste à rajouter de l'information sur f en choisissant un espace de recherche $\mathcal{E}$ plus restreint. Une première approche classique est de contrôler la dimension en décomposant f dans une base de fonctions (ex : polynômes pour la régression polynomiale, splines pour les splines de régression). Une seconde approche, appelée régularisation, consiste à ajouter au critère (3.10) un terme de pénalité afin d'imposer une régularité de la solution (Bertero [16], Tikhonov et Arsenin [163]). On cherche alors à minimiser le critère des moindres carrés pénalisés suivant :

$$\min_{f \in \mathcal{E}} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda J(f), \tag{3.11}$$

où $J(f)$ est un terme de pénalité et $\lambda > 0$ est le paramètre de lissage. Généralement, on utilise la pénalité $J(f) = \int_a^b (f^{(m)}(x))^2 dx$ où $f \in \mathcal{C}^m[a,b]$ comme mesure de courbure de la fonction de lissage, et plus particulièrement $J(f) = \int_a^b (f^{(2)}(x))^2 dx$, appelée pénalité de la rugosité ("*roughness penalty*", Green et Silverman [69]), qui traduit la notion de courbure moyenne d'une fonction $f \in \mathcal{C}^2[a,b]$. Le critère des moindres carrés pénalisés devient alors

$$\min_{f \in \mathcal{C}^m[a,b]} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx, \tag{3.12}$$

où le paramètre λ contrôle le compromis entre la fidélité aux données mesurée par le 1^{er} terme, et la régularité de la fonction mesurée par le $2^{\text{ème}}$ terme. Le paramètre λ permet ainsi de contrôler le degré de lissage désiré : plus λ est grand, plus la courbe est lisse, et inversement. En particulier, lorsque $\lambda \to 0$, on se ramène au critère classique (3.10) des moindres carrés et la solution tend vers une fonction interpolant les données. Au contraire, lorsque $\lambda \to \infty$, la solution est une régression polynomiale de degré $(m-1)$ annulant la pénalisation.

3.2.2.2 Existence et unicité des splines de lissage

Supposons que la fonction de lissage f du modèle (3.11) soit régulière, autrement dit que $f \in W^m[a,b]$ où $W^m[a,b]$ est l'espace de Sobolev d'ordre m défini par :

$$W^m[a,b] = \{f : f^{(j)} \text{ absolument continues } j = 0, \ldots, m-1;\ f^{(m)} \in \mathcal{L}_2[a,b]\}, \tag{3.13}$$

où $\mathcal{L}_2[a,b]$ est l'ensemble des fonctions de carré intégrable sur l'intervalle $[a,b]$ (Adams et Fournier [3]). Dans cet espace, on peut "mesurer" la régularité d'une fonction f par $J(f) = \int_a^b (f^{(m)}(x))^2 dx$. Alors, si $m \leq n$, l'unique fonction $\widehat{f}$ minimisant le critère

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)}(x))^2 dx \tag{3.14}$$

dans l'ensemble des fonctions de $W^m[a,b]$ est une spline polynomiale naturelle d'ordre $2m$ ayant pour noeuds les points d'observation $x_1, \ldots, x_n$ (voir démonstration dans Eubank [53] section 5.2). Un tel estimateur est appelé spline de lissage ("*smoothing spline*"). En particulier, lorsque $m = 2$, l'estimateur minimisant le critère (3.14) dans l'espace des fonctions de $W^2[a,b]$ est une spline cubique naturelle avec pour noeuds les points $x_1, \ldots, x_n$. Contrairement aux splines de régression où l'on supposait que f était une spline, ici la forme spline de f résulte d'un problème d'optimisation. De plus, avec les splines de lissage, le problème du choix de nombre de noeuds et de leur position ne se pose plus puisque les noeuds correspondent aux points d'observation. Pour une étude des propriétés asymptotiques de l'estimateur spline cubique de lissage, nous renvoyons le lecteur à Simonoff [157] (section 5.6.2) et Eubank [53] (section 5.5). Silverman [156] (1984) a notamment montré que, asymptotiquement,

un estimateur spline est un estimateur à noyau. De plus, les estimateurs splines ont la caractéristique de pouvoir être interprétés comme des estimateurs bayésiens (voir Eubank [53] section 5.6, Kimeldorf et Wahba [87] et Kimeldorf et Wahba [85]). Pour plus d'informations sur les splines de lissage, nous invitons le lecteur à consulter les articles de Besse et Thomas-Agnan [18] et Silverman [155], ainsi que les ouvrages de Eubank [53], Green et Silverman [69] et l'excellent ouvrage de Wahba [173].

Remarque 3.1. Le résultat garantissant l'existence et l'unicité des splines de lissage s'applique aussi en interpolation. En effet, l'unique solution du problème de minimisation contrainte :

$$\begin{cases} \min\limits_{f\in W^m[a,b]} \int_a^b (f^{(m)}(x))^2 dx \\ \text{sous les contraintes : } f(x_i) = y_i,\ i = 1,\dots,n, \end{cases} \tag{3.15}$$

est une spline cubique polynomiale naturelle d'ordre $2m$ ayant pour noeuds les points $x_1,\dots,x_n$, et appelée spline d'interpolation.

Remarque 3.2. Le critère des moindres carrés pénalisés (3.14) peut être étendu à un critère des moindres carrés pondérés par l'ajout d'un vecteur de poids $w = (w_1,\dots,w_n)$ sur les observations (Ramsay et Silverman [130] chap.5).

3.2.2.3 Calcul des splines de lissage

La technique la plus courante de calcul des splines de lissage consiste à utiliser une base de splines naturelles. En effet, l'estimateur $\widehat{f}$ étant une spline naturelle telle que $\widehat{f} \in \mathcal{NS}_{2m}(x_1,\dots,x_n)$, alors il peut s'écrire sous la forme

$$\widehat{f}(x) = \sum_{j=1}^{n} \widehat{\beta}_j B_j(x), \tag{3.16}$$

où les B_j forment une base de splines naturelles prenant souvent la forme de B-splines (De Boor [37], Eubank [53]). Ainsi, le problème de minimisation du critère (3.14) dans l'espace de fonctions $W^m[a,b]$ se ramène à un problème de minimisation dans un espace de dimension finie. On cherche alors le vecteur de coefficients $\widehat{\beta} = (\widehat{\beta}_1,\dots,\widehat{\beta}_n)$ qui minimise le critère

$$(y - B\beta)^T(y - B\beta) + \lambda\beta^T\Omega\beta, \tag{3.17}$$

où B est la matrice $n \times n$ contenant les $B_j(x_i)$, $i = 1,\dots,n$, $j = 1,\dots,n$, et Ω est la matrice $n \times n$ contenant les éléments $\int_a^b B_i^{(m)}(x)B_j^{(m)}(x)dx$. La solution s'écrit alors

$$\widehat{\beta} = (B^TB + \lambda\Omega)^{-1}B^Ty. \tag{3.18}$$

On en déduit que l'estimateur spline de lissage est linéaire, où la matrice de lissage S définie en (3.2) est donnée par

$$S = B(B^TB + \lambda\Omega)^{-1}B^T. \tag{3.19}$$

Une autre méthode de calcul des splines de lissage consiste à utiliser la représentation $g-\gamma$ et l'algorithme de Reinsch [135] (1967) comme indiqué dans Green et Silverman [69]. Il est également possible d'utiliser la représentation par les noyaux reproduisants comme nous le verrons à la section 3.3.

3.2.2.4 Choix du paramètre λ

Nous avons vu à la section 3.2.2.1 que le paramètre de lissage λ contrôlait le lissage de la fonction, à savoir le compromis entre un bon ajustement des données et l'obtention d'une fonction lisse. Ainsi, plus le paramètre λ est proche de 0, plus la fonction de lissage s'approche d'une fonction interpolant les données, et inversement, plus λ est grand, plus la fonction de lissage s'approche d'un polynôme de régression. A titre d'exemple, la figure 3.3 représente en rouge un estimateur par spline cubique de lissage (i.e. obtenue en prenant $m = 2$ dans le critère (3.14)) pour différentes valeurs de λ. La fonction de régression que l'on cherche à estimer est la fonction $f(x) = x + \frac{1}{6\pi} sin(6\pi x)$ (en bleu), et les données (en vert) sont générées par $y = f(x) + \varepsilon$ pour $n = 50$ valeurs de x uniformément réparties sur l'intervalle $[0, 1]$ et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma = 0.05$.

a) $\lambda = 10^{-8}$ b) $\lambda = 10^{-4}$ c) $\lambda = 10^{5}$

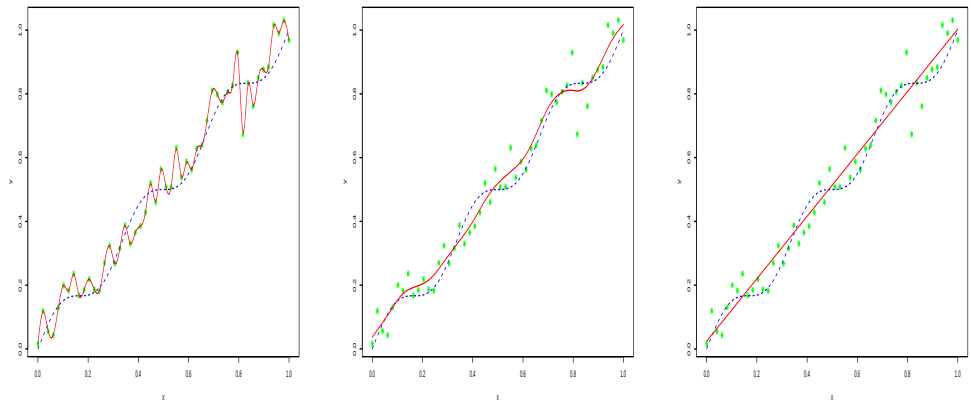


Figure 3.3 - Estimation par spline cubique de lissage (en rouge) pour différentes valeurs du paramètre λ. L'estimateur est représenté en rouge. Les données (en vert) sont générées par $y = f(x) + \varepsilon$ pour $n = 50$ valeurs de x uniformément réparties sur l'intervalle $[0, 1]$ avec $f(x) = x + \frac{1}{6\pi} sin(6\pi x)$ (en bleu) et $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ avec $\sigma = 0.05$.

Le choix du paramètre λ est donc important et l'objectif est de déterminer le paramètre de lissage réalisant le meilleur compromis biais-variance. On cherche alors la valeur optimale du paramètre minimisant l'espérance de l'erreur quadratique moyenne,

MASE ("*Mean Average Squared Error*") :

$$MASE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\widehat{f}(x,\lambda) - f(x)]^2, \tag{3.20}$$

ou l'intégrale de l'erreur quadratique moyenne, MISE ("*Mean Integrated Squared Error*") :

$$MISE(\lambda) = \int \mathbb{E}[\widehat{f}(x,\lambda) - f(x)]^2 dx. \tag{3.21}$$

Cependant la fonction de régression f étant en pratique inconnue, différentes méthodes ont été proposées afin d'estimer la valeur optimale du paramètre λ par la minimisation d'un critère spécifique.

L'un des critères couramment utilisé est le critère de validation croisée, CV ("*cross-validation*"), initialement proposé par Allen [5] puis Stone [160] dans le cadre de la régression, et par Wahba et Wold [176] dans le cadre des splines de lissage. Le principe consiste à minimiser le critère suivant :

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{f}^{-i}(x_i))^2, \tag{3.22}$$

où $\widehat{f}^{-i}$ est l'estimateur de f construit sans la i$^{\text{ème}}$ observation. On peut montrer que $\mathbb{E}[CV(\lambda)] \approx \sigma^2 + MASE(\lambda)$. La minimisation de $CV(\lambda)$ nécessite d'évaluer cette fonction pour un grand nombre de valeur de λ et il est donc important de disposer d'une méthode de calcul efficace de $CV(\lambda)$, sans avoir à recalculer n fois l'estimateur pour chaque valeur de λ. Il a alors été démontré que, dans le cas des estimateurs linéaires, le critère de validation croisée pouvait s'écrire comme une fonction des valeurs estimées $\widehat{f}(x_i)$:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f}(x_i)}{1 - S_{ii}(\lambda)}\right)^2, \tag{3.23}$$

où les $S_{ii}(\lambda)$ sont les éléments diagonaux de la matrice chapeau $S(\lambda)$ (Wahba [173]).

Une autre méthode couramment utilisée est la méthode de validation croisée généralisée, GCV ("*generalized cross-validation*"), présentée par Craven et Wahba [33]. Le critère de validation croisée généralisée est une version pondérée du critère de validation croisée défini en (3.23), et consiste à remplacer $S_{ii}(\lambda)$ dans (3.23) par $\frac{1}{n}trace(S(\lambda))$. Le critère GCV consiste donc à minimiser la fonction suivante :

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f}(x_i)}{1 - n^{-1}trace(S(\lambda))}\right)^2. \tag{3.24}$$

La figure 3.4 reprend l'exemple de la figure 3.3 et représente les valeurs du critère GCV pour différentes valeurs du paramètre λ. A titre de comparaison, les valeurs

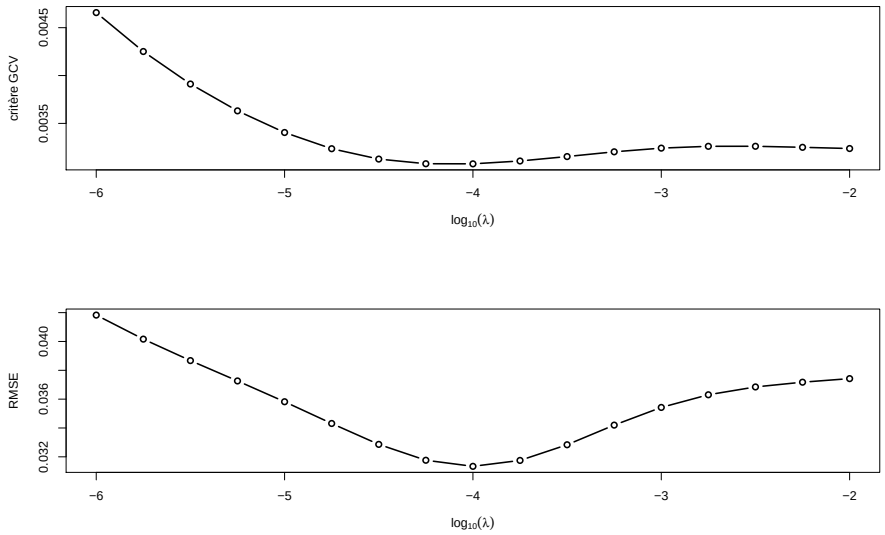


Figure 3.4 - Valeurs du critère GCV et valeurs de la RMSE pour différentes valeurs du paramètre λ.

de la racine carrée de l'erreur quadratique moyenne, RMSE ("*Root Mean Squared Error*") :

$$RMSE(\lambda) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}, \tag{3.25}$$

où $\widehat{y}_i = \widehat{f}(x_i, \lambda)$, sont également représentées. On observe alors que le critère GCV est minimal pour $\lambda = 10^{-4}$, et que cette valeur de λ correspond également à la valeur minimale de la RMSE. L'estimateur par spline de lissage obtenu avec ce paramètre λ optimal est représenté à la figure 3.3 b. La méthode GCV est généralement préférée à la méthode de validation croisée car elle est plus simple à mettre en oeuvre et plus adapté aux données réparties de manière non uniforme (Simonoff [157] section 5.6.3). De plus, le critère $GCV(\lambda)$ est invariant par rotation de la base (Wahba [173] section 4.3).

Enfin, lorsque σ^2 est connu, on peut obtenir un estimateur sans biais du paramètre λ basé sur le critère de Mallows (Wahba [173] section 4.7 et Wang [182] section 3.3). Il est également possible, sous des hypothèses de normalité, de construire des estimateurs du maximum de vraisemblance de λ basé sur un modèle bayésien (Wahba [173] section 4.8 et Wang [182] section 3.6).

3.2.3 Splines pénalisées

Nous avons vu précédemment deux méthodes de lissage utilisant les fonctions splines. La méthode des splines de régression nécessite dans un premier temps une sélection du nombre de noeuds et de leur position, puis une fois que la suite de noeuds est fixé, on estime les coefficients de la spline par minimisation des moindres carrés. Cependant, le fait de n'ajouter aucune contrainte sur les coefficients de la spline entraîne un sur-lissage des données et donc l'obtention d'une courbe non lisse dès que le nombre de noeuds est important. Au contraire, avec la méthode des splines de lissage on place un noeud en chaque point d'observation et on contrôle le phénomène de sur-lissage par l'ajout d'une pénalité de la rugosité dans la minimisation des moindres carrés. Cependant, les splines de lissage peuvent présenter des difficultés numériques dès que le nombre d'observations est trop important.

Une solution consiste à combiner les deux approches, c'est-à-dire choisir une suite de noeuds de longueur K différentes des points d'observation avec $K < n$, et estimer les coefficients de la spline par minimisation des moindres carrés pondérés (3.14). Cette approche, appelée splines hybrides, a été introduite par Kelly et Rice [83], puis développée par Luo et Wahba [100] qui proposent la procedure HAS ("*Hybrid Adaptive Procedure*") avec une sélection automatique des noeuds.

Une généralisation de l'approche par splines hybrides est la méthode des spline pénalisées, ou P-splines, développée par Eilers et Marx [50] d'après les travaux de O'Sullivan [119]. Le principe général des splines pénalisées consiste à choisir un nombre de noeuds K relativement grand mais inférieur au nombre n d'observations, et de contraindre leur influence par l'ajout d'un terme de pénalité plus général que le terme de pénalité de la rugosité. Ainsi, si on note $f(x) = \sum_{j=0}^{p+K} \beta_j B_j$ où les $(B_0, \ldots, B_{p+K})$ sont une base de fonctions d'une spline de degré p ayant pour noeuds $\xi_1, \ldots, \xi_K$, et $\beta = (\beta_1, \ldots, \beta_K)$ est le vecteur des coefficients, et si on note B la matrice des $B_j(.)$, alors le principe des splines pénalisées consiste à minimiser le critère des moindres carrés pénalisés

$$(y - B\beta)^T (y - B\beta) + \lambda \beta^T D \beta, \tag{3.26}$$

où D est une matrice symétrique, semi-définie positive, appelée matrice de pénalité (dans le cas simple, D est la matrice identité). Par exemple, Eilers et Marx [50] proposent de prendre des noeuds équirépartis et de pénaliser sur les différences d'ordre $p-1$ des coefficients dans une base de B-splines (i.e. une pénalisation correspondant à $\lambda \sum_{j=p}^{p+K} (\Delta^k \beta_j)^2$). Un autre exemple proposé par Ruppert [138] et développé dans Ruppert *et al.* [140], consiste à placer les noeuds sur les quantiles de la variable X, à prendre $D = diag(0_{p+1}, 1_K)$, et à remplacer λ par λ^{2p}. Dans ce cas, l'estimateur obtenu correspond à un cas d'estimateur de régression ridge. De plus, Ruppert *et al.* [140] (section 4.9) démontrent que l'estimateur spline pénalisée peut s'écrire comme l'estimateur optimal (BLUP) d'un modèle mixte.

Enfin, citons également l'article de Pearce et Wand [120] qui fait le lien entre les splines pénalisées et la théorie des espaces de Hilbert à noyau reproduisant (RKHS) que nous développerons à la section 3.3.

3.2.4 Application : estimation de profils temporels de vitesse et d'accélération

Afin, de montrer l'intérêt de l'approche fonctionnelle et plus précisément des splines de lissage pour l'estimation des profils de vitesse, nous proposons de comparer plusieurs méthodes de lissage sur un jeu de données réelles issues d'un véhicule équipé d'un GPS, et d'évaluer les performances de ces méthodes pour l'estimation de profils temporels de vitesse et d'accélération. En effet, nous avons cité à la section 1.3.4 du chapitre 1, les principaux inconvénients des méthodes actuellement utilisées pour l'estimation des profils de vitesse et d'accélération.

Les données utilisées sont issues d'une expérimentation réalisée sur les pistes de Satory (Versailles, France) avec un véhicule Renault Clio III équipé d'un boîtier enregistreur connecté au bus CAN et à deux GPS : un GPS GlobalSat BR-355 intégrant la puce SiRF Star III, et un GPS différentiel RTK ("*Real Time Kinematic*") modèle Thalès Sagitta (voir annexe A pour plus de détails). Cependant, le GPS RTK ne fournissant que des mesures de position, nous ne l'utiliserons pas pour cette étude et nous utiliserons les mesures de vitesse issues du GPS GlobalSat BR-355. Ce GPS fournit des mesures de position du véhicule avec une fréquence d'échantillonnage de 1 Hz (soit 1 mes/s) et une précision de 10 m, et fournit également des mesures de vitesse estimées par effet Doppler, et donc indépendantes des mesures de position, avec une précision de 0.1 m/s. Les mesures de distance parcourue issues de l'odomètre et les mesures de vitesses qui en dérivent sont collectées sur le bus CAN du véhicule avec une fréquence d'échantillonnage de 10 Hz. Le trajet réalisé est illustré à la figure 3.5. Ce trajet correspond à la réalisation de deux tours de piste (en bleu sur la figure 3.5) à partir de la position D0, chaque tour ayant une longueur d'environ 2 km.

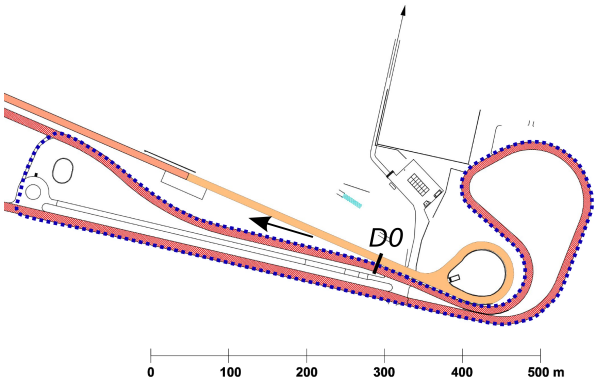


Figure 3.5 - Trajet réalisé sur les pistes de Versailles-Satory (en bleu). Le point D0 correspond au point de départ.

On cherche à estimer la vitesse et l'accélération du véhicule sur ce trajet à partir des mesures de vitesse issues du GPS. Le problème se modélise alors sous la forme

du modèle de régression non paramétrique suivant :

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.27}$$

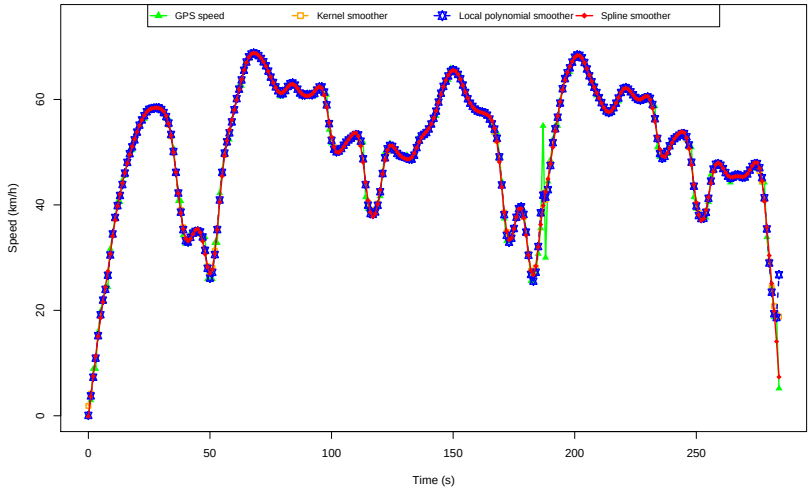
où les y_i sont les mesures de vitesse issues du GPS et les ε_i sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ^2, correspondant aux erreurs de mesures. On cherche alors dans un premier temps un estimateur du profil de vitesse $f(t)$, puis dans un second temps un estimateur du profil d'accélération $f'(t)$. Notons que l'on s'intéresse ici aux profils de vitesse en fonction du temps et non en fonction de la position. L'estimation des profils spatiaux de vitesse sera détaillée au chapitre suivant. Nous proposons de comparer les trois estimateurs suivants : l'estimateur de Nadaraya-Watson, l'estimateur par polynômes locaux et l'estimateur par spline de lissage. Cependant, les mesures de vitesse issues du GPS étant collectées sur une piste dans des conditions presque idéales (pas de masquage des satellites), celles-ci ne présentent pas de grosses erreurs de mesures. Nous avons donc ajouté volontairement deux valeurs aberrantes de vitesse aux temps $t_i = 187$ et $t_i = 188$ afin de tester la robustesse des trois estimateurs étudiés. L'objectif étant d'estimer à la fois la vitesse et l'accélération du véhicule, nous avons choisi pour chacun des estimateurs les paramètres suivants :

- Pour l'estimateur à noyau, nous avons choisi un noyau gaussien et une fenêtre de lissage de largeur 2 s. L'estimation de la dérivée est réalisée par différentiation numérique en utilisant une méthode de type "central-difference", i.e. $f'(t_i) = (f(t_{i+1}) - f(t_{i-1}))/(t_{i+1} - t_{i-1})$. La mise en oeuvre numérique est effectuée avec le logiciel R en utilisant la fonction `locpoly` du package `KernSmooth`.
- Pour l'estimateur par polynômes locaux, nous avons choisi d'utiliser des polynômes de degré 3, un noyau gaussien et une fenêtre de lissage de largeur 3 s. La mise en oeuvre numérique est effectuée avec la fonction `locpoly` du package `KernSmooth`.
- Pour l'estimateur par spline de lissage, nous avons choisi d'utiliser une spline d'ordre 6 (spline quintique, i.e. de degré 5), ce qui revient à prendre $m = 3$ dans le critère 3.14. Le paramètre de lissage λ a été déterminé par le critère GCV ($\lambda = 3.162$). La mise en oeuvre numérique est effectuée avec les fonctions du package `fda` (voir Ramsay *et al.* [131] pour plus de détails).

La figure 3.6 montre les résultats obtenus pour l'estimation du profil de vitesse avec les trois estimateurs. Si les trois estimateurs semblent fournir des résultats similaires sur l'ensemble du trajet, le zoom sur la section où sont présents les deux outliers montre une meilleure robustesse de l'estimateur par spline de lissage. En effet, on observe que les deux outliers ont tendance à affecter l'estimation des vitesses obtenues par la méthode du noyau et la méthode des polynômes locaux, ce qui n'est pas le cas avec la méthode des splines de lissage.

Afin d'avoir une idée de la précision des valeurs estimées obtenues avec les trois méthodes proposées, nous avons comparé ces valeurs avec les mesures de vitesse collectées sur le bus CAN du véhicule qui dérivent des mesures de l'odomètre et qui

a) Trajet complet



b) Zoom sur la section avec outliers

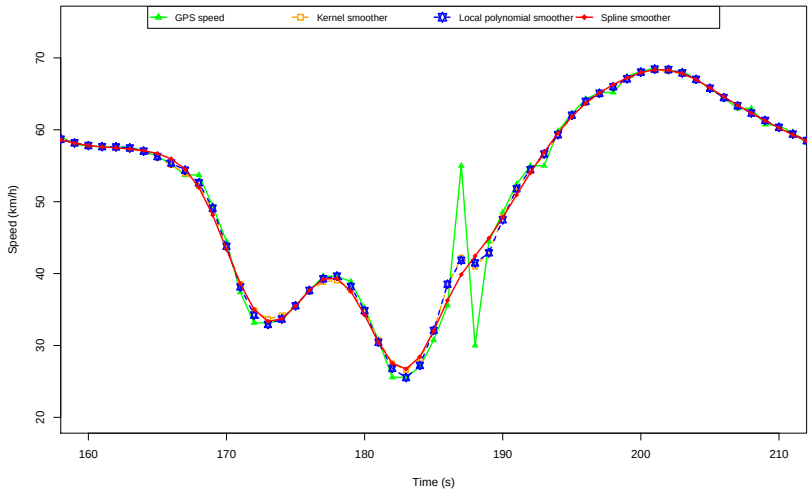


Figure 3.6 - Estimation du profil de vitesse à partir des mesures de vitesse issues du GPS (en vert). Comparaison entre 3 estimateurs : estimateur à noyau (en orange), estimateur par polynômes locaux (en bleu) et estimateur par spline de lissage (en rouge).

	GPS speed	Kernel smoother	Local polynomial smoother	Spline smoother
RMSE speed (km/h)	3.046	1.412	1.273	1.272

Table 3.1 - RMSE des trois estimateurs par rapport à la vitesse CAN pour l'estimation des vitesses.

sont relativement précises, même si dans l'absolu ces valeurs ne peuvent pas être considérées comme des valeurs de référence. Nous avons alors calculé sur la section de 50 s illustré à la figure 3.6.b, la racine carrée de l'erreur quadratique moyenne définie par :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{f}(t_i) - v_{CAN}(t_i))^2}. \tag{3.28}$$

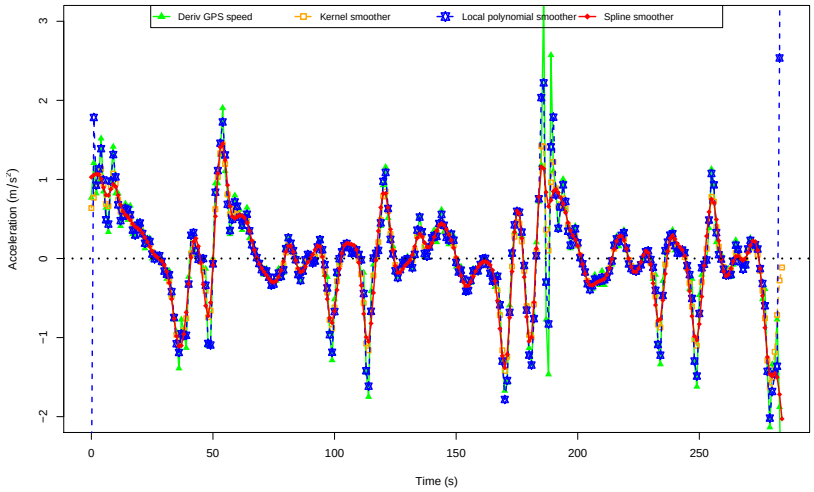
Les résultats obtenus sont résumés dans le tableau 3.1. On observe que, en moyenne, les estimateurs par polynômes locaux et par spline de lissage donnent des résultats similaires même si l'estimateur par spline de lissage est plus robuste.

On s'intéresse maintenant à l'estimation du profil d'accélération du véhicule, i.e. à la dérivée des profils de vitesse estimés obtenus à la figure 3.6. La figure 3.7 montre les résultats obtenus pour les trois estimateurs. Les mesures d'accélération issues du GPS ont été obtenues par différentiation numérique comme pour l'estimateur à noyau. On observe que l'estimateur par polynômes locaux conserve bien les pics d'accélération mais est fortement influencé par les outliers et présente des problèmes aux bords de l'intervalle. Les estimateurs à noyau et par splines de lissage ont tendance à aplatir les pics d'accélération mais sont moins sensibles aux outliers.

Comme pour la vitesse, nous avons également calculé les RMSE des valeurs estimées d'accélération en prenant comme valeurs de référence, la dérivée des mesures de vitesses issues du bus CAN ("central difference approach"). Les résultats figurant dans le tableau 3.2 montrent que l'estimateur par spline de lissage présente l'erreur d'estimation la plus faible contrairement à l'estimateur par polynômes locaux. Cependant, pour confirmer ces résultats, il pourrait être intéressant de comparer les valeurs estimées d'accélération obtenues avec les trois estimateurs, avec les mesures d'accélération issues d'une centrale inertielle.

Pour conclure, cette étude a montré les bonnes performances de l'estimateur par spline de lissage pour l'estimation des profils de vitesse et d'accélération. Nous avons notamment montré sa robustesse aux outliers contrairement à l'estimateur à noyau et

a) Trajet complet



b) Zoom sur la section avec outliers

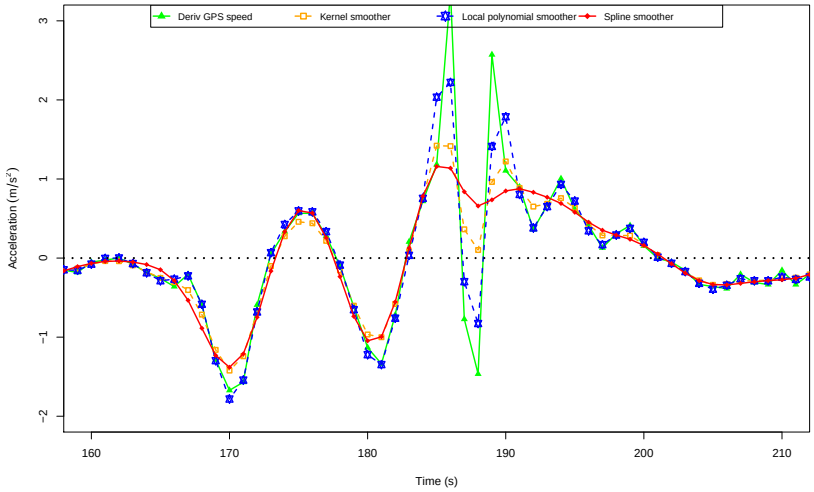


Figure 3.7 - Estimation du profil d'accélération à partir des mesures de vitesse issues du GPS. Comparaison entre 3 estimateurs : estimateur à noyau, estimateur par polynômes locaux et estimateur par spline de lissage.

	GPS speed	Kernel smoother	Local polynomial smoother	Spline smoother
RMSE accel (m/s^2)	0.574	0.216	0.391	0.180

Table 3.2 - RMSE des trois estimateurs par rapport à la dérivée de la vitesse CAN pour l'estimation des accélérations.

à l'estimateur par polynômes locaux. Le principal avantage de la méthode des splines de lissage est que sa mise en oeuvre ne nécessite le choix que d'un unique paramètre, le paramètre de lissage λ. Si dans cette étude, la valeur du paramètre λ obtenue par le critère GCV semble satisfaisante, il est possible de modifier cette valeur en fonction de l'objet de l'étude. Par exemple, si l'on souhaite obtenir un profil d'accélération un peu moins lissé, on peut choisir une valeur de λ plus faible. Un autre avantage des splines de lissage est que les données brutes de nature vectorielle sont transformées en un objet fonctionnel et que l'on a une expression explicite de la fonction obtenue. Cette approche fonctionnelle facilite énormément le calcul des dérivées et permet notamment en choisissant des splines de degré plus élevé, de calculer des dérivées d'ordre supérieur comme la dérivée seconde (correspondant au jerk). Dans le cas de l'estimation par polynômes locaux, on se limite généralement au choix de polynômes de degré 3 car si le choix de degré plus élevé permet de diminuer le biais, il augmente cependant la variabilité de l'estimateur du fait d'une augmentation du nombre de paramètres locaux (Fan et Gijbels [55] section 3.3). Enfin, notons également que le caractère fonctionnel de l'estimateur par spline de lissage facilite l'évaluation de l'estimateur en tout point de l'intervalle étudié. Nous verrons plus en détails aux chapitres suivants l'intérêt des splines de lissage pour l'estimation des profils spatiaux de vitesse ainsi que les avantages d'une approche fonctionnelle pour le calcul d'un profil de vitesse de référence (profil moyen ou profil V85).

3.3 Généralisation de la notion de splines : la régression régularisée dans un espace de Hilbert à noyau reproduisant

3.3.1 Introduction

Dans cette section, nous allons introduire la notion de splines dans le contexte plus large des méthodes de régularisation et des espaces de Hilbert à noyau reproduisant. Cette approche basée sur les propriétés d'optimalité des splines polynomiales, présente une notion plus abstraite des splines définies comme la solution d'un pro-

blème variationnel dans un espace de Hilbert. L'approche variationnelle des splines, dont les splines polynomiales naturelles définies à la section 3.2.2.2 sont un cas particulier, a été introduite par les mathématiciens français Atteia et Laurent (Atteia [12], Atteia [11], Laurent [92]), et fournit un cadre unificateur de traitement de divers types de splines comme les splines périodiques, les splines plaques minces (qui généralisent les splines polynomiales naturelles en dimension supérieure à 1), et bien d'autres (Wahba [173], Berlinet et Thomas-Agnan [15]).

Nous avons vu à la section 3.2.2.1 que la méthode de régularisation, qui consiste à restreindre l'espace fonctionnel de recherche par l'ajout d'un terme de pénalité, permettait de résoudre des problèmes mal posés. Ainsi, la méthode de régularisation permet de reformuler le problème de régression sous la forme du problème variationnel suivant :

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\|f\|^2_{\mathcal{H}}, \tag{3.29}$$

où $L(y, f(x))$ est une fonction de perte représentant le prix à payer lorsque l'on prédit $f(x)$ à la place de y, $\|\cdot\|_{\mathcal{H}}$ est la norme associée à l'espace de fonctions $\mathcal{H}$, et λ est un réel positif contrôlant le compromis entre les deux termes. Cette écriture très générale permet de traiter de manière uniforme une grande variété de techniques liées à la régression et à la théorie de l'apprentissage ("*Support Vector Machines*") par le choix de la fonction de perte L et de l'espace $\mathcal{H}$. Cependant la principale difficulté est le choix d'une norme traduisant la notion de régularité de la solution. Le concept de noyaux reproduisants introduit par Aronszajn [10] permet de pallier cette difficulté, la régularité de la solution étant contrôlée par le choix du noyau. Nous nous plaçons donc dans le cas où l'espace de fonctions $\mathcal{H}$ est un espace de Hilbert caractérisé par un noyau qui reproduit, par l'intermédiaire d'un produit scalaire, chaque fonction de l'espace. Un tel espace est appelé espace de Hilbert à noyau reproduisant ou plus couramment RKHS pour "*Reproducing Kernel Hilbert Space*". Grace Wahba (Wahba [173]) a beaucoup contribué à la théorie des RKHS dans le cadre de la théorie des splines entre les années 1970 et 1990. Plus récemment les RKHS sont apparus dans la théorie de l'apprentissage et plus particulièrement dans la littérature des séparateurs à vaste marge (en anglais "*Support Vector Machine*", SVM) (voir par exemple Schölkopf *et al.* [146], Evgeniou *et al.* [54] et Pearce et Wand [120]).

3.3.2 Définition et principales propriétés des RKHS

La théorie des espaces hilbertiens à noyau reproduisant est très classique en analyse fonctionnelle. Les premières études sont apparues au début du XX^{e} siècle mais la théorie générale s'est essentiellement développée après 1950 avec les travaux d'Aronszajn [10]. Nous donnons ici les définitions et principales propriétés. Plus de détails liés à la théorie des RKHS sont donnés dans Aronszajn [10], Wahba [173], Gu [72] et Berlinet et Thomas-Agnan [15]. Pour les définitions de base sur les espaces de Hilbert et leurs applications, le lecteur pourra se référer à Dieudonné [45], Dudley

[47] et Rudin [137].

On commence par rappeler la définition d'une fonctionnelle d'évaluation.

Définition 3.2. *On appelle fonctionnelle d'évaluation sur un espace de Hilbert $\mathcal{H}$ une forme linéaire $\delta_t : \mathcal{H} \longrightarrow \mathbb{R}$ qui évalue chaque fonction de l'espace au point t, i.e.*

$$\delta_t(f) = f(t), \quad \textit{pour tout } f \in \mathcal{H}.$$

On donne alors une première définition des espaces de Hilbert à noyau reproduisant.

Définition 3.3. *Un espace de Hilbert $\mathcal{H}$ est un espace de Hilbert à noyau reproduisant (RKHS) si les fonctionnelles d'évaluation sont continues.*

Rappelons que les fonctionnelles d'évaluation sont continues si et seulement si elles sont bornées, i.e. si pour tout t, il existe un réel $M > 0$ tel que :

$$|\delta_t(f)| = |f(t)| \leq M\|f\|_{\mathcal{H}}, \quad \text{pour tout } f \in \mathcal{H}.$$

La définition 3.3 signifie également que si deux fonctions f et g appartenant à un RKHS sont proches au sens de la norme associée au produit scalaire, alors pour tout x, les valeurs $f(x)$ et $g(x)$ sont également proches. Par exemple, l'espace de fonctions $L^2(\mathbb{R}^n)$ n'est pas un RKHS (car les formes linéaires d'évaluation n'y sont pas continues). Cependant, en pratique, la définition 3.3 est difficile à utiliser. On introduit alors la notion de noyau reproduisant. Avant d'en donner la définition, commençons par introduire la propriété suivante qui se déduit de la définition 3.3 en appliquant le théorème de représentation de Riesz :

Théorème 3.1. *Soit $\mathcal{X}$ un ensemble quelconque. Si $\mathcal{H}$ est un RHKS, alors pour tout $t \in \mathcal{X}$, il existe une fonction $K_t \in \mathcal{H}$, appelée représentant de l'évaluation en t, avec la propriété de reproduction :*

$$\delta_t(f) = f(t) = \langle K_t, f\rangle_{\mathcal{H}}, \quad \textit{pour tout } f \in \mathcal{H}.$$

La propriété de reproduction permet de représenter toute fonctionnelle d'évaluation par un produit scalaire. D'autre part, comme K_t est une fonction de $\mathcal{H}$, alors d'après la propriété de reproduction on a pour tout $s \in \mathcal{X}$:

$$K_t(s) = \langle K_t, K_s\rangle_{\mathcal{H}}.$$

On peut maintenant donner la définition d'un noyau reproduisant.

Définition 3.4. *Une fonction $K(s,t) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau reproduisant d'un espace de Hilbert $\mathcal{H}$ si :*

a) $\forall t \in \mathcal{X},\ K(\cdot, t) \in \mathcal{H}$,

b) $\forall t \in \mathcal{X},\ \forall f \in \mathcal{H},\ \langle f, K(\cdot, t)\rangle = f(t)$ *(i.e. f vérifie la propriété de reproduction).*

On en déduit alors une nouvelle définition des RKHS équivalente à la définition 3.3.

Définition 3.5. *Un espace de Hilbert $\mathcal{H}$ qui possède un noyau reproduisant est appelé un espace de Hilbert à noyau reproduisant (RKHS).*

Remarque 3.3. Si on considère f comme une fonction de $\mathcal{X}$ dans $\mathcal{Y}$, alors le calcul de $f(t)$ est généralement une opération non linéaire. Par contre, si l'on considère f comme un élément de $\mathcal{H}$ de noyau K, alors d'après la propriété de reproduction, on peut calculer $f(t)$ via $\langle f, K(\cdot, t)\rangle$ qui est une opération linéaire. De plus, d'après la propriété de reproduction, on a : $\forall (s,t) \in \mathcal{X}\times\mathcal{X},\ K(s,t) = K_t(s) = \langle K(\cdot,t), K(\cdot,s)\rangle$. Ainsi, le produit scalaire dans $\mathcal{H}$ peut être calculé simplement en utilisant la fonction noyau K. Cette propriété, appelée astuce du noyau ("*kernel trick*"), est à la base du développement des méthodes à noyaux ("*kernel methods*") qui se sont considérablement développées depuis le milieu des années 1990 et qui sont très en vogue en théorie de l'apprentissage (Schölkopf et Smola [144], Shawe-Taylor et Cristianini [154]). Les méthodes à noyau permettent de trouver des fonctions de décision non linéaires, tout en s'appuyant fondamentalement sur des méthodes linéaires (ex : Kernel PCA, Kernel Discriminant Analysis, SVM...).

Citons quelques exemples de RKHS.

Exemple 3.1.
Soit $\mathcal{H}$ un espace vectoriel de dimension finie et soit $(e_1, \ldots, e_n)$ une base orthonormale de $\mathcal{H}$. On montre facilement que $\mathcal{H}$ a pour noyau reproduisant

$$K(x,y) = \sum_{i=1}^{n} e_i(x)e_i(y).$$

Ainsi, tout espace vectoriel de dimension finie est un RKHS.

Exemple 3.2.
Soient $\mathcal{X} = [0,1]$ et $\mathcal{H} = \{f \mid f(0) = 0,\ f$ est absolument continue et $f' \in L^2[0,1]\}$. $\mathcal{H}$ est un espace de Hilbert muni du produit scalaire :

$$\langle f, g\rangle_{\mathcal{H}} = \int_0^1 f'(x)g'(x)dx,$$

et appartient à la classe des espaces de Sobolev (Adams et Fournier [3]. $\mathcal{H}$ a pour noyau reproduisant $K(x,y) = min(x,y)$. En effet, la dérivée de $min(.,y)$ étant la fonction $\mathbb{1}_{(0,y)}$, on a :

$$\langle f, K(.,y)\rangle_{\mathcal{H}} = \int_0^1 f'(x)K'(.,y)dx = \int_0^y f'(x)dx = f(y).$$

Exemple 3.3.
Soient $\mathcal{X} = \mathbb{R}$ et

$$\mathcal{H} = \mathcal{H}^1(\mathbb{R}) = \{f \mid f(0) = 0,\ f \text{ est absolument continue, } f \text{ et } f' \text{ sont dans } L^2(\mathbb{R})\}.$$

$\mathcal{H}$ est un espace de Hilbert muni du produit scalaire :

$$\langle f, g\rangle_{\mathcal{H}} = \int_{\mathbb{R}} f(x)g(x) + f'(x)g'(x)dx,$$

et appartient à la classe des espaces de Sobolev (Adams et Fournier [3]). Une simple intégration par parties montre que $\mathcal{H}$ a pour noyau reproduisant :

$$K(x, y) = \frac{1}{2}\exp(-|x - y|).$$

On s'intéresse maintenant à la caractérisation des noyaux reproduisants : quand est-ce qu'une fonction K à valeurs réelles (ou complexes) définie sur $\mathcal{X} \times \mathcal{X}$ est un noyau reproduisant ? Nous allons énoncer dans ce qui suit le lien entre les noyaux reproduisants et les fonctions définies positives.

Définition 3.6. *Une fonction $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est appelée fonction définie positive si :*

$$\forall n \geq 1, \forall (a_1, \ldots, a_n) \in \mathbb{R}^n, \forall (x_1, \ldots, x_n) \in \mathcal{X}^n, \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) \geq 0. \quad (3.30)$$

Remarque 3.4. La condition (3.30) est équivalente au fait que la matrice $K = (K(x_i, x_j))_{1\leq i,j\leq n}$, appelée matrice de noyau ou matrice de Gram, soit définie positive.

Lemme 3.1. *Tout noyau reproduisant est une fonction définie positive.*

Ce lemme se démontre très facilement. En effet, si $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau reproduisant, alors pour tout $(a_1, \ldots, a_n) \in \mathbb{R}^n$ et $(x_1, \ldots, x_n) \in \mathcal{X}^n$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K(x_i, x_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \langle K_{x_i}, K_{x_j}\rangle = \|\sum_{i=1}^{n} K_{x_i}\|^2.$$

En fait, le théorème fondamental qui suit montre que la réciproque est vraie.

Théorème 3.2 (Moore-Aronszajn, 1950)**.** *Soit K une fonction symétrique, définie positive sur $\mathcal{X} \times \mathcal{X}$. Alors il existe un unique espace de Hilbert H de fonctions sur $\mathcal{X}$ ayant pour noyau reproduisant K.*

Ce théorème montre que l'ensemble des fonctions définies positives et l'ensemble des noyaux reproduisants sur $\mathcal{X} \times \mathcal{X}$ sont identiques et qu'il existe donc une bijection entre les RKHS et les fonctions définies positives. Ainsi, on peut parler du "noyau d'un RKHS" ou du "RKHS d'un noyau".

L'idée de la preuve est que, pour toute fonction définie positive K, on peut construire un unique RKHS $\mathcal{H}_K$ de noyau reproduisant K où $\mathcal{H}_K$ est la complétion de l'espace vectoriel engendré par les combinaisons linéaires des $K_t = K(t, \cdot)$, i.e. $\mathcal{H}_K = \overline{vect\{K_t, t \in \mathcal{X}\}}$, muni du produit scalaire défini comme suit :

$$\langle \sum_{i=1}^{n} a_i K_{t_i}, \sum_{i=j}^{m} b_j K_{t_j} \rangle_{\mathcal{H}_K} = \sum_{i=1}^{n} \sum_{i=j}^{m} a_i b_j K(t_i, t_j).$$

Pour une démonstration complète du théorème, nous renvoyons le lecteur à Berlinet et Thomas-Agnan [15].

Exemple 3.4 (Exemples de noyaux reproduisants)**.**

- Noyau linéaire : $K(x, y) = x \cdot y$.
- Noyau gaussien : $K(x, y) = \exp(-\frac{\|x-y\|^2}{\sigma^2}), \quad \sigma > 0$.
- Noyau polynomial : $K(x, y) = (x \cdot y + 1)^d, \quad d \in \mathbb{N}$.

Enfin, nous terminons cette section en énonçant une propriété importante sur la somme de RKHS.

Théorème 3.3 (somme de RKHS)**.** *Soit $\mathcal{H}$ un RKHS sur $\mathcal{X}$ dont le noyau K peut se décomposer en $K = K_0 + K_1$, avec K_0 et K_1 tous deux définis positifs, tels que $K_0(x, \cdot) \in \mathcal{H}$ et $K_1(x, \cdot) \in \mathcal{H}$ pour tout $x \in \mathcal{X}$, et $\langle K_0(x, \cdot), K_1(x, \cdot) \rangle_{\mathcal{H}} = 0, \ \forall x, y \in \mathcal{X}$. Alors $\mathcal{H}$ admet la décomposition $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ où $\mathcal{H}_0$ et $\mathcal{H}_1$ sont les espaces associés respectivement aux noyaux K_0 et K_1.*
Réciproquement, si $\mathcal{H}_0$ et $\mathcal{H}_1$ sont des RKHS de noyaux respectifs K_0 et K_1 et si $\mathcal{H}_0 \bigcap \mathcal{H}_1 = \{0\}$, alors $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ est un RKHS ayant pour noyau reproduisant $K = K_0 + K_1$.

Nous renvoyons le lecteur à Gu [72] (Th. 2.5) pour une démonstration de ce théorème.

3.3.3 Régularisation et RKHS

Revenons au problème de régularisation (3.29) énoncé en introduction mais en supposant que l'espace de recherche est un RKHS. On cherche alors à résoudre le problème d'optimisation suivant :

$$\min_{f \in \mathcal{H}_K} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_K}, \tag{3.31}$$

où λ est un réel positif, $\mathcal{H}_K$ est un RKHS de noyau reproduisant $K(\cdot,\cdot)$, et $L(\cdot,\cdot)$ est une fonction de perte. L'intérêt de se placer dans un RKHS est que la norme contrôle la régularité de la solution. En effet, d'après l'inégalité de Cauchy-Schwarz, on a pour toute fonction $f \in \mathcal{H}_K$ et tout point $x, y \in \mathcal{X}$:

$$|f(x) - f(y)| = |\langle f, K_x - K_y\rangle_{\mathcal{H}}| \leq \|f\|_{\mathcal{H}}\|K_x - K_y\|_{\mathcal{H}}.$$

La norme du RKHS contrôle donc la constante de Lipschitz de f pour la métrique $d_K(x,y) = \|K_x - K_y\|_{\mathcal{H}}$. On se demande alors si le problème de minimisation (3.31) admet une solution et si cette solution est unique. La section qui suit va nous permettre de répondre à cette question.

3.3.3.1 Théorème du représentant

Le théorème du représentant (Kimeldorf et Wahba [86]) permet d'obtenir une forme explicite de la solution du problème (3.31). Nous commençons par énoncer une version plus générale de ce théorème présentée par Schölkopf *et al.* [145].

Théorème 3.4 (dit du représentant)**.** *Soient $\mathcal{X}$ un ensemble, $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ un noyau défini positif, $\mathcal{H}_K$ le RKHS associé, et $\mathcal{S} = (x_1, \ldots, x_n) \subset \mathcal{X}$ un ensemble fini de points de $\mathcal{X}$. Soit $\Psi : \mathbb{R}^{n+1} \to \mathbb{R}$ une fonction strictement croissante par rapport à sa dernière variable. Alors toute solution au problème d'optimisation*

$$\min_{f \in \mathcal{H}_K} \Psi(f(x_1), \ldots, f(x_n), \|f\|_{\mathcal{H}_K}), \tag{3.32}$$

admet une représentation de la forme

$$\forall x \in \mathcal{X}, \quad f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) \quad (\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n. \tag{3.33}$$

Démonstration. On note $\mathcal{H}_{\mathcal{S}} := vect\{K(x_i, .), i = 1, \ldots, n\}$. $\mathcal{H}_{\mathcal{S}}$ étant un sous-espace vectoriel de dimension finie de $\mathcal{H}_K$, on en déduit que $\mathcal{H}_{\mathcal{S}}$ est un sous-espace vectoriel fermé de $\mathcal{H}_K$ et donc admet un supplémentaire orthogonal $\mathcal{H}_{\mathcal{S}}^{\perp}$. Ainsi, toute fonction $f \in \mathcal{H}_K$ peut s'écrire de manière unique sous la forme $f = f_{\mathcal{S}} + f_{\perp}$, avec $f_{\mathcal{S}} \in \mathcal{H}_{\mathcal{S}}$ et $f_{\perp} \in \mathcal{H}_{\mathcal{S}}^{\perp}$. $\mathcal{H}_K$ étant un RKHS, on a :

$$\forall i = 1, \ldots, n, \ f_{\perp}(x_i) = \langle f_{\perp}, K(x_i, .)\rangle_{\mathcal{H}_K} = 0.$$

Par conséquent,

$$\forall i = 1, \ldots, n, \ f(x_i) = f_{\mathcal{S}}(x_i).$$

On note $\xi(f, \mathcal{S})$ la fonctionnelle à minimiser dans l'équation (3.32). D'après le théorème de Pythagore, on a :

$$\|f\|_{\mathcal{H}_K}^2 = \|f_{\mathcal{S}}\|_{\mathcal{H}_K}^2 + \|f_{\perp}\|_{\mathcal{H}_K}^2.$$

Par conséquent, $\xi(f, \mathcal{S}) \geq \xi(f_{\mathcal{S}}, \mathcal{S})$ avec égalité si et seulement si $\|f_{\perp}\|_{\mathcal{H}_K} = 0$, i.e. si $f \in \mathcal{H}_{\mathcal{S}}$. La fonction Ψ étant supposée croissante en la norme $\|f\|_{\mathcal{H}_K}$, le problème (3.32) admet donc nécessairement sa ou ses solution(s) dans $\mathcal{H}_{\mathcal{S}}$. □

Le théorème du représentant s'applique notamment au problème de minimisation (3.31). En fait, dans la pratique, la fonction Ψ est souvent de la forme

$$\Psi(f(x_1),\dots,f(x_n),\|f\|_{\mathcal{H}_K}) = c(f(x_1),\dots,f(x_n)) + \lambda\Omega(\|f\|_{\mathcal{H}_K}), \tag{3.34}$$

où c est un critère de coût mesurant l'adéquation de f à un problème donné (régression, classification, ...) et $\Omega : [0,\infty[\to\mathbb{R}$ une fonction strictement croissante.

Ce théorème fondamental montre que même si l'espace $\mathcal{H}_K$ est de dimension infinie, la solution du problème (3.31) appartient à l'espace de dimension finie engendrée par les $K(x_i,\cdot)$. Notons que si la fonction de perte $L(\cdot,\cdot)$ est convexe en f, alors la fonctionnelle

$$\Phi(f) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda\|f\|^2_{\mathcal{H}_K}$$

est strictement convexe et coercive[1], ce qui garantit l'unicité de la solution. On peut citer par exemple la fonction de perte quadratique $L(y,f(x)) = (y-f(x))^2$ (régression régularisée) et la fonction de perte "hinge loss function" $L(y,f(x)) = max(0,1-yf(x))$ (classification par SVM) qui sont des fonctions convexes. En revanche, la fonction de perte 0-1 $L(y,f(x)) = \mathbb{1}_{(y\neq f(x))}$ n'est pas convexe.

3.3.3.2 Modèle général des splines de lissage

Le modèle classique des splines de lissage est le suivant :

$$y_i = f(t_i) + \varepsilon_i\ ,\quad i = 1,\dots,n, \tag{3.35}$$

où $t\in\mathcal{X} = [a,b]$, $f\in W^m[a,b]$ où $W^m[a,b]$ est l'espace de Sobolev d'ordre m, et les ε_i sont des erreurs aléatoires indépendantes de moyenne 0 et de variance constante σ^2. Nous avons montré à la section 3.2.2.2 que le problème classique des splines de lissage correspondait à la minimisation du critère suivant :

$$\min_{f\in W^m[a,b]} \sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda\int_a^b (f^{(m)}(t))^2 dt. \tag{3.36}$$

On s'intéresse ici au problème général des splines de lissage (appelé "*general spline smoothing problem*" dans Wahba [173]) dont le modèle associé est

$$y_i = \mathcal{L}_i f + \varepsilon_i\ ,\quad i = 1,\dots,n, \tag{3.37}$$

où les ε_i sont définies comme dans le modèle classique (3.35) mais où $\mathcal{X}$ est un ensemble quelconque, $f\in\mathcal{H}_R$ un RKHS sur $\mathcal{X}$ de noyau $R(s,t)$, et les $\mathcal{L}_i$ sont des formes linéaires bornées sur $\mathcal{H}_R$ (par exemple, $\mathcal{L}_i f = f'(t_i)$ ou $\mathcal{L}_i f = \int w_i(u)f(u)du$ où les w_i sont des fonctions connues). Le modèle classique (3.35) est un cas particulier

1. Une fonction $J:\mathcal{H}\to\mathbb{R}$, où $\mathcal{H}$ est un espace de Hilbert, est coercive si $\lim_{\|x\|\to+\infty} J(x) = +\infty$.

de ce modèle où les $\mathcal{L}_i$ sont les fonctionnelles d'évaluation aux points d'observation définies par $\mathcal{L}_i f = f(t_i)$. $\mathcal{H}_R$ est supposé admettre une décomposition en somme directe

$$\mathcal{H}_R = \mathcal{H}_0 \bigoplus \mathcal{H}_1,$$

où $\mathcal{H}_0$ est un espace de dimension finie $p \leq n$. Le problème général des splines de lissage revient alors à trouver $f \in \mathcal{H}_R$ qui minimise

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathcal{L}_i f)^2 + \lambda \|P_1 f\|^2_{\mathcal{H}_R}, \tag{3.38}$$

où P_1 est la projection orthogonale sur $\mathcal{H}_1$. Le paramètre de lissage λ contrôle le compromis entre la qualité de l'ajustement mesuré par le 1$^{\text{er}}$ terme, et l'éloignement par rapport à l'espace $\mathcal{H}_0$ mesuré par $\|P_1 f\|^2_{\mathcal{H}_R}$. L'espace $\mathcal{H}_0$ est généralement appelé espace nul et est constitué de fonctions qui ne sont pas pénalisées puisque $\|P_1 f\|^2_{\mathcal{H}_R} = 0$ quand $f \in \mathcal{H}_0$. Ainsi, toute fonction $f \in \mathcal{H}_R$ peut s'écrire sous la forme $f = f_0 + f_1$ où $f_0 \in \mathcal{H}_0$ et $f_1 \in \mathcal{H}_1$, la composante f_0 représentant un modèle de régression linéaire dans $\mathcal{H}_0$, et la composante f_1 représentant les variations non expliquées par f_0. Notons également que d'après la propriété énoncée au théorème 3.3, on montre que $\mathcal{H}_0$ et $\mathcal{H}_1$ sont également des RKHS de noyaux respectifs R_0 et R_1 avec $R = R_0 + R_1$.

Comme les $\mathcal{L}_i$ sont des formes linéaires bornées, alors par le théorème de Riesz il existe un représentant $\eta_i \in \mathcal{H}_R$ tel que $\mathcal{L}_i f = \langle \eta_i, f \rangle_{\mathcal{H}_R}$. D'après les propriétés du noyau reproduisant, on a :

$$\forall s \in \mathcal{X},\ \eta_i(s) = \langle \eta_i, R_s \rangle = \mathcal{L}_i R_s = \mathcal{L}_{i(t)} R(s,t),$$

où $\mathcal{L}_{i(t)}$ signifie que $\mathcal{L}_i$ est appliqué à ce qui suit et qui est considéré comme une fonction de t. Par exemple, si $\mathcal{L}_i f = f(t_i)$, alors $\eta_i(s) = R(s,t_i)$, et si $\mathcal{L}_i f = f'(t_i)$, alors $\eta_i(s) = \frac{\partial}{\partial t} R(s,t)|_{t=t_i}$. Le critère (3.38) que l'on cherche à minimiser peut alors s'écrire sous la forme :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \eta_i, f \rangle)^2 + \lambda \|P_1 f\|^2_{\mathcal{H}_R}. \tag{3.39}$$

La variante donnée ci-dessous du théorème du représentant (théorème 3.4) donne une solution explicite au problème de minimisation dans $\mathcal{H}_R$ du critère (3.39).

Théorème 3.5 (Kimeldorf et Wahba, 1971)**.** *Soit $\phi_1, \ldots, \phi_p$ une base de l'espace nul $\mathcal{H}_0$ et soit T une matrice $n \times p$ de plein rang définie par :*

$$T = \{\mathcal{L}_i \phi_\nu\}_{i=1\ \nu=1}^{n\ \ p}.$$

Alors le critère (3.39) a un unique minimum donné par

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i \xi_i(t), \tag{3.40}$$

où

$$\begin{aligned} \xi_i &= P_1\eta_i, \\ \Sigma &= \{\langle \xi_i, \xi_j \rangle\}_{i,j=1}^n, \\ M &= \Sigma + n\lambda I, \\ d = (d_1, \ldots, d_p)^T &= (T^T M^{-1} T)^{-1} T^T M^{-1} y, \\ c = (c_1, \ldots, c_n)^T &= M^{-1}\{I - T(T^T M^{-1} T)^{-1} T^T M^{-1}\} y. \end{aligned} \tag{3.41}$$

Démonstration. De la même manière que dans la preuve du théorème 3.4 du représentant, on peut affirmer l'existence d'un unique élément $\rho \in \mathcal{H}_R$ orthogonal aux $\{\phi_\nu,\ \nu = 1, \ldots, p\}$ et aux $\{\xi_i,\ i = 1, \ldots, n\}$ tel que l'estimateur $\widehat{f}_\lambda$ s'écrive sous la forme

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i \xi_i(t) + \rho.$$

En utilisant la décomposition de tout élément $g \in \mathcal{H}_R$ de la forme $g = P_0 g + P_1 g$ par somme directe, et en utilisant le caractère auto-adjoint de P_0, on obtient :

$$\begin{aligned} \forall i = 1, \ldots, n,\ \langle \rho, \eta_i \rangle &= \langle \rho, P_0\eta_i \rangle + \langle \rho, P_1\eta_i \rangle \\ &= \langle \underbrace{P_0\rho}_{0}, \eta_i \rangle + \underbrace{\langle \rho, \xi_i \rangle}_{0} \\ &= 0. \end{aligned}$$

Ainsi, le critère (3.39) à minimiser peut s'écrire sous la forme matricielle suivante :

$$\frac{1}{n}\|y - Td - \Sigma c\|^2 + \lambda(c^T \Sigma c + \|\rho\|^2).$$

Il est alors clair que ce critère est minimal quand $\rho = 0$, ce qui démontre la forme (3.40) de l'estimateur.
L'unicité de la solution découle du fait que la fonctionnelle $L(f) = \sum_{i=1}^n (y_i - \mathcal{L}_i f)^2$ est strictement convexe sur $\mathcal{H}_0$ si la matrice T est de plein rang, et donc également strictement convexe sur $\mathcal{H}_R$. On en déduit alors que si la matrice T est de plein rang, $L(f) + \lambda\|P_1 f\|^2_{\mathcal{H}_R}$ est strictement convexe sur $\mathcal{H}_R$, ce qui implique l'unicité de la solution (voir Théorème 2.9 dans Gu [72]).
Il reste alors à estimer les coefficients $c = (c_1, \ldots, c_n)^T$ et $d = (d_1, \ldots, d_p)^T$ qui minimisent le critère

$$\frac{1}{n}\|y - Td - \Sigma c\|^2 + \lambda c^T \Sigma c.$$

En dérivant par rapport à c puis par rapport à d, on obtient les équations suivantes :

$$(\Sigma + n\lambda I)\Sigma c + \Sigma T d = \Sigma y, \tag{3.42}$$

$$T^T \Sigma c + T^T T d = T^T y. \tag{3.43}$$

On montre alors facilement que les équations (3.42) et (3.43) sont équivalentes aux équations suivantes :

$$\begin{aligned} Mc + Td &= y, && (3.44)\\ T^T c &= 0. && (3.45) \end{aligned}$$

On en déduit alors que

$$\begin{aligned} d &= (T^T M^{-1} T)^{-1} T^T M^{-1} y, && (3.46)\\ c &= M^{-1}\{I - T(T^T M^{-1} T)^{-1} T^T M^{-1}\} y. && (3.47) \end{aligned}$$

□

Ce théorème montre que l'estimateur par splines de lissage $\widehat{f}$ appartient à un espace de dimension finie et s'exprime comme une combinaison linéaire de la base de $\mathcal{H}_0$ et des représentants de $\mathcal{H}_1$. Notons que l'estimateur $\widehat{f}_\lambda$ dépend du paramètre λ même si cette dépendance n'est pas explicite. Généralement, les équations (3.46) et (3.47) ne sont pas appropriées au calcul numérique et on utilise plutôt les équations (3.44) et (3.45) pour le calcul des coefficients c et d. En effet, ces équations sont équivalentes à

$$\begin{pmatrix} M & T \\ T^T & 0 \end{pmatrix} \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} y \\ 0 \end{pmatrix}, \tag{3.48}$$

qui est un système linéaire de $n+p$ équations de la forme $Ax = b$ avec A symétrique et de plein rang. Il est alors possible d'utiliser des algorithmes de calcul efficaces tels que l'algorithme de Cholesky (Gu [72], Wahba [173]).

Il est également possible de calculer c et d en utilisant la décomposition QR de T :

$$T = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R \\ 0 \end{pmatrix},$$

où Q_1, Q_2 et R sont respectivement des matrices $n \times p$, $n \times (n-p)$ et $p \times p$, $Q = (Q_1\ Q_2)$ est une matrice orthogonale, et R est une matrice triangulaire supérieure, avec $T^T Q_2 = 0_{p\times(n-p)}$. On peut alors montrer que les coefficients c et d vérifient les équations suivantes :

$$\begin{aligned} c &= Q_2 (Q_2^T M Q_2)^{-1} Q_2^T y, && (3.49)\\ Rd &= Q_1^T (y - Mc). && (3.50) \end{aligned}$$

Nous renvoyons le lecteur à Wahba [173] et Wang [182] pour plus de détails sur l'obtention de ces équations.

Remarque 3.5. On a $\xi_i = P_1 \eta_i$ la projection de η_i sur H_1. Comme $R(s,t) = R_0(s,t) + R_1(s,t)$ où R_0 et R_1 sont les noyaux respectifs de H_0 et H_1, et comme P_1 est auto-adjoint, on a :

$$\xi_i(s) = \langle \xi_i, R_s \rangle = \langle P_1 \eta_i, R_s \rangle = \langle \eta_i, P_1 R_s \rangle = \mathcal{L}_{i(t)} R_1(s,t).$$

Cette équation montre que le représentant ξ_i peut être obtenu en appliquant l'opérateur au noyau R_1. De plus, on a :

$$\langle \xi_i, \xi_j \rangle = \mathcal{L}_{i(s)} \xi_j(s) = \mathcal{L}_{i(s)} \mathcal{L}_{j(t)} R_1(s,t).$$

On en déduit donc que :

$$\Sigma = \{\mathcal{L}_{i(s)} \mathcal{L}_{j(t)} R_1(s,t)\}_{i,j=1}^n.$$

Notons que dans le cas particulier où les $\mathcal{L}_i$ sont les fonctionnelles d'évaluation aux points d'observation définies par $\mathcal{L}_i f = f(t_i)$ (correspondant au modèle classique (3.35)), on a :

$$\xi_i(t) = R_1(t, t_i) \quad \text{et} \quad \Sigma = \{R_1(t_i, t_j)\}_{i,j=1}^n.$$

Ainsi, le vecteur des valeurs estimées peut s'écrire sous la forme :

$$(\mathcal{L}_1 \widehat{f}, \ldots, \mathcal{L}_n \widehat{f})^T = Td + \Sigma c. \tag{3.51}$$

De plus, en utilisant les équations (3.44) et (3.49), on en déduit que

$$(\mathcal{L}_1 \widehat{f}, \ldots, \mathcal{L}_n \widehat{f})^T = Td + \Sigma c = y - n\lambda c = A(\lambda) y, \tag{3.52}$$

où

$$A(\lambda) = I - n\lambda Q_2 (Q_2^T M Q_2)^{-1} Q_2^T \tag{3.53}$$

est la matrice chapeau. Notons que l'expression (3.52) permet de faciliter le calcul numérique des valeurs estimées $(\mathcal{L}_1 \widehat{f}, \ldots, \mathcal{L}_n \widehat{f})$ mais que cette expression ne permet pas de calculer la valeur de l'estimateur en tout point t contrairement à l'expression (3.40).

Finalement, la résolution du problème général des splines de lissage nécessite de suivre les étapes suivantes :

i) choisir un RKHS $\mathcal{H}$ comme espace du modèle pour f ;

ii) choisir une décomposition de l'espace en deux sous-espaces $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$ où $\mathcal{H}_0$ est un ensemble de fonctions non pénalisées ;

iii) choisir une pénalité $\|P_1 f\|^2$.

Différents choix peuvent être effectués concernant l'espace du modèle, sa décomposition et la pénalité, ce qui rend la méthode des splines de lissage très flexible. En effet, ce modèle est à la base de nombreux types de splines : les splines polynomiales ou D^m splines (détaillées à la section suivante), les splines périodiques, les splines de type plaque mince ("thin plate spline")... D'autres exemples de splines sont donnés dans Wahba [173], Berlinet et Thomas-Agnan [15], Gu [72] et Wang [182]).

3.3.3.3 Un exemple important : les splines polynomiales

Revenons au modèle classique des splines de lissage présenté au début de la section précédente mais on se place ici sur l'intervalle $[0,1]$ (en pratique, on peut toujours se ramener à cette situation par un recalage linéaire des données originales). Le modèle considéré est alors le suivant :

$$y_i = f(t_i) + \varepsilon_i \ , \quad i = 1, \ldots, n, \tag{3.54}$$

où $t \in \mathcal{X} = [0,1]$ et $f \in W^m[0,1]$ où $W^m[0,1]$ est l'espace de Sobolev défini par :

$$W^m[0,1] = \{f : [0,1] \to \mathbb{R} \text{ t.q. } f, f', \ldots, f^{(m-1)} \text{ abs. cont., et } \ f^{(m)} \in L^2([0,1])\}.$$

Alors on peut montrer que l'espace de Sobolev $W^m[0,1]$ muni de la norme

$$\|f\|^2 = \sum_{\nu=0}^{m-1} (f^{(\nu)}(0))^2 + \int_0^1 (f^{(m)}(t))^2 dt \tag{3.55}$$

est un RKHS, et que son noyau reproduisant est défini par :

$$R(s,t) = R_t(s) = \sum_{\nu=0}^{m-1} \frac{s^\nu t^\nu}{[\nu!]^2} + \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{[(m-1)!]^2} du. \tag{3.56}$$

Ce résultat est essentiellement basé sur la formule de Taylor avec reste intégral (formule de Taylor-Laplace). On rappelle que, si f est une fonction de $[0,1]$ dans $\mathbb{R}$, $m-1$ fois continûment dérivable et telle que $f^{(m)} \in L^2([0,1])$, on a :

$$\forall t \in [0,1], \quad f(t) = \sum_{\nu=0}^{m-1} \frac{t^\nu}{\nu!} f^{(\nu)}(0) + \int_0^1 \frac{(t-u)_+^{m-1}}{(m-1)!} f^{(m)}(u) du, \tag{3.57}$$

où $(x)_+ = max(0,x)$. Or, on peut noter que $R_t^{(\nu)}(0) = t^\nu/\nu!$, $\nu = 0, \ldots, m-1$ et $R_t^{(m)}(s) = (t-s)_+^{m-1}/(m-1)!$. Ainsi, si on considère le produit scalaire associé à la norme (3.55) :

$$\langle f, g \rangle = \sum_{\nu=0}^{m-1} f^{(\nu)}(0) g^{(\nu)}(0) + \int_0^1 f^{(m)}(t) g^{(m)}(t) dt, \tag{3.58}$$

et si on pose $g = R_t$, on obtient (3.57), et donc on en déduit que $\langle R_t, f \rangle = f(t)$. De plus, les deux termes du noyau $R(s,t)$ défini en (3.56),

$$R_0(s,t) = \sum_{\nu=0}^{m-1} \frac{s^\nu t^\nu}{[\nu!]^2}, \tag{3.59}$$

et

$$R_1(s,t) = \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{[(m-1)!]^2} du, \tag{3.60}$$

étant tous deux définis positifs, on peut montrer facilement que les conditions du théorème 3.3 sont vérifiées. En effet, on montre que :

$$\mathcal{H}_0 = vect\{\phi_\nu,\ \nu = 1, \dots, m\} \text{ avec } \phi_\nu(t) = \frac{(t-a)^{\nu-1}}{(\nu-1)!}, \tag{3.61}$$

muni de la norme $\|\phi\|^2 = \sum_{\nu=0}^{m-1}(\phi^{(\nu)})^2$, est un espace de Hilbert de dimension finie m, que $\{\phi_1, \dots, \phi_m\}$ forme une base orthonormale de cet espace, et que $\mathcal{H}_0$ est un RKHS de noyau R_0. Si on note $\mathcal{B}_m$ l'ensemble des fonctions f, $(m-1)$ fois dérivables, telles que $f^{(\nu)}(0) = 0,\ \nu = 0, \dots, m-1$, on montre également que :

$$\mathcal{H}_1 = \{f : [0,1] \to \mathbb{R} \text{ t.q. } f \in \mathcal{B}_m,\ f, f', ..., f^{(m-1)} \text{ abs. cont.}, \text{ et } f^{(m)} \in L^2([0,1])\}, \tag{3.62}$$

muni de la norme $\|f\|^2 = \int_0^1 (f^{(m)}(t))^2 dt$, est un espace de Hilbert, et que $\mathcal{H}_1$ est un RKHS de noyau R_1. Pour plus de détails, nous renvoyons le lecteur à Wahba [173].

Ainsi, on obtient la décomposition suivante

$$W^m[0,1] = \mathcal{H}_0 \bigoplus \mathcal{H}_1, \tag{3.63}$$

c'est-à-dire que toute fonction $f \in W^m[0,1]$ peut s'écrire de manière unique sous la forme $f = f_0 + f_1$ où $f_0 \in \mathcal{H}_0$ et $f_1 \in \mathcal{H}_1$. Enfin, notons que, en plus d'être en somme directe, les espaces $\mathcal{H}_0$ et $\mathcal{H}_1$ sont orthogonaux lorsque $W^m[0,1]$ est muni de la norme définie en (3.55), et donc en appliquant le théorème 3.3, on en déduit bien que le noyau $R(s,t)$ défini en (3.56) est le noyau reproduisant de $W^m[0,1]$.
Finalement, nous avons vu à la section 3.2.2 que l'estimateur par spline de lissage $\widehat{f}_\lambda$, solution de la minimisation du critère des moindres carrés pénalisés :

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(t))^2 dt$$

dans l'espace de Sobolev $W^m[0,1]$, est une spline polynomiale naturelle d'ordre $2m$ ayant pour noeuds les points d'échantillonnage. En utilisant la norme définie en (3.55) associé à l'espace de Sobolev $W^m[0,1]$, on retrouve le caractère géométrique de la pénalité, c'est-à-dire $\int_0^1 (f^{(m)}(t))^2 dt = \|P_1 f\|^2_{\mathcal{H}_R}$. On peut alors appliquer le théorème du représentant de Kimeldorf et Wahba (théorème 3.5) et exprimer l'estimateur par spline de lissage $\widehat{f}_\lambda$ comme une combinaison linéaire de la base de $\mathcal{H}_0$ et des représentants de $\mathcal{H}_1$:

$$\widehat{f}_\lambda(t) = \sum_{\nu=0}^{m-1} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i R_1(t_i, t). \tag{3.64}$$

Remarque 3.6. Dans cet exemple, nous avons muni l'espace de Sobolev $W^m[0,1]$ de la norme définie en (3.55). Plusieurs autres normes équivalentes (au sens topologique) peuvent être utilisées dans cet espace, chaque norme déterminant son propre noyau reproduisant. On peut citer notamment la norme

$$\|f\|^2 = \sum_{\nu=0}^{m-1} (\int_0^1 f^{(\nu)}(t)dt)^2 + \int_0^1 (f^{(m)}(t))^2 dt. \tag{3.65}$$

On montre alors que l'espace de Sobolev $W^m[0,1]$ muni cette norme est un RKHS. De plus, soit $k_r(t) = B_r(t)/r!$ les polynômes de Bernoulli normalisés où les $B_r,\ r = 1, 2, \ldots$ sont définis récursivement par $B_0(t) = 1$, $B_r'(t) = rB_{r-1}(t)$ et $\int_0^1 B_r(t)dt = 0$ pour $r = 1, 2, \ldots$ (voir Abramowitz et Stegun [2]). Alors on a $W^m[0,1] = \mathcal{H}_0 \bigoplus \mathcal{H}_1$, où

$$\mathcal{H}_0 = vect\{k_0(x), k_1(x), \ldots, k_{m-1}(x)\} \tag{3.66}$$

et

$$\mathcal{H}_1 = \{f : [0,1] \to \mathbb{R} \text{ t.q. } \int_0^1 f^{(\nu)}(x)dx = 0,\ \nu = 0, ..., m-1, \quad f^{(m)} \in L^2([0,1])\} \tag{3.67}$$

sont des RKHS de noyaux reproduisants respectifs

$$R_0(s,t) = \sum_{\nu=0}^{m-1} k_\nu(s)k_\nu(t) \tag{3.68}$$

et

$$R_1(s,t) = k_m(s)k_m(t) + (-1)^{m-1}k_{2m}(|s-t|). \tag{3.69}$$

Cette nouvelle construction des splines polynomiales de lissage est détaillée dans Craven et Wahba [33] et dans Gu [72] section 2.3.3. Notons que les coefficients c_i et d_ν de l'estimateur spline de lissage seront alors différents de ceux obtenus en utilisant les ϕ_ν et R_1 definis en (3.61) et (3.60), mais que l'estimateur

$$\widehat{f}_\lambda(t) = \sum_{\nu=0}^{m-1} d_\nu\phi_\nu(t) + \sum_{i=1}^{n} c_i R_1(t_i, t)$$

sera le même.

3.4 Conclusion

Ce chapitre a montré l'intérêt des méthodes de lissage utilisant les fonctions splines par rapport aux méthodes du noyau ou des polynômes locaux. En effet, l'utilisation des splines permet de garder le caractère fonctionnel des données et de fournir une expression explicite de la fonction de lissage facilitant ainsi l'évaluation

en tout point de l'intervalle étudié et le calcul des dérivées. Les splines de lissage résultant d'un problème d'optimisation dans l'espace de Sobolev sont particulièrement attractives puisque le degré de lissage de la fonction de régression est contrôlé par un unique paramètre. Les performances de l'estimateur par spline de lissage ont notamment été montré pour l'estimation des profils temporels de vitesse et d'accélération. Nous avons entre autre montré sa robustesse aux valeurs aberrantes contrairement aux estimateurs à noyau et par polynômes locaux.

Nous avons également présenté le principe de régularisation qui consiste à restreindre l'espace fonctionnel de recherche par l'ajout d'un terme de pénalité et qui permet de résoudre des problèmes dits mal posés. Un cas particulier important des problèmes de régularisation est celui où l'on se place dans un espace de Hilbert à noyau reproduisant (RKHS). Dans ce cas, le terme de pénalité est défini en terme de noyau défini positif, et la régularité de la solution est contrôlée par le choix du noyau. Cette généralisation du problème des splines de lissage permet de traiter de manière unifiée de nombreux types de splines de part le choix du RKHS comme espace du modèle, de sa décomposition en somme directe, et de la pénalisation, ce qui rend cette méthode très flexible. De plus, nous avons vu que l'estimateur par spline de lissage appartient à un espace de dimension finie et s'exprime comme une combinaison linéaire des fonctions noyaux. Nous verrons au chapitre suivant que le cadre général des splines polynomiales de lissage permet notamment de traiter le cas de l'estimation d'une fonction de lissage en utilisant de l'information sur sa dérivée, et est particulièrement approprié à l'estimation des profils spatiaux de vitesse.

Chapitre 4

Lissage sous contraintes de profils spatiaux de vitesse

Ce chapitre est consacré à l'étape de lissage de profils spatiaux de vitesse à partir de données bruitées de position et de vitesse. En effet, nous avons vu à la section 1.3.4 que la précision des données de vitesse et de position était très importante dans l'étude des profils de vitesse. Cependant, les données issues de capteurs n'étant pas toujours très précises, il est nécessaire de chercher à minimiser les erreurs de mesures en calculant un estimateur du "vrai" profil spatial de vitesse.

Nous avons vu au chapitre précédent que cette étape de lissage consistait à se ramener à un problème de régression non paramétrique où l'on cherche à estimer la fonction de régression. Nous avons notamment montré que les méthodes de lissage basées sur les splines permettaient de convertir les données brutes de nature vectorielle en objet fonctionnel, et de se placer ainsi dans le cadre de l'analyse des données fonctionnelles. De plus, nous avons montré les bonnes performances des splines de lissage pour l'estimation des profils temporels de vitesse et d'accélération.

Dans ce chapitre, on s'interesse uniquement aux profils spatiaux de vitesse pour lesquels une modélisation fonctionnelle a été définie au chapitre 2. Dans une première section, nous proposons d'améliorer la qualité des mesures de position en construisant plusieurs estimateurs fusionnant les informations issues des deux principaux capteurs de localisation, à savoir l'odomètre et le GPS.

Après cette étape optionnelle de pré-traitement des données de position, nous nous intéressons à l'étape de lissage de profils spatiaux de vitesse à partir d'observations bruitées de la position et de la vitesse du véhicule. Cependant, nous verrons dans la deuxième section de ce chapitre que, suite à différentes contraintes difficiles à prendre en compte dans l'estimation directe d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse, il est préférable dans un premier temps de changer d'espace d'étude, et de se ramener à l'estimation de la distance parcourue en fonction du temps $F(t)$ à partir de ces mêmes mesures bruitées.

Nous verrons alors que ce changement d'espace d'étude nous amène à un problème

de régression non paramétrique sous les deux contraintes suivantes :

(C_1) Estimer la fonction de régression $F(t)$ à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).

(C_2) Une contrainte de monotonie sur F.

La prise en compte de ces deux contraintes est l'objet de la troisième section pour la contrainte (C_1), et de la quatrième section pour la contrainte (C_2).

4.1 Étape de pré-traitement des données : estimation par fenêtre glissante de la position du véhicule à partir de l'odomètre et du GPS

Nous proposons dans cette section une étape optionnelle de pré-traitement des données, afin d'améliorer la qualité des mesures de position lorsque l'on dispose à la fois des mesures GPS et des mesures odométriques. En effet, si l'estimation des profils spatiaux de vitesse nécessite d'estimer à la fois la position et la vitesse du véhicule en chaque instant, en pratique, les mesures de vitesses sont généralement plus précises que les mesures de position. Or, il est nécessaire de connaître précisément la position du véhicule à un instant donné car une erreur de positionnement peut avoir de graves conséquences dans une étude sur les profils de vitesse. La précision des mesures de position varie en fonction des capteurs utilisés. Par exemple, la figure 4.1 montre un exemple de profil spatial de vitesse obtenu à partir de différents capteurs : un GPS différentiel RTK (mesures de position avec une précision centimétrique), un GPS classique et un odomètre (voir annexe A pour plus d'informations sur ces capteurs). Cette figure met en évidence la dérive de l'odomètre (mesures collectées sur le bus CAN) au cours du trajet due à une accumulation d'erreurs dans le calcul de la distance parcourue. Ainsi, sur un trajet de longueur totale 4 km, on observe un décalage de 32 m de la distance parcourue mesurée par l'odomètre par rapport à celle mesurée par le GPS RTK à la fin du trajet.

Afin de pallier ce problème, nous proposons dans une étape de pré-traitement des données précédant l'étape de lissage, d'améliorer la qualité des mesures de position en proposant un estimateur ponctuel de la position du véhicule en chaque instant d'échantillonnage à partir des deux principaux capteurs de localisation, à savoir l'odomètre et le GPS, et tenant compte des avantages et des inconvénients de ces deux capteurs. Nous avons en fait construit plusieurs estimateurs basés sur un modèle statistique simple des erreurs de ces deux capteurs, et permettant une généralisation de ce modèle à d'autres applications. Le détail sur la construction de ces estimateurs ainsi que leur étude sur des données simulées et réelles a donné lieu à un article (actuellement soumis à la revue "Navigation : Journal of the Institute of Navigation" et en attente d'acceptation) et dont l'original est présenté dans l'an-

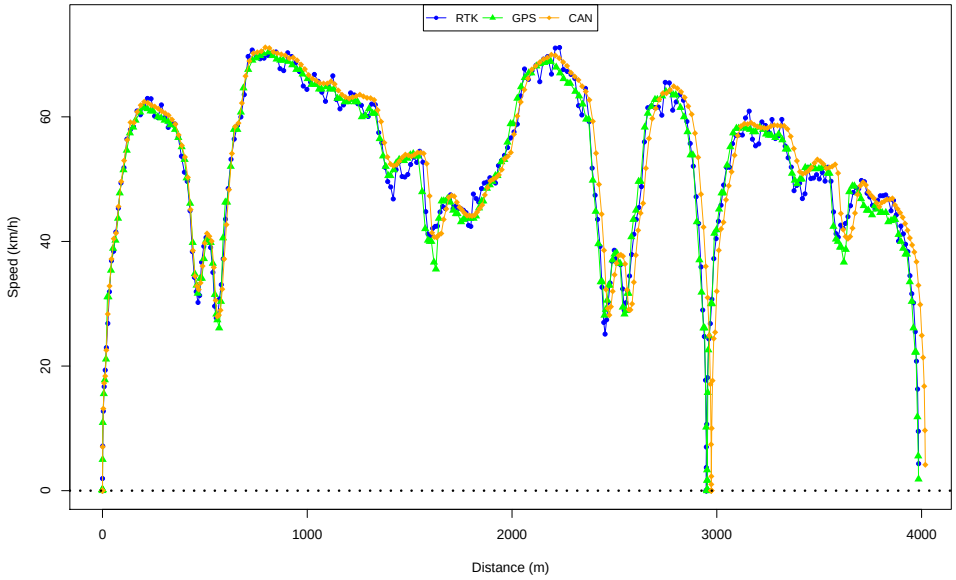


Figure 4.1 - Mise en évidence de la dérive de l'odomètre sur un profil spatial de vitesse.

nexe B. Nous donnons ici un simple résumé des résulats présentés en annexe.

Dans un premier temps, nous nous sommes placés dans une approche temps réel. Nous avons construit un estimateur sans biais de variance asymptotique minimale correspondant à un filtre de Kalman avec un gain fixe qui est optimal à l'infini. Si cet estimateur récursif présente une moins bonne précision que le filtre de Kalman au début de l'intervalle étudié, sa vitesse de convergence est relativement rapide et son temps de calcul est deux fois plus rapide que le filtre de Kalman. Puis nous avons construits deux estimateurs utilisant uniquement les mesures situées dans une fenêtre glissante de largeur fixé. Le choix de la largeur de la fenêtre de lissage rend ces estimateurs très flexibles par rapport à un estimateur récursif comme le filtre de Kalman. En outre, des études sur données simulées et réelles ont montrés leurs bonnes performances dans l'estimation de la position du véhicule.

Dans un second temps, nous nous sommes placés en phase de post-traitement des données pour l'estimation de la position à un instant t et nous avons étendu les deux estimateurs à fenêtre glissante précédents en ajoutant les observations obtenues après l'instant t. Les résultats montrent une amélioration dans la précision de l'estimation de la position par rapport au cas temps réel.

Notons que dans la suite de ce manuscrit, nous n'avons pas utilisé cette étape de

pré-traitement des données. En effet, l'étude de profils spatiaux de vitesse nécessite une seule source de mesures de position qui peut être soit un odomètre, soit un GPS, soit une fusion des ces deux capteurs (et qui est l'objet de l'étape de pré-traitement décrite dans l'annexe B). Cependant, en pratique, on dispose généralement d'un seul capteur de position. Ainsi, par la suite nous avons choisi d'utiliser uniquement les mesures GPS (mesures disponibles sur un smartphone) en terme de données brutes de position utilisées pour le calcul de l'estimateur. Lorsqu'elles sont disponibles, les mesures odométriques seront simplement données à titre d'information.

4.2 Contraintes sur les profils spatiaux de vitesse

4.2.1 Modèle de régression non paramétrique

Nous avons vu au chapitre 1, l'importance de la connaissance des vitesses pratiquées sur l'ensemble du réseau. Cette information est accessible avec la généralisation des véhicules traceurs qui sont capables de transmettre leur position et leur vitesse au cours du temps. Cependant, ces mesures de position et de vitesse étant généralement entachées d'erreurs, il est nécessaire avant toute étude de procéder à une étape de lissage. Cette étape consiste à minimiser les erreurs de mesures en calculant un estimateur du "vrai" profil spatial de vitesse. On se ramène alors à un problème de régression non paramétrique, dont le modèle peut s'écrire sous la forme suivante :

$$y_i = v_S(x_i) + \varepsilon_{v,i}, \quad i = 1, \dots, n, \tag{4.1}$$

où les y_i sont des mesures bruitées de vitesse issues de capteurs (mesures dérivées de l'odomètre, ou mesures issues du GPS et calculées par effet Doppler) correspondant respectivement aux positions x_i du véhicule, et les $\varepsilon_{v,i}$ sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ_v^2, correspondant aux erreurs de mesures. La fonction v_S est la fonction de régression que l'on cherche à estimer.

Cependant, les positions x_i du véhicule étant également des mesures issues de capteurs et donc entachées d'erreurs, les observations de la variable explicative X sont en fait les mesures w_i obtenues aux instants t_i, $i = 1, \dots, n$, telles que

$$w_i = x(t_i) + \varepsilon_{x,i}, \quad i = 1, \dots, n, \tag{4.2}$$

où les $\varepsilon_{x,i}$ sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ_x^2, correspondant aux erreurs de mesures de la position du véhicule. L'objectif est donc d'estimer le vrai profil spatial de vitesse v_s à partir des observations (y_i, w_i), $i = 1, \dots, n$, de vitesse et de position.

Une méthode classique consiste dans un premier temps à débruiter les observations x_i de position à l'aide d'une technique de lissage, puis à revenir au problème de régression non paramétrique (4.1). Nous avons vu aux deux chapitres précédents, l'intérêt de se placer dans un cadre fonctionnel et d'utiliser des méthodes de lissage telles que les splines pour estimer la fonction de régression v_S. Cependant, l'étude

des propriétés des profils spatiaux de vitesse réalisée au chapitre 2, nous amène à prendre en compte certaines contraintes dans l'estimation de la fonction de régression v_S. En effet, nous avons vu à la section 2.2.2.3 que les profils spatiaux de vitesse n'étaient pas dérivables aux points où la dérivée s'annule. Cette propriété doit donc être vérifiée par l'estimateur $\widehat{v}_S$ du vrai profil v_S. Cette propriété est une contrainte difficile à prendre en compte dans le lissage et pose le problème de la modélisation des arrêts du véhicule dans l'espace `vitesse`×`distance`. Par exemple, l'utilisation de fonctions splines dans l'espace `vitesse`×`distance` n'est pas appropriée pour estimer un profil spatial de vitesse puisque dans ce cas, l'estimateur obtenu $\widehat{v}_S$ ne vérifiera pas cette propriété. On propose alors de pallier cette difficulté en changeant d'espace d'étude, et en se ramenant à l'espace `distance`×`temps` afin d'estimer la fonction $F(t)$ représentant la distance parcourue au cours du temps.

4.2.2 Changement d'espace d'étude

Nous avons vu au chapitre 2 qu'un profil spatial de vitesse était une succession de mesures horodatées de position et de vitesse, et pouvait donc être manipulé dans les 3 espaces `distance`×`temps`, `vitesse`×`temps` ou `vitesse`×`distance`, comme l'illustre la figure 4.2.

On propose dans un premier temps de se placer dans l'espace `distance`×`temps` et de chercher à estimer la fonction $F(t)$ représentant la distance parcourue par le véhicule au cours du temps. On se ramène alors au problème de régression non paramétrique suivant :

$$y_i = F(t_i) + \varepsilon_{x,i}, \quad i = 1, \dots, n, \tag{4.3}$$

où les y_i sont les mesures bruitées de la distance parcourue du véhicule (issues d'un odomètre ou d'un GPS après map-matching (voir annexe A)) obtenues en chaque instant t_i, et les $\varepsilon_{x,i}$ sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ_x^2, correspondant aux erreurs de mesures. Notons que ce modèle est en fait une réécriture du modèle (4.2) énoncé à la section précédente. L'avantage de se placer dans l'espace `distance`×`temps` est que la seule contrainte de forme à prendre en compte dans l'estimation de F est une contrainte de monotonie. En effet, la fonction F représentant la distance parcourue en fonction du temps est une fonction croissante. Nous verrons dans la suite de ce chapitre que de nombreuses méthodes de lissage sous contrainte de monotonie ont été développées, et que cette contrainte de monotonie est moins difficile à prendre en compte que la contrainte de non dérivabilité aux points où la fonction s'annule.

De plus, après avoir effectué cette étape de lissage dans l'espace `distance`×`temps`, il est très simple de se ramener dans l'espace `vitesse`×`distance` et d'en déduire une estimation du profil spatial de vitesse. En effet, supposons que $\widehat{F}$ soit un estimateur de F tel que $\widehat{F}$ soit une fonction $\mathcal{C}^2$ et croissante, alors on en déduit par dérivation un estimateur $\widehat{F}'$ du profil de vitesse F' en fonction du temps (espace `vitesse`×`temps`).

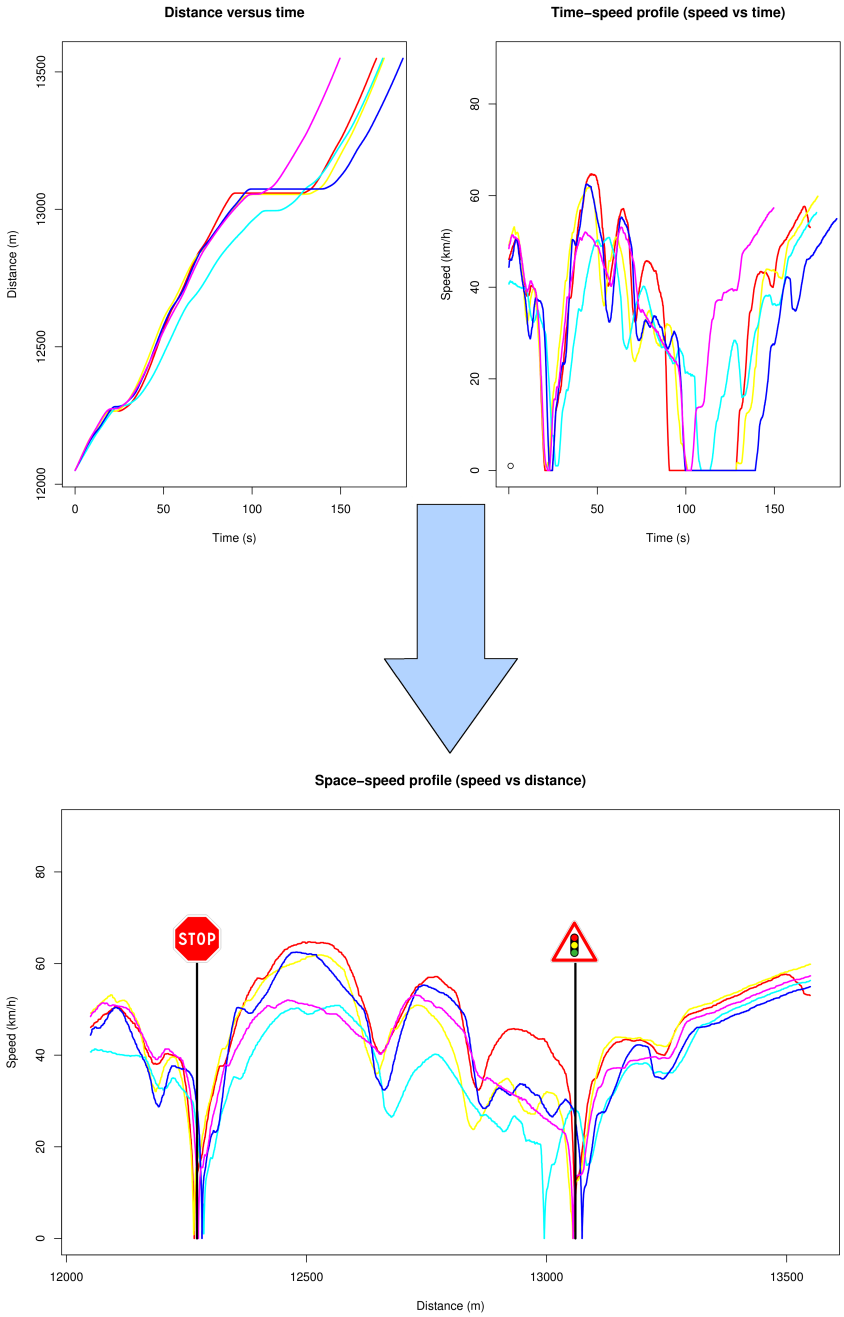


FIGURE 4.2 - Lien entre les trois espaces : [`distance`×`temps`, `vitesse`×`temps`] et [`vitesse`×`distance`].

Puis, d'après la définition 2.1, on en déduit un estimateur $\widehat{v_S}$ du profil spatial de vitesse v_S par la transformation suivante : $\widehat{v_S} = \widehat{F}' \circ \widehat{F}^{-1}$. On garde ainsi la cohérence entre les trois espaces illustrés à la figure 4.2.

Cependant, si cette méthode consistant à trouver un estimateur $\widehat{F}$ de F à partir des mesures bruitées y_i de position, puis à en déduire un estimateur $\widehat{v_S} = \widehat{F}' \circ \widehat{F}^{-1}$ du profil spatial de vitesse v_S est correcte, celle-ci ne tient pas compte des mesures de vitesse également disponibles. En effet, en pratique, on dispose également de mesures de vitesse du véhicule qui peuvent être indépendantes des mesures de position. Par exemple, les GPS actuels fournissent à la fois des mesures de la position du véhicule (latitude, longitude, altitude), mais également des mesures de la vitesse du véhicule calculées par effet Doppler (et donc indépendantes des mesures de position). De plus, si l'on dispose d'un boîtier enregistreur connecté au bus CAN du véhicule, on peut obtenir des mesures de la vitesse du véhicule qui dérivent des mesures de distance de l'odomètre et qui sont donc indépendantes des positions GPS. Plus de détails sur les différents capteurs relatifs à la position et à la vitesse du véhicule sont données dans l'annexe A. Ainsi, lorsque l'on dispose à la fois de mesures de position et de vitesse et que ces mesures sont indépendantes l'une de l'autre, il est nécessaire d'utiliser ces deux sources d'information pour estimer au mieux le profil spatial de vitesse. De plus, la précision des mesures de vitesse est en pratique meilleure que celle des mesures de position.

On propose donc d'estimer la fonction F représentant la distance parcourue par le véhicule au cours du temps en utilisant à la fois des observations bruitées de la position et de la vitesse du véhicule. Cependant, la vitesse d'un mobile en un instant t correspondant à la dérivée de sa position en t, le problème revient à estimer la fonction de régression F du modèle (4.3) en utilisant de l'information sur sa dérivée F'. On se ramène donc au modèle de régression non paramétrique (4.3) sous les deux contraintes suivantes :

(C_1) Estimer la fonction de régression $F(t)$ à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).

(C_2) Une contrainte de monotonie sur F.

La prise en compte de ces deux contraintes est l'objet de la suite de ce chapitre.

4.3 Lissage en utilisant les observations sur la dérivée

L'objet de cette section est la prise en compte de la contrainte (C_1) dans le modèle (4.3). On cherche à estimer la distance parcourue, notée $F(t)$, d'un véhicule en fonction du temps à partir de deux types d'observations :

- d'une part, des observations bruitées y_i $(i = 1, \ldots, n)$ de la distance parcourue F du véhicule aux instants t_i $(i = 1, \ldots, n)$ fournies par un capteur de position

(odomètre ou GPS), ou par l'un des estimateurs de position construits à la section 4.1 ;
– d'autre part, des observations bruitées v_i $(i = 1, \ldots, n)$ de la vitesse du véhicule (i.e. la dérivée $F'(t)$ de la position par rapport au temps) aux instants t_i $(i = 1, \ldots, n)$ indépendantes des observations y_i (par exemple, vitesse Doppler indépendante de la position GPS).

La fonction représentant la vitesse en fonction du temps étant la dérivée première de la fonction représentant la distance parcourue au cours du temps, ceci revient au problème de régression non paramétrique suivant : estimer une fonction de régression à partir d'observations bruitées de celle-ci et de sa dérivée.

Le problème d'estimation d'une fonction de régression à partir d'observations bruitées de cette fonction et d'observations bruitées de ses dérivées apparaît dans divers domaines d'applications comme l'économie (Hall et Yatchew [74]) ou la biologie moléculaire (Calderon *et al.* [27]). Si Hall et Yatchew [74] propose d'utiliser un estimateur à noyau, l'utilisation des splines pour résoudre ce type de problème a également été abordée dans la littérature. Ainsi, Cox [32] propose une analyse asymptotique d'estimateurs obtenus par méthode de régularisation ("*Method of Regularization (MOR) estimators*"), et étudie différents exemples dont le cas des splines de lissage lorsque l'on dispose d'observations sur la dérivée (exemple 0.4 de l'article). L'objectif de cet article est essentiellement d'obtenir des approximations pour les premier et second moments de la norme de l'erreur d'estimation. Mardia *et al.* [106] utilisent les splines d'interpolation en ajoutant des contraintes linéaires sur les dérivées et se basent sur le lien entre spline et krigeage. Une application au problème de la déformation par landmarks dans $\mathbb{R}^2$ et $\mathbb{R}^3$ à partir d'information sur les dérivées telle que les tangentes ou les courbures est donnée. Enfin, plus récemment, Calderon *et al.* [27] proposent une méthode basée sur les splines pénalisées, "*the P-splines using Derivative Information (PuDI) method*", et l'applique à l'estimation d'équations différentielles stochastiques non linéaires caractérisant la dynamique d'une molécule.

Nous proposons de résoudre ce problème en utilisant les splines de lissage. L'avantage de l'estimation par splines de lissage est que celle-ci nécessite de déterminer un unique paramètre, à savoir le paramètre de lissage λ, contrairement aux méthodes basées sur les splines pénalisées, telles que celle développée par Calderon *et al.* [27], pour lesquelles le choix du nombre de noeuds peut être difficile. Nous proposons de nous placer dans le cadre d'un problème de régression régularisée dans un espace de Hilbert à noyau reproduisant (RKHS), et de se ramener à un cas particulier du modèle général des splines de lissage présenté au chapitre 3. La forme de l'estimateur est donnée dans Mardia *et al.* [106] dans le cas des splines d'interpolation, mais son écriture utilise la théorie du krigeage et le lien avec la théorie des splines. Les auteurs évoquent qu'il est possible de réécrire l'estimateur en utilisant la théorie des RKHS et c'est ce que nous proposons dans cette section. Notons que Cox [32] se place également dans le cadre de la théorie des RKHS, mais donne une écriture très générale

de l'estimateur par spline de lissage sous forme d'une combinaison linéaire des observations (estimateur linéaire), et explicite donc uniquement la matrice chapeau. Nous allons voir que la théorie des RKHS permet d'obtenir une forme explicite de l'estimateur et que celui-ci peut s'écrire comme une combinaison linéaire de fonctions de base et de fonctions noyaux.

4.3.1 Estimateur par spline de lissage

4.3.1.1 Forme de l'estimateur

On cherche à estimer une fonction f à partir d'observations bruitées de f et de sa dérivée f'. On modélise alors le problème sous la forme suivante :

$$\begin{cases} y_{1,i} = f(t_i) + \varepsilon_{1,i}, & i = 1, \dots, n \\ y_{2,i} = f'(t_i) + \varepsilon_{2,i}, & i = 1, \dots, n \end{cases}, \tag{4.4}$$

où les $\varepsilon_{1,1}, \dots, \varepsilon_{1,n}, \varepsilon_{2,1} \dots, \varepsilon_{2,n}$ sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle, et telles que $Var(\varepsilon_{1,i}) = \sigma_1^2$ et $Var(\varepsilon_{2,i}) = \sigma_2^2$ pour $i = 1, \dots, n$. Afin de simplifier le problème, on suppose que les observations $y_{1,i}$ et $y_{2,i}$ sont obtenues aux mêmes instants t_i. Dans le cas contraire, on procédera à un rééchantillonnage des données pour se ramener à ce cas.

On suppose que le domaine de f est $\mathcal{X} = [a, b]$ et que $f \in W^m[a, b]$ où $W^m[a, b]$ est l'espace de Sobolev d'ordre m (m étant un entier strictement positif) sur $[a, b]$. Nous avons vu au chapitre 3 que l'espace de Sobolev était un espace de Hilbert à noyau reproduisant (RKHS). On note $\mathcal{H}_R = W^m[0, T]$ et $R(s, t)$ le noyau reproduisant associé. Le modèle (4.4) est alors un cas particulier du modèle général des splines de lissage énoncé à la section 3.3.3.2 et peut se réécrire sous la forme

$$y_j = \mathcal{L}_j f + \varepsilon_j \ , \quad j = 1, \dots, 2n, \tag{4.5}$$

où

- les observations y_j sont définies par :
  $\begin{cases} y_j = y_{1,i} \ \text{ avec } \ i = j \ \text{ pour } \ j = 1, \dots, n \\ y_j = y_{2,i} \ \text{ avec } \ i = j - n \ \text{ pour } \ j = n + 1, \dots, 2n \end{cases}$,
- les $\mathcal{L}_j$ sont des formes linéaires bornées sur $W^m[a, b]$ définies par :
  $\begin{cases} \mathcal{L}_j f = f(t_i) \ \text{ avec } \ i = j \ \text{ pour } \ j = 1, \dots, n \\ \mathcal{L}_j f = f'(t_i) \ \text{ avec } \ i = j - n \ \text{ pour } \ j = n + 1, \dots, 2n \end{cases}$,
- les ε_j sont définies par :
  $\begin{cases} \varepsilon_j = \varepsilon_{1,i} \ \text{ avec } \ i = j \ \text{ pour } \ j = 1, \dots, n \\ \varepsilon_j = \varepsilon_{2,i} \ \text{ avec } \ i = j - n \ \text{ pour } \ j = n + 1, \dots, 2n \end{cases}$.

L'hypothèse selon laquelle les $\mathcal{L}_j$ sont des formes linéaires bornées sur $W^m[a, b]$ peut

se réécrire plus simplement en utilisant le théorème suivant énoncé dans Wahba et Wendelberger [175] et Berlinet et Thomas-Agnan [15] (Th. 133) :

Théorème 4.1. *Soit Ω un sous-ensemble ouvert de $\mathbb{R}^d$, $d > 0$, et soit $H^m(\Omega)$ l'espace de Sobolev d'ordre m, $m > 0$, sur Ω. L'opérateur $\mathcal{L}$ défini par $\mathcal{L}u = \frac{\partial^k u}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_d} x_d}$ pour $\alpha_1 + \ldots + \alpha_d = k$ $(k, \alpha_1, \ldots, \alpha_d \in \mathbb{N})$ est une forme linéaire continue sur $H^m(\Omega)$ si et seulement si $2m - 2k - d > 0$.*

Dans notre cas, $d = 1$ et $k = 0$ pour $j = 1, \ldots, n$ (puisque $\mathcal{L}_j f = f(t_i)$ pour $j = 1, \ldots, n$), et $k = 1$ pour $j = n+1, \ldots, 2n$ (puisque $\mathcal{L}_j f = f'(t_i)$ pour $j = n+1, \ldots, 2n$). On en déduit que les $\mathcal{L}_j$ sont des formes linéaires bornées sur $W^m[a, b]$ si et seulement si $m > 1$. On supposera donc dans la suite de ce chapitre que $m > 1$.

On se place ensuite dans le cadre d'un problème de régression régularisée dans un RKHS. On obtient alors un estimateur de f par minimisation du critère suivant dans l'espace de Sobolev $\mathcal{H}_R = W^m[a, b]$:

$$\frac{1}{2n} \sum_{j=1}^{2n} \sigma_j^{-2} (y_j - \mathcal{L}_j f)^2 + \lambda \int_a^b (f^{(m)}(t))^2 dt, \tag{4.6}$$

où σ_j^2 est la variance des ε_j pour $j = 1, \ldots, 2n$, et où $\lambda \in \mathbb{R}^+$ est le paramètre de lissage. On suppose ici que $m > 1$ et $\lambda > 0$ sont fixés. Le choix de ces deux paramètres sera abordé dans la suite de ce chapitre.

En utilisant les notations du modèle (4.5), on peut réécrire ce critère sous la forme :

$$\frac{1}{2n} \{\sigma_1^{-2} \sum_{i=1}^{n} (y_{1,i} - f(t_i))^2 + \sigma_2^{-2} \sum_{i=1}^{n} (y_{2,i} - f'(t_i))^2\} + \lambda \int_a^b (f^{(m)}(t))^2 dt. \tag{4.7}$$

On décompose l'espace de Sobolev $\mathcal{H}_R = W^m[a, b]$ en somme directe de la forme :

$$\mathcal{H}_R = \mathcal{H}_0 \bigoplus \mathcal{H}_1$$

où $\mathcal{H}_0$ est un espace de dimension finie $m \leq n$, et on note R_0 et R_1 les noyaux reproduisants associés respectivement à H_0 et H_1. Alors la proposition suivante, qui découle du théorème de Kimeldorf et Wahba (Th. 3.5), donne une solution explicite au problème de minimisation dans $\mathcal{H}_R$ du critère (4.6).

Proposition 4.1. *Soit $\phi_1, \ldots, \phi_m$ une base de l'espace nul $\mathcal{H}_0$ et soit T une matrice $2n \times m$ de plein rang définie par :*

$$T = \{\mathcal{L}_j \phi_\nu\}_{j=1}^{2n} {}_{\nu=1}^{m}.$$

Alors le critère (4.6) a un unique minimum dans $W^m[a,b]$ donné par

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i R_1(t_i,t) + \sum_{i=1}^{n} c'_i \frac{\partial}{\partial s} R_1(s,t)|_{s=t_i} \tag{4.8}$$

où

$$\begin{aligned} \Sigma &= \{\mathcal{L}_{j(s)}\mathcal{L}_{k(t)}R_1(s,t)\}_{j,k=1}^{2n}, \\ W &= diag(\sigma_1^{-2},\ldots,\sigma_{2n}^{-2}), \\ M &= \Sigma + 2n\lambda W^{-1}, \\ d = (d_1,\ldots,d_m)^T &= (T^TM^{-1}T)^{-1}T^TM^{-1}y, \\ c = (c_1,\ldots,c_n,c'_1,\ldots,c'_n)^T &= M^{-1}\{I - T(T^TM^{-1}T)^{-1}T^TM^{-1}\}y. \end{aligned} \tag{4.9}$$

Démonstration. Le critère (4.6) peut se réécrire sous la forme du critère des moindres carrés pénalisés ("*Penalized Weighted Least Squares*", PWLS) (voir Wang [182] section 5.2.1) suivant :

$$\frac{1}{2n}\sum_{j=1}^{2n} w_j (y_j - \mathcal{L}_j f)^2 + \lambda \|P_1 f\|^2_{\mathcal{H}_R} \tag{4.10}$$

où $\begin{cases} w_j = \sigma_j^{-2} = \sigma_1^{-2} & \text{pour } \ j = 1,\ldots,n \\ w_j = \sigma_j^{-2} = \sigma_2^{-2} & \text{pour } \ j = n+1,\ldots,2n \end{cases}$, et P_1 est la projection orthogonale sur $\mathcal{H}_1$. Soit η_j le représentant de $\mathcal{L}_j$ dans $\mathcal{H}_R$ tel que $\mathcal{L}_j f = \langle \eta_j, f\rangle_{\mathcal{H}_R}$. Alors, comme $T = \{\mathcal{L}_j\phi_\nu\}_{\substack{j=1\\ \nu=1}}^{\substack{2n\\ m}}$ est de rang plein, on en déduit par application du théorème de Kimeldorf et Wahba (Th. 3.5), que le critère (4.6) a un unique minimum donné par

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{j=1}^{2n} c_j \xi_j(t) \tag{4.11}$$

où

$$\begin{aligned} \xi_j &= P_1\eta_j, \\ \Sigma &= \{\langle \xi_j, \xi_k\rangle\}_{j,k=1}^{2n}, \\ W &= diag(w_1,\ldots,w_{2n}), \\ M &= \Sigma + 2n\lambda W^{-1}, \\ d = (d_1,\ldots,d_m)^T &= (T^TM^{-1}T)^{-1}T^TM^{-1}y, \\ c = (c_1,\ldots,c_{2n})^T &= M^{-1}\{I - T(T^TM^{-1}T)^{-1}T^TM^{-1}\}y. \end{aligned}$$

Or, nous avons montré à la remarque 3.5 que $\Sigma = \{\langle \xi_j,\xi_k\rangle\}_{j,k=1}^{2n} = \{\mathcal{L}_{j(s)}\mathcal{L}_{k(t)}R_1(s,t)\}_{j,j=1}^{2n}$, et que $\xi_i(s) = \mathcal{L}_{i(t)}R_1(s,t)$ où $\mathcal{L}_{i(t)}$ signifie que $\mathcal{L}_i$ est appliqué à ce qui suit et qui est considéré comme une fonction de t. Le minimum (4.11) peut donc se réécrire sous la forme :

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{j=1}^{2n} c_j \mathcal{L}_{j(s)} R_1(s,t)$$

Et comme $\begin{cases} \mathcal{L}_j f = f(t_i) \quad \text{avec} \quad i = j \quad \text{pour} \quad j = 1, \ldots, n \\ \mathcal{L}_j f = f'(t_i) \quad \text{avec} \quad i = j - n \quad \text{pour} \quad j = n+1, \ldots, 2n \end{cases}$, on en déduit l'écriture (4.8). □

Remarque 4.1. Nous avons supposé ici l'indépendance des erreurs ε_j pour $j = 1, \ldots, 2n$, cependant le modèle général des splines de lissage peut être généralisé au cas où les erreurs sont corrélées (voir Wang [182] (chap. 5) pour plus de détails).

4.3.1.2 Sélection du paramètre λ

Nous avons supposé à la section précédente que le paramètre de lissage λ était fixé. En pratique, si le paramètre de lissage λ peut être obtenu par essais et erreurs ("*trial and error method*"), il peut également être obtenu par des méthodes automatiques basées sur la minimisation d'un critère spécifique (voir section 3.2.2.4). Les critères les plus couramment utilisés sont :

- le critère UBR ("*UnBiased Risk*") qui est une extension du critère du C_p de Mallows ;
- le critère GCV ("*Generalized Cross-Validation*") qui est une version pondérée du critère de validation croisée ;
- le critère GML ("*Generalized Maximum Likelihood*") basé sur un modèle bayésien.

On se place dans le cadre du modèle général des splines de lissage décrit à la section 3.3.3.2 et on note $A(\lambda)$ la matrice chapeau telle que

$$(\mathcal{L}_1 \widehat{f}, \ldots, \mathcal{L}_n \widehat{f})^T = A(\lambda) y.$$

Les méthodes UBR, GCV et GML d'estimation du paramètre de lissage λ consistent respectivement à minimiser la fonction UBR

$$UBR(\lambda) = \frac{1}{n} y^T (I - A(\lambda))^2 y + \frac{2\sigma^2}{n} trA(\lambda),$$

la fonction GCV

$$GCV(\lambda) = \frac{\frac{1}{n} y^T (I - A(\lambda))^2 y}{[\frac{1}{n} tr(I - A(\lambda))]^2},$$

et la fonction GML

$$GML(\lambda) = \frac{y^T (I - A(\lambda)) y}{[det^+ \{(I - A(\lambda))\}]^{\frac{1}{n-p}}},$$

où det^+ correspond au produit des valeurs propres non nulles, et p est la dimension de l'espace nul H_0.

Notons que la méthode UBR nécessite la connaissance de la variance σ^2 des erreurs. De manière générale, les méthodes GCV et GML présentent des résultats

similaires en terme de performance dans l'estimation du paramètre λ. On peut simplement noter que la méthode GCV peut conduire à une estimation du paramètre λ proche de zéro, et donc à une interpolation des données, lorsque la taille de l'échantillon est petite (Wahba et Wang [174]). Au contraire, dans le cas où la taille de l'échantillon est grande, la méthode GCV présente de meilleures performances que la méthode GML. Une comparaison des méthodes GCV et GML est donnée dans Wahba [171].

Remarque 4.2. Dans le cas où les erreurs aléatoires sont de variances inégales et/ou sont corrélées, i.e. $Cov(\varepsilon) = \sigma^2 W^{-1}$, des versions étendues des méthodes UBR, GCV et GML ont été développées dans Wang [181].

4.3.1.3 Le problème numérique du calcul des vecteurs de coefficients c et d

Le calcul de l'estimateur de l'estimateur $\widehat{f}_\lambda(t)$ énoncé en (4.8) est essentiellement basé sur le calcul des vecteurs de coefficients c et d définis par

$$d = (d_1, \ldots, d_m)^T = (T^T M^{-1} T)^{-1} T^T M^{-1} y, \tag{4.12}$$

$$c = (c_1, \ldots, c_n, c'_1, \ldots, c'_n)^T = M^{-1}\{I - T(T^T M^{-1} T)^{-1} T^T M^{-1}\} y. \tag{4.13}$$

Cependant, nous avons vu au chapitre précédent que ces équations n'étaient pas appropriées au calcul numérique et que l'on utilisait plutôt les équations

$$(\Sigma + 2n\lambda W^{-1})c + Td = y \tag{4.14}$$

$$T^T c = 0. \tag{4.15}$$

Divers algorithmes peuvent alors être utilisés pour résoudre ce système. Les routines RKPACK développées en FORTRAN par Gu [71] (et disponibles à l'adresse `http://www.stat.purdue.edu/~chong/software.html`) font figure de référence pour la résolution de ce système et l'estimation du paramètre λ par les critères GCV ou UBR (Gu [72] section 3.4). Wang [180] propose une généralisation de cet ensemble de fonctions, nommée GRKPACK, au cas de données issues d'une famille exponentielle (disponible à l'adresse `http://www.pstat.ucsb.edu/faculty/yuedong/software.html`). Plus récemment, Wang et Ke [183] ont développés un ensemble de fonctions pour le logiciel R (R Development Core Team [123]) regroupées dans le package `assist` (disponible à l'adresse `http://cran.r-project.org`) et adapté à divers modèles de splines (voir Wang [182] pour plus de détails). La fonction `ssr` de ce package est notamment adaptée au modèle général des splines de lissage et permet de calculer les vecteurs de coefficients c et d, et plus généralement l'estimateur $\widehat{f}_\lambda$ du théorème de Kimeldorf et Wahba. Cette fonction prend en argument les paramètres suivants :

- le vecteur y des observations ;
- la matrice T ;
- la matrice Σ.

Dans notre cas, ces paramètres sont de la forme suivante :

– $y = (y_1, \ldots, y_n, v_1, \ldots, v_n)^T$;

– $T = \{\mathcal{L}_j \phi_\nu\}_{j=1}^{2n} {}_{\nu=1}^{m} = \begin{bmatrix} \phi_1(t_1) & \cdots & \phi_m(t_1) \\ \vdots & \vdots & \vdots \\ \phi_1(t_n) & \cdots & \phi_m(t_n) \\ \phi_1'(t_1) & \cdots & \phi_m'(t_1) \\ \vdots & \vdots & \vdots \\ \phi_1'(t_n) & \cdots & \phi_m'(t_n) \end{bmatrix}$;

–

$$\begin{aligned} \Sigma &= \{\mathcal{L}_{j(s)} \mathcal{L}_{k(t)} R_1(s,t)\}_{j,k=1}^{2n} \\ &= \begin{bmatrix} \{R_1(t_i,t_j)\}_{i=1}^{n} {}_{j=1}^{n} & \{\frac{\partial}{\partial t} R_1(t_i,t)|_{t=t_j}\}_{i=1}^{n} {}_{j=1}^{n} \\ \{\frac{\partial}{\partial s} R_1(s,t_j)|_{s=t_i}\}_{i=1}^{n} {}_{j=1}^{n} & \{\frac{\partial}{\partial s}\{\frac{\partial}{\partial t} R_1(s,t)|_{t=t_j}\}|_{s=t_i}\}_{i=1}^{n} {}_{j=1}^{n} \end{bmatrix}. \end{aligned}$$

Le calcul des matrices T et Σ dépend donc de l'expression des fonctions $\phi_1, \ldots, \phi_m$ qui forment une base de l'espace nul $\mathcal{H}_0$, ainsi que de l'expression du noyau reproduisant R_1 associée à $\mathcal{H}_1$. Plus généralement, les calculs dépendent du choix de la norme associée à l'espace de Sobolev $\mathcal{H}_R = W^m[a,b]$. En effet, nous avons vu à la remarque 3.6 que l'espace de Sobolev pouvait être muni de plusieurs choix de normes, toutes équivalentes d'un point de vue topologique, et que chaque norme déterminait son propre noyau reproduisant $R = R_0 + R_1$. Nous allons voir dans la section suivante que le choix de la norme est déterminant pour le calcul des vecteurs des coefficients c et d.

4.3.1.4 Le choix de la norme : quelques exemples

Plusieurs choix de norme peuvent être associés à l'espace de Sobolev $W^m[a,b]$. L'ouvrage de Berlinet et Thomas-Agnan [15] contient une collection d'exemples d'espaces avec la norme et le noyau reproduisant associés. Nous détaillons aux tables 4.1 et 4.2 deux exemples de choix de norme associé à l'espace de Sobolev $W^m[0,1]$ et abordés au chapitre précédent dans le cas du modèle classique des splines de lissage. En pratique, on peut toujours se ramener à l'intervalle $[0,1]$ par une transformation appropriée. Pour chaque choix de norme, nous donnons l'expression du noyau reproduisant associé, ainsi que l'expression des fonctions $\phi_1, \ldots, \phi_m$ et du noyau R_1 dans les cas $m = 2$ (splines cubiques) et $m = 3$ (splines quintiques). On observe sur ces deux exemples que l'expression du noyau $R_1(s,t)$ est généralement complexe, et en particulier dès que $m > 2$, ce qui pose des problèmes numériques dans le calcul des vecteurs de coefficients c et d. En effet, un mauvais choix de norme associée à l'espace de Sobolev $W^m[0,T]$, et donc un mauvais choix de noyau reproduisant R_1, peut conduire à l'inversion d'une matrice Σ mal conditionnée, ce qui pose des difficultés

Espace et norme	L'espace de Sobolev $W^m[0,1]$ est muni de la norme $\|f\|^2 = \sum_{\nu=0}^{m-1}(f^{(\nu)}(0))^2 + \int_0^1 (f^{(m)}(t))^2 dt$
Noyau reproduisant	$R(s,t) = \sum_{\nu=0}^{m-1} \frac{s^\nu t^\nu}{[\nu!]^2} + \int_0^1 \frac{(s-u)_+^{m-1}(t-u)_+^{m-1}}{[(m-1)!]^2} du$
Cas $m=2$	$\mathcal{H}_0 = span\{1,\ t\}$ et $R_0(s,t) = 1+st$ $R_1(s,t) = (s\wedge t)^2\{3(s\vee t) - s\wedge t\}/6$ où $s\wedge t = min(s,t)$ et $s\vee t = max(s,t)$
Cas $m=3$	$\mathcal{H}_0 = span\{1,\ t,\ \frac{t^2}{2}\}$ et $R_0(s,t) = 1+st+\frac{s^2t^2}{2!^2}$ $R_1(s,t) =$ $\frac{s^5-(s-s\wedge t)^5}{20} + \frac{t-s}{8}(s^4-(s-s\wedge t)^4) + \frac{(t-s)^2}{12}(s^3-(s-s\wedge t)^3)$ où $s\wedge t = min(s,t)$

Table 4.1 - Exemple 1 de norme associée à l'espace de Sobolev $W^m[0,1]$ et noyau reproduisant associé.

dans le calcul numérique de l'estimateur $\widehat{f}_\lambda$. Une solution consiste alors à remplacer le noyau reproduisant R_1 par un semi-noyau pour le calcul de l'estimateur.

4.3.1.5 L'utilisation du semi-noyau

Nous avons vu à la section 3.3 que, d'un point de vue général, une spline est la solution d'un problème variationnel dans un espace de Hilbert, et que la forme du critère de minimisation détermine le type de spline obtenu. Les splines plaque mince ("*thin plate splines*" en anglais) permettent notamment d'étendre les splines de lissage classiques au cas où la dimension du domaine d'étude est supérieure à 1. L'intérêt d'utiliser la théorie des splines plaque mince dont les splines de lissage sont un cas particulier, est une écriture de l'estimateur utilisant un semi-noyau et simplifiant considérablement les calculs numériques. En effet, sous certaines conditions, il est possible d'obtenir un minimiseur du critère (4.6) ne faisant pas intervenir de noyaux reproduisants mais uniquement un semi-noyau. Les bases théoriques pour les splines plaque mince ont été posées notamment par Duchon [46] et Meinguet [109], et les ouvrages de Wahba [173] (section 2.4) et Gu [72] (section 4.4) traitent également de ce sujet.

Espace et norme	L'espace de Sobolev $W^m[0,1]$ est muni de la norme $\|f\|^2 = \sum_{\nu=0}^{m-1}(\int_0^1 f^{(\nu)}(t)dt)^2 + \int_0^1 (f^{(m)}(t))^2 dt$
Noyau reproduisant	$R(s,t) = \sum_{\nu=0}^{m-1} k_\nu(s)k_\nu(t) + k_m(s)k_m(t) + (-1)^{m-1}k_{2m}(\|s-t\|)$ où les $k_r(t) = B_r(t)/r!$ sont les polynômes de Bernoulli normalisés avec $B_r,\ r = 1, 2, \ldots$ définis récursivement par $B_0(t) = 1$, $B_r'(t) = rB_{r-1}(t)$ et $\int_0^1 B_r(t)dt = 0$ pour $r = 1, 2, \ldots$
Cas $m = 2$	$\mathcal{H}_0 = span\{1,\ k_1(t)\}$ et $R_0(s,t) = 1 + k_1(s)k_1(t)$ $R_1(s,t) = k_2(s)k_2(t) + k_4(\|s-t\|)$ où $k_1(t) = t - 1/2$ et $k_2(t) = \frac{1}{2}[k_1^2(t) - \frac{1}{12}]$
Cas $m = 3$	$\mathcal{H}_0 = span\{1,\ k_1(t),\ k_2(t)\}$ et $R_0(s,t) = 1 + k_1(s)k_1(t) + k_2(s)k_2(t)$ $R_1(s,t) = k_3(s)k_3(t) + k_6(\|s-t\|)$ où $k_1(t) = t - 1/2$ et $k_2(t) = \frac{1}{2}[k_1^2(t) - \frac{1}{12}]$

Table 4.2 - Exemple 2 de norme associée à l'espace de Sobolev $W^m[0,1]$ et noyau reproduisant associé.

Nous donnons ici une généralisation du modèle des splines plaque mince énoncé par Wahba et Wendelberger [175] et correspondant à une extension du modèle général des splines de lissage vu à la section 3.3.3.2 sur $\mathbb{R}^d$, $d \geq 1$.

Soient $t = (x_1, \ldots, x_d)$ et $t_i = (x_1(i), \ldots, x_d(i))$. On considère le modèle suivant :

$$y_i = \mathcal{L}_i f + \varepsilon_i \ , \quad i = 1, \ldots, n \tag{4.16}$$

où les $\mathcal{L}_i$ sont des formes linéaires continues de f et les ε_i sont des erreurs indépendantes de moyenne nulle avec $E[\varepsilon_i^2] = \sigma_i^2$. On introduit la pénalité suivante :

$$J_m^d(f) = \sum_{\alpha_1 + \ldots + \alpha_d = m} \frac{m!}{\alpha_1! \ldots \alpha_d!} \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \left(\frac{\partial^m f}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}}\right)^2 dx_1 \ldots dx_d. \tag{4.17}$$

La pénalité $J_m^d(f)$ permet de quantifier le degré de lissage d'une fonction f à d variables $x_1, \ldots, x_d$. Lorsque $d = 1$ et $m = 2$, on retrouve la pénalité de la rugosité vue à la section 3.2.2.1 dans le cas du modèle des splines cubiques de lissage. $J_m^d(f)$ est invariante par rotation et est donc particulièrement adaptée aux données spatiales. On considère l'espace de Hilbert $\mathcal{H} = \{f : J_m^d(f) < \infty\}$ muni de $J_m^d(f)$ comme carré de la semi-norme [1]. D'après le théorème 4.1, pour que $\mathcal{H}$ soit un RKHS, il est nécessaire que $2m - d > 0$ (condition pour que les fonctionnelles d'évaluation $\mathcal{L}f = f(t)$ soient continues). On suppose que cette condition est vérifiée et on cherche la fonction $f \in \mathcal{H}$ qui minimise le critère suivant :

$$\frac{1}{n} \sum_{i=1}^{n} \sigma_i^{-2} (\mathcal{L}_i f - y_i)^2 + \lambda J_m^d(f). \tag{4.18}$$

où $\lambda > 0$ est le paramètre de lissage. L'espace nul $\mathcal{H}_0 = \{f : J_m^d(f) = 0\}$ est l'espace vectoriel de dimension $p = \binom{d+m-1}{d}$ engendré par les polynômes de degré inférieur ou égal à $m - 1$. On note $\{\phi_\nu\}_{\nu=1}^p$ ces polynômes. Alors le théorème suivant donne la forme de l'estimateur minimisant le critère (4.18) :

Théorème 4.2 (Wahba et Wendelberger, 1980)**.** *Si les deux hypothèses suivantes sont vérifiées :*

1. *les formes linéaires continues $\mathcal{L}_1, \ldots, \mathcal{L}_n$ sont linéairement indépendantes ;*
2. *pour tout $k = 1, 2, \ldots, n$, $\mathcal{L}_k \sum_{\nu=1}^p a_\nu \phi_\nu = 0$ implique que tous les a_ν sont nuls ;*

alors le critère (4.18) admet un unique minimum dans $\mathcal{H}$ donné par

$$f_\lambda(t) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i \xi_i(t), \tag{4.19}$$

1. Une semi-norme vérifie les mêmes propriétés qu'une norme excepté le fait que la semi-norme d'un vecteur non nul n'est pas nécessairement non nulle.

où

$$\xi_i(t) = \mathcal{L}_{i(s)} E_m(t,s), \tag{4.20}$$

$\mathcal{L}_{i(s)}$ *signifiant que la forme linéaire* $\mathcal{L}_i$ *est appliqué à ce qui suit et qui est considéré comme une fonction de* s*, et*

$$E_m(s,t) = E(|s-t|), \tag{4.21}$$

$|s-t|$ *correspondant à la distance euclidienne, avec*

$$E(u) = \begin{cases} \theta_{m,d}|u|^{2m-d} log|u|, & d \ \ pair \\ \theta_{m,d}|u|^{2m-d}, & d \ \ impair \end{cases} \tag{4.22}$$

où

$$\theta_{m,d} = \begin{cases} \frac{(-1)^{d/2+m+1}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!}, & d \ \ pair \\ \frac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!}, & d \ \ impair. \end{cases} \tag{4.23}$$

Les coefficients $c = (c_1, \ldots, c_n)^T$ *et* $d = (d_1, \ldots, d_p)^T$ *sont solutions du système linéaire suivant :*

$$(K + n\lambda W^{-1})c + Td = y \tag{4.24}$$

$$T^T c = 0 \tag{4.25}$$

où

$$\begin{aligned} K &= \{\mathcal{L}_{i(s)}\mathcal{L}_{j(t)} E_m(s,t)\}_{i,j=1}^n, \\ T &= \{\mathcal{L}_j \phi_\nu\}_{i=1}^{n} \ {}_{\nu=1}^{p}, \\ W &= diag(\sigma_1^{-2}, \ldots, \sigma_n^{-2}). \end{aligned}$$

La démonstration de ce théorème est donnée dans l'appendice A de l'article de Wahba et Wendelberger [175]. La fonction $E_m(s,t)$ est appelée semi-noyau et joue dans l'expression (4.19) de l'estimateur le même rôle que le noyau reproduisant $R_1(s,t)$ associé à l'espace $\mathcal{H}_1$ lorsque l'on a la décomposition $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$. Cependant, la fonction E_m n'est pas un noyau reproduisant car elle n'est pas définie positive. Néanmoins, E_m est conditionnellement définie positive dans le sens où $T^T c = 0 \Rightarrow c^T K c > 0$, ce qui est une propriété suffisante. Ainsi, le théorème 4.2 montre que pour calculer l'estimateur minimisant le critère (4.18), il suffit de connaître le semi-noyau, qui a généralement une forme plus simple que le noyau reproduisant R_1. L'expression du "véritable" noyau reproduisant R_1 est donnée dans Wahba et Wendelberger [175], ainsi que dans Gu [72] section 4.4.2 dans le cas du modèle classique des splines plaque mince où les $\mathcal{L}_i$ sont les fonctionnelles d'évaluation.

Le modèle général des splines de lissage, dont le modèle (4.4) étudié est un cas particulier, correspond au modèle général des splines plaque mince dans le cas $d = 1$. En effet, dans le cas $d = 1$, la pénalité (4.17) s'écrit :

$$J_m^1 = \int_{-\infty}^{+\infty} (f^{(m)})^2 dt, \tag{4.26}$$

et le critère (4.6) que l'on cherche à minimiser correspond donc au critère (4.18) en prenant $d = 1$, $\mathcal{X} = [a, b]$ comme domaine d'étude, $\mathcal{H} = W^m[a, b]$ et les paramètres énoncés en (4.5). On peut donc appliquer le théorème 4.2 sous condition que les deux hypothèses sur les $\mathcal{L}_j$, $j = 1, \ldots, 2n$ soient vérifiées. Or, pour le modèle (4.4) étudié, ces deux hypothèses sont vérifiées si les points d'échantillonnage $t_1, \ldots, t_n$ sont distincts (la matrice T est alors de rang plein). La proposition suivante fournit une réécriture de l'estimateur $\widehat{f}_\lambda$ énoncé en (4.8) et minimisant le critère (4.6) en utilisant le semi-noyau.

Proposition 4.2. *Soit $\phi_1, \ldots, \phi_m$ une base de l'espace nul $\mathcal{H}_0$ et soit T une matrice $n \times m$ de plein rang définie par*

$$T = \{\mathcal{L}_j \phi_\nu\}_{j=1}^{2n} {}_{\nu=1}^{m}.$$

Alors le critère (4.6) a un unique minimum donné par

$$\widehat{f}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i E_m(t_i, t) + \sum_{i=1}^{n} c'_i \frac{\partial}{\partial s} E_m(s, t)|_{s=t_i} \tag{4.27}$$

où

$$\begin{aligned} E_m(s,t) &= \theta_{m,1} |s-t|^{2m-1}, \\ \theta_{m,1} &= \frac{\Gamma(1/2-m)}{2^{2m}\pi^{1/2}(m-1)!}. \end{aligned}$$

Les coefficients $c = (c_1, \ldots, c_n, c'_1, \ldots, c'_n)^T$ et $d = (d_1, \ldots, d_m)^T$ sont solutions du système linéaire suivant :

$$(K + n\lambda W^{-1})c + Td = y \tag{4.28}$$

$$T^T c = 0 \tag{4.29}$$

où

$$\begin{aligned} K &= \{\mathcal{L}_{j(s)} \mathcal{L}_{k(t)} E_m(s,t)\}_{j,k=1}^{2n}, \\ T &= \{\mathcal{L}_j \phi_\nu\}_{j=1}^{2n} {}_{\nu=1}^{m}, \\ W &= diag(\sigma_1^{-2}, \ldots, \sigma_{2n}^{-2}). \end{aligned}$$

m	Spline	ϕ_ν	E_m
2	Cubique	$1,\ t$	$\theta_{2,1}\lvert s-t\rvert^3$ où $\theta_{2,1}=\frac{\Gamma(-3/2)}{2^4\pi^{1/2}}$
3	Quintique	$1,\ t,\ t^2$	$\theta_{3,1}\lvert s-t\rvert^5$ où $\theta_{3,1}=\frac{\Gamma(-5/2)}{2^6\pi^{1/2}2!}$

Table 4.3 - Expression des fonctions engendrant l'espace nul et du semi-noyau dans les cas $m=2$ et $m=3$.

Le tableau 4.3 donne l'expression des fonctions ϕ_ν engendrant l'espace nul et du semi-noyau dans les cas $m=2$ et $m=3$. On observe que les expressions du semi-noyau sont beaucoup plus simples que les expressions du noyau reproduisant R_1 énoncés dans les deux exemples de choix de norme donnés aux tables 4.1 et 4.2. Ainsi, l'utilisation du semi-noyau permet de simplifier l'écriture de la matrice K qui joue ici le même rôle que la matrice Σ et facilite la résolution numérique du système linéaire (4.28), et donc le calcul de l'estimateur.

Cependant, afin de pouvoir reporter l'expression (4.27) de l'estimateur $\widehat{f}_\lambda(t)$ dans le critère (4.6), il faut que le semi-noyau $E_m(s,t)$ et que sa dérivée partielle par rapport à s, $\frac{\partial}{\partial s}E_m(s,t)$, soient dérivables jusqu'à l'ordre m (ordre de la pénalité). Or dans le cas $m=2$ (spline cubique), $\frac{\partial}{\partial s}E_m(s,t)$ n'est pas dérivable à l'ordre 2, donc on choisira pour la suite $m=3$ (spline quintique, i.e. de degré 5).

4.3.1.6 Application numérique sur données simulées

On propose de tester la méthode présentée sur un exemple de données simulées où les observations sur f sont très bruitées, alors que les observations sur la dérivée f' le sont peu. On considère le modèle (4.4) avec $n=50$ points d'échantillonnage t_i uniformément répartis sur l'intervalle $[0,1]$. Les erreurs $\varepsilon_{1,i}$ et $\varepsilon_{2,i}$ sont simulées par des distributions gaussiennes d'écart-types respectifs $\sigma_1=0.2$ et $\sigma_2=0.01$. La vraie fonction de régression f que l'on cherche à estimer est

$$f(t)=\frac{1}{2}(2t-1)^3+\frac{1}{2},$$

dont la dérivée est

$$f'(t)=3(2t-1)^2.$$

La particularité de cette fonction (étudiée notamment dans Dette *et al.* [43]) est qu'elle est strictement croissante avec un plateau. La forme de cette fonction est ainsi très proche de la forme d'une fonction représentant la distance parcourue au cours du temps avec la présence d'un arrêt représenté par le plateau.

Dans un premier temps, nous avons estimé la fonction de régression f uniquement à partir des observations bruitées $y_{1,i}$, $i=1,\dots,n$, en utilisant un estimateur par

a) Graphe de f.

b) Graphe de f'.

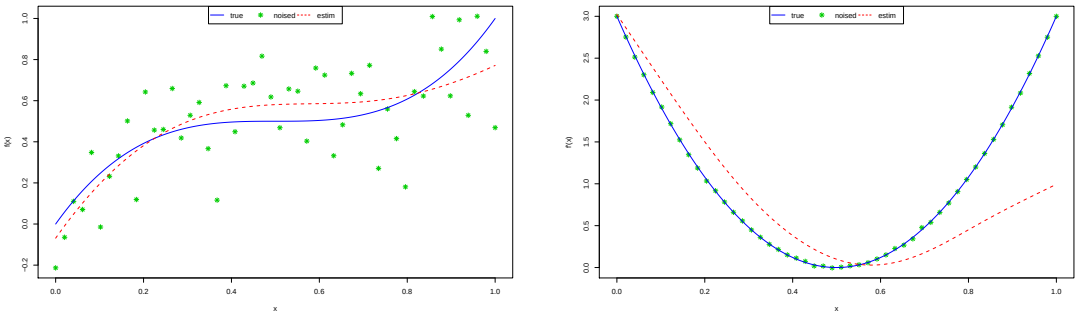


FIGURE 4.3 - Estimation de f par spline de lissage ($m = 3$) sans utiliser les observations sur la dérivée. Les observations bruitées sont représentées en vert, la fonction de référence est en bleu, et l'estimateur est en rouge.

spline de lissage d'ordre 6 ($m = 3$, spline quintique). Le paramètre de lissage λ a été choisi à l'aide du critère GCV ($\lambda = 5.74 \times 10^{-6}$). Les graphes de la vraie fonction f et de l'estimateur obtenu ainsi que les observations bruitées de f sont représentés à la figure 4.3.a, et les graphes de leur dérivée première sont représentés à la figure 4.3.b.

Dans un deuxième temps, nous avons estimé la fonction de régression f à partir des observations bruitées $y_{1,i}$, $i = 1, \ldots, n$, de f, et également des observations $y_{2,i}$, $i = 1, \ldots, n$, de f'. L'estimateur par spline de lissage est toujours d'ordre 6 ($m = 3$) et le paramètre de lissage λ a été choisi à l'aide du critère GCV ($\lambda = 1.38 \times 10^{-3}$). La mise en oeuvre numérique a été réalisée avec la fonction `ssr` du package R `assist` et l'utilisation du semi-noyau. Les graphes de l'estimateur obtenu et de sa dérivée première sont représentés respectivement aux figures 4.4.a et 4.4.b.

Ces résultats montrent que l'utilisation de l'information sur la dérivée, lorsqu'elle est disponible, permet d'améliorer considérablement l'estimation de la fonction de régression f.

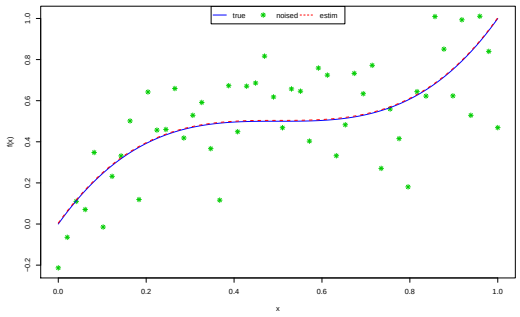
4.3.2 Application à l'estimation des profils spatiaux de vitesse

4.3.2.1 Forme de l'estimateur

Revenons à notre problème initial, à savoir estimer la distance parcourue, notée $F(t)$, d'un véhicule en fonction du temps à partir de deux types d'observations :

- d'une part, des observations bruitées y_i ($i = 1, \ldots, n$) de la distance parcourue F du véhicule aux instants t_i ($i = 1, \ldots, n$) ;
- d'autre part, des observations bruitées v_i ($i = 1, \ldots, n$) de la vitesse du véhicule (i.e. la dérivée $F'(t)$ de la position par rapport au temps) aux instants t_i ($i = 1, \ldots, n$) indépendantes des observations y_i.

a) Graphe de f.



b) Graphe de f'.

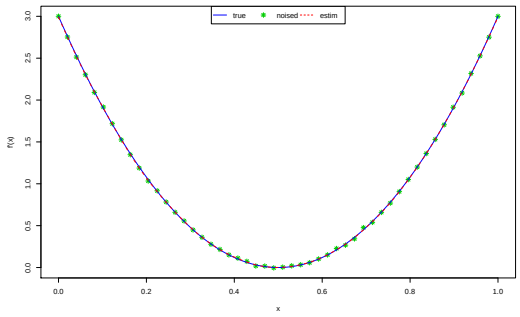


FIGURE 4.4 - Estimation de f par spline de lissage ($m = 3$) en utilisant les observations sur la dérivée. Les observations bruitées sont représentées en vert, la fonction de référence est en bleu, et l'estimateur est en rouge.

On suppose que le domaine de F est $\mathcal{X} = [0, T]$, où T est un réel positif, et que $F \in W^m[0, T]$ avec $m > 1$. Le modèle s'écrit alors de manière analogue au modèle (4.4) sous la forme :

$$\begin{cases} y_i = F(t_i) + \varepsilon_{x,i}, & i = 1, \dots, n \\ v_i = F'(t_i) + \varepsilon_{v,i}, & i = 1, \dots, n \end{cases} \tag{4.30}$$

où :

- les y_i sont des mesures bruitées de la distance parcourue par le véhicule en chaque instant t_i ;
- les $\varepsilon_{x,i}$ sont des erreurs aléatoires non corrélées de moyenne nulle et de variance σ_x^2 ;
- les v_i sont des mesures bruitées de la vitesse du véhicule en chaque instant t_i ;
- les $\varepsilon_{v,i}$ sont des erreurs aléatoires non corrélées de moyenne nulle et de variance σ_v^2, indépendantes des $\varepsilon_{x,i}$.

Alors, d'après la proposition 4.1, si les points d'échantillonnage $t_1, \dots, t_n$ sont distincts, le critère

$$\frac{1}{2n}\{\sigma_x^{-2}\sum_{i=1}^{n}(y_i - F(t_i))^2 + \sigma_v^{-2}\sum_{i=1}^{n}(v_i - F'(t_i))^2\} + \lambda\int_0^T (F^{(m)}(t))^2 dt. \tag{4.31}$$

admet un unique minimiseur de la forme

$$\widehat{F}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu \phi_\nu(t) + \sum_{i=1}^{n} c_i R_1(t_i, t) + \sum_{i=1}^{n} c_i' \frac{\partial}{\partial s} R_1(s, t)|_{s=t_i} \tag{4.32}$$

où

$$\begin{aligned} T &= \{\mathcal{L}_j\phi_\nu\}_{j=1\ \nu=1}^{2n\ \ m}, \\ \Sigma &= \{\mathcal{L}_{j(s)}\mathcal{L}_{k(t)}R_1(s,t)\}_{j,k=1}^{2n}, \\ W &= diag(\sigma_x^{-2},\ldots,\sigma_x^{-2},\sigma_v^{-2},\ldots,\sigma_v^{-2}), \\ M &= \Sigma + 2n\lambda W^{-1}, \\ d=(d_1,\ldots,d_m)^T &= (T^TM^{-1}T)^{-1}T^TM^{-1}y, \\ c=(c_1,\ldots,c_n,c'_1,\ldots,c'_n)^T &= M^{-1}\{I-T(T^TM^{-1}T)^{-1}T^TM^{-1}\}y. \end{aligned}$$

D'après la proposition 4.2, l'estimateur $\widehat{F}_\lambda(t)$ peut également s'écrire en remplaçant le noyau R_1 par le semi-noyau E_m. L'estimateur $\widehat{F}_\lambda(t)$ peut alors se réécrire sous la forme

$$\widehat{F}_\lambda(t) = \sum_{\nu=1}^{m} d_\nu\phi_\nu(t) + \sum_{i=1}^{n} c_i E_m(t_i,t) + \sum_{i=1}^{n} c'_i \frac{\partial}{\partial s}E_m(s,t)|_{s=t_i} \tag{4.33}$$

où

$$\begin{aligned} E_m(s,t) &= \theta_{m,1}|s-t|^{2m-1}, \\ \theta_{m,1} &= \frac{\Gamma(1/2-m)}{2^{2m}\pi^{1/2}(m-1)!}. \end{aligned}$$

Les coefficients $c=(c_1,\ldots,c_n,c'_1,\ldots,c'_n)^T$ et $d=(d_1,\ldots,d_m)^T$ sont alors solutions du système linéaire suivant :

$$(K+2n\lambda W^{-1})c + Td = y \tag{4.34}$$
$$T^Tc = 0 \tag{4.35}$$

où

$$K = \{\mathcal{L}_{j(s)}\mathcal{L}_{k(t)}E_m(s,t)\}_{j,k=1}^{2n}.$$

On peut ensuite, par dérivation, en déduire un estimateur $\widehat{F'}$ du profil de vitesse F' en fonction du temps (espace `vitesse`×`temps`). Puis, d'après la définition 2.1, on en déduit un estimateur $\widehat{v_S}$ du profil spatial de vitesse v_S par la transformation suivante : $\widehat{v_S} = \widehat{F'} \circ \widehat{F}^{-1}$.

4.3.2.2 Estimation des variances des capteurs

L'écriture (4.32) de l'estimateur $\widehat{F}_\lambda(t)$ suppose de connaître les variances σ_x^2 et σ_v^2 des mesures de position et de vitesse. Cependant, en pratique, ces variances ne sont

pas connues. Il est alors possible d'utiliser un estimateur de chacune de ces variances.

On se place dans le cadre du modèle classique des splines de lissage

$$y_i = f(t_i) + \varepsilon_i \ , \quad i = 1, \dots, n$$

où $f \in W^m[a,b]$ et les ε_i sont des erreurs aléatoires indépendantes de moyenne 0 et de variance constante σ^2. On note $\widehat{f}_\lambda$ l'estimateur par spline de lissage associé (spline polynomiale d'ordre $2m$) pour le paramètre λ donné, et on note $A(\lambda)$ la matrice chapeau définie par

$$(\widehat{f}_\lambda(t_1), \dots, \widehat{f}_\lambda(t_n))^T = A(\lambda)y.$$

Alors, de manière analogue à la régression paramétrique,

$$\widehat{\sigma}^2 = \frac{y^T(I - A(\lambda))^2 y}{n - trA(\lambda)}$$

est un estimateur de σ^2. Cet estimateur est clairement dépendant du paramètre de lissage λ choisi et est donc lié au critère de sélection du paramètre λ utilisé (voir section 4.3.1.2). Par exemple, si $\lambda = \lambda_{gcv}$ est l'estimateur du paramètre λ minimisant le critère GCV, l'estimateur de la variance σ^2 associé est :

$$\widehat{\sigma}^2_{gcv} = \frac{y^T(I - A(\lambda_v))^2 y}{tr(I - A(\lambda_v))}. \tag{4.36}$$

Cet estimateur est consistant sous certaines conditions (Gu [72]). Si $\lambda = \lambda_{gml}$ est l'estimateur du paramètre λ minimisant le critère GML, l'estimateur de la variance σ^2 associé est :

$$\widehat{\sigma}^2_{gml} = \frac{y^T(I - A(\lambda_m))y}{n - m}. \tag{4.37}$$

On propose alors de calculer l'estimateur par spline de lissage de F associé au modèle

$$y_i = F(t_i) + \varepsilon_{x,i}, \quad i = 1, \dots, n \tag{4.38}$$

où $F \in W^m[0,T]$ pour un paramètre de lissage λ sélectionné par la méthode GCV ou GML, puis d'en déduire une estimation de la variance σ_x^2 des erreurs de mesure de la distance parcourue. De même, on propose de calculer l'estimateur par spline de lissage de F' associé au modèle

$$v_i = F'(t_i) + \varepsilon_{v,i}, \quad i = 1, \dots, n \tag{4.39}$$

où $F' \in W^m[0,T]$ pour un paramètre de lissage λ sélectionné par la méthode GCV ou GML, puis d'en déduire une estimation de la variance σ_v^2 des erreurs de mesure de vitesse.
On peut alors calculer l'estimateur $\widehat{F}_\lambda(t)$ défini en (4.32) en remplaçant les variances σ_x^2 et σ_v^2 inconnues par leurs estimateurs $\widehat{\sigma}_x^2$ et $\widehat{\sigma}_v^2$.

4.3.2.3 Application sur données réelles

On utilise les données obtenues lors d'une expérimentation réalisée sur les pistes de Satory et décrite à la section 3.2.4 du chapitre précédent. Le trajet réalisé correspond toujours à la réalisation de deux tours de piste (en bleu sur la figure 3.5) à partir de la position D0, avec en plus un arrêt de 5 s lors du second tour, avant l'arrêt final au point D0. On propose d'estimer le profil spatial de vitesse v_S du véhicule sur ce trajet à partir des mesures de position issues du GPS (GPS GlobalSat BR-355 intégrant la puce SiRF Star III) et des mesures de vitesse issues de ce même GPS et calculées par effet Doppler (donc indépendantes des mesures de position). On rappelle que les mesures sont obtenues avec une fréquence d'échantillonnage de 1 Hz (soit 1 mes/s).

Dans un premier temps, on estime les variances σ_x^2 et σ_v^2 des mesures de position et de vitesse. Pour chacun des modèles (4.38) et (4.39), nous avons calculé l'estimateur par spline de lissage associé. Nous avons choisi dans les deux cas de prendre $m = 3$ (spline quintique de degré 5), et le paramètre de lissage λ a été sélectionné par minimisation du critère GML. Une étude rapide des résidus a montré qu'aucune tendance ne se dégageait aussi bien pour les mesures de position que pour les mesures de vitesse, et que l'hypothèse d'une variance constante était cohérente. Les résultats obtenus pour l'estimation des écart-types σ_x et σ_v sont les suivantes :

$$\widehat{\sigma}_x = 2.21 \ m, \tag{4.40}$$

et

$$\widehat{\sigma}_v = 0.42 \ m/s. \tag{4.41}$$

On se place ensuite dans le cadre du modèle (4.30), i.e. on cherche à estimer la fonction de régression $F(t)$ représentant la distance parcourue par le véhicule au cours du temps à partir des mesures de position y_i et des mesures de vitesse v_i issues du GPS. On calcule alors l'estimateur $\widehat{F}_\lambda(t)$ défini en (4.33) (utilisation du semi-noyau) et minimisant le critère (4.31) en utilisant les variances $\widehat{\sigma}_x^2$ et $\widehat{\sigma}_v^2$ estimées. Nous avons choisi de prendre $m = 3$ (spline quintique), et le paramètre de lissage λ a été sélectionné par minimisation du critère GML ($\lambda_{gml} = 0.0143$). La méthode GCV a également été testée, mais le paramètre λ_{gcv} obtenue n'est pas satisfaisant et conduit à une interpolation des données ($\lambda_{gcv} = 2.02 \times 10^{-6}$). Les résultats obtenus sont illustrés aux figures 4.5.a et 4.5.b, ainsi qu'aux figures 4.6.a et 4.6.b.

La figure 4.5.a, et la figure 4.5.b qui est un zoom de la zone encadrée en noir sur la figure 4.5.a, représentent l'espace `distance`×`temps`, et l'on distingue :

– en vert, les mesures de position issues du GPS et utilisées dans le calcul de l'estimateur ;

	Mesures GPS	Estimateur $\widehat{F}_\lambda(t)$
RMSE distance (m)	3.13	2.04

Table 4.4 - RMSE de l'estimateur $\widehat{F}_\lambda(t)$.

- en rouge, l'estimateur $\widehat{F}_\lambda(t)$ par spline de lissage calculé à partir des mesures de position et de vitesse issues du GPS ;
- en bleu, les mesures de position issues du GPS RTK et pouvant être considérées comme les mesures de référence (i.e. la vraie fonction $F(t)$ que l'on cherche à estimer) ;
- en orange, les mesures de l'odomètre collectées sur le bus CAN du véhicule et données à titre d'information.

La figure 4.6.a, et la figure 4.6.b qui est un zoom de la zone encadrée en noir sur la figure 4.6.a (même section que la figure 4.5), représentent l'espace `vitesse`×`temps`, et l'on distingue :

- en vert, les mesures de vitesse issues du GPS (calculées par effet Doppler) et utilisées dans le calcul de l'estimateur ;
- en rouge, la dérivée première $\widehat{F}'_\lambda(t)$ de l'estimateur $\widehat{F}_\lambda(t)$ par spline de lissage calculé à partir des mesures de position et de vitesse issues du GPS ;
- en bleu, la dérivée numérique (calculée par différence centrale) des mesures de position issues du GPS RTK ;
- en orange, les mesures de vitesse collectées sur le bus CAN du véhicule (et obtenues à partir des mesures de l'odomètre) et données à titre d'information.

Le tableau 4.4 donne les RMSE définies de la manière suivante :

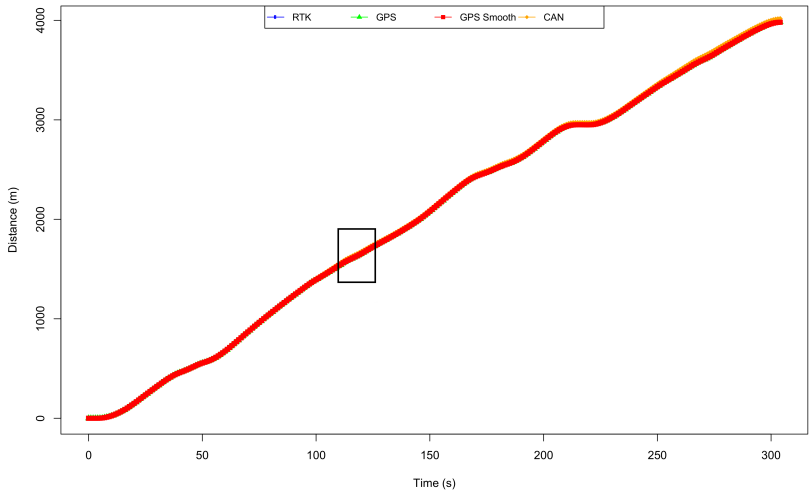
- pour les mesures GPS, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y(t_i) - y_{rtk}(t_i))^2}$ où les $y(t_i)$ sont les mesures de positions issues du GPS et les $y_{rtk}(t_i)$ sont les mesures de positions issues du GPS RTK pris comme mesures de référence ;
- pour l'estimateur $\widehat{F}_\lambda(t)$, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{F}_\lambda(t_i) - y_{rtk}(t_i))^2}$.

Ces résultats montrent que même si les mesures de position issues du GPS étaient déjà relativement précises (3 m), l'erreur d'estimation de l'estimateur $\widehat{F}_\lambda(t)$ est encore plus faible (2 m), ce qui montre la performance de cet estimateur.

Les performances de l'estimateur $\widehat{F}_\lambda(t)$ sont confirmées par l'étude des erreurs sur la vitesse, comme le montre le tableau 4.5 où sont données les RMSE définies de la manière suivante :

- pour les mesures GPS, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v(t_i) - y'_{rtk}(t_i))^2}$ où les $v(t_i)$ sont les mesures de vitesse issues du GPS et les $y'_{rtk}(t_i)$ sont les mesures de vitesse

a) Trajet complet.



b) Zoom sur la section encadrée en noir.

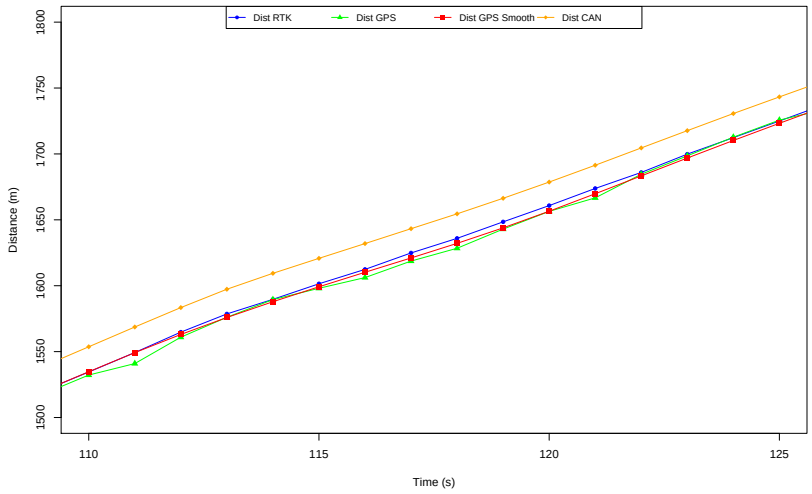
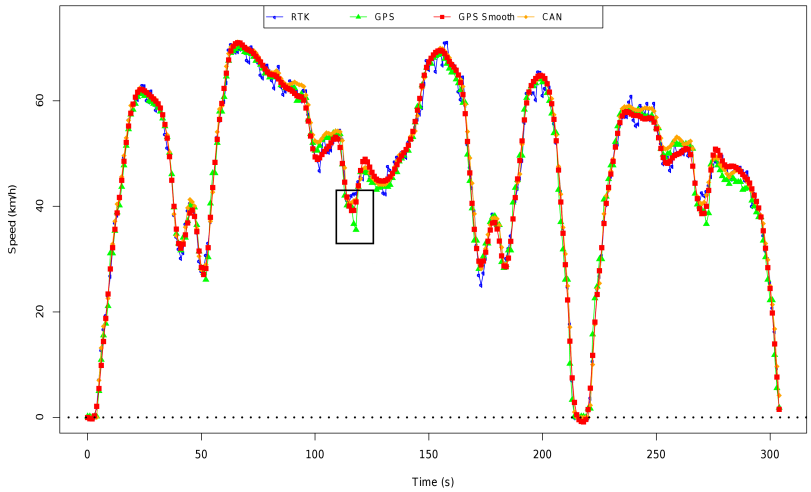


FIGURE 4.5 - Espace `distance×temps` : estimateur $\widehat{F}_\lambda(t)$ par spline de lissage de la distance parcourue au cours du temps obtenu à partir des mesures de position et de vitesse issues du GPS.

a) Trajet complet.



b) Zoom sur la section encadrée en noir.

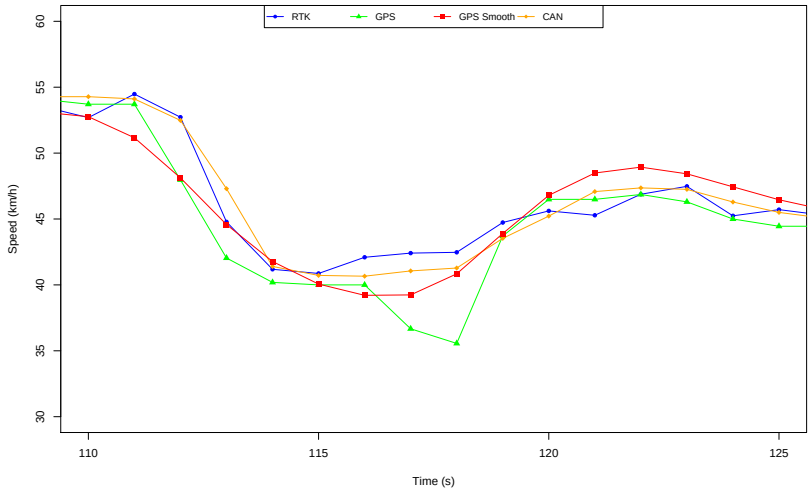


Figure 4.6 - Espace `vitesse`×`temps` : dérivée de l'estimateur $\widehat{F}_\lambda(t)$ par spline de lissage de la distance parcourue au cours du temps obtenu à partir des mesures de position et de vitesse issues du GPS.

	Mesures GPS	Dérivée de l'estimateur $\widehat{F}_\lambda(t)$
RMSE vitesse (km/h)	2.14	1.60

Table 4.5 - RMSE de la dérivée de l'estimateur $\widehat{F}_\lambda(t)$.

obtenues par dérivation numérique des mesures de position issues du GPS RTK ;

– pour l'estimateur $\widehat{F}_\lambda(t)$, $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{F}'_\lambda(t_i) - y'_{rtk}(t_i))^2}$.

Notons que la dérivée des mesures de position issues du GPS RTK ne peut être considérée dans l'absolu comme la vraie fonction $F'(t)$ puisque la dérivation numérique (par différence centrale) introduit des erreurs. Cependant, il n'existe pas de capteurs de vitesse fournissant des mesures "parfaites", donc ce choix nous semble cohérent avec le fait que les mesures de position du GPS RTK sont d'une précision centimétriques et donc quasi-parfaites.

Notons également que les mesures de position issues du GPS RTK étant des mesures discrètes (données brutes), i.e. représentées comme un vecteur de $\mathbb{R}^n$, ceci pose le problème du calcul de la dérivée. Nous avons ici choisi de calculer les mesures de vitesse par dérivation numérique en utilisant l'approche par différence centrale, mais cette opération introduit des erreurs dans le calcul de la dérivée. Ce problème est évidemment le même pour le calcul de la dérivée des mesures de position issues du GPS classique. Cependant, le calcul de l'estimateur $\widehat{F}_\lambda(t)$ par spline de lissage permet de se ramener à des données de nature fonctionnelle, ce qui facilite le calcul de la dérivée $\widehat{F}'_\lambda(t)$. On observe ainsi sur la figure 4.6 que l'estimateur $\widehat{F}'_\lambda(t)$ (en rouge) est bien lisse contrairement aux mesures de vitesse obtenues par dérivation numérique des mesures de position issues du RTK (en bleu).

Enfin, la figure 4.7 représente l'espace `vitesse`×`distance` où l'estimateur du profil spatial de vitesse, défini par $\widehat{v_S} = \widehat{F}'_\lambda \circ \widehat{F}_\lambda^{-1}$, est représenté en rouge. Cependant, si l'estimateur $\widehat{v_S}$ déduit de l'estimateur $\widehat{F}_\lambda$ présente de bons résultats, cet estimateur n'est pas un "vrai" profil spatial de vitesse. En effet, nous n'avons imposé aucune contrainte de monotonie sur l'estimateur $\widehat{F}_\lambda(t)$, alors que d'après la définition 2.1, $\widehat{v_S}$ est un profil spatial de vitesse, si $\widehat{F}_\lambda$ est une fonction croissante. Un zoom sur l'arrêt, effectué sur l'intervalle de temps $[210, 225]$, est représenté dans les 3 espaces (`distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`) à la figure 4.8. On observe dans l'espace `distance`×`temps`, que l'estimateur $\widehat{F}_\lambda(t)$ décroît légèrement entre les instants 215 et 220, ce qui se traduit par des valeurs négatives de vitesse dans l'espace `vitesse`×`temps`, et un retour en arrière (boucle avec valeurs négatives) dans l'espace `vitesse`×`distance`. L'estimateur $\widehat{v_S}$ est donc irréaliste sans cette contrainte

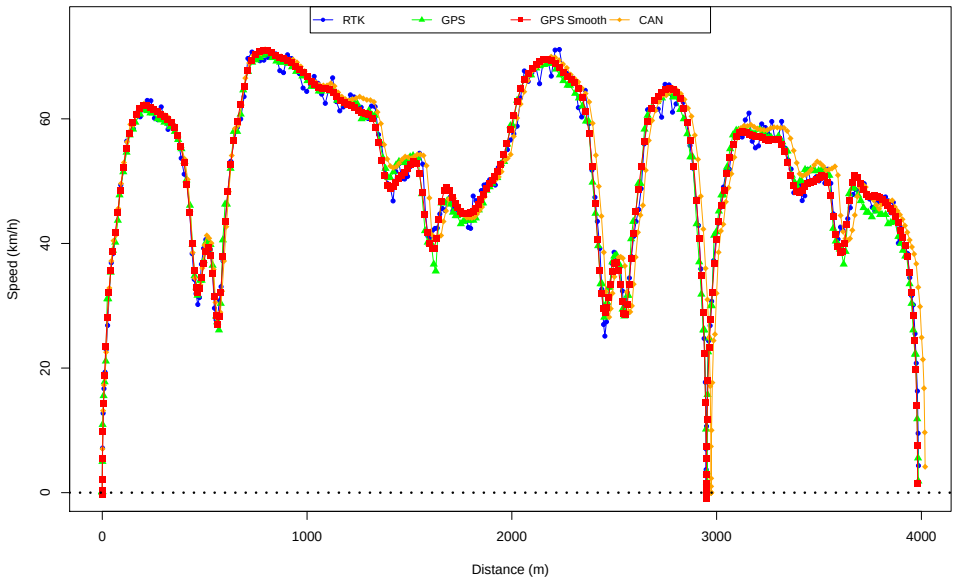


Figure 4.7 - Espace `vitesse`×`distance` : Estimateur du profil spatial de vitesse défini par $\widehat{v_S} = \widehat{F}' \circ \widehat{F}^{-1}$.

de monotonie sur l'estimateur $\widehat{F}_\lambda$.

Afin de résoudre le problème de l'estimation du profil spatial de vitesse au niveau des arrêts, nous avons essayé de mettre un poids très important sur les données de vitesse par rapport aux données de position dans le critère (4.31), en modifiant les variances σ_x^2 et σ_v^2 au niveau des arrêts. Cependant, cette approche ne permet pas non plus d'obtenir un estimateur $\widehat{F}(t)$ monotone. De plus, cette approche ne garantit par la monotonie de l'estimateur lorsque le véhicule n'est pas à l'arrêt. Nous proposons donc d'effectuer une deuxième étape de lissage afin d'ajouter la contrainte de monotonie à notre estimateur de $F(t)$. L'ajout de cette contrainte est l'objet de la section suivante.

4.4 Lissage sous contrainte de monotonie

L'objet de cette section est la prise en compte de la contrainte (C_2), à savoir une contrainte de monotonie sur l'estimateur $\widehat{F}(t)$ de $F(t)$. Nous commençons par exposer un état de l'art des différentes méthodes de lissage sous une telle contrainte de forme.

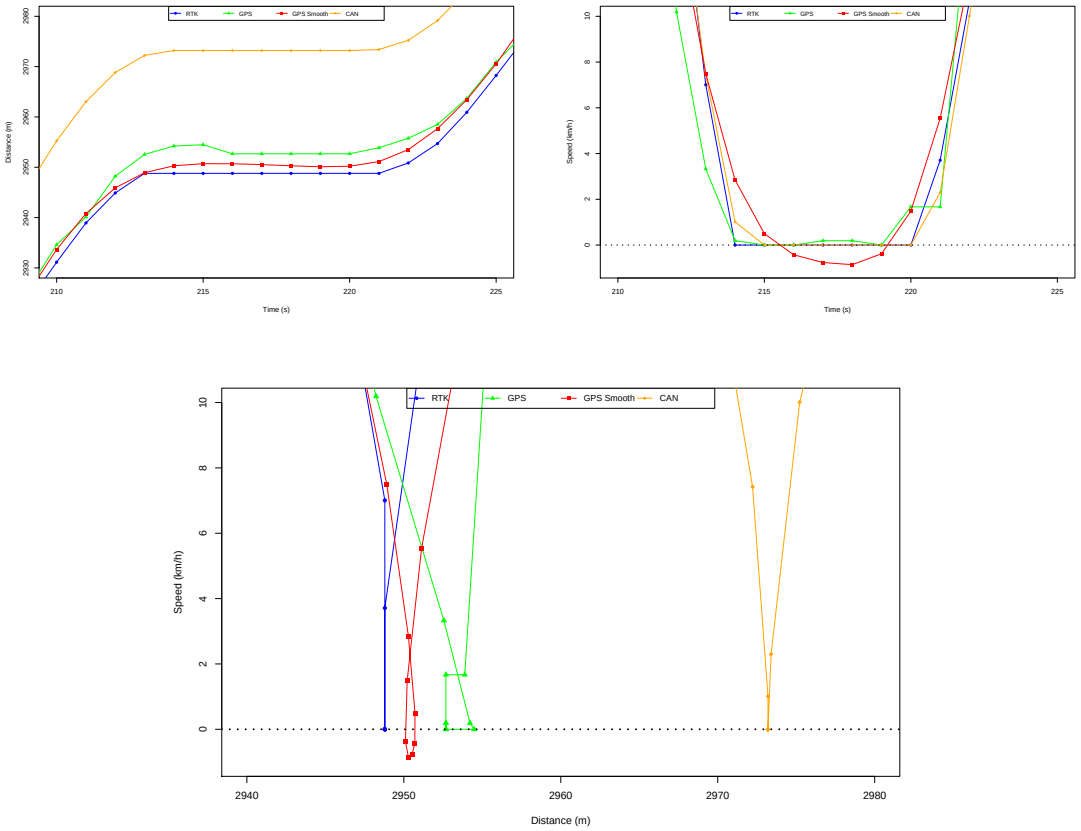


Figure 4.8 - Zoom sur l'arrêt (sur l'intervalle de temps $[210, 225]$). Représentation dans les 3 espaces : [`distance`×`temps`, `vitesse`×`temps`] et [`vitesse`×`distance`].

4.4.1 Revue des différentes méthodes

La prise en compte de contraintes de forme telles que la positivité, la monotonie ou la convexité, dans un problème de régression non paramétrique, apparaît dans de nombreuses applications comme en économie (Matzkin [107]) ou en biométrie (Kelly et Rice [83]). De nombreuses approches ont été développées, la plupart étant basées sur des estimateurs à noyau ou des estimateurs splines (Delecroix et Thomas-Agnan [38]). Nous présentons dans cette section un aperçu des différentes méthodes proposées dans le cas d'une contrainte de monotonie.

La première méthode que l'on peut citer est la régression isotonique, initialement proposée par Brunk [25], qui consiste à minimiser la somme des carrés des résidus sous contrainte de monotonie. L'estimateur est alors une fonction constante par morceaux. Afin de pallier ce problème, Friedman et Tibshirani [63] proposent de lisser cet estimateur classique de régression isotonique en utilisant une moyenne mobile, alors que Mukerjee [113] et Mammen [102] proposent d'effectuer deux étapes : une étape de lissage utilisant un estimateur à noyau, puis une étape d'isotonisation.

Une autre famille de méthodes généralisant ces procédures en deux étapes est la famille des méthodes de projection, qui consiste à projeter un estimateur non contraint dans un sous-espace de fonctions contraintes (ex : fonctions monotones, convexes, ...), lequel est généralement un ensemble convexe. La méthode de projection proposée par Delecroix *et al.* [39] consiste à projeter un estimateur non paramétrique initial sur une version discrétisée d'un cône convexe fermé décrivant les contraintes de formes. Cette méthode est basée sur la théorie des RKHS et sur des techniques d'optimisation convexe. Mammen et Thomas-Agnan [103] et Mammen *et al.* [104] proposent également de diviser le problème de régression monotone en deux étapes :

1. obtenir un estimateur non contraint (généralement, estimateur à noyau ou estimateur spline) ;
2. monotoniser l'estimateur non contraint, i.e. le projeter dans un sous-espace de fonctions monotones.

Si ce type de méthode présente l'avantage de s'adapter à tout type d'estimateur non contraint, Gijbels [68] note que la plupart de ces estimateurs monotones apparaissent moins lisses que l'estimateur sans contraintes initial.

Parmi les méthodes basées sur une procédure en deux étapes, on peut également citer les travaux de Hall et Huang [73] qui proposent une méthode de monotonisation des estimateurs à noyau classiques, tel que l'estimateur de Nadaraya-Watson, basée sur une modification des poids de l'estimateur à noyau de telle sorte que la fonction obtenue soit monotone. Dette *et al.* [43] proposent de combiner des techniques d'estimation de densité et de régression afin d'obtenir une estimation monotone de l'inverse f^{-1} de la fonction de régression, et comparent leur approche à celle de Hall et Huang [73] dans Dette et Pilz [42]. Enfin, plus récemment, Bigot et Gadat [21] proposent

une méthode de splines homéomorphes, basée sur des techniques développées dans le contexte de la déformation d'images pour la construction de difféomorphismes en deux et trois dimensions. Cette méthode consiste à définir toute fonction monotone comme la solution d'une équation différentielle ordinaire (ODE) régie par un champ de vecteur approprié dépendant du temps, puis à déterminer cette solution en utilisant les techniques classiques d'estimation par splines de lissage.

Une dernière grande famille de méthodes consiste à construire un estimateur spline sous contrainte de monotonie. Dans le cas des splines de régression (voir section 3.2.1), le principe consiste à utiliser une base de spline de fonctions monotones et à considérer une combinaison linéaire positive de ces fonctions de base. Ainsi, Ramsay [133] propose d'utiliser la base des I-splines, obtenues par intégration des B-splines dont les coefficients sont positifs. He et Shi [77] proposent une méthode de lissage par B-splines monotones basée sur une optimisation $\mathcal{L}_1$. Plus récemment, Meyer [111] propose d'étendre la méthode de Ramsay [133] limitée aux splines quadratiques monotones (degré 2), au cas des splines cubiques monotones.

L'étude des splines de lissage sous contraintes de forme a également été abordée dans la littérature. Le problème consiste alors, étant donné un échantillon (t_i, y_i), avec $t_i \in [a, b]$ pour $i = 1, \ldots, n$, à résoudre le problème suivant :

$$\begin{aligned} \text{minimiser} \quad & \sum_{i=1}^{n} \{(y_i - f(t_i))^2\} + \lambda \int_a^b (f^{(m)}(t))^2 dt, \\ \text{avec} \quad & f^{(r)}(t) \geq 0 \quad t \in [a, b]. \end{aligned} \tag{4.42}$$

Utreras [167] propose une caractérisation de l'estimateur par splines de lissage monotone (cas $r = 1$) solution de (4.42) dans le cas $m = 2$. Dans le cas $m > 2$, de nombreux auteurs proposent des algorithmes permettant d'obtenir une approximation de la solution de (4.42) (Schwetlick et Kunert [150], Turlach [166], et Villalobos et Wahba [169] dans le cas multivarié). Ces algorithmes consistent à remplacer le nombre infini de contraintes par un nombre fini de contraintes convenablement choisies. Cependant, la plupart de ces algorithmes conduisent à un nombre important d'inégalités de contraintes, et donc à des temps de calcul relativement longs.

Enfin, Kelly et Rice [83], et plus récemment Meyer [110], proposent de combiner les splines de régression et les splines de lissage, en utilisant les splines pénalisées (voir section 3.2.3), et en restreignant la forme de la spline par l'ajout de contraintes sur les coefficients des B-splines.

4.4.2 Estimation de fonctions lisses monotones (Ramsay, 1998)

Nous détaillons ici une méthode développée par Ramsay [132] qui permet de ramener le problème d'estimation d'une fonction monotone à un problème d'estimation d'une fonction sans contraintes.

Soit f une fonction strictement monotone définie sur $[0,T]$, $T>0$. On note $D^m f$, $m>0$, la dérivée d'ordre m de f. L'idée de cette méthode est que toute fonction f strictement monotone a une dérivée Df strictement positive, et qui peut donc s'écrire comme l'exponentielle d'une fonction sans contraintes. Ainsi, toute fonction f strictement monotone est solution de l'équation différentielle suivante :

$$D^2 f(t) = w(t) Df(t), \tag{4.43}$$

où w est une fonction de carré intégrable ($w \in \mathcal{L}^2$). On montre alors facilement que la solution de cette équation est de la forme :

$$f(t) = \beta_0 + \beta_1 \int_{t_0}^{t} exp[\int_{t_0}^{u} w(v)dv]du \tag{4.44}$$

où β_0 et β_1 sont des constantes telles que $f(t_0)=\beta_0$ et $Df(t_0)=\beta_1$.

On considère alors le modèle de régression non paramétrique suivant :

$$y_i = f(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.45}$$

où les ε_i sont des erreurs aléatoires non corrélées entre elles, de moyenne nulle et de variance σ^2, et où f est supposée strictement croissante et définie sur $[0,T]$, $T>0$. On cherche alors à minimiser le critère suivant :

$$\sum_{i=1}^{n} [y_i - \beta_0 - \beta_1 h(t_i)]^2 + \lambda \int_0^T w^{(m)}(t)dt, \tag{4.46}$$

où $h(t) = \int_0^t exp[\int_0^u w(v)dv]du$, et $m>0$ est le degré de la pénalité. Le problème revient donc à estimer les coefficients β_0 et β_1, et la fonction sans contraintes w. Or, la fonction w étant sans contraintes, elle peut s'écrire sous la forme d'une combinaison linéaire de fonctions de base appropriées ϕ_k, $k=1,\ldots,K$ (ex : B-splines). Ainsi, si on note $\phi = (\phi_1, \ldots, \phi_K)^T$ le vecteur des fonctions de base, et $c = (c_1, \ldots, c_K)^T$ le vecteur des coefficients, on peut écrire $w(t)$ sous la forme matricielle $w(t) = c^T \phi(t)$. Par intégration, on obtient $W(t) = c^T \Phi(t)$, où $W(t) = \int_0^t w(u)du$, et $\Phi = (\Phi_1, \ldots, \Phi_K)^T$ avec $\Phi_k = \int \phi_k$. L'estimateur de f s'écrit alors sous la forme :

$$\widehat{f}(t) = \beta_0 + \beta_1 \int_0^t exp[c^T \Phi(u)]du, \tag{4.47}$$

où les coefficients β_0 et β_1 (vérifiant $\widehat{f}(0)=\beta_0$ et $\widehat{f}'(0)=\beta_1$), et le vecteur c des coefficients, sont estimés par minimisation du critère (4.46). L'estimation de ces coefficients est alors effectuée par l'algorithme itératif suivant :

- Initialisation :
  - on choisit un estimateur initial $c^{(0)}$ (ce vecteur est soit un vecteur de zéros, soit obtenu par un lissage spline sans contraintes) ;

  - on estime $\beta_0^{(0)}$ et $\beta_1^{(0)}$ par régression linéaire.
- Pour chaque itération $\nu > 0$, on effectue les deux étapes suivantes :
  1. On obtient $c^{(\nu)}$ en minimisant le critère (4.46) par rapport à c. On utilise alors une procédure de Gauss-Jordan ou une procédure du score pour les problèmes de moindres carrés non linéaires.
  2. On calcule $\beta_0^{(\nu)}$ et $\beta_1^{(\nu)}$ par régression linéaire.

Cet algorithme est notamment réalisé par la fonction `smooth.monotone` du package R `fda` (Ramsay *et al.* [131] chapitre 5). Notons que la vitesse de convergence de cet algorithme est généralement peu rapide, ce qui implique une augmentation importante du temps de calcul par rapport à un lissage spline classique, même si ce temps de calcul reste acceptable.

4.4.3 Application à l'estimation des profils spatiaux de vitesse

4.4.3.1 Methodologie

Différentes méthodes exposées à la section 4.4.1 ont été envisagées afin de "monotoniser" l'estimateur $\widehat{F}(t)$ obtenu à partir des mesures de position et de vitesse (section 4.3.2.1). Si les méthodes de projection nous ont semblé intéressantes, nous avons rencontré des problèmes d'implémentation numérique. Nous avons notamment testé la méthode des splines homéomorphes développée par Bigot et Gadat [21] qui présentait de bons résultats pour la monotonisation de l'estimateur $\widehat{F}(t)$ (voir Andrieu *et al.* [8]), mais posait des difficultés dans le calcul numérique de la dérivée de l'estimateur (fort bruit au niveau des arrêts du véhicule, correspondant à une vitesse nulle). Le manque de temps ne nous a pas permis de trouver une solution permettant de résoudre ce problème. Nous avons également étudié la famille des estimateur splines sous contraintes. Cependant, nous souhaitions une méthode demandant peu de paramètres, c'est pourquoi nous avons rapidement exclu les méthodes basées sur les splines de régression ou les splines pénalisées qui nécessitent de choisir le nombre de noeuds de la spline, choix difficile en pratique (voir section 3.2.1.4). Quant aux splines de lissage sous contrainte de monotonie, il n'existe pas de solution exacte dans le cas $m > 2$, et les différentes approximations proposées dans ce cas nous ont semblé difficiles à mettre en oeuvre numériquement.

Nous avons donc choisi d'utiliser la méthode de Ramsay [132] décrite à la section précédente, qui présente l'avantage d'être relativement simple à mettre en oeuvre numériquement grâce au package R `fda` développé par Ramsay (Ramsay *et al.* [131]). De plus, dans le cas où l'on place un noeud en chaque point d'échantillonnage, il ne reste plus qu'un seul paramètre à déterminer, à savoir le paramètre de lissage λ. On propose donc d'appliquer cette méthode aux données estimées $\widehat{F}(t_i)$ obtenues à partir des mesures de position et de vitesse, dont le calcul a été décrit à la section 4.3. L'idée est de calculer un nouvel estimateur $\widehat{F}_{mono}(t)$ de la fonction $F(t)$ représentant la distance parcourue au cours du temps, qui soit le plus proche possible de l'estimateur

$\widehat{F}(t)$ obtenu à la section 4.3, et qui soit croissant afin d'obtenir un estimateur de $F(t)$ satisfaisant les conditions de la définition 2.1. On se ramène ainsi à une étape de monotonisation similaire aux méthodes de projection. Le nouveau modèle de régression est alors le suivant :

$$y_i = F(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{4.48}$$

où les $y_i = \widehat{F}(t_i)$ sont les valeurs de l'estimateur $\widehat{F}(t)$ obtenues à partir des mesures de position et de vitesse (section 4.3.2.1) et où les ε_i sont les erreurs associées. On obtient alors par la méthode de Ramsay [132] décrite à la section précédente, un estimateur $\widehat{F}_{mono}(t)$ strictement croissant de $F(t)$, minimisant le critère (4.46), de la forme :

$$\widehat{F}_{mono}(t) = \beta_0 + \beta_1 \int_0^t exp[W(u)]du. \tag{4.49}$$

La fonction $w(t)$ telle que $W(t) = \int_0^t w(u)du$ est décomposée dans une base de B-splines telle que $w(t) = c^T\phi(t)$ en reprenant les notations de la section 4.4.2 et avec $K = n$. Les coefficients de régression β_0 et β_1, et le vecteur c des coefficients de la base des B-splines sont estimées à l'aide de l'algorithme décrit à la section 4.4.2.

4.4.3.2 Application sur données réelles

On reprend l'expérimentation décrite à la section 4.3.2.3 et l'estimateur $\widehat{F}(t)$ construit à partir des mesures de position et de vitesse du GPS. On applique alors la méthode de Ramsay en prenant $m = 3$ comme degré pour la pénalité, et en prenant $\lambda = 10^{-1}$ pour le paramètre de lissage. La valeur de λ a été choisi après avoir testé plusieurs valeurs (méthode par essais et erreurs). L'algorithme permettant d'estimer les coefficients β_0 et β_1, et le vecteur c des coefficients de la base des B-splines, converge en 110 itérations. On observe généralement une mauvaise estimation du paramètre initial $\beta_0^{(0)}$, ce qui implique un nombre d'itérations relativement grand de l'algorithme. L'estimateur strictement croissant $\widehat{F}_{mono}(t)$ de $F(t)$ est représenté en violet à la figure 4.9 dans l'espace `distance×temps`. Sa dérivée ainsi que le profil spatial de vitesse associé sont également représentés à la figure 4.9 (espaces `vitesse×temps` et `vitesse×distance`). Les RMSE associées et calculées de manière similaire à l'estimateur $\widehat{F}(t)$ sont données dans le tableau 4.6.

On constate que ce nouvel estimateur $\widehat{F}_{mono}(t)$ est très satisfaisant par rapport à l'estimateur $\widehat{F}(t)$ construit à partir des mesures de position et de vitesse du GPS car il présente les deux avantages suivants :

- d'une part, $\widehat{F}_{mono}(t)$ est monotone et permet donc d'obtenir un estimateur du profil spatial de vitesse qui appartient bien à l'espace des profils spatiaux de vitesse défini en 2.1 ;
- d'autre part, $\widehat{F}_{mono}(t)$ présente des résultats similaires à l'estimateur $\widehat{F}(t)$ en terme d'erreurs, ce qui montre que ce second lissage ne dégrade pas les bonnes performances du premier lissage.

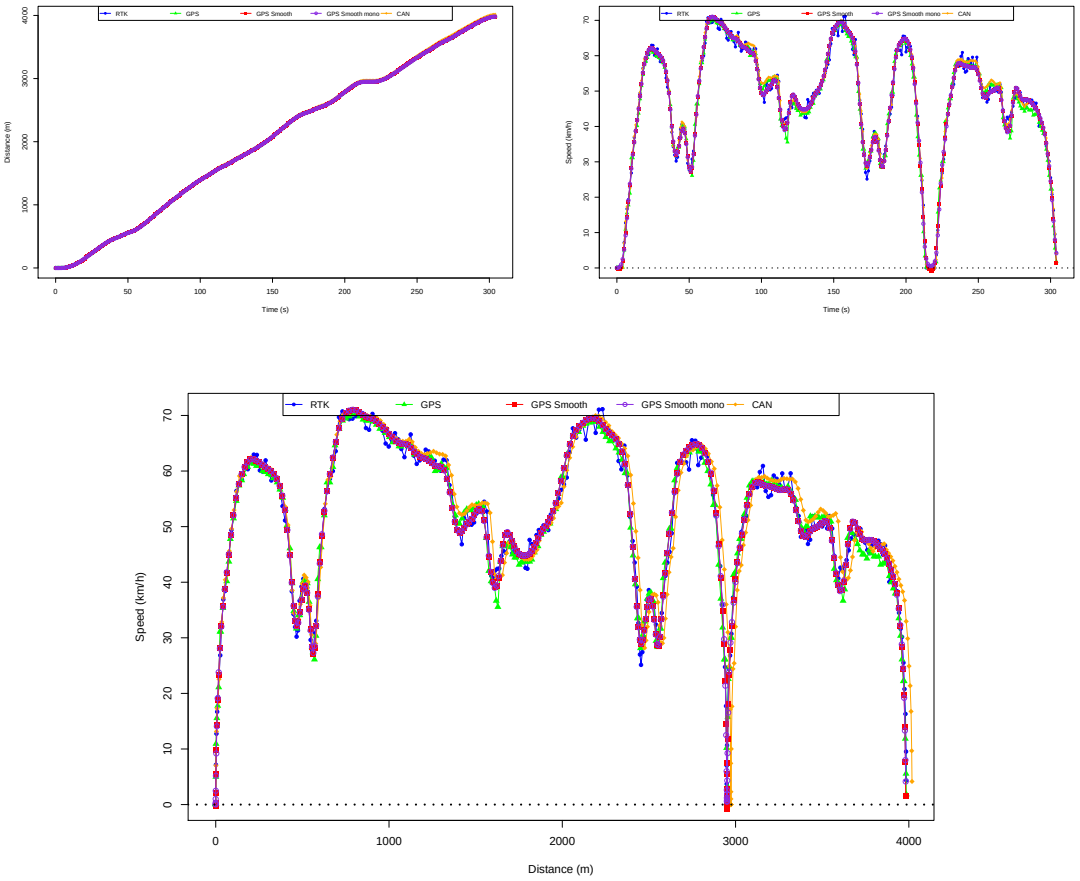


Figure 4.9 - Estimateur $\widehat{F}_{mono}(t)$ (en violet) obtenu par lissage monotone des $\widehat{F}_{\lambda}(t_i)$. Représentation dans les 3 espaces : [`distance`×`temps`, `vitesse`×`temps`] et [`vitesse`×`distance`].

	Mesures GPS	Estimateur non monotone	Estimateur monotone
RMSE distance (m)	3.13	2.04	2.05
RMSE vitesse (km/h)	2.14	1.60	1.64

Table 4.6 - Tableau des RMSE pour les données GPS, l'estimateur $\widehat{F}(t)$ et l'estimateur $\widehat{F}_{mono}(t)$.

Enfin, la figure 4.10 représente un zoom sur l'arrêt (intervalle de temps $[210, 225]$) dans les trois espaces `distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`. On observe que l'estimateur $\widehat{F}_{mono}(t)$ (en violet) est bien monotone, et que sa dérivée est positive. L'estimateur du profil spatial de vitesse est donc un "vrai" profil de vitesse.

Notons tout de même, un inconvénient majeur de la méthode de Ramsay, à savoir qu'elle permet de construire un estimateur strictement monotone. Or, dans notre cas, la fonction $F(t)$ représentant la distance parcourue d'un véhicule au cours du temps est strictement croissante lorsque le véhicule ne s'arrête pas, mais est constante lorsque le véhicule est à l'arrêt (vitesse nulle). La méthode de Ramsay ne permet donc pas de bien représenter les arrêts. Nous n'avons actuellement pas trouvé de solution pour résoudre le problème d'une bonne estimation d'un profil spatial de vitesse lorsque le véhicule est à l'arrêt. En effet, ce problème difficile revient à trouver un estimateur $\widehat{F}_{mono}(t)$ de $F(t)$ vérifiant les deux conditions suivantes :

- $\widehat{F}_{mono}(t)$ doit être strictement monotone lorsque le véhicule n'est pas à l'arret (vitesse non nulle) ;
- $\widehat{F}_{mono}(t)$ doit être monotone constant lorsque le véhicule est à l'arrêt (vitesse nulle).

La difficulté consiste dans l'estimation de ces plateaux de la fonction $F(t)$ caractérisant les arrêts du véhicule. Cependant, même si la méthode proposée ne permet pas une bonne estimation de ces plateaux, on constate sur la figure 4.10 que la vitesse est inférieure à 2 km/h pendant les 5 s d'arrêt du véhicule, ce qui reste un seuil acceptable.

4.5 Conclusion

Dans ce chapitre, nous avons décrit une méthodologie permettant de construire un estimateur $\widehat{v_S}$ d'un profil spatial de vitesse v_S comme défini à la définition 2.1, à partir de mesures bruitées de position et de vitesse. Nous avons vu que, suite à différentes contraintes difficiles à prendre en compte dans l'estimation directe d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse, il était préférable de changer d'espace d'étude, et de construire un estimateur de la distance parcourue en fonction du temps $F(t)$ à partir de ces mêmes mesures bruitées. En effet, l'estimation directe d'un profil spatial de vitesse dans l'espace `vitesse`×`distance` correspond à un problème de régression non paramétrique avec une variable explicative bruitée, et une contrainte de non dérivabilité aux points où la fonction de régression s'annule. Le changement d'espace d'étude permet de se ramener à un problème de régression non paramétrique sous les deux contraintes suivantes :

(C_1) Estimer la fonction de régression $F(t)$ à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).

(C_2) Une contrainte de monotonie sur F.

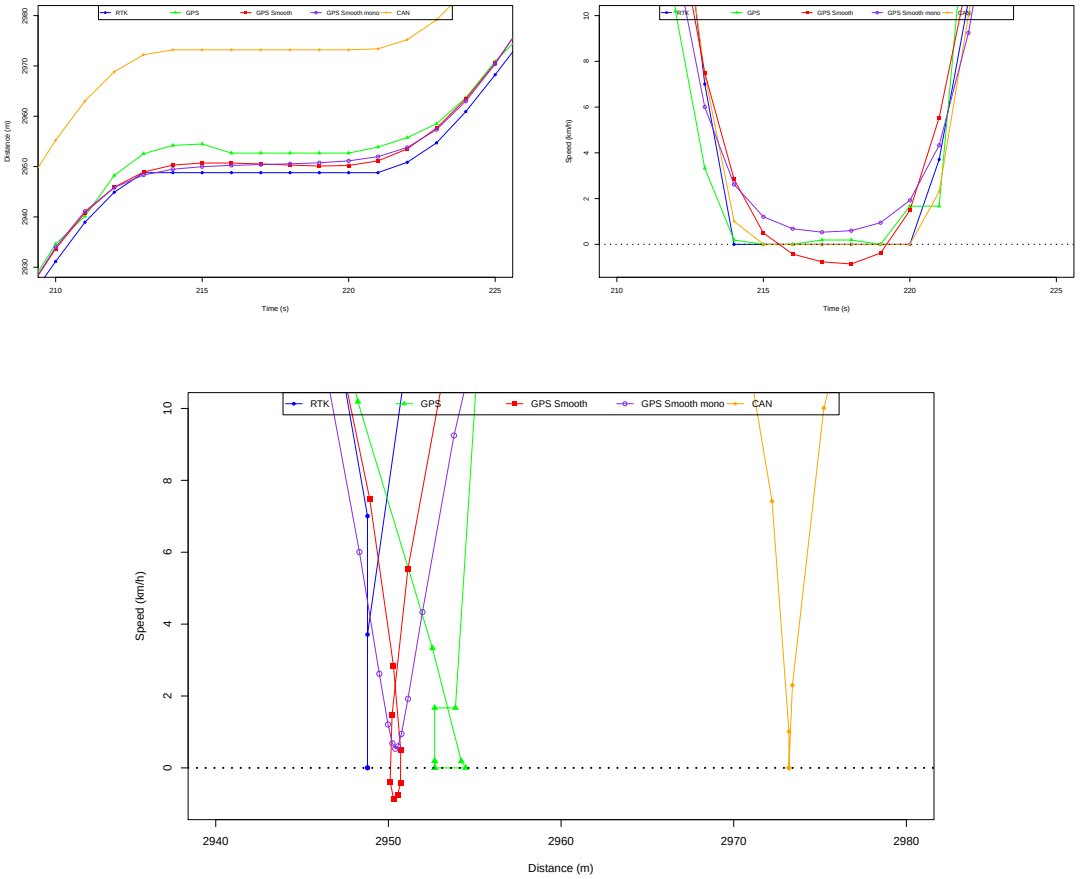


Figure 4.10 - Estimateur $\widehat{F}_{mono}(t)$ (en violet) obtenu par lissage monotone des $\widehat{F}_{\lambda}(t_i)$. Zoom sur l'arrêt (sur l'intervalle de temps $[210, 225]$). Représentation dans les 3 espaces : [`distance`×`temps`, `vitesse`×`temps`] et [`vitesse`×`distance`].

La prise en compte de ces deux contraintes a été traitée séparément. Ainsi, dans un premier temps, nous avons construit un estimateur par spline de lissage de $F(t)$ utilisant l'information disponible sur la dérivée $F'(t)$, et nous avons donné son expression explicite. Par ailleurs, nous avons montré que ce problème général d'estimation d'une fonction de régression à partir d'observations bruitées de celle-ci et de sa dérivée, était un cas particulier du modèle général des splines de lissage décrit au chapitre précédent. Nous avons notamment montré l'importance du choix de la norme associée à l'espace d'étude (ici l'espace de Sobolev $W^m[0,T]$) dans le calcul de l'estimateur, et l'intérêt d'utiliser un semi-noyau à la place du noyau reproduisant R_1 afin de faciliter les calculs numériques. Les résultats numériques sur données simulées et réelles ont montré les bonnes performances de ce premier estimateur. Puis, dans un second temps, nous avons ajouté la contrainte de monotonie sur l'estimateur de $F(t)$ en se ramenant en quelque sorte à une étape de monotonisation de l'estimateur construit à l'étape précédente. Cette deuxième étape nous a permis de construire un estimateur strictement monotone $\widehat{F}_{mono}(t)$ de $F(t)$ et d'en déduire un estimateur $\widehat{v_S} = \widehat{F}'_{mono} \circ \widehat{F}^{-1}_{mono}$ du profil spatial de vitesse v_S. Si l'application sur données réelles a montré les bonnes performances de cet estimateur, nous avons cependant souligné le problème de l'estimation des arrêts du véhicule pour lesquels la fonction $F(t)$ est constante, et que nous n'avons pas réussi à résoudre. Les performances et les limites de la méthodologie de lissage décrite dans ce chapitre seront également montrés au chapitre suivant sur un jeu de données réelles constitué de 78 profils de vitesse.

En conclusion, la méthodologie décrite dans ce chapitre, afin de construire un estimateur $\widehat{v_S}$ d'un profil spatial de vitesse v_S à partir de mesures bruitées de position et de vitesse, se décompose en différentes étapes :

0 **Une étape optionnelle de pré-traitement dans le cas où l'on dispose de mesures de position issues à la fois d'un GPS et de l'odomètre.**
Méthode : construction d'un estimateur ponctuel par fenêtre glissante de la position du véhicule en chaque instant d'échantillonnage à partir des deux principaux capteurs de localisation, à savoir l'odomètre et le GPS.

1 **Une première étape de lissage utilisant l'information sur la dérivée.**
Méthode : Construction d'un estimateur $\widehat{F}(t)$ par spline de lissage de la fonction de régression $F(t)$, à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).

2 **Une seconde étape de lissage sous contrainte de monotonie.**
Méthode : Construction d'un estimateur strictement monotone $\widehat{F}_{mono}(t)$ qui est une "monotonisation" de l'estimateur $\widehat{F}(t)$ construit à l'étape précédente.

Une perspective de recherche est la fusion des étapes 1 et 2 afin de simplifier la procédure. Une solution qui nous a été suggérée est l'intégration de l'estimation des plateaux où $F(t)$ est constante, et correspondants aux arrêts du véhicule, dans l'étape 1 en utilisant les "*mixed splines*" (Berlinet et Thomas-Agnan [15] section

2.3) qui combinent splines d'interpolation et splines de lissage. Ainsi, si on note J la réunion des intervalles de temps correspondants aux instants d'arrêts, le problème s'écrit de la manière suivante :

$$\begin{aligned} \text{minimiser} \quad & \frac{1}{2n}\{\sigma_x^{-2}\sum_{i\notin J}(y_i - F(t_i))^2 + \sigma_v^{-2}\sum_{i\notin J}(v_i - F'(t_i))^2\} + \lambda\int_0^T (F^{(m)}(t))^2 dt, \\ \text{avec} \quad & y_i = F(t_i) \ \text{ pour } i \in J. \end{aligned}$$

Cette approche suppose que les mesures y_i de position des arrêts sont constantes sur chaque sous-intervalle de J correspondant aux instants d'arrêts, ce qui peut être imposé en pré-traitement (voir Andrieu *et al.* [7]). Cependant, cette approche ne garantit pas une stricte monotonie de l'estimateur lorsque le véhicule n'est pas à l'arrêt.

Chapitre 5

Construction de profils agrégés et d'enveloppes de vitesse. Application à l'éco-conduite

Ce dernier chapitre est consacré à la construction de divers profils agrégés représentatifs d'un ensemble de profils spatiaux de vitesse individuels, et permettant ainsi d'estimer les vitesses pratiquées par les usagers de la route le long d'un itinéraire. En effet, nous avons décrit au chapitre précédent, une méthodologie permettant d'obtenir un estimateur d'un profil spatial de vitesse à partir de données bruitées de position et de vitesse. Dans le cas de gros volume de données, on obtient alors un ensemble de profils de vitesse individuels qui sont riches d'information sur le comportement des usagers du réseau, mais qu'il est nécessaire de résumer par des profils agrégés (ex : moyenne, médiane) afin d'en faciliter l'étude. Notons que l'objectif de ce chapitre n'est pas de construire un unique profil de référence qui serait le plus représentatif, au sens d'un certain critère, d'un ensemble de profils individuels, mais plutôt de proposer une méthodologie de construction de divers profils agrégés tels que le profil moyen ou le profil médian. Nous ne chercherons donc pas à comparer ces différents profils agrégés, ce choix étant subjectif et dépendant de l'objectif de l'étude.

Dans une première section, nous rappelons les enjeux et les problématiques de la connaissance des vitesses pratiquées à travers l'étude des profils de vitesse, et nous présentons les contributions proposées dans la suite de ce chapitre. Nous décrivons également le jeu de données utilisé comme exemple d'étude afin d'illustrer les différentes méthodes proposées dans ce chapitre. Ce jeu de données réelles relativement conséquent est constitué de 78 courbes représentant les profils de vitesse obtenus lors du parcours d'une section de 1100 m.

Dans une seconde section, nous proposons une méthodologie de construction d'un profil moyen représentatif d'un ensemble de profils individuels. Dans un premier temps, on se ramène à cadre fonctionnel en utilisant la méthode de lissage décrite au chapitre 4 afin de convertir les données brutes de position et de vitesse

(de nature vectorielle) en profils de vitesse de nature fonctionnelle. Puis, le problème des variations de phase entre les profils individuels, notamment au niveau des arrêts, est mis en évidence.

Dans une troisième section, nous proposons de pallier ce problème de recalage d'un ensemble de profils de vitesse individuels par l'utilisation de la méthode d'alignement par landmarks, fondée sur une mise en correspondance d'un ensemble de points caractéristiques pré-définis. Nous proposons alors la construction de divers profils moyens, chacun étant adapté à une situation de conduite (dans notre cas, l'état du feu de signalisation).

Dans une quatrième section, nous proposons un outil graphique fondé sur l'extension des boîtes à moustaches, classiquement utilisées en statistique univariée, aux données fonctionnelles, et particulièrement adapté à l'étude de la variabilité des vitesses pratiquées selon les individus. Après avoir défini la notion de profondeur statistique, différentes mesures de profondeur sont comparées. Puis nous utilisons une mesure de profondeur convenablement choisie pour construire des enveloppes de vitesse reflétant la dispersion des vitesses le long d'un itinéraire.

Enfin, nous concluons ce chapitre avec une application à l'éco-conduite. Nous proposons d'utiliser les différents outils proposés dans les sections précédentes pour analyser les effets de l'éco-conduite par rapport à une conduite "normale", à travers l'étude de profils de vitesse obtenus avec ces deux styles de conduite.

5.1 L'estimation des vitesses pratiquées : problématiques et étude de cas

5.1.1 L'analyse des profils de vitesse : enjeux et problématiques

Comme nous l'avons vu au chapitre 1, les profils de vitesse sont riches d'information sur le comportement individuel des usagers de la route, et plus particulièrement sur les vitesses pratiquées. Cette information contenue dans les profils de vitesse est rendue accessible grâce au développement des smartphones qui permettent de transformer tout individu en capteur mobile (véhicule traceur), et ainsi d'accroître le nombre de "traces" numériques laissées par les véhicules. Cependant l'estimation des vitesses pratiquées à partir de ces traces constituées de mesures bruitées de vitesse et de position, nécessite l'utilisation de méthodes appropriées.

La méthodologie de lissage développée au chapitre précédent permet de convertir ces données brutes de nature vectorielle en un profil de vitesse de nature fonctionnelle. Cette étape de lissage permet à partir d'une série de mesures d'obtenir une bonne estimation d'un "vrai" profil de vitesse individuel tout en tenant compte de la structure sous-jacente continue de ce type de données. La répétition de cette étape pour plusieurs séries de mesures permet de constituer un ensemble de profils de vitesse individuels le long d'un itinéraire, mais qu'il est nécessaire de résumer afin d'en déduire une estimation des vitesses pratiquées par les usagers. Ceci nous amène à la problématique suivante : déterminer un profil agrégé tel que la moyenne ou la

médiane qui soit représentatif de cet ensemble de profils individuels.

De nombreuses études basées sur l'analyse des profils de vitesse utilisent le profil moyen comme profil agrégé. Par exemple, Hyden et Varhelyi [80] étudient l'impact d'une modification de l'infrastructure sur les vitesses pratiquées en comparant le profil moyen obtenu le long d'un itinéraire avant et après l'ajout de rond-points. De même, Várhelyi *et al.* [168] étudient les effets d'un système ISA actif agissant sur la pédale d'accélérateur en comparant le profil de vitesse moyen sans et avec le système. Cependant, si le profil moyen est un profil représentatif des vitesses pratiquées sur un itinéraire donné, sa construction n'est pas si simple et nécessite l'utilisation de méthodes adaptées. Par exemple, dans leur étude sur le comportement des conducteurs à une intersection lorsque le véhicule tourne à gauche, Laureshyn *et al.* [95] distinguent différentes classes de comportements types et construisent le profil moyen associé. Cependant, suite à des variations de phase (variations horizontales) entre les profils individuels, on observe que le profil moyen obtenu n'est pas un profil type de chacune des classes. Il est donc nécessaire avant de calculer un profil moyen, de synchroniser les profils de vitesse afin de supprimer ces variations de phase, comme par exemple dans les travaux de Kerper *et al.* [84]. Nous proposons aux sections 5.2 et 5.3 une méthodologie plus élaborée, tenant compte de la nature fonctionnelle des profils estimés lors de l'étape de lissage, permettant de pallier ce problème de décalage des profils de vitesse, et conduisant ainsi à l'obtention d'un profil moyen représentatif d'un ensemble de profils de vitesse individuels.

L'étude de la variabilité des vitesses pratiquées entre les individus sur une section de route donnée est également un sujet important à traiter. Par exemple, Violette *et al.* [170] proposent une méthodologie de détermination d'un profil de vitesse V85 le long d'un itinéraire, fondée sur la fusion de mesures de vitesses issues de dispositifs bord de voie et d'un véhicule instrumenté, et l'applique ensuite à l'étude de la sévérité de chocs latéraux en fonction de la vitesse pratiquée. Boonsiripant [22] considère également les variations de vitesse comme indicateur de risque et propose divers indicateurs agrégés, calculés à partir des profils moyens et centiles (ex : profil V85 ou V50), permettant de résumer la variabilité des vitesses entre les individus par un indicateur global. Nous proposons à la section 5.4 un autre outil, tenant compte de la nature fonctionnelle des données, et permettant la construction d'enveloppes de vitesse reflétant la dispersion des vitesses pratiquées sur une section donnée.

5.1.2 Exemple d'étude : description du jeu de données

Dans la suite de ce chapitre, nous proposons d'illustrer la méthodologie proposée sur un jeu de données issu d'une expérimentation réalisée par le LIVIC entre septembre et octobre 2012, et dont l'objectif est d'identifier les conséquences de la pratique de l'éco-conduite en termes d'économie de carburant et de comportements à risque en milieu urbain. Dans le cadre de cette expérimentation, 39 participants ont effectué deux fois un trajet de référence : une première fois sans consignes parti-

culières (conduite normale), puis une deuxième fois après avoir suivi une initiation à l'éco-conduite (présentation des principales règles d'éco-conduite et mise en oeuvre sur un trajet de 10 min en suivant les conseils d'un expérimentateur). Le trajet de référence d'une longueur d'environ 17 km est essentiellement situé dans la commune de Versailles, et est constitué de segments de route de type urbain et inter-urbain. Pour des raisons de logistique, deux véhicules ont été utilisés pour cette expérimentation : une Renault Clio III équipée d'un GPS Garmin 16x LVC, et une Renault Modus équipée d'un GPS GlobalSat BR-355. Parmi les 39 participants, 20 ont conduit le véhicule Renault Clio et 19 ont conduit le véhicule Renault Modus. Les deux véhicules ont également été équipés de deux radars (un à l'avant et un à l'arrière) permettant de mesurer les inter-distances, et d'une caméra à l'avant permettant de filmer la route et d'analyser ainsi les situations de conflits rencontrées.

Dans la suite de ce chapitre, nous nous focalisons uniquement sur une petite section de ce trajet de référence. La section étudiée d'une longueur d'environ 1100 m, est de type urbain-résidentiel et est constituée d'un stop, de deux rond-points, et d'un feu de signalisation. Cette section est illustrée à la figure 5.1 et correspond au trajet allant du point A au point B. On peut noter que cette section correspond à une zone résidentielle avec généralement peu de trafic, une bonne visibilité au niveau des intersections, et pas de zones dangereuses (pas de priorité à droite sans signalement).

La suite de ce chapitre est consacrée à l'étude des profils spatiaux de vitesse issus de ce jeu de données. Par conséquent, nous utiliserons uniquement les mesures de position et de vitesse, et plus particulièrement les mesures issues du GPS. La section étudiée étant située dans un environnement bien dégagé (pas de zones de "canyon urbain"), nous n'avons pas observé de problème de réception du signal GPS. Ainsi, les coordonnées du GPS ayant toujours correspondues aux coordonnées géométriques de la route, aucun algorithme de map-matching n'a été utilisé. La distance parcourue a donc été calculée en sommant directement les segments reliant chacun des points GPS. Enfin, les mesures de position et de vitesse issues du GPS ont été ré-échantillonnées à 1 Hz, soit 1 mes/s.

5.2 Construction d'un profil moyen

Dans un premier temps, nous proposons d'étudier les profils spatiaux de vitesse issus du jeu de données décrit à la section 5.1.2 sans distinguer le type de conduite (normale ou éco-conduite). Les 39 participants ayant réalisés chacun deux fois le trajet de référence, on obtient ainsi 78 profils spatiaux individuels de vitesse sur la section de 1100 m décrite à la figure 5.1. Notons que dans la suite de l'étude, nous ne tiendrons pas compte de la corrélation entre les deux trajets d'un même conducteur. Nous nous plaçons ainsi dans le cadre général d'une étude de mesures issues de véhicules traceurs, pour laquelle on observe un ensemble de traces laissées

Figure 5.1 - Cartographie de la section étudiée.

par des véhicules sans aucune information sur le conducteur ou le type de véhicule. L'objectif de cette section est de proposer une méthodologie de construction de divers profils moyens représentatifs de cet ensemble de profils, chacun étant adapté à une situation de conduite (ex : feu rouge ou vert).

5.2.1 Des données brutes aux courbes : l'étape de lissage

Les données brutes de position et de vitesse issues du GPS pour les 78 trajets effectués sur la section décrite à la figure 5.1 sont représentées à la figure 5.2. Ces données sont représentées dans les trois espaces `distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`.

Ces mesures de position et de vitesse étant observées en des instants discrétisés, les données brutes représentées à la figure 5.2 sont de nature vectorielle. La première étape de cette étude consiste donc à convertir les données brutes en objet fonctionnel par une étape de lissage. Pour ce faire, nous appliquons la méthode de lissage que nous avons développée au chapitre 4, constituée des deux étapes suivantes :

1. **Une première étape de lissage utilisant l'information sur la dérivée.**
   Méthode : Construction d'un estimateur $\widehat{F}(t)$ par spline de lissage de la fonction de régression $F(t)$, à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).
2. **Une seconde étape de lissage sous contrainte de monotonie.**
   Méthode : Construction d'un estimateur strictement monotone $\widehat{F}_{mono}(t)$ qui est une "monotonisation" de l'estimateur $\widehat{F}(t)$ construit à l'étape précédente.

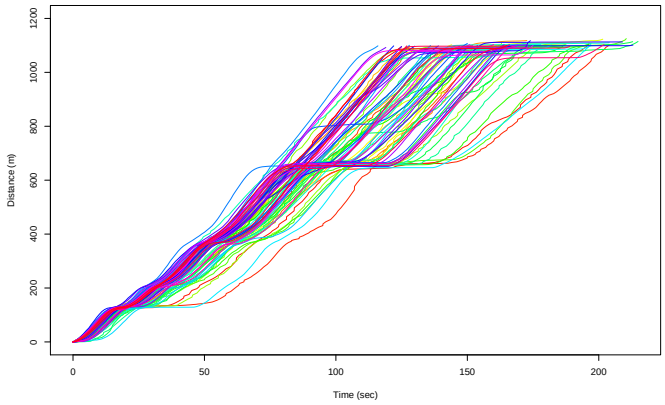
A l'étape 1, nous construisons un estimateur par spline de lissage $\widehat{F}_{\lambda_i}(t)$ de chacune des fonctions $F_i(t)$ représentant la distance parcourue en fonction du temps correspondant au trajet i, pour $i = 1, \ldots, 78$, avec les paramètres suivants :

- une estimation des variances $\sigma^2_{x,i}$ et $\sigma^2_{v,i}$ des mesures de position et de vitesse ;
- $m = 3$ (spline quintique de degré 5) ;
- une sélection automatique du paramètre de lissage λ_i par minimisation du critère GML.

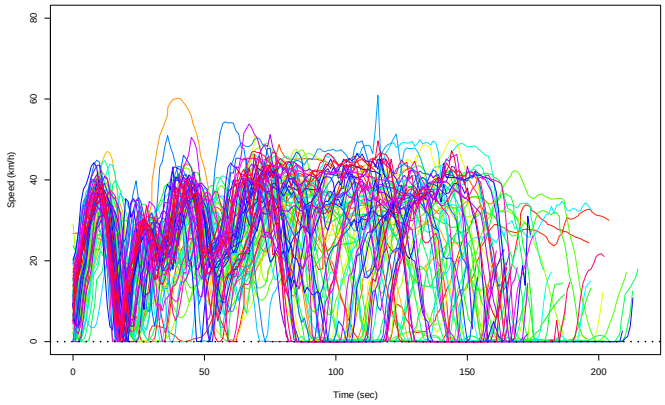
Par dérivation, on en déduit un estimateur $\widehat{F}'_{\lambda_i}(t)$ de chacune des fonctions $F'_i(t)$ représentant la vitesse en fonction du temps (espace `vitesse`×`temps`). Puis, on en déduit un estimateur de chacun des profils spatiaux de vitesse par l'application $\widehat{F}'_{\lambda_i}(t) \circ \widehat{F}^{-1}_{\lambda_i}(t)$. Les résultats obtenus après l'étape 1 de cette procédure de lissage sont illustrés à la figure 5.3.

A l'étape 2, nous "monotonisons" chacun des $\widehat{F}_{\lambda_i}(t)$ obtenus à l'étape 1, pour $i = 1, \ldots, 78$, en utilisant la méthode décrite à la section 4.4.3.1 avec les paramètres

a) Distance parcourue en fonction du temps.



b) Vitesse en fonction du temps.



c) Profils spatiaux de vitesse (Vitesse en fonction de la distance parcourue).

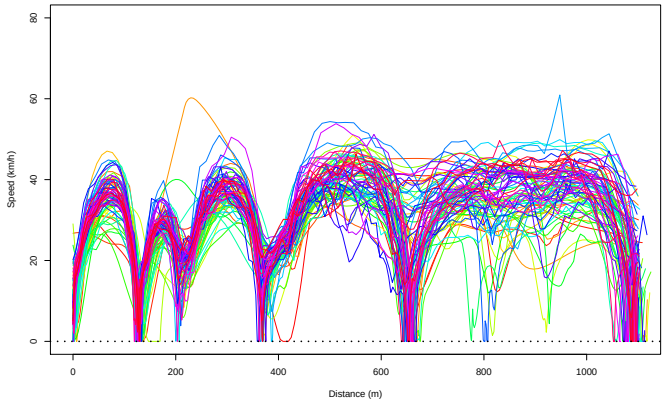
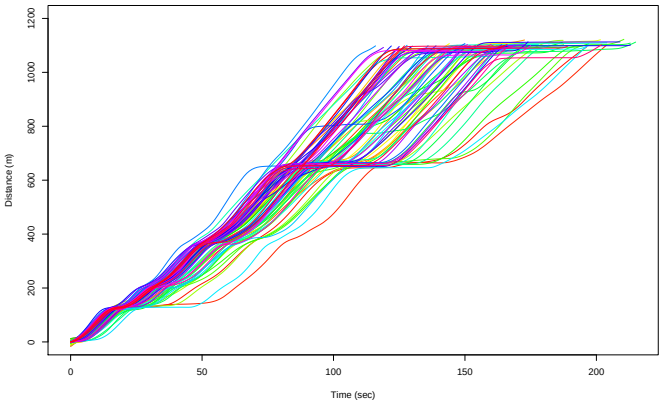
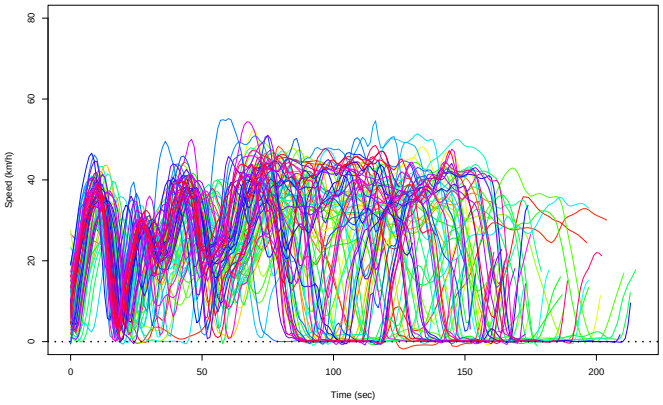


FIGURE 5.2 - Données brutes. Représentation dans les 3 espaces : `distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`.

a) Distance parcourue en fonction du temps.



b) Vitesse en fonction du temps.



c) Profils spatiaux de vitesse (Vitesse en fonction de la distance parcourue).

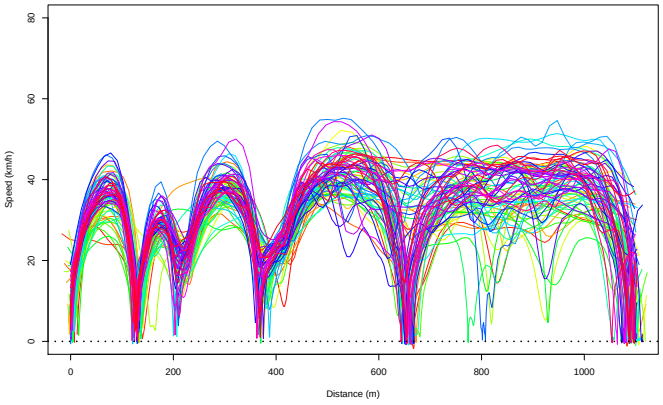


Figure 5.3 - Lissage des données : Etape 1 (utilisation de l'information sur la dérivée). Représentation dans les 3 espaces : `distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`.

suivants :

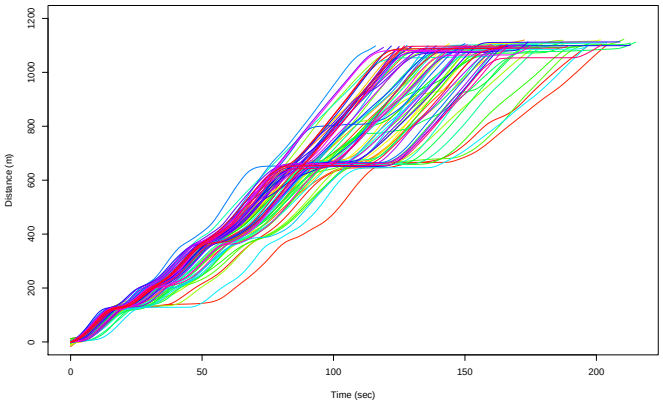
- $m = 3$ comme degré pour la pénalité ;
- une sélection du paramètre λ par essais et erreurs (λ compris entre 10^{-4} et 10^{-1} selon les profils estimés).

Les résultats obtenus après cette étape de monotonisation sont illustrés à la figure 5.4.

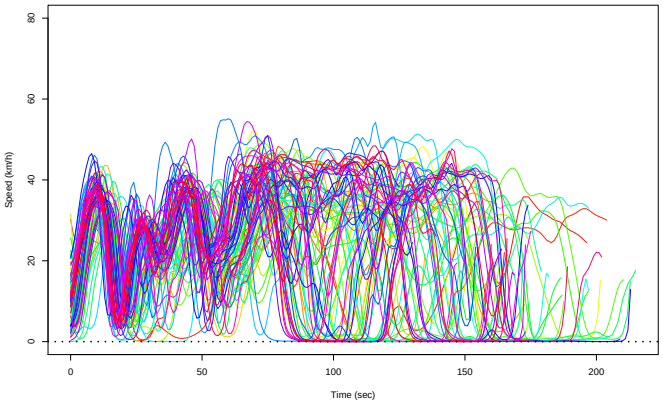
De manière générale, on observe sur les figures 5.3 et 5.4 que la forme globale des profils de vitesse est bien conservée après les deux étapes de lissage, quel que soit l'espace étudié (`distance`×`temps`, `vitesse`×`temps` ou `vitesse`×`distance`). On observe également une atténuation de certains pics présents dans les données brutes (courbes orange et bleue sur la figure 5.2.c) et qui semblent correspondre à des valeurs aberrantes. Notons également que ce lissage a permis de corriger certaines valeurs manquantes présentes dans les données brutes. Enfin, on observe sur la figure 5.3.b des vitesses négatives (entre 120 et 150 s) qui sont corrigées lors de l'étape de monotonisation (figure 5.4.b).

Afin d'observer plus précisément les résultats des deux étapes de lissage lorsque la vitesse s'annule, nous proposons d'effectuer un zoom sur certaines zones de la section étudiée correspondant aux arrêts des véhicules. Ainsi, les figures 5.5 et 5.6 représentent les données brutes correspondant aux profils spatiaux de vitesse, ainsi que les courbes lissées obtenues après chacune des deux étapes de lissage, au niveau des deux principaux arrêts imposés par l'infrastructure, à savoir le stop (situé entre 110 et 150 m) et le feu de signalisation (situé entre 630 et 680 m). On observe sur les données brutes des erreurs de positionnement puisque le véhicule se déplace alors que sa vitesse est nulle (par exemple, courbe turquoise sur la figure 5.5.a entre 120 et 125 m, ou courbe fuschia sur la figure 5.6.a autour de 640 m). Ces erreurs sont dues à une imprécision des données de position du GPS. On observe que ces erreurs de positionnement sont corrigées lors de l'étape 1 du lissage mais que l'on obtient des vitesses négatives puisqu'aucune hypothèse de monotonie n'ait faite dans cette première étape. L'étape 2 de monotonisation permet de corriger ces vitesses négatives, mais a tendance à sur-estimer la vitesse au niveau de l'arrêt du véhicule correspondant au minimum local de chaque courbe. Ce problème déjà évoqué au chapitre précédent (section 4.4.3.2) est lié à la difficulté d'estimer les "plateaux" de la fonction $F(t)$ lorsque le véhicule est à l'arrêt, et au fait que notre estimateur soit strictement monotone. Cette difficulté est d'autant plus importante lorsque l'arrêt est très court. En effet, on observe que cette sur-estimation des vitesses au niveau des arrêts du véhicule est particulièrement marquée pour l'arrêt au stop (figure 5.5.c) qui est un arrêt court d'une durée moyenne inférieure à 5 s, alors que les arrêts au feu d'une durée moyenne d'environ 15 s sont bien estimés (figure 5.6.c).

a) Distance parcourue en fonction du temps.



b) Vitesse en fonction du temps.



c) Profils spatiaux de vitesse (Vitesse en fonction de la distance parcourue).

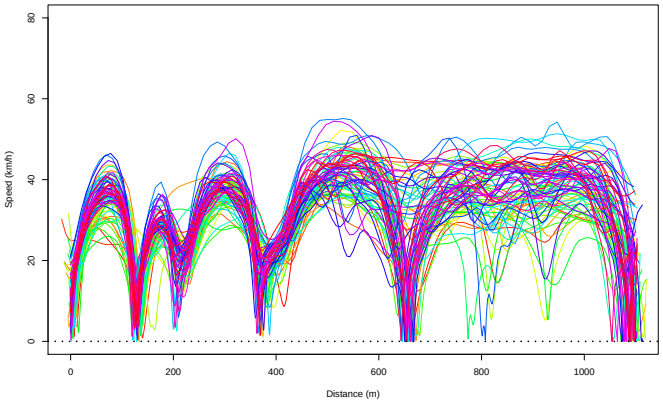


Figure 5.4 - Lissage des données : Etape 2 (monotonisation). Représentation dans les 3 espaces : `distance`×`temps`, `vitesse`×`temps` et `vitesse`×`distance`.

a) Données brutes. b) Etape 1 du lissage. c) Etape 2 du lissage.

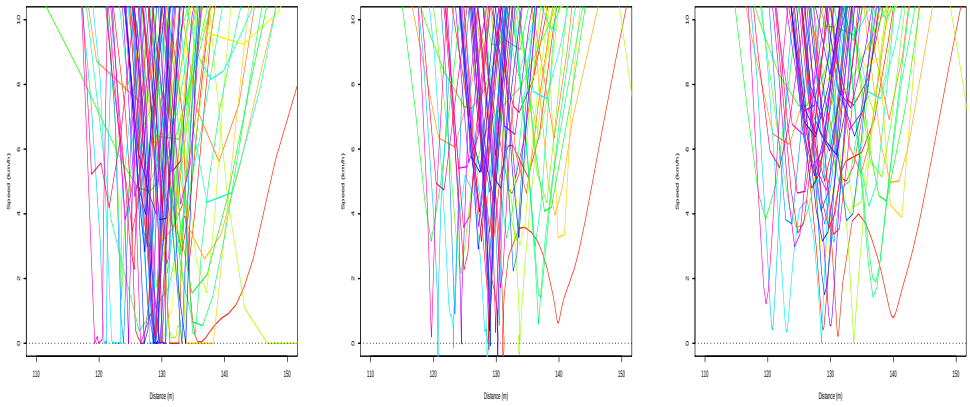


Figure 5.5 - Profils spatiaux de vitesse : données brutes et courbes lissées obtenues après chacune des deux étapes de lissage. Zoom sur l'arrêt au stop situé entre 110 et 150 m.

a) Données brutes. b) Etape 1 du lissage. c) Etape 2 du lissage.

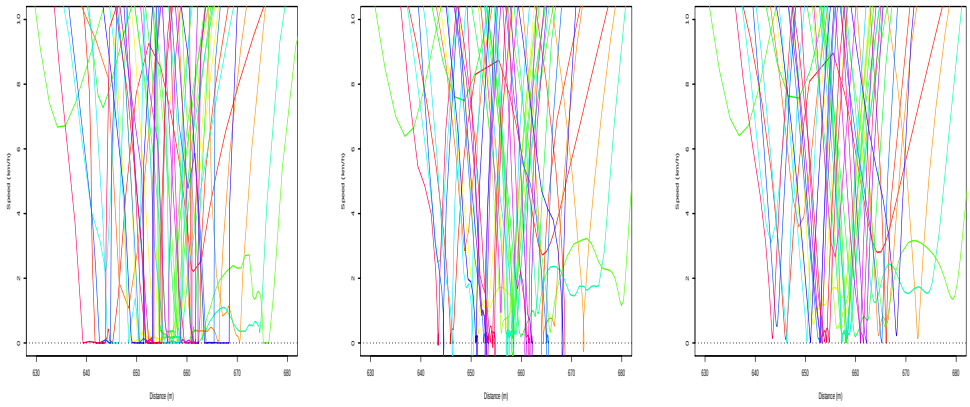


Figure 5.6 - Profils spatiaux de vitesse : données brutes et courbes lissées obtenues après chacune des deux étapes de lissage. Zoom sur l'arrêt au feu situé entre 630 et 680 m.

5.2.2 Le problème du décalage

On s'intéresse à l'ensemble des 78 profils spatiaux de vitesse lissés obtenus à la figure 5.4.c, et on cherche à construire un profil moyen représentatif de cet ensemble de profils (profil de vitesse de référence). La section étudiée comprenant un feu de signalisation situé à environ 650 m, nous proposons de décomposer cet ensemble de profils en deux groupes en distinguant les cas où le feu était rouge et les cas où le feu était vert, afin de se ramener à des situations de conduite comparables. On obtient ainsi deux sous-ensembles de profils, chacun correspondant à une situation de conduite :

- 36 profils spatiaux de vitesse de la section étudiée dans le cas où le feu était rouge. On note $\mathcal{E}_{red}$ =\{indices des véhicules qui se sont arrêtés au feu parmi les 78 étudiés\} et $N_{red} = card(\mathcal{E}_{red}) = 36$.
- 42 profils spatiaux de vitesse de la section étudiée dans le cas où le feu était vert. On note $\mathcal{E}_{green}$ =\{indices des véhicules qui ne se sont pas arrêtés au feu parmi les 78 étudiés\} et $N_{green} = card(\mathcal{E}_{green}) = 42$.

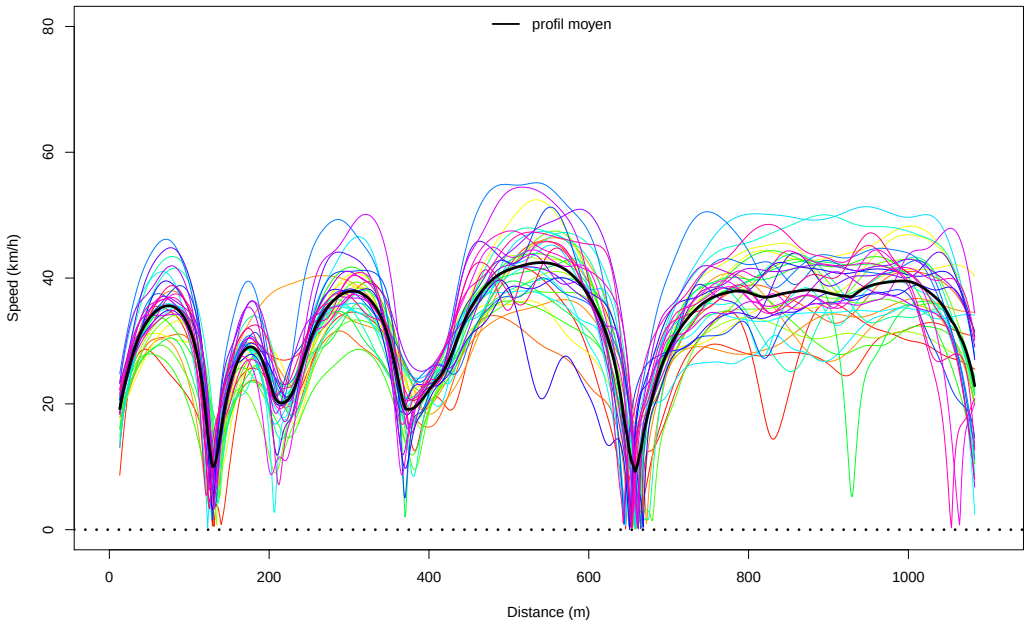
Notons que dans le cadre de l'expérimentation dont est issu l'ensemble des profils de vitesse étudiés, l'information concernant l'état du feu de signalisation a été enregistré (présence d'une caméra filmant la route), ce qui nous a permis d'effectuer cette classification. Lorsque cette information n'est pas disponible, il est alors possible d'utiliser des méthodes de classification non supervisée afin de discriminer les états feu rouge/feu vert caractérisés par un passage à zéro de la vitesse dans le premier cas, et le maintien d'une vitesse constante dans le second cas.

La figure 5.7 représente les deux sous-ensembles de profils de vitesse ainsi que les profils moyens associés (en noir) définis de la manière suivante :

- dans le cas où le feu est rouge, le profil moyen $\overline{v_{S,red}} = N_{red}^{-1} \sum_{j=1}^{N_{red}} v_{S,j}(x)$, où $j \in \mathcal{E}_{red}$, correspond à la moyenne des différents profils $v_{S,j}$ en chaque point x du domaine étudié ;
- dans le cas où le feu est vert, le profil moyen $\overline{v_{S,green}} = N_{green}^{-1} \sum_{j=1}^{N_{green}} v_{S,j}(x)$, où $j \in \mathcal{E}_{green}$, correspond à la moyenne des différents profils $v_{S,j}$ en chaque point x du domaine étudié.

Pour chacun des deux sous-ensembles, le profil moyen est représenté avec un pas d'échantillonnage de 1 m. Les figures 5.7.a et 5.7.b mettent en évidence un problème majeur dans la forme des profils moyens : les variations locales de la position des arrêts au feu et au stop conduisent à l'obtention de profils moyens non représentatifs et irréalistes au niveau du stop et du feu de signalisation. En effet, si le stop et le feu de signalisation sont des éléments fixes de l'infrastructure, chaque véhicule ne s'arrête pas exactement au même endroit (ex : file de véhicules), ce qui entraîne une variation locale de la position de l'arrêt et biaise donc le calcul d'un profil agrégé tel que le profil moyen. Il est donc nécessaire avant de calculer un profil moyen, de recaler les profils de vitesse en certains points caractéristiques tels que les arrêts afin d'obtenir un profil moyen représentatif de l'ensemble des profils.

a) Cas où le feu est rouge (36 courbes).



b) Cas où le feu est vert (42 courbes).

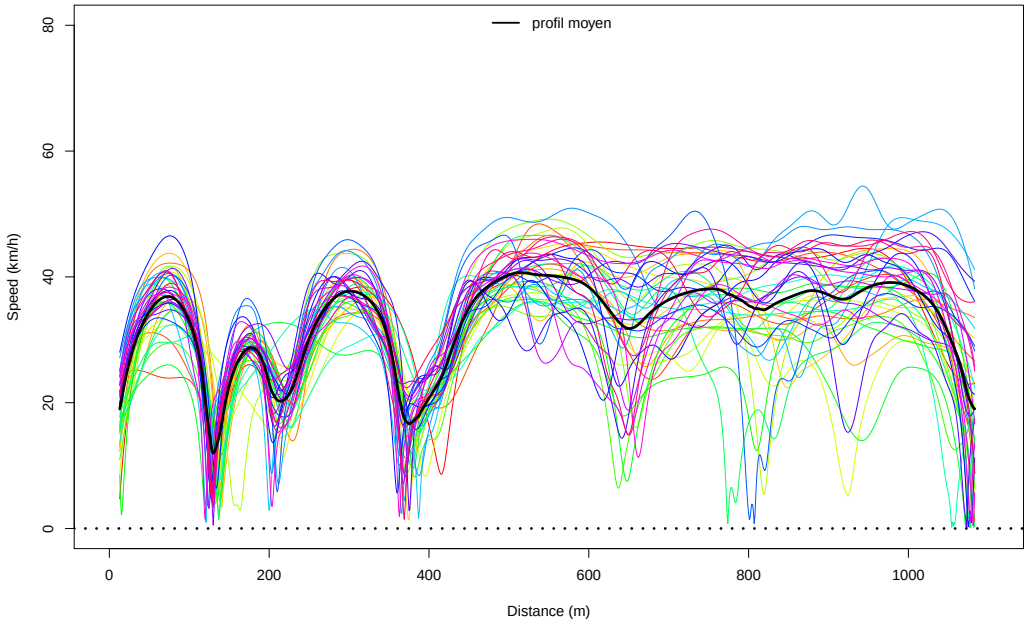


Figure 5.7 - Profils spatiaux de vitesse lissés : distinction feu rouge/feu vert. Le profil moyen est représenté en noir.

5.3 Recalage des profils de vitesse

Le problème du recalage (ou de la synchronisation) d'un ensemble de signaux apparaît dans de nombreux domaines tels que la biologie, la météorologie, la reconnaissance de formes... (Ramsay et Li [134], Bigot [20]). En effet, lorsque l'on compare un ensemble de signaux, on observe deux types de variation : la variation d'amplitude ("*amplitude variation*") correspondant à la variation verticale, et la variation de phase ("*phase variation*") correspondant à la variation horizontale (Ramsay et Silverman [130], Ramsay et Silverman [128]). Afin de construire un signal représentatif de cet ensemble et présentant des caractéristiques similaires à cet ensemble de signaux, il est alors nécessaire de supprimer ces différences de phase entre les signaux en les déformant. Ce problème de recalage ou d'alignement de signaux (appelé "*curve registration*" en anglais) est un sujet sur lequel la littérature est relativement vaste. On citera notamment les articles de Sakoe et Chiba [142], Kneip et Gasser [90], Ramsay et Li [134] et Wang et Gasser [179].

Dans le cadre de l'étude des profils de vitesse, le problème du recalage apparaît notamment dans Violette *et al.* [170] et Kerper *et al.* [84]. Dans les deux cas, les auteurs proposent de translater les profils initiaux afin de les aligner ("*shift registration*"). Ainsi, si on note $v_1, \ldots, v_m$ un ensemble de m profils de vitesse, les profils obtenus après recalage sont de la forme :

$$\widetilde{v}_i(x) = v_i(x + \delta_i), \quad i = 1, \ldots, m, \tag{5.1}$$

où δ_i est le paramètre de translation ("*shift parameter*"). Cependant si une simple translation est adéquat pour aligner des signaux en un point, celle-ci n'est pas adaptée dès que l'on cherche à aligner plusieurs points caractéristiques. Il est alors nécessaire de construire des fonctions de déformations h_i plus complexes (fonctions non linéaires), afin que les points caractéristiques des courbes déformées

$$\widetilde{v}_i(x) = v_i[h_i(x)], \quad i = 1, \ldots, m, \tag{5.2}$$

soient alignés. La méthode d'alignement dite par landmarks permet de traiter ce type de problème et est l'objet de la section suivante.

5.3.1 Principe de l'alignement à partir de landmarks

Le principe de l'alignement par landmarks consiste à déterminer des transformations qui mettent en correspondance un ensemble de points caractéristiques, appelés landmarks, qui sont communs à l'ensemble des signaux devant être recalés. Les points désignés comme landmarks sont généralement les positions des maximas, des minimas, des points d'inflexion, ou les passages par zéro. Après déformation, les landmarks des signaux se trouvent ainsi à la même position.

La méthode d'alignement par landmarks de m signaux $f_1, \ldots, f_m$ définis sur [0,X] se décompose en plusieurs étapes :

1. Définition des points caractéristiques qui seront utilisés comme landmarks (ex : minima, maxima, passages par zéro...).
2. Extraction des landmarks $x_{i,1}, \ldots, x_{i,K}$ à partir d'une série d'observations de chaque signal f_i, $i = 1 \ldots, m$. Les signaux observés étant bruités, les landmarks $x_{i,1}, \ldots, x_{i,K}$ sont généralement extraits à partir d'un estimateur $\widehat{f_i}$ de f_i.
3. Détermination des landmarks de référence $x_{0,1}, \ldots, x_{0,K}$, i.e. les points en lesquels les courbes doivent se correspondre.
4. Calcul des transformations $h_1, \ldots, h_m$ qui alignent les landmarks devant être mis en correspondance, i.e. pour tout $i = 1, \ldots, m$, $h_i(x_{0,j}) = x_{i,j}$, $j = 1, \ldots, K$.
5. Déformation des signaux à l'aide des transformations obtenues à l'étape précédente. Les signaux déformés $\widetilde{f_i}(x) = f_i[h_i^{-1}(x)]$, $i = 1, \ldots, m$, sont alors alignés aux points $x_{0,1}, \ldots, x_{0,K}$.

Les fonctions de déformations $h_i(x)$, $i = 1 \ldots, m$, appelées "*warping functions*" en anglais, doivent vérifier les propriétés suivantes :

- Conditions aux bornes : $h_i(0) = 0$, $h_i(X) = X$.
- Alignement des landmarks : $h_i(x_{0,j}) = x_{i,j}$.
- Stricte monotonie : $x_1 < x_2$ implique $h_i(x_1) < h_i(x_2)$.

La condition de monotonie permet de respecter le séquencement des points.

Notons que la méthode d'alignement par landmarks peut être étendue à l'ensemble des points du signal. Ces méthodes d'alignement, dit global, consistent à déterminer des transformations suffisamment régulières qui alignent l'ensemble des points d'un signal au sens d'une certaine distance entre les signaux à aligner. Si l'on souhaite aligner deux signaux f_1 et f_2 définis sur l'ensemble Ω, le principe consiste alors à trouver une transformation u telle que :

- $f_1(x) \approx f_2(u(x))$, pour tout $x \in \Omega$,
- u soit suffisamment lisse (afin de ne pas trop déformer le signal f_2).

Nous renvoyons le lecteur à Ramsay et Li [134] et Wang et Gasser [179] pour plus de détails sur ces méthodes.

Remarque 5.1.
Si la détermination des landmarks devant être mis en correspondance se fait généralement manuellement, en essayant de repérer le nombre de landmarks communs à tous les signaux, cette étape peut être rendue difficile par la présence de landmarks erronés ou l'absence de landmarks sur certains des signaux. Afin de pallier ce problème, des méthodes automatiques de détermination des points caractéristiques communs à un ensemble de signaux ont été développées. On citera notamment les travaux de Gasser et Kneip [67] et la méthode développée par Bigot [20] basée sur la notion d'intensité structurelle.

5.3.2 Calcul de profils moyens après recalage

On considère l'ensemble des 78 profils spatiaux de vitesse lissés obtenus à la figure 5.4.c et l'on cherche à construire un profil moyen qui soit représentatif de cet ensemble et qui ne tienne pas compte des variations de phase locales. Notons que le fait de s'être ramené à un cadre fonctionnel lors de l'étape de lissage nous permet de tenir compte de ce type de caractéristique fonctionnelle liée au déphasage de courbes et donc d'utiliser des méthodes de recalage de signaux. S'il est possible de procéder à un alignement global, la difficulté consiste à déterminer le profil de référence servant de profil cible pour l'alignement de l'ensemble des profils. Nous avons donc choisi d'utiliser un alignement par landmarks qui nécessite uniquement la détermination de points de référence. La difficulté consiste alors à déterminer l'ensemble des landmarks qui doivent être mis en correspondance.

Comme nous l'avons vu à la section précédente, la première étape de la méthode d'alignement par landmarks consiste à définir les points caractéristiques qui seront utilisés comme landmarks. Dans notre étude, plusieurs choix peuvent être faits : les positions des deux minima locaux correspondants au franchissement des deux rond-points, les passages à zéro correspondants aux arrêts du véhicule... Nous avons choisi de définir comme landmarks les positions des deux éléments de l'infrastructure imposant un arrêt du véhicule, à savoir le stop et le feu de signalisation lorsque celui-ci est rouge.

La seconde étape consiste alors à extraire ces landmarks de notre jeu de données. Pour chacun des profils estimés $\widehat{v}_{S,i}(x)$, $i = 1, \ldots, 78$, nous avons déterminé le passage par zéro $x_{i,stop}$ $(i = 1, \ldots, 78)$ au niveau du stop, et pour les 36 véhicules qui se sont arrêtés au feu, nous avons déterminé le passage par zéro $x_{j,feu}$ au niveau du feu où $j \in \mathcal{E}_{red}$. Notons que si le stop impose normalement un arrêt du véhicule et donc un passage par zéro du profil de vitesse, dans cette étude la plupart des véhicules ne se sont pas arrêtés au stop (voir figure 5.5.a). De plus, l'estimateur proposé ayant tendance à sur-estimer les vitesses au niveau des arrêts, nous avons caractérisé les landmarks $x_{i,stop}$ $(i = 1, \ldots, 78)$ et $x_{j,feu}$ $(j \in \mathcal{E}_{red})$ par les minima locaux au niveau du stop et du feu de signalisation.

La troisième étape de l'alignement par landmarks consiste à déterminer les landmarks de référence en lesquels les profils seront alignés. Nous avons choisi de définir comme landmarks de référence la position moyenne des arrêts au stop et au feu, i.e.

$$x_{0,stop} = \frac{1}{78}\sum_{i=1}^{78} x_{i,stop} = 129\ m \quad \text{et} \quad x_{0,feu} = \frac{1}{36}\sum_{j\in\mathcal{E}_{red}} x_{j,feu} = 656\ m.$$

La quatrième étape consiste à déterminer les transformations $h_1, \ldots, h_{78}$ qui alignent les landmarks devant être mis en correspondance. Nous avons choisi de

prendre une spline cubique d'interpolation monotone h_i vérifiant les propriétés suivantes :

- Conditions aux bornes : pour $i = 1, \ldots, 78$, $h_i(0) = 0$ et $h_i(1100) = 1100$.
- Alignement des landmarks : pour $i = 1, \ldots, 78$, $h_i(x_{0,stop}) = x_{i,stop}$, et pour $j \in \mathcal{E}_{red}$, $h_j(x_{0,feu}) = x_{j,feu}$.

Cependant, afin de ne pas trop déformer les profils dans le voisinage des arrêts et afin que les profils déformés soient de vrais profils (i.e. appartiennent à l'espace $\mathcal{E}_{SSP}$ défini en 2.1), nous avons choisi d'imposer une condition supplémentaire sur les transformations h_i, à savoir d'appliquer une transformation linéaire de pente égale à 1 au voisinage des arrêts au stop et au feu. Nous avons choisi de prendre un voisinage de longueur 100 m afin de conserver la forme des profils sur les 50 m précédant l'arrêt (phase de décélération) et les 50 m suivant l'arrêt (phase d'accélération). On ajoute ainsi les landmarks suivants :

$$x_{i,stop} - d \text{ et } x_{i,stop} + d \text{ pour } i = 1, \ldots, 78, \text{ et } x_{j,feu} - d \text{ et } x_{j,feu} + d \text{ pour } j \in \mathcal{E}_{red},$$

et les landmarks de référence associés :

$$x_{0,stop} - d,\ x_{0,stop} + d,\ x_{0,feu} - d,\ x_{0,feu} + d,$$

où $d = 50$ correspond à la moitié de la longueur du voisinage autour de chacun des arrêts. Cette contrainte supplémentaire se traduit alors par les propriétés suivantes sur les fonctions de transformation :

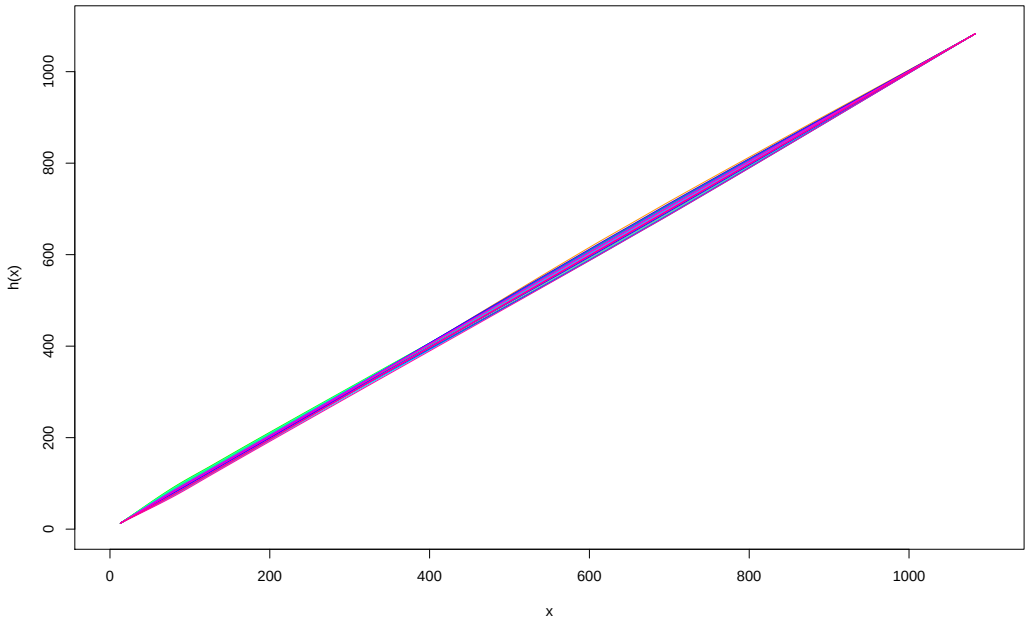
- $h_i(x) = x + |x_{i,stop} - x_{0,stop}|$ sur l'intervalle $[x_{0,stop} - d, x_{0,stop} + d]$ pour $i = 1, \ldots, 78$,
- $h_j(x) = x + |x_{j,feu} - x_{0,feu}|$ sur l'intervalle $[x_{0,feu} - d, x_{0,feu} + d]$ pour $j \in \mathcal{E}_{red}$.

Enfin les profils de vitesse $\widehat{v}_{S,i}(x)$ $(i = 1, \ldots, 78)$ sont déformés à l'aide de ces transformations. Les profils déformés $\widetilde{v}_{S,i}(x) = \widehat{v}_{S,i}[h_i^{-1}(x)]$ sont alors alignés au point $x_{0,stop}$ uniquement dans le cas où le feu est vert, et aux points $x_{0,stop}$ et $x_{0,feu}$ dans le cas où le feu est rouge.

Les fonctions de transformation $h_i(x)$ sont représentées à la figure 5.8 en distinguant les cas feu rouge/feu vert. Pour simplifier la mise en oeuvre numérique, nous avons utilisé la fonction `splinefun` du logiciel R avec l'option "monoH.FC" qui permet de construire une fonction cubique $\mathcal{C}^1$ monotone d'interpolation utilisant les polynômes d'Hermite et basée sur une méthode développée par Fritsch et Carlson [64]. Cette fonction présente l'avantage d'être disponible immédiatement et de construire une bonne approximation des fonctions de transformation $h_i(x)$ au voisinage des arrêts.

Les profils de vitesse obtenus après recalage sont représentés à la figure 5.9. Les profils moyens (en noir) ont été calculés à partir de ces nouveaux profils déformés avec un pas d'échantillonnage de 1 m. On constate que le calcul de ces profils agrégés

a) Cas où le feu est rouge (36 courbes).



b) Cas où le feu est vert (42 courbes).

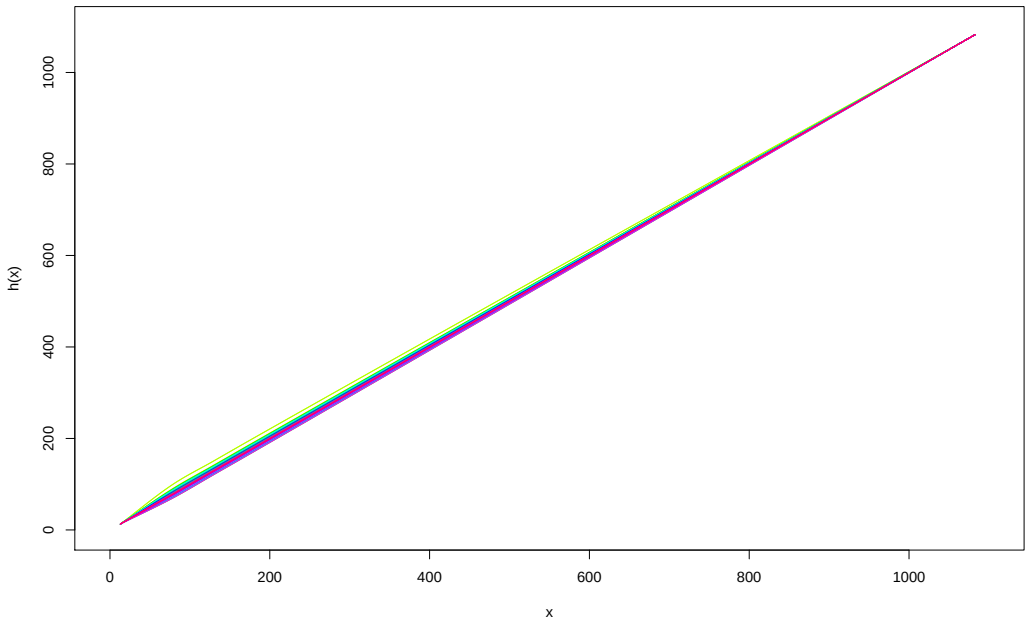


Figure 5.8 - Fonctions de transformation $h_i(x)$: distinction feu rouge/feu vert.

après recalage donne une bien meilleure représentation d'un profil représentatif de l'ensemble des profils que ceux obtenus sans recalage (figure 5.7).

Notons également que, en théorie, le profil moyen obtenu après recalage des profils au niveau des arrêts devrait être un "vrai" profil de vitesse appartenant à l'espace $\mathcal{E}_{SSP}$ (définition 2.1). En effet, nous avons vu au chapitre 2 que, suite à la propriété de non dérivabilité des profils spatiaux de vitesse aux points où la vitesse est nulle, faire la moyenne arithmétique de plusieurs profils spatiaux de vitesse n'avait de sens que dans deux cas :

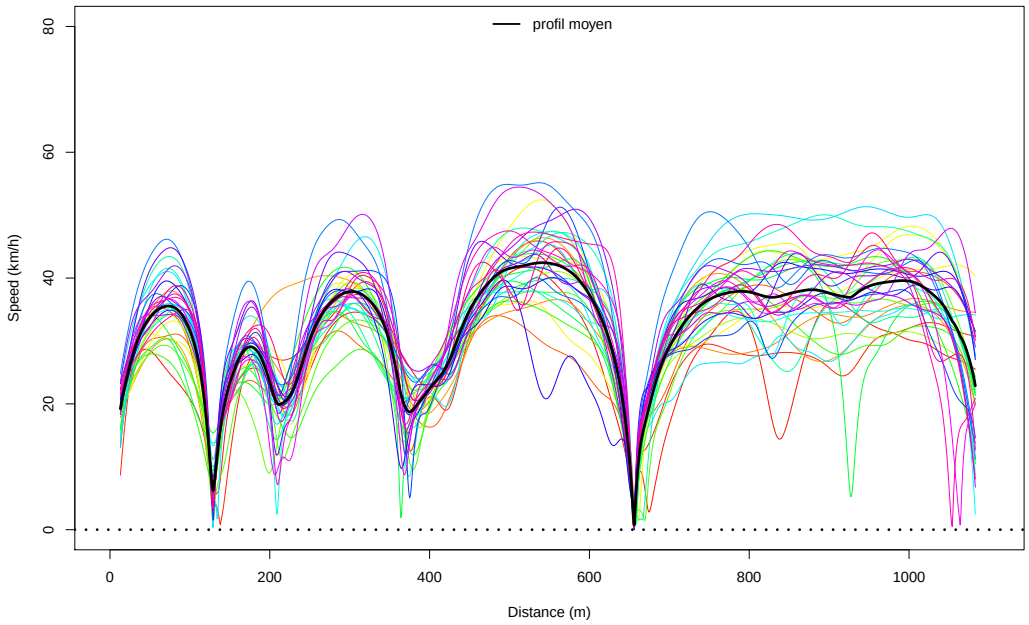
- soit lorsque tous les profils sont strictement positifs (pas d'arrêts),
- soit lorsque tous les profils passent par zéro aux mêmes points (i.e. tous les véhicules s'arrêtent exactement aux mêmes endroits).

Donc le fait de distinguer les cas feu rouge/feu vert, puis de recaler les profils au niveau des arrêts imposés par l'infrastructure permet en théorie de se ramener à l'un de ces deux cas. Cependant, en pratique, les profil moyens que nous avons construit aux figures 5.9.a et 5.9.b ne sont pas des "vrais" profils au sens de la définition 2.1. En effet, nous avons vu que tous les véhicules ne se sont pas arrêtés au stop, et donc qu'il aurait fallu distinguer les cas arrêt au stop/pas d'arrêt au stop avant de calculer la moyenne des profils. De plus, les arrêts des véhicules peuvent également être imposés par le trafic. Par exemple, on observe sur les figures 5.9.a et 5.9.b que certains véhicules se sont arrêtés au niveau des rond-points ou entre le feu et le point d'arrivée. Ainsi, le calcul d'un "vrai" profil moyen nécessite au préalable de distinguer les véhicules qui s'arrêtent de ceux qui ne s'arrêtent pas dès qu'il y a un arrêt d'un véhicule, ce qui semble difficilement réalisable en pratique sauf si l'objectif est de se focaliser sur une petite section d'étude ou un élément précis de l'infrastructure (ex : étude du comportement des conducteurs à une intersection, impact de l'ajout d'un rond-point ou d'un ralentisseur). Dans le cadre de notre étude, même si les profils moyens obtenus sur les figures 5.9.a et 5.9.b ne sont pas de "vrais" profils au sens de la définition 2.1, ces profils nous semblent être représentatifs de chacune des situations de conduite distinguant les cas feu rouge/feu vert. De plus, le fait que dans les deux situations de conduite, le profil moyen ne passe pas par zéro au niveau du stop, permet de mettre en évidence le fait que la plupart des conducteurs ne respectent pas cet arrêt imposé (la plupart de conducteurs "glissent" au niveau du stop).

5.4 Enveloppes de vitesse : extension des boîtes à moustaches aux données fonctionnelles

La construction d'un profil de vitesse agrégé tel que le profil moyen permet d'obtenir un profil de vitesse représentatif des vitesses moyennes pratiquées sur la section étudiée. Cependant, un tel profil agrégé ne reflète pas la variabilité des vitesses entre

a) Cas où le feu est rouge (36 courbes).



b) Cas où le feu est vert (42 courbes).

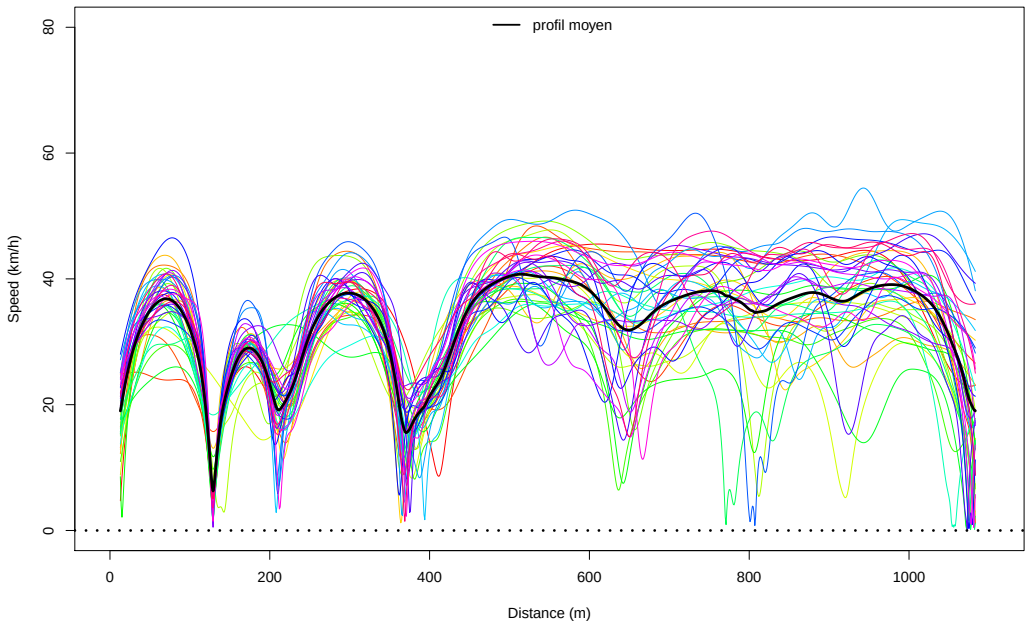


Figure 5.9 - Profils spatiaux de vitesse lissés après recalage : distinction feu rouge/feu vert. Le profil moyen est représenté en noir.

les différents conducteurs. Nous proposons dans cette section de construire des enveloppes de vitesse représentant la dispersion des vitesses entre les individus sur une section de route donnée.

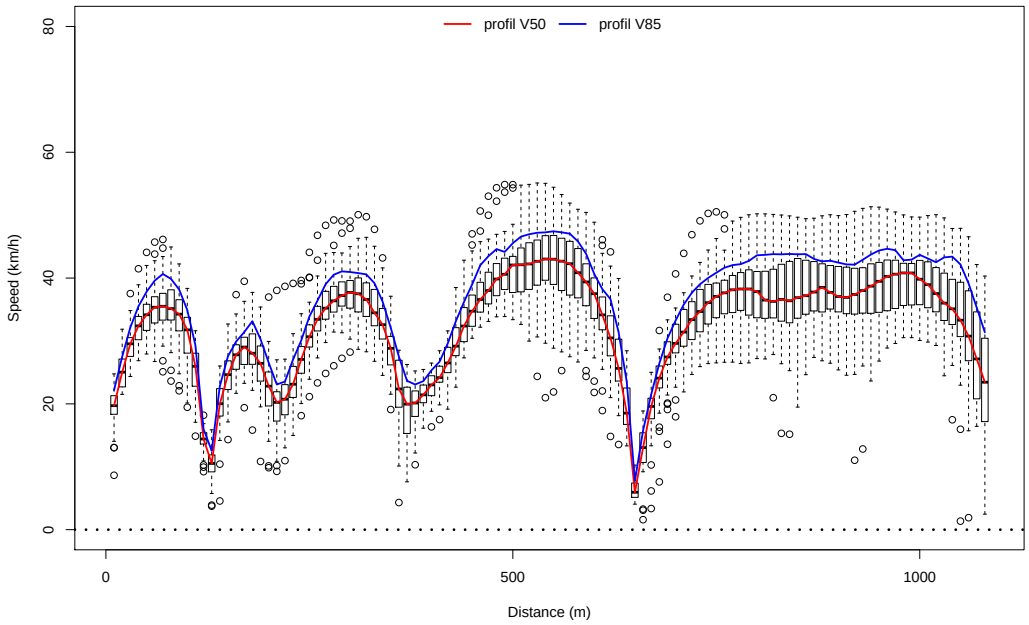
5.4.1 Une représentation classique de la dispersion des vitesses : la boîte à moustaches

Un outil graphique classiquement utilisé en statistique univariée pour représenter la distribution des valeurs d'une série statistique est la boîte à moustaches (ou "boxplot" en anglais) proposée par Tukey [164] (1977). Cet outil peut être utilisé pour représenter la dispersion des vitesses entre les individus en un point donné. La figure 5.10 représente un ensemble de boîtes à moustaches ponctuelles calculées tous les 10 m sur l'ensemble de la section étudiée à partir des valeurs des profils de vitesse recalés au niveau des arrêts. Une interpolation linéaire entre les points médians ($50^{\text{ème}}$ centile) et les $85^{\text{ème}}$ centiles nous permet de représenter les profils V50 et V85, avec un pas d'échantillonnage de 10 m, de chacun des sous-ensembles de profils distinguant les cas feu rouge/feu vert. Cependant, cette représentation ne permet pas de garder le caractère fonctionnel des données, et nécessite de discrétiser chacun des profils de vitesse fonctionnels afin de représenter la dispersion des vitesses en un point donné même si le pas d'échantillonnage peut être petit. Le profil médian (ou profil V50) n'est donc pas une courbe appartenant à l'ensemble des profils étudiés et n'est pas non plus un "vrai" profil de vitesse, au sens de la définition 2.1, contrairement au profil moyen. De nombreux auteurs se sont donc intéressés à l'extension des boîtes à moustaches aux données fonctionnelles. Cependant, la première étape dans la construction d'une boîte à moustache est d'ordonner les valeurs de la série étudiée. Si la notion d'ordre est simple dans le cas univariée, elle l'est beaucoup moins dans le cas fonctionnel. En effet, selon quels critères peut-on ordonner un ensemble de courbes ? Ces questions ont conduit à l'apparition du concept de profondeur statistique.

5.4.2 Notion d'ordre entre courbes : exemples de mesures de profondeur fonctionnelle

La notion de profondeur statistique a été introduite initialement afin de généraliser la notion d'ordre au cas multivariée. Ainsi, étant donné une loi de probabilité $P \in \mathbb{R}^d$, on associe à chaque point $x \in \mathbb{R}^d$ un réel positif $D(x, P)$, appelée profondeur de x, et mesurant la centralité du point x par rapport à un nuage de points générés à partir de la loi P. Les points les plus centraux sont associés aux plus grandes valeurs de profondeur. Plusieurs fonctions de profondeur statistique ont été proposées dans la littérature, comme par exemple la profondeur de demi-espace ("*halfspace depth*") introduite par Tukey [165] (1975) et ayant conduit à une version bivariée de la boîte à moustache appelée "bagplot" (Rousseeuw *et al.* [136]). Nous renvoyons le lecteur à

a) Cas où le feu est rouge (36 courbes).



b) Cas où le feu est vert (42 courbes).

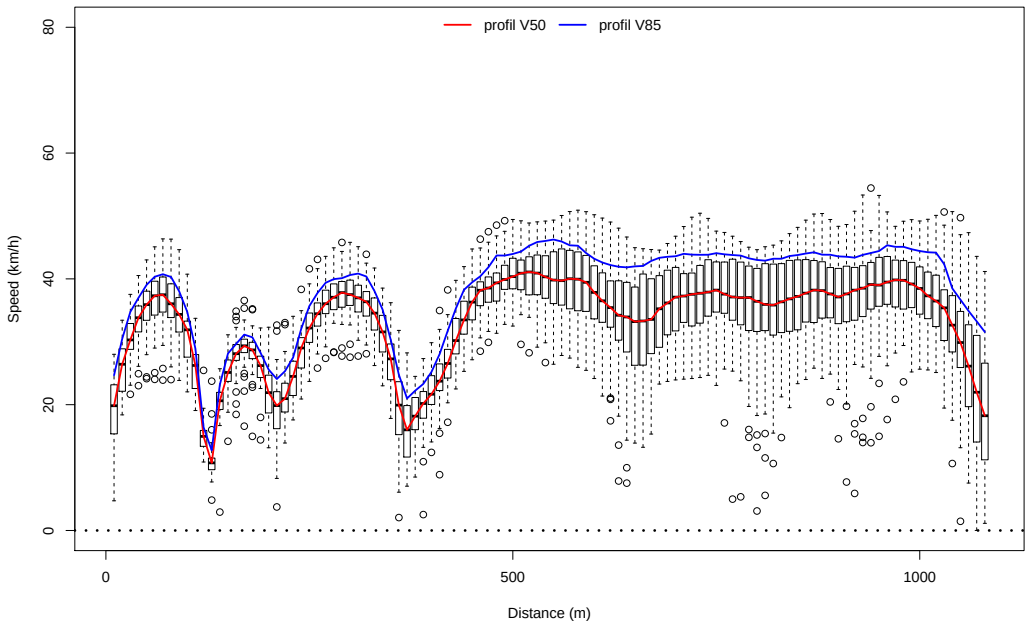


Figure 5.10 - Boîtes à moustaches ponctuelles représentées tous les 10 m : distinction feu rouge/feu vert. Le profil V50 est représenté en rouge, et le profil V85 en bleu.

l'article de Zuo et Serfling [186] pour une analyse détaillée des différentes profondeurs utilisées en statistique multivariée.

Plus récemment, la notion de profondeur a été étendue aux cas de données fonctionnelles et consiste à mesurer la centralité d'une courbe donnée par rapport à un ensemble de courbes (réalisations d'un processus stochastique). La courbe ayant la plus grande valeur de profondeur est alors la courbe médiane. Plusieurs mesures de profondeur fonctionnelle ont été proposées dans la littérature, parmi lesquelles on citera les quatre principales mesures suivantes :

1. La profondeur de Fraiman et Muniz (en abrégé, FMD).
   Fraiman et Muniz [61] (2001) ont été les premiers a introduire une mesure de profondeur fonctionnelle. Cette mesure de profondeur est basée sur une intégration d'une mesure univariée. Ainsi, si on note $F_{n,t}$ la fonction de répartition empirique de l'échantillon $x_1(t), \ldots, x_n(t)$ où $t \in [a, b]$, la profondeur de Fraiman et Muniz d'une courbe x_i est donnée par :

   $$FMD_n(x_i) = \int_a^b D_n(x_i(t))dt \tag{5.3}$$

   où $D_n(x_i(t))$ est une mesure de profondeur univariée du point $x_i(t)$ définie par $D_n(x_i(t)) = 1 - |\frac{1}{2} - F_{n,t}(x_i(t))|$.
2. La profondeur modale (en abrégé, hMD pour "*h-mode depth*").
   Cette profondeur fonctionnelle introduite par Cuevas *et al.* [34] (2006) est basée sur le concept de mode. Les auteurs définissent la notion de mode fonctionnel comme la fonction la plus entourée par les autres courbes du point de vue d'une certaine densité. La profondeur modale d'une courbe x_i par rapport à un ensemble de courbes $x_1, \ldots, x_n$ est ainsi définie par :

   $$MD_n(x_i, h) = \sum_{k=1}^{n} K(\frac{\|x_i - x_k\|}{h}) \tag{5.4}$$

   où $\|.\|$ est une norme appropriée, K est une fonction noyau, et h est un paramètre de largeur de bande ("*bandwidth*"). En pratique, on utilise la norme L^2, un noyau gaussien tronqué, et le $15^{\text{ème}}$ centile de la distribution empirique des $\{\|x_i - x_k\|,\ i, k = 1, \ldots, n\}$.
3. La profondeur par projection aléatoire (en abrégé, RMD pour "*random projection depth*").
   Cette profondeur fonctionnelle introduite par Cuevas *et al.* [35] (2007) tient compte de l'information sur la dérivée. Le principe consiste à projeter chaque courbe ainsi que sa dérivée dans une direction aléatoire afin de se ramener à un point de $\mathbb{R}^2$, puis à utiliser une mesure de profondeur bivariée afin d'ordonner les points projetés. L'étape de projection est réalisée selon un grand nombre de directions aléatoires et la valeur moyenne des profondeurs des points projetés définit la profondeur fonctionnelle.

4. La profondeur de bande (en abrégé, BD pour "*band depth*").
   Cette profondeur introduite par López-Pintado et Romo [97] (2009) est basée sur la notion graphique de bande délimitée par un ensemble de courbes $x_{i_1}, \ldots, x_{i_k}$ et définie dans $\mathbb{R}^2$ par :

$$B(x_{i_1}, \ldots, x_{i_k}) = \{(t, y) : \ t \in I, \ \min_{r=1,\ldots,k} x_{i_r}(t) \leq y \leq \max_{r=1,\ldots,k} x_{i_r}(t)\} \tag{5.5}$$

   Généralement, on se limite aux bandes définies par deux courbes, i.e. la région du plan délimitée par deux courbes données, et on définit la profondeur de bande d'une courbe x_i comme le rapport entre le nombre de bandes contenant la courbe x_i et le nombre total de bandes. López-Pintado et Romo [97] ont également introduit une définition plus flexible de la profondeur de bande, appelée profondeur de bande modifiée et notée MBD ("*modified band depth*") qui prend en compte le "pourcentage de temps" que la courbe x_i appartient à chaque bande.

Une comparaison entre les trois premières mesures de profondeur est donnée dans Febrero *et al.* [56] dans un objectif de détection des niveaux anormalement élevés d'un polluant ("outliers").

Nous proposons de comparer quatre des mesures de profondeurs citées sur notre jeu de données de profils de vitesse dans le cas où le feu est rouge (36 courbes). Les résultats sont donnés à la figure 5.11 avec :

- en 5.11.a, l'utilisation de la mesure de profondeur de Fraiman et Muniz (FMD) ;
- en 5.11.b, l'utilisation de la mesure de profondeur modale (hMD) ;
- en 5.11.c, l'utilisation de la mesure de profondeur de bande (BD) ;
- en 5.11.d, l'utilisation de la mesure de profondeur de bande modifiée (MBD) ;

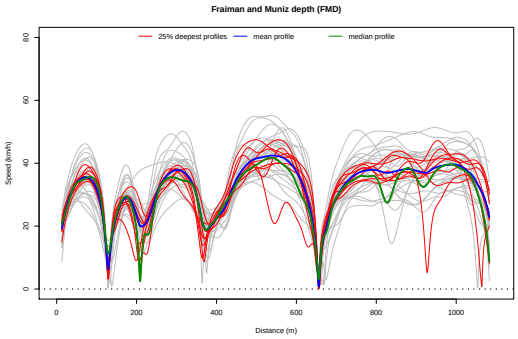
Le calcul des profondeurs a été fait à l'aide de la fonction R `fdepth` du package `rainbow` et de la fonction `fbplot` du package `fda`. Dans chaque cas, les 25% de courbes présentant les plus grandes valeurs de profondeurs sont représentées en rouge, la courbe médiane (courbe avec la valeur de profondeur la plus élevée) est représentée en vert, et la courbe moyenne est représentée en bleu. On observe que le profil médian obtenu avec les profondeurs FM et MBD est le même mais que ce profil médian ne semble pas représentatif de l'ensemble au niveau du 1$^{\text{er}}$ rond-point (forte décélération). De même le profil médian obtenu avec la profondeur BD n'est pas représentatif de l'ensemble entre le feu et le point d'arrivée. En revanche, le profil médian obtenu avec la profondeur hBD semble bien représentatif de l'ensemble des profils et l'on peut souligner qu'il est relativement proche du profil moyen. La profondeur modale nous semble donc être la profondeur la plus appropriée à notre étude.

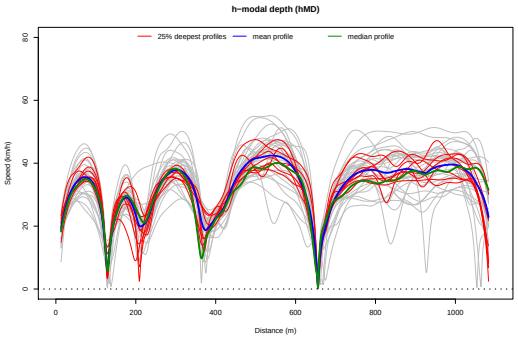
5.4.3 Extension des boîtes à moustaches aux données fonctionnelles : vers la notion d'enveloppes de vitesse

La notion de profondeur fonctionnelle introduite à la section précédente permet d'ordonner un ensemble de courbes de la plus centrale à la plus extrême. Il est alors

a) Profondeur de Fraiman et Muniz (FMD).

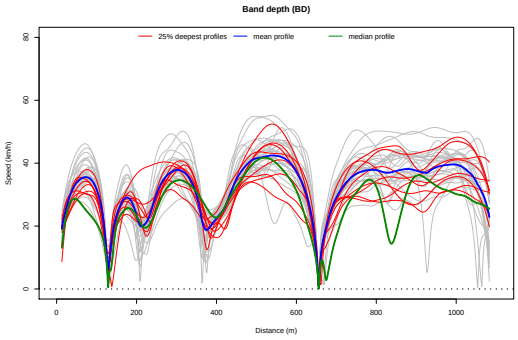
b) Profondeur modale (hMD).





c) Profondeur de bande (BD).

d) Profondeur de bande modifiée (MBD).



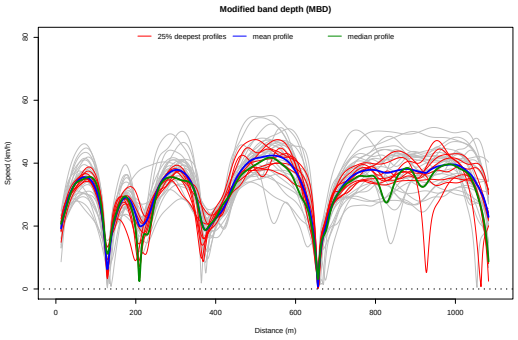


Figure 5.11 - Comparaison entre différentes mesures de profondeur fonctionnelle sur l'ensemble des 36 profils de vitesse dans le cas où le feu est rouge. Les 25% de courbes présentant les plus grandes valeurs de profondeurs sont représentées en rouge, la courbe médiane (valeur de profondeur la plus élevée) est représentée en vert, et la courbe moyenne est représentée en bleu.

possible d'étendre la notion classique de boîte à moustaches aux données fonctionnelles et de représenter ainsi un ensemble de courbes par des enveloppes caractérisant la distribution des courbes. Cette notion de boîte à moustache fonctionnelle (ou "*functional boxplot*" en anglais) a été introduite récemment par Sun et Genton [161] (2011) et peut être implémenté à l'aide de la fonction R `fbplot` du package `fda`. Cette représentation graphique est basée sur la notion de région centrale d'ordre α, avec $0 < \alpha < 1$, représentant la bande (voir équation (5.5)) délimitée par la proportion α des courbes présentant les plus grandes valeurs de profondeur, et définie par :

$$C_\alpha = \{(t, y(t)) : \ t \in I, \ \min_{r=1,\dots,\lceil n\alpha \rceil} y_{[r]}(t) \leq y(t) \leq \max_{r=1,\dots,\lceil n\alpha \rceil} y_{[r]}(t)\}, \tag{5.6}$$

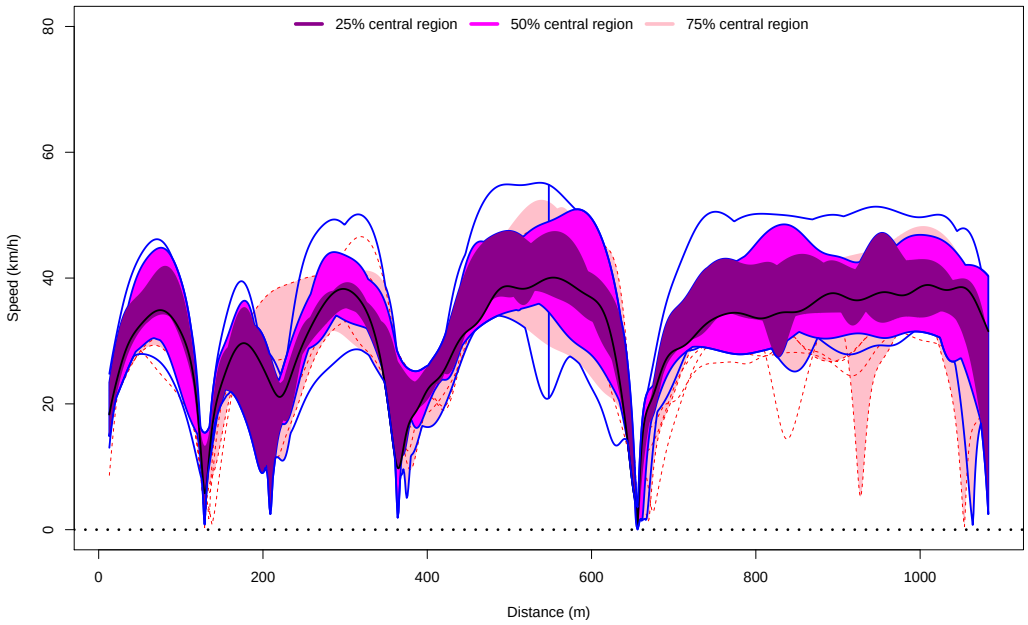
où $\lceil n\alpha \rceil$ est le plus petit entier supérieur à $n\alpha$, et les $y_{[1]}, \dots, y_{[n]}$ sont les courbes ordonnées par ordre décroissant de profondeur. Ainsi la région centrale à 50% est l'analogue de l'intervalle inter-quartile défini dans le cas univariée.

La figure 5.12 représente une version étendue des boîtes à moustaches fonctionnelles appliquées aux profils de vitesse dans les cas feu rouge/feu vert. Cette version étendue ("enhanced functional boxplot" dans l'article de Sun et Genton [161]) représente la région centrale à 25% en rose foncée, la région centrale à 50% en magenta, et la région centrale à 75% en rose pale. Si Sun et Genton [161] utilisent les profondeurs de bande (BD) et de bande modifiée (MBD), nous avons choisi d'utiliser la profondeur modale (hMD) au vu des résultats obtenus à la figure 5.11. Le profil médian associé à la profondeur la plus grande est représenté en noir. Notons que contrairement au profil médian ponctuel représenté à la figure 5.10, celui-ci est un "vrai" profil de vitesse puisque c'est l'un des profils de vitesse de l'ensemble. Les "moustaches" de la boîte sont représentées par des barres verticales bleues délimitées par l'enveloppe dite maximale dont les frontières sont colorées en bleu. Les frontières de l'enveloppe maximale sont obtenues en augmentant la largeur de la région centrale à 50% d'un facteur de 1.5 comme dans le cas univariée. Toute courbe qui passe en dehors de cette enveloppe maximale est alors considérée comme atypique ("outliers") et est représentée en pointillés rouges sur la figure 5.10.

L'avantage de cette représentation graphique est qu'elle permet de représenter la variabilité des vitesses pratiquées selon les individus sur une section de route donnée. Ces enveloppes de vitesse permettent de distinguer les zones pour lesquelles la variabilité des vitesses est la plus grande, et celles pour lesquelles les vitesses pratiquées sont plus homogènes. De plus, cet outil graphique permet également d'extraire un profil de vitesse représentatif de l'ensemble des profils étudiés (profil médian) au sens de la profondeur statistique choisi, et pouvant être utilisé comme alternative au profil moyen selon l'usage.

Remarque 5.2.
La notion de profondeur permet également d'étendre le concept de moyenne tronquée

a) Cas où le feu est rouge (36 courbes).



b) Cas où le feu est vert (42 courbes).

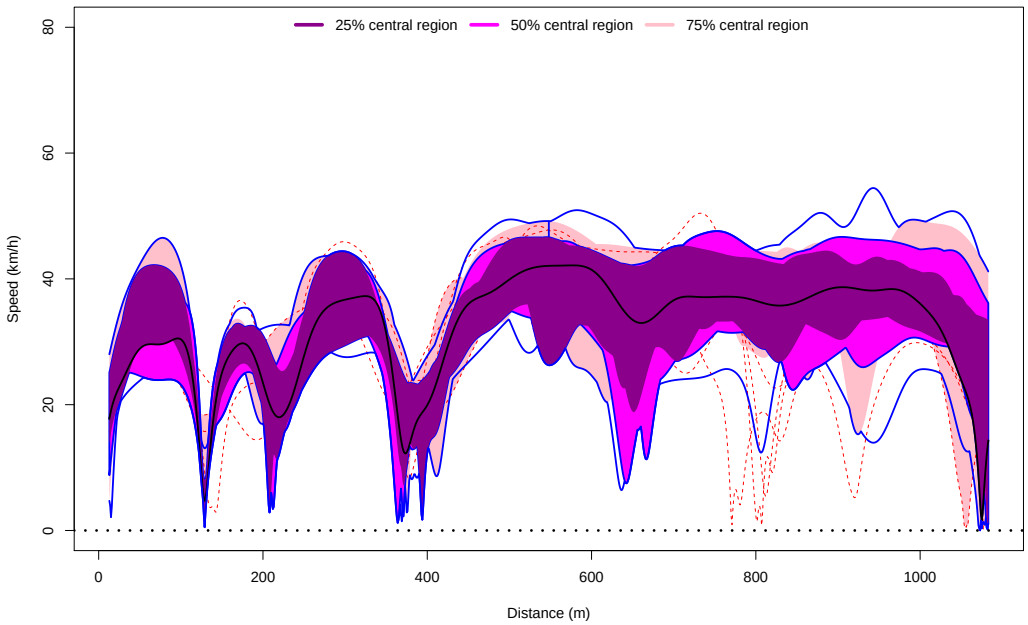


Figure 5.12 - Boîtes à moustaches fonctionnelles des profils de vitesse à partir de la profondeur modale hMD : distinction feu rouge/feu vert. La région centrale à 25% est en rose foncé, la région centrale à 50% est en magenta, et la région centrale à 75% est en rose pale. La courbe noire est la médiane et les courbes en pointillés rouges sont les courbes considérées comme atypiques (outliers).

("*trimmed mean*") aux données fonctionnelles. La moyenne tronquée d'ordre α ($0 < \alpha < 1$) d'un ensemble $x_1(t), \ldots, x_n(t)$ de courbes est ainsi définie comme la moyenne des $n - [n\alpha]$ courbes ayant les valeurs de profondeurs les plus grandes, soit :

$$\overline{x}_\alpha(t) = \frac{1}{n - [n\alpha]} \sum_{i=1}^{n-[n\alpha]} x_{[i]}(t)$$

où les $x_{[1]}, \ldots, x_{[n]}$ sont les courbes ordonnées par ordre décroissant de profondeur. Cette moyenne tronquée présente l'avantage d'être moins sensible aux valeurs aberrantes que la moyenne classique et d'être donc plus robuste.

5.5 Application à l'éco-conduite

On propose dans cette section de revenir à l'objectif premier de l'expérimentation dont est issu le jeu de données de 78 profils de vitesse utilisé dans ce chapitre et décrit à la section 5.1.2, à savoir identifier les conséquences de la pratique de l'éco-conduite. Pour plus de détails sur la notion d'éco-conduite, nous renvoyons le lecteur à l'annexe C. Afin de simplifier l'étude, on s'intéresse uniquement au cas où les véhicules se sont arrêtés au feu (soit 36 profils de vitesse) mais l'on distingue les trajets réalisés sans consignes particulières, que l'on appellera "trajets normaux", des trajets réalisées après avoir suivi une formation à l'éco-conduite, que l'on appellera "trajets économiques". Les 36 trajets contenant un arrêt au feu se divisent donc en deux catégories :

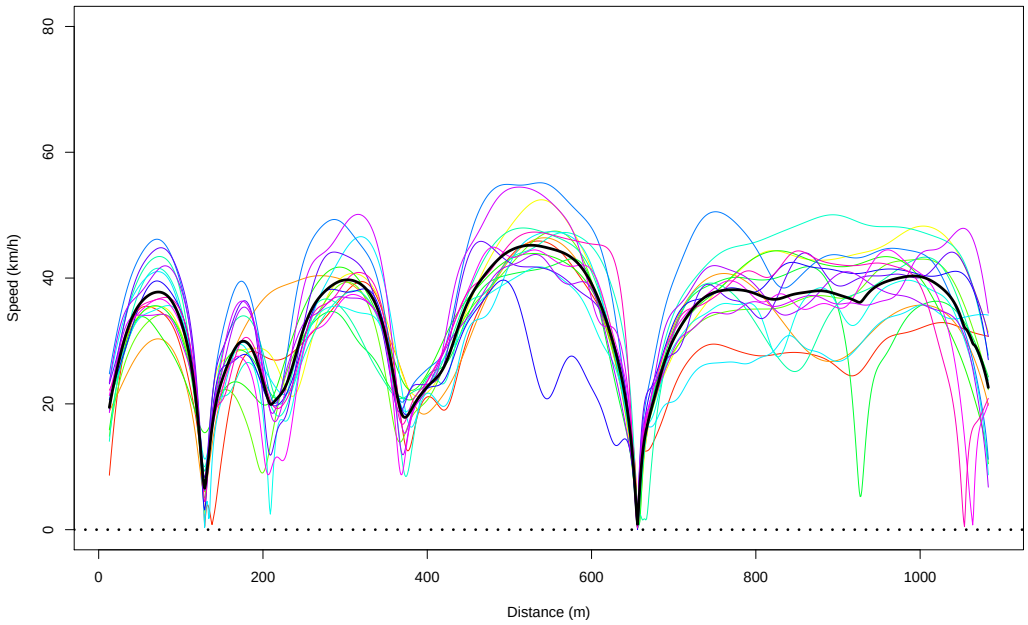
- 17 trajets normaux ;
- 19 trajets économiques.

Les profils de vitesse lissés obtenus après recalage au niveau du stop et du feu sont représentés à la figure 5.13 en distinguant les trajets normaux et économiques.

Afin de comparer les profils de vitesse entre les trajets normaux et économiques, on compare dans un premier temps les profils moyens obtenus dans les deux cas. Ces profils moyens sont représentés à la figure 5.14 avec en bleu le profil moyen des trajets normaux, et en vert le profil moyen des trajets économiques. On observe que dans le cas des trajets économiques, les vitesses moyennes sont plus faibles (inférieures à 40 km/h) et que les pics sont plus aplatis traduisant des accélérations et décélérations moins fortes. Ces résultats sont en accord avec l'une des principales règles de l'éco-conduite qui consiste à adopter une conduite souple et à anticiper le trafic afin de limiter les fortes accélérations et les forts freinages (voir annexe C).

Dans un second temps, on s'intéresse à la variabilité des vitesses entre les individus selon le type de conduite pratiquée (normal ou éco-conduite) grâce à la représentation graphique sous forme de boîtes à moustaches fonctionnelles. Nous avons utilisé

a) Trajets normaux (17 courbes).



b) Trajets économiques (19 courbes).

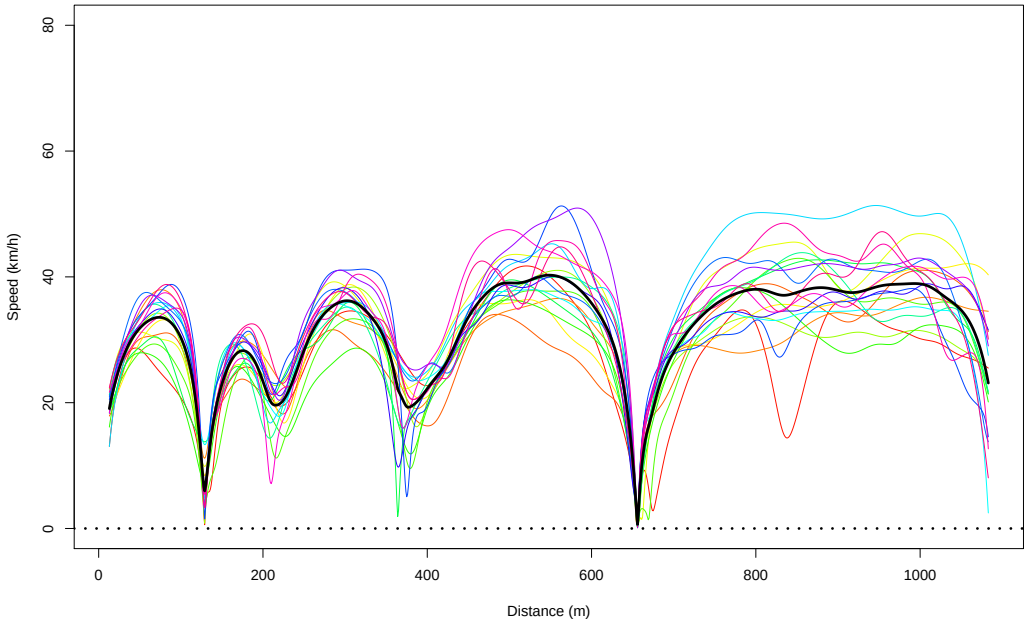


Figure 5.13 - Profils spatiaux de vitesse lissés après recalage au niveau du stop et du feu (cas où le feu est rouge) : distinction trajets normaux/trajets économiques. Le profil moyen est représenté en noir.

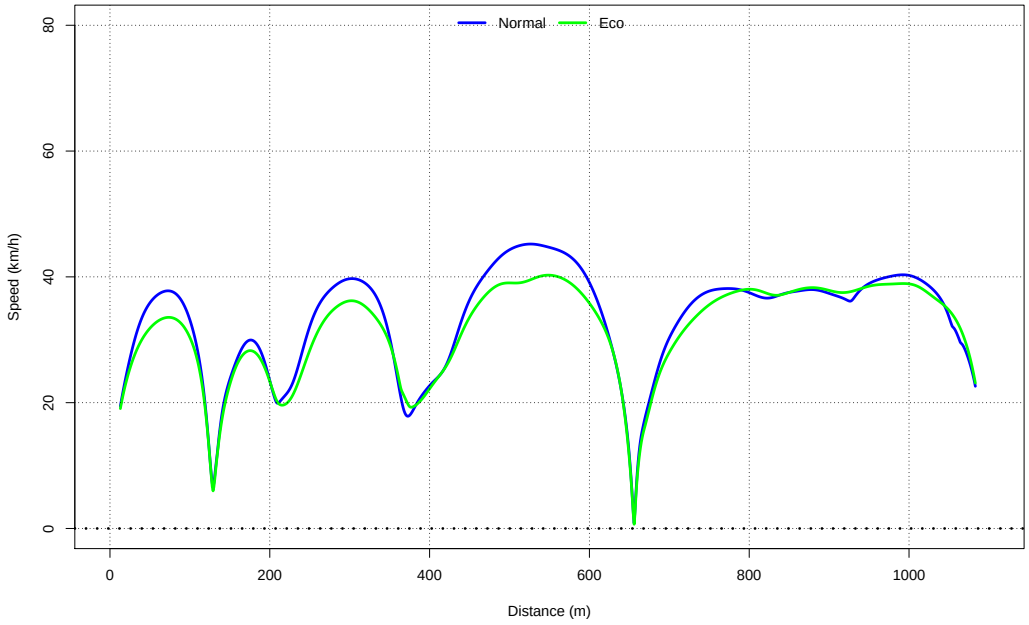


Figure 5.14 - Profils de vitesse moyens des trajets normaux (en bleu) et des trajets économiques (en vert) dans le cas où le feu est rouge.

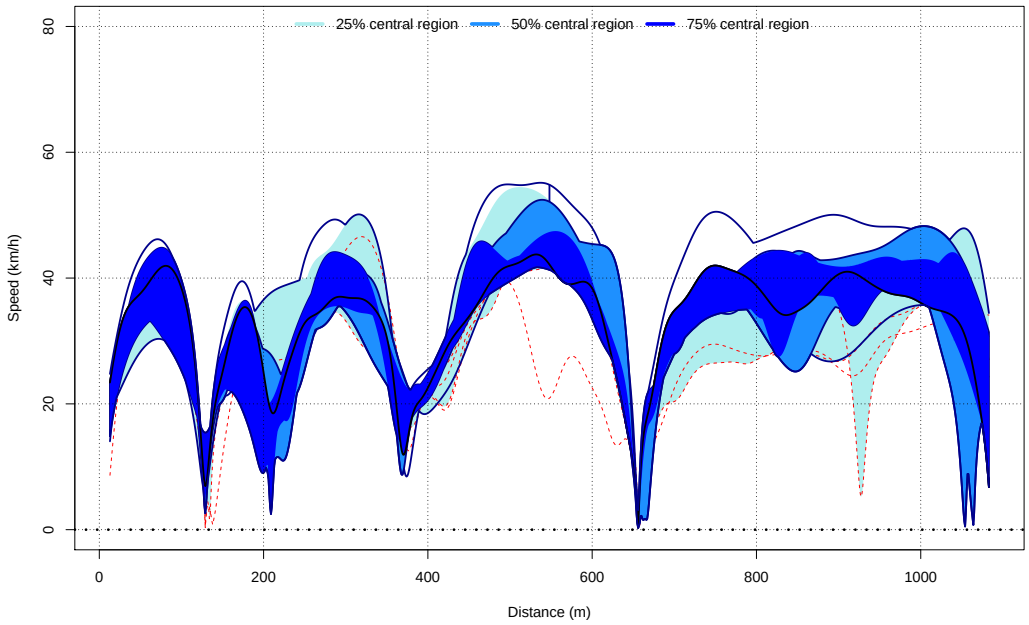
comme précédemment la mesure de profondeur modale. Les résultats obtenus sont représentés à la figure 5.15. Les régions centrales à 25%, 50% et 75% sont représentées du bleu foncé au bleu clair dans le cas d'une conduite normale, et du vert foncé au vert clair dans le cas d'une conduite économique. Dans les deux cas, le profil médian (courbe avec la valeur de profondeur la plus grande) est représenté en noir et les profils atypiques sont en pointillés rouges. Comme dans l'étude des profils moyens, on observe que les vitesses pratiquées en éco-conduite sont plus faibles. On observe notamment que la région centrale à 25% est en dessous du seuil des 40 km/h dans le cas d'une conduite économique, ce qui n'est pas le cas avec une conduite normale. On observe également une plus grande variabilité des vitesses pratiquées en conduite normale qu'en éco-conduite au niveau du franchissement des deux rond-points.

5.6 Conclusion

Dans ce chapitre, nous avons décrit une méthodologie permettant de construire un profil moyen ou d'extraire un profil médian à partir d'un ensemble de profils. Nous avons vu notamment que la méthodologie de lissage développée au chapitre précédent donnait de bons résultats, excepté pour l'estimation des arrêts de courte durée pour lesquels l'étape de monotonisation a tendance à entraîner une sur-estimation de la vitesse (pas de passage par zéro). Nous avons également mis en évidence le problème du décalage des profils notamment au niveau des arrêts et conduisant à un profil moyen irréaliste et non représentatif de l'ensemble. Ce problème a été corrigé par l'utilisation d'un alignement par landmarks correspondant aux points caractéristiques que l'on souhaite mettre en correspondance, soit dans notre cas les arrêts. Cependant, nous avons également vu les limites de cette méthode dans le cas de l'étude d'une section trop longue. En effet, si en théorie le calcul d'un "vrai" profil moyen (au sens de la définition 2.1) nécessite au préalable de distinguer les véhicules qui s'arrêtent de ceux qui ne s'arrêtent pas dès qu'il y a un arrêt d'un véhicule, cela semble difficilement réalisable en pratique. Nous avons donc choisi dans notre étude de nous limiter à la distinction de situations de conduite caractéristiques telles que l'état des feux de signalisation.

Nous avons également introduit la notion d'ordre entre un ensemble de courbes grâce au concept de profondeur statistique fonctionnelle. Cette notion de profondeur permet de mesurer le degré de centralité d'une courbe parmi un ensemble de courbes, et permet ainsi d'adapter les notions de médiane et de quantiles aux données fonctionnelles. Le profil médian, correspondant au profil de vitesse associé à la plus grande valeur de profondeur, peut ainsi être considéré comme un profil de référence représentatif d'un ensemble de profils de vitesse. De plus, nous avons vu que la notion de profondeur permettait d'étendre la représentation graphique sous forme de boîtes à moustaches aux données fonctionnelles, conduisant ainsi à la construc-

a) Trajets normaux (17 courbes).



b) Trajets économiques (19 courbes).

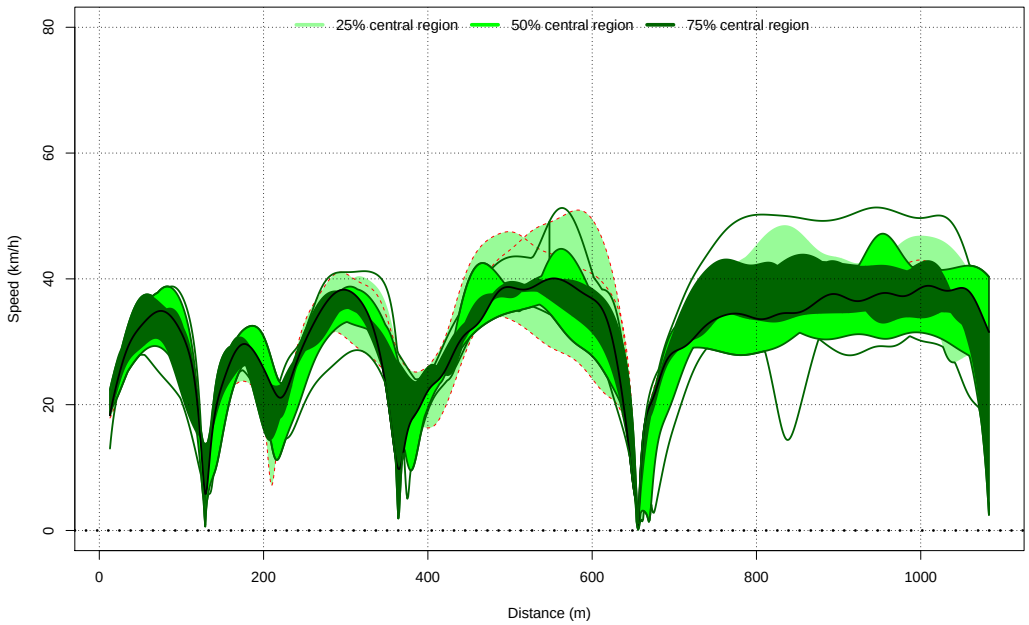


Figure 5.15 - Boîtes à moustaches fonctionnelles des profils de vitesse à partir de la profondeur modale hMD dans le cas où le feu est rouge : distinction trajets normaux/trajets économiques. La région centrale à 25% est en bleu foncé (normal) ou vert foncé (éco), la région centrale à 50% est en bleu (normal) ou vert (éco), et la région centrale à 75% est en bleu clair (normal) ou vert clair (éco). La courbe noire est la médiane et les courbes en pointillés rouges sont les courbes considérées comme atypiques (outliers).

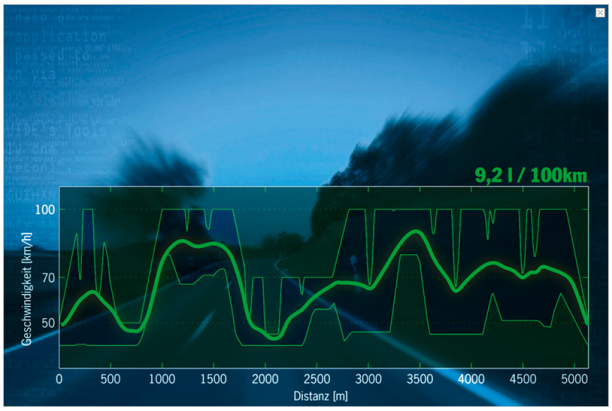


Figure 5.16 - "Driving corridor" extrait du système Porsche Innodrive.

tion de régions centrales caractérisant la distribution des courbes. Dans le cas de l'étude des profils de vitesse, la construction de boîtes à moustaches fonctionnelles permet d'obtenir des enveloppes de vitesse reflétant la variabilité des vitesses pratiquées sur une section de route donnée. Ces enveloppes de vitesse peuvent notamment être utilisées dans le cadre d'un système embarqué d'aide à la conduite fournissant ainsi au conducteur une vitesse maximale et minimale à ne pas dépasser tout au long de son trajet. Ainsi, en se basant sur les vitesses pratiquées, la région centrale à 25% avec éventuellement un écrêtement à la vitesse réglementaire, pourrait représenter une enveloppe de vitesse sécuritaire. On peut également imaginer un système d'aide à l'éco-conduite fournissant au conducteur une enveloppe de vitesse "économique" permettant de réduire sa consommation de carburant et d'adopter une conduite plus souple. On peut citer à ce sujet l'un des futurs système d'aide à la conduite de type ACC ("Adaptative Cruise Control") développé par Porsche et intitulé "InnoDrive" (http://www.porsche.com/international/aboutporsche/responsibility/environment/technology/porscheinnodrive/). Ce régulateur de vitesse adaptatif calcule un profil de vitesse optimal tenant compte du style de conduite désiré par le conducteur : "eco", "comfort" ou "dynamic". La première étape de construction de ce profil de vitesse optimal consiste à déterminer une enveloppe de vitesse appelée "driving corridor" en fonction de la géométrie de la route et du style de conduite choisi par le conducteur. Le profil optimal est ensuite calculé sous la contrainte d'appartenance à ce "driving corridor" comme l'illustre la figure 5.16. Cet exemple d'application montre l'intérêt de disposer d'enveloppes de vitesse, en plus de disposer d'un profil de vitesse de référence.

Conclusion générale

Conclusion

La connaissance des vitesses pratiquées est essentielle afin de mesurer et d'évaluer l'usage réel du réseau routier. Cette connaissance est rendue accessible grâce au développement des véhicules traceurs et des smartphones qui permettent de collecter les mesures de position et de vitesse des usagers sur l'ensemble du réseau. En effet, nous avons vu au chapitre 1 que les profils de vitesse étaient riches d'information sur le comportement des conducteurs et étaient utilisés dans de nombreuses études réalisées dans le domaine de la prévision de trafic, de la sécurité, de l'environnement... Cependant, le développement des technologies des capteurs permet de collecter ces données de vitesse et de position sur des grilles temporelles de plus en plus fines, conduisant à l'obtention de vecteurs de grande dimension pour lesquels les méthodes usuelles de statistique multivariée ne sont pas adaptées.

Dans cette thèse, nous avons proposé des méthodes statistiques appropriées au traitement des profils de vitesse en se ramenant à des données fonctionnelles. Cette approche fondée sur l'Analyse des Données Fonctionnelles (en anglais, "Functional Data Analysis") qui s'est considérablement développée depuis une vingtaine d'années, et qui est utilisée dans de nombreux domaines tels que la biologie, la chimiométrie ou la climatologie, est à notre connaissance encore peut utilisée dans le domaine routier et en particulier dans l'étude des profils de vitesse. En effet, l'état de l'art des méthodes de traitement des profils de vitesse actuellement utilisées a montré les limites de celles-ci, et la nécessité de développer des méthodes plus adaptées à ce type de données. L'avantage de l'approche fonctionnelle est notamment de tenir compte de la structure sous-jacente continue de ce type de données, et ainsi de pouvoir prendre en compte certaines de leurs caractéristiques fonctionnelles (dérivées, régularité, déphasage de courbes...).

Dans un premier temps, nous avons proposé au chapitre 2 une modélisation fonctionnelle des profils spatiaux de vitesse (i.e. la vitesse en fonction de la position du véhicule) préservant la cohérence physique entre vitesse et position (et implicitement temps). L'étude des propriétés de ces fonctions a montré la difficulté d'utiliser une méthode de lissage appropriée afin de convertir les données brutes de position et de

vitesse (de nature vectoriel) en objet fonctionnel. En effet, nous avons montré que les fonctions affines par morceaux n'étaient pas des profils spatiaux de vitesse, et que les profils spatiaux de vitesse étaient des fonctions non dérivables au point où la vitesse s'annule, ce qui exclut l'utilisation de méthodes classiques telles que l'interpolation linéaire ou les splines. Une partie importante de cette thèse a donc été consacrée au développement d'une méthode de lissage appropriée afin d'estimer au mieux un "vrai" profil de vitesse à partir d'observations bruitées.

Au chapitre 3, nous avons passé en revue les différentes méthodes de lissage fondées sur l'utilisation des splines. Nous avons notamment présenté en détails les splines de régression, les splines de lissage et les splines pénalisées, et nous avons montré l'efficacité des splines de lissage pour l'estimation des profils temporels de vitesse et d'accélération. Nous avons également présenté une généralisation des splines de lissage fondée sur la résolution d'un problème de régression régularisée dans un espace de Hilbert à noyau reproduisant (RKHS), et qui permet de traiter de manière unifiée de nombreux type de splines. Nous avons donc choisi d'utiliser les splines de lissage afin d'obtenir un estimateur d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse.

Cependant, si le problème de régression non paramétrique d'estimation d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse est difficile dans l'espace `vitesse`×`distance`, nous avons montré que ce problème pouvait être résolu en changeant d'espace d'étude. Ainsi, nous avons proposé au chapitre 4, de nous placer dans un premier temps dans l'espace `distance`×`temps`, et de chercher à estimer la fonction $F(t)$ représentant la distance parcourue par le véhicule au cours du temps. On en déduit alors facilement un estimateur du profil spatial de vitesse $v_S(x) = F' \circ F^{-1}(x)$ en remplaçant $F(t)$ par l'estimateur obtenu. Ce changement d'espace d'étude permet de se ramener à un problème de régression non paramétrique sous les deux contraintes suivantes :

(C_1) Estimer la fonction de régression $F(t)$ à partir d'observations bruitées de cette fonction (mesures de position) et également d'observations bruitées de sa dérivée $F'(t)$ (mesures de vitesse).

(C_2) Une contrainte de monotonie sur F.

La première contrainte est un problème mathématique apparaissant dans de nombreux domaines d'application. Nous avons proposé de le résoudre grâce à un estimateur par spline de lissage. Nous avons notamment montré que l'utilisation de la théorie des espaces de Hilbert à noyau reproduisant (RKHS) permettait d'obtenir une expression explicite de l'estimateur s'écrivant comme une combinaison linéaire de fonctions de base et de fonctions noyaux, et dont le degré de lissage était contrôlé par un unique paramètre. Nous avons également proposé une réécriture de notre estimateur ne faisant pas intervenir de noyaux reproduisants mais uniquement un semi-noyau, et facilitant la mise en oeuvre numérique.

La seconde contrainte a été ajoutée dans une seconde étape basée sur une monotonisation de l'estimateur construit à l'étape précédente. Cependant, nous n'avons pas réussi à résoudre le problème de l'estimation des arrêts du véhicule pour lesquels la fonction $F(t)$ à estimer est constante, et nous avons donc proposé un estimateur strictement monotone. Si la stricte monotonie de notre estimateur au niveau des arrêts est un inconvénient de notre approche, les résultats obtenus sur données réelles montrent cependant que seuls les arrêts courts (d'une durée inférieure à 5 s) ne sont pas bien estimés (sur-estimation de la vitesse), et que les performances de notre estimateur restent globalement satisfaisantes.

Enfin, la dernière partie de cette thèse a été consacrée au développement d'une méthodologie de construction de divers profils agrégés, tels que le profil moyen ou le profil médian, représentatifs d'un ensemble de profils spatiaux de vitesse individuels et adaptés à une situation de conduite (dans notre étude, feu rouge/feu vert). Nous avons notamment montré au chapitre 5 que l'approche fonctionnelle permettait de corriger le déphasage entre les profils de vitesse, notamment au niveau des arrêts. Ainsi, nous avons montré l'intérêt d'aligner les profils de vitesse en certains points caractéristiques appelés landmarks et définis au préalable afin de construire un profil agrégé réaliste. Enfin, nous avons utilisé la notion de profondeur statistique afin d'extraire le profil médian d'un ensemble de profils de vitesse. Cette notion de profondeur nous a permis également de construire des enveloppes de vitesse reflétant la dispersion des vitesses pratiquées sur une section de route donnée, et fondée sur une extension des boîtes à moustaches aux données fonctionnelles.

Perspectives

Un premier axe de recherche concerne l'amélioration de la méthode de lissage proposée. En effet, une première piste de réflexion est l'utilisation d'une méthode plus appropriée pour l'étape de monotonisation afin de résoudre le problème de l'estimation des arrêts du véhicule. Le principe consiste alors à construire un estimateur de $F(t)$ qui soit strictement monotone lorsque la vitesse (i.e. la dérivée $F'(t)$) est non nulle, et constant lorsque la vitesse est nulle. Une deuxième piste de réflexion est la prise en compte des deux contraintes, à savoir l'utilisation de l'information sur la dérivée et la monotonie, dans une unique étape de lissage afin d'éviter la mise en oeuvre de deux lissages successifs. Une troisième piste de réflexion est l'ajout de mesures issues d'équipements statiques (ex : boucles magnétiques, radars) afin d'améliorer la précision des mesures de position et de vitesse. Si Louah [98] propose de fusionner des mesures issues de dispositifs de bord de voie avec des mesures obtenues en continu avec un véhicule instrumenté en procédant par une simple translation du profil de vitesse, il serait intéressant d'appliquer la méthode de lissage proposée au chapitre 4 en ajoutant des poids différents selon la précision des mesures.

Un deuxième axe de recherche est l'utilisation de méthodes de classification non

supervisée afin de distinguer des classes de profil de vitesse associées aux conditions de trafic (trafic libre/trafic contraint) ou aux spécificités de l'infrastructure (état des feux). En effet, on peut faire l'hypothèse qu'il existe un nombre fini de comportements types sur une section donnée, et que chaque profil de vitesse peut donc être associé à une classe correspondant à une situation de conduite. La classification de profils de vitesse a notamment déjà été abordée dans les travaux de Allain [4] et de Kerper *et al.* [84], et ceux ci ont montré que la distance dynamique par déformation temporelle, appelée "Dynamic Time Warping distance" (DTW), semblait être une distance appropriée.

Un troisième axe de recherche est l'utilisation de méthodes d'analyse factorielle, et plus particulièrement de l'analyse en composantes principales fonctionnelles, afin d'étudier les variations entre un ensemble de profils de vitesse individuels. Cette méthode de réduction de dimension est une des premières méthodes classiques adaptées aux données fonctionnelles, et correspond à une représentation linéaire optimale (au sens des moindres carrées) d'un ensemble de données fonctionnelles dans un espace de dimension finie. On pourra citer par exemple les travaux de Hyndman et Shang [81] qui proposent d'appliquer la version bivariée de la boîte à moustache appelée "bagplot" initialement proposée par Rousseeuw *et al.* [136] aux données fonctionnelles, en utilisant les scores des deux premières composantes principales fonctionnelles.

Un quatrième axe de recherche est la caractérisation de certains éléments de l'infrastructure. En effet, les profils de vitesse sont le reflet de l'adaptation des conducteurs à l'environnement et fournissent à ce titre une information sur l'infrastructure elle-même. Il est donc possible d'obtenir une "signature" de certains éléments de l'infrastructure par l'observation des vitesses pratiquées. Par exemple :

- un stop sera caractérisé par un passage à zéro de tous (ou une majorité) les profils de vitesse ;
- un feu de signalisation sera caractérisé par un passage bref ou prolongé à zéro d'une partie des profils de vitesse, et par le maintien d'une vitesse de croisière pour les autres ;
- un passage piéton ou un ralentisseur de type dos d'âne seront caractérisés par un fort ralentissement avec présence de quelques arrêts brefs.

Il serait donc intéressant d'utiliser les traces numériques rendues disponibles grâce au développement des véhicules traceurs (smartphones) pour développer un modèle permettant de calculer la probabilité de présence de différents éléments de l'infrastructure, et d'enrichir ainsi les cartes numériques.

Enfin, un dernier axe de recherche est d'appliquer l'approche fonctionnelle proposée dans ce manuscrit pour l'étude des profils de vitesse, à l'étude des profils d'accélération et de jerks pour caractériser les situations de presque-accidents. En effet, plusieurs études ont montré que la présence de pics dans les profils d'accélération ou de jerks étaient de bons indicateurs des comportements à risque (Nygard

[116], Bagdadi et Várhelyi [13]). Une piste de réflexion est donc l'utilisation d'un débruitage par ondelettes des signaux (les ondelettes étant des fonctions de base plus adaptées que les splines aux fonctions composées de pics), puis de détecter les pics en utilisant par exemple la notion d'intensité structurelle développée par Antoniadis *et al.* [9].

Annexe A

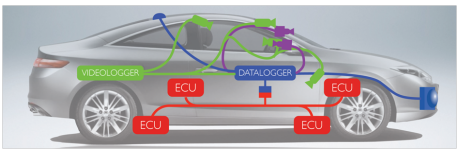
Véhicules traceurs : instrumentation et capteurs

A.1 Instrumentation des véhicules et capteurs

La généralisation des véhicules traceurs permet de recueillir les traces numériques laissées par ces véhicules qui peuvent être considérés comme des capteurs mobiles explorant le réseau routier en continu. Suivant les objectifs, l'instrumentation de ces véhicules peut être lourde, permettant ainsi l'enregistrement de dizaines de paramètres relatifs au véhicule, au conducteur ou à son environnement, mais elle peut au contraire être très légère grâce au développement des smartphones qui rendent facilement accessibles les informations essentielles que sont la position et la vitesse du véhicule (figure A.1).

Dans le cas d'une instrumentation lourde, la plupart des paramètres relatifs au véhicule (vitesse instantanée, régime moteur, distance parcourue...) et aux actions du conducteur (position des pédales de frein et d'accélérateur, état des feux...) sont

a) Instrumentation lourde



b) Instrumentation légère

Figure A.1 - Exemple de véhicules traceurs (source CEESAR).

collectés à l'aide d'un enregistreur de données connecté au bus CAN du véhicule. Le bus CAN est le réseau qui permet à tous les capteurs et actionneurs du véhicule de s'échanger leurs informations. Dans un câblage traditionnel, un capteur est relié à un autre (liaison point à point) ce qui fait un câblage très complexe. Dans un réseau CAN, tous les capteurs sont reliés entre eux par un même fil (le câblage est donc très simplifié) et les informations sont échangées dans des trames avec une gestion de la priorité des trames et un contrôle d'erreurs poussé. Notons que les données transitant sur le bus CAN peuvent également être recueillies par un smartphone relié (par USB ou Bluetooth) à ce réseau à l'aide d'un adaptateur branché sur la prise OBD-II ("On-Boards Diagnostics") du véhicule.

D'autres capteurs peuvent également être ajoutés au véhicule. On distingue généralement deux catégories de capteurs :

- les capteurs proprioceptifs, qui permettent de mesurer des grandeurs intrinsèques du véhicule (ex : odomètre, centrale inertielle...) ;
- les capteurs extéroceptifs, qui permettent de situer le véhicule par rapport à son environnement (ex : GPS, radar, caméra...).

Dans le cadre de notre étude, nous nous intéresserons uniquement aux capteurs permettant de récolter des données relatives à l'obtention de profils spatiaux de vitesse, à savoir la position et la vitesse du véhicule.

A.2 Capteurs relatifs à la position du véhicule

A.2.1 L'odomètre

Principe :	Mesure incrémentale de la distance parcourue par le véhicule depuis une position initiale connue, via le comptage de fractions de tour de roues. Technique de localisation dite à l'estime.
Avantages :	– Bonne précision à court terme. – Fréquence d'échantillonnage élevée (de 10 à 50 Hz, soit de 10 à 50 mes/s).
Inconvénients :	– Accumulation des erreurs de mesure au cours du temps causées par des glissements et une incertitude sur le diamètre de la roue ⇒ mauvaise précision à moyen et long terme. – Localisation relative (fournit uniquement une abscisse curviligne).

A.2.2 Le GPS ("Global Positioning System")

Principe : Système permettant d'obtenir une localisation 3D (latitude, longitude, altitude) à partir d'un récepteur, n'importe où sur la Terre, grâce à une constellation de satellites dédiés. Le principe consiste à mesurer la distance (appelée pseudodistance) entre le récepteur GPS et un certain nombre (au moins trois) de satellites de positions connues, grace au calcul du temps de propagation d'une onde entre chacun des satellites et le récepteur GPS, puis d'en déduire la position du récepteur par triangulation. Localisation dite absolue.

Avantages :
- Pas d'accumulation des erreurs de mesure.
- Localisation 3D absolue.

Inconvénients :
- Mauvaise précision (généralement de 10 à 15 m) due à de nombreuses sources d'erreurs : erreurs des satellites (horloge, éphéméride), retards atmosphériques, multi-trajets (réflexion du signal sur des obstacles), qualité du récepteur.
- Localisation très dépendante de l'environnement (masquage des satellites, multi-trajets).
- Fréquence d'échantillonnage peu élevée (généralement 1 Hz, soit 1 mes/s, et 10 Hz, soit 10 mes/s, pour les modèles haut de gamme).

Remarque A.1. Le GPS fournit une mesure tridimensionnelle (latitude, longitude, altitude) de la position du véhicule. Cependant, il est possible de se ramener à une mesure unidimensionnelle comparable à celle fournie par l'odomètre, en utilisant un algorithme de map-matching. Le principe du map-matching consiste à déterminer la position d'un véhicule par rapport à une cartographie routière numérique. Divers algorithmes ont été développés dont les principaux sont décrits dans l'article de Quddus *et al.* [121]. La distance parcourue par le véhicule peut alors être obtenue en sommant les segments reliant chacun des points map-matchés.

A.2.3 Technique d'amélioration du GPS : le RTK ("Real Time Kinematic")

Principe : Correction en temps réel du signal GPS par l'intermédiaire d'une base terrestre dont la position est connue précisément (station de référence). La base fixe calcule la correction, i.e. le différentiel entre la position connue de la base et la position calculée en temps réel par les satellites, puis la transmet au récepteur mobile sous forme d'ondes radio.

Avantages :
- Précision de 2 à 30 cm.

Inconvénients :
- Coût très élevé : entre 10 000 et 15 000 euros.
- Portée limitée (le récepteur ne doit pas être à plus de 10 km de la base).
- Sensibilité aux obstacles (ex : forêt, tunnel).

Remarque A.2. Le GPS différentiel, ou DGPS, repose sur le même principe de correction différentiel que le RTK, mais la correction se fait par l'intermédiaire d'un satellite géostationnaire.

A.3 Capteurs relatifs à la vitesse du véhicule

A.3.1 Vitesse lue sur le bus CAN du véhicule (vitesse CAN)

Principe : Un calculateur délivre la vitesse du véhicule à partir des vitesses de rotation de chaque roue. Les capteurs utilisés sont communs avec l'odomètre.

Avantages :
- Données toujours disponibles.

Inconvénients :
- Précision et mode de calcul dépendant du constructeur (non documenté).
- Dépendant des mesures odométriques puisque les capteurs sont communs.

A.3.2 Vitesse calculée par effet Doppler

Principe : Estimation de la vitesse de déplacement du récepteur GPS par effet Doppler. Ce phénomène physique correspond au décalage de fréquence d'une onde entre la mesure à l'émission et la mesure à la réception lorsque la distance entre l'émetteur et le récepteur varie au cours du temps.

Avantages :
- Très bonne précision (de l'ordre de 0.1 m/s).
- Mesures indépendantes des position GPS.

Inconvénients :
- Sensibilité aux obstacles (ex : forêt, tunnel).

Annexe B

Étape de pré-traitement des données : estimation par fenêtre glissante de la position du véhicule à partir de l'odomètre et du GPS

A moving fixed-interval filter/smoother for estimation of vehicle position using odometer and map-matched GPS

Cindie Andrieu [1], Guillaume Saint Pierre [1] and Xavier Bressaud [2]

[1] *IFSTTAR, IM, LIVIC, 14, route de la Minière, 78000 Versailles, France.*
E-mail : cindie.andrieu@ifsttar.fr and guillaume.saintpierre@ifsttar.fr
[2] *Université Paul Sabatier, Institut de Mathématiques de Toulouse, F-31062 Toulouse Cedex 9, France.*
E-mail: bressaud@math.univ-toulouse.fr

Abstract. This paper presents some optimal real-time and post-processing estimators of vehicle position using odometer and map-matched GPS measurements. These estimators were based on a simple statistical error model of the odometer and the GPS which makes the model generalizable to other applications. Firstly, an asymptotically minimum variance unbiased estimator and two optimal moving fixed interval filters which are more flexibles are exposed. Then, the post-processing case leads to the construction of two moving fixed interval smoothers. These estimators are tested and compared with the classical Kalman filter with simulated and real data, and the results show a good accuracy of each of them.

1 Introduction

The development of Field Operational Test (FOT) and Naturalistic Driving Study (NDS) allow to collect large databases that provide a wealth of information regarding driving behavior and more generally the interactions between driver, vehicle and/or environment factors (e.g. the SHRP 2 NDS with about 3000 vehicles in the United States for 2 years [1], and the EuroFOT project with about 1000 vehicles in Europe for 1 year [2]). These mass data, generally collected from Floating Car Data (FCD), can be used both to study global effects by calculating aggregated indicators such as mean or median, and also effects at a more local scale by studying individual speed or accelerate profiles. Some studies have shown that space-speed profiles (speed versus vehicle position) are very informative to study driver behavior and the effects of some infrastructure elements (for example, behavioral studies at a signalised intersection [3], or effects of traffic calming measures such as speed humps and speed cushions [4]). Such studies require relatively accurate location information.

Global navigation satellite systems (GNSS), such as the Global Positioning System (GPS), are commonly used for vehicle positioning and are based on measurements of the propagation time of a signal between each visible satellites and the receiver. However, GNSS performance is highly dependent on the environment, and in urban environments the signal is affected by many errors due to satellite masking and multipath. A common solution is to use additional sensors such to overcome the weaknesses of GNSS. In practice, the reliability of vehicle positioning is obtained by the coupling of GNSS that provide absolute positioning, with dead reckoning (DR) system, such as odometer and gyroscope, that provides vehicle's position relative to an initial position [5, 6, 7]. However, in the long term the performance of DR systems is poor due to the accumulation of measurements errors over time. Thus, positioning information from GNSS and DR systems are complementary.

Many methods exist for multi-sensor vehicle navigation (e.g. neural networks [8], fuzzy logic [9], particle filter [10]) but Kalman filtering/smoothing techniques are the most used for their speed and ease of implementation [6, 11, 12, 5, 7]. The Kalman filter/smoother is a recursive algorithm to estimate a signal from noisy measurements, based on a compromise between a predictive dynamic model and a measurement model. The Kalman filter is a real-time estimator that uses only the past observations $y(k)$ $(0 \leq k \leq t)$ to estimate the state vector $x(t)$ at the time t. The basic Kalman filter ([13]), based on least squares approach, is an optimal estimator under the assumptions of linearity of the system and the gaussian distribution of the errors. Some ex-

tensions algorithms have also been developed, such as the Extended Kalman Filter (EKF) and the Unscented Kalman Filter (UKF), in the case of nonlinear systems. However, in Naturalistic Driving Studies, data are usually post-processed and it is desirable to dispose all the measurement data of the experiment in order to achieve better estimation accuracy. Estimators that take into account both past and future observations are often called smoothers. Fixed-Interval Smoothing (FIS) algorithms, based on Kalman filtering/smoothing theory, involve measurements over a given fixed time interval $[0, T]$ and use all the measurements $y(k)$ $(0 \leq k \leq T,\ T > t)$ to estimate the state vector $x(t)$. Fixed-interval smoothers are generally two-filter smoothers based on a combination of a forward and a backward estimate : a forward pass that processes a Kalman filter, and a backward pass that operates backward in time by using the measurements after the time t. The most popular fixed-interval smoothing algorithms are the Rauch-Tung-Striebel (RTS) smoother [14], the Main-Fraser smoother [15, 16] and the Wall-Willsky-Sandell smoother [17]. The main drawback of these smoothers is that they require the operation of two filters.

This paper presents some optimal real-time and post-processing estimators of the distance traveled by a vehicle on a road segment relative to an initial position, using odometer and map-matched GPS measurements. The main contributions of this paper are to propose a simple error model of the sensors which makes the model generalizable to other applications while being efficient, and to propose moving fixed interval filter/smoother which allow flexibility of use. In section 2, the statistical model is explained and the construction of the estimators are developed. Firstly, two real-time estimators are exposed: an asymptotically minimum variance unbiased estimator and an optimal moving fixed interval filter. Then, a generalization of the two previous filters in the post-processing case, leads to the construction of two moving fixed interval smoothers. The effectiveness of these estimators is tested and a comparison with the Kalman filter is performed in section 3 with simulated and real data. Finally, a discussion about the results is proposed.

2 Methodology

2.1 Statistical modelisation

The aim of this study is to estimate the vehicle position $x(t_i)$ at time t_i on a road segment. We denote $\{X(t) : t \in [0, T]\}$ the continuous random process representing the vehicle position on the time interval $[0, T]$, and $\{X(t_i) : i = 1, ..., n\}$ the sampled process. Let $\{x(t_1), ..., x(t_n)\}$ a realization of this random process. Our aim is to estimate this realization from odometer and GPS noisy data.

Let n and m, two integers with $m \leq n$, the number of measurements respectively provided by the odometer and the GPS. In the remainder of this paper, it is assumed that the GPS measurements are map-matched, so that the vehicle is positioned on the correct road segment. Many map-matching algorithms have been developed to identify the correct road segment on which the vehicle is travelling. Most of these algorithms use navigation data from GPS and digital spatial road network data and current map-matching algorithms are described in [18], but the choice of the correct road segment is not the subject of this study. We suppose that the correct road segment have been identified and we search to determine the vehicle location on that segment. For example, Taylor et al. (2006) developed in [19] a map-matching algorithm called OMMGPS that combine GPS pseudorange observations and odometer positions to provide a vehicle position at 1s epochs.

In this study, the map-matched GPS measurements denoted $(y_{gps}(t_0), \ldots, y_{gps}(t_m))$ represent the curvilinear abscissa of the vehicle on the studied road segment (absolute location). GPS position data are affected by many errors including atmospheric and ionospheric errors, satellite orbit errors, satellite clock errors, and multipath errors. We represent these errors by a white Gaussian noise which is a classical hypothesis especially in Kalman filtering ([20], [21]), even if current studies have shown that noises are non centered Gaussian distributions in urban environments but rather Gaussian mixture ([22]). The odometer measurements denoted $(y_{od}(t_0), \ldots, y_{od}(t_n))$ represent the distance traveled by the vehicle from the initial position $x(t_0)$ (relative location). Odometer data are affected by many errors which are divided into two categories: systematic errors related to the properties of the

vehicle (mainly unequal wheel diameters and uncertainty about the wheelbase) and nonsystematic errors related to the environment (mainly wheel slippage due to slippery roads, over-acceleration, ...)([23]). Nonsystematic errors are very difficult to estimate because any unexpected irregularity can introduce a huge error, while systematic errors accumulate constantly over time. In our study, we propose a simple modeling of odometer errors and we represented them by a cumulative sum of centered Gaussian distributions. So the discretized observation model of these two sensors can be written as the following system:

$$\begin{cases} y_{od}(t_i) = x(t_i) - x(t_0) + \sum_{k=1}^{i} \varepsilon_{od,k} \\ \qquad \text{with } t_i = \frac{iT}{n},\ i = 1, ..., n \\ y_{gps}(t'_j) = x(t'_j) + \varepsilon_{gps,j} \\ \qquad \text{with } t'_j = \frac{jT}{m},\ j = 1, ..., m \end{cases} \tag{2.1}$$

where $\varepsilon_{od,i}$ and $\varepsilon_{gps,j}$ are independent gaussian centered errors with respective variance σ^2_{od} and σ^2_{gps}. To simplify the model, we assume that the initial position $x(t_0)$ is zero. Thus, the odometer model and the GPS model described in (2.1) differ only by measurement errors and sampling rate: a high sampling rate with accumulating errors for the odometer, and generally a lower sampling rate without accumulating errors for the GPS. The construction of an estimator $\widehat{x}(t_i)$ of the vehicle position $x(t_i)$ at the sampling time $t_i,\ i = 1, \ldots, n$ from noisy measurements of GPS and odometer will take into account advantages and disadvantages of these two sensors.
Later in the paper, we will denote $\lambda = \frac{f_{od}}{f_{gps}}$ the ratio between the odometer and GPS sampling frequencies (in practice, $\lambda \geq 1$) and we will assume that $\lambda \in \mathbb{N}^*$, i.e. that for some time t_i we have both odometer and GPS measurements.

2.2 Real-time estimator

In this section, vehicle position is estimated in real-time, i.e. the position $x(t_i)$ at a given sampling time t_i is estimated using only measurements obtained up to time t_i.

2.2.1 Asymptotically minimum variance unbiased estimator

The main idea is to use odometer measurements, which has the advantage of having a high sampling rate and provide good accuracy in the short term, and to readjust with the GPS measurements, when they are available, to compensate for the accumulation of positional errors. Our estimator is then defined as follows:

Definition 2.1. *Let $\lambda = \frac{f_{od}}{f_{gps}} \in \mathbb{N}^*$. The estimator $\widehat{x}^{\infty}_{RT}(t_i)$ is a real-time estimator of the vehicle position at the given sampling time $t_i,\ i = 1, \ldots, n$, defined recursively as follows: For $i = 1, \ldots, n$,*

$$\begin{cases} \widehat{x}^{\infty}_{RT}(t_i) = w_1\ [\widehat{x}^{\infty}_{RT}(t_{i-1}) + y_{od}(t_i) - y_{od}(t_{i-1})] \\ \quad + w_2\ y_{gps}(t_i) \quad \textit{if} \quad i \equiv 0\ (mod\ \lambda) \\ \textit{Otherwise,} \\ \widehat{x}^{\infty}_{RT}(t_i) = \widehat{x}^{\infty}_{RT}(t_{i-1}) + y_{od}(t_i) - y_{od}(t_{i-1}) \end{cases} \tag{2.2}$$

where $w_1 + w_2 = 1$.

In practice, the initial position is unknown, so we suppose that:

$$\widehat{x}^{\infty}_{RT}(t_0) = \begin{cases} y_{gps}(t_0) & \text{if } y_{gps}(t_0) \text{ is available} \\ y_{od}(t_0) & \text{otherwise} \end{cases}$$

Theorem 2.1. *The real-time estimator $\widehat{x}^{\infty}_{RT}$ defined in definition 2.1 with the following weights: $w_1 = \frac{\lambda\, r + 2 - \sqrt{\lambda\, r(\lambda\, r+4)}}{2}$ and $w_2 = \frac{-\lambda\, r + \sqrt{\lambda\, r(\lambda\, r+4)}}{2}$ where $r = \frac{\sigma^2_{od}}{\sigma^2_{gps}}$ is the ratio between the odometer and GPS variances, is an asymptotically minimum variance unbiased estimator. The asymptotic variance can be written as follows:*

$$Var[\widehat{x}^{\infty}_{RT}(t_i)] \to \sigma^2_{od}\ \frac{\lambda\ w_1^2 + \frac{1}{r} w_2^2}{1 - w_1^2} \quad \textit{when} \quad t_i \to \infty \tag{2.3}$$

The recursive definition of the estimator given in definition 2.1 has the advantage of being simple to compute. However, in general, recursive algorithms require more computational resource than iterative algorithms. So, we give a non-recursive expression of the real-time estimator with asymptotically minimum variance $\widehat{x}^{\infty}_{RT}$ defined in definition 2.1.

Definition 2.2. *Let $\lambda = \frac{f_{od}}{f_{gps}} \in \mathbb{N}^*$ and $\lfloor x \rfloor$ the floor function. The estimator $\widehat{x}^{\infty}_{RT}(t_i)$ is a real-time estimator, with asymptotically minimum variance, of the vehicle position at the given sampling time $t_i,\ i = 1, \ldots, n$, defined as follows:*

$$\widehat{x}^{\infty}_{RT}(t_i) = \sum_{j=1}^{N} \widetilde{w}_j^{-}\, \widehat{x}_j^{-}(t_i) \quad \textit{with} \quad N = \lfloor \frac{i}{\lambda} \rfloor + 1 \tag{2.4}$$

where for $j = 1, ..., N$,

$$\widehat{x}_j^-(t_i) = y_{gps}(t_{g_i^-(j)}) + \sum_{k=g_i^-(j)+1}^{i} (y_{od}(t_k) - y_{od}(t_{k-1})) \quad (2.5)$$

$$with \;\; g_i^-(j) = \lambda \lfloor \frac{i}{\lambda} \rfloor - \lambda(j-1)$$

and the weights $\widetilde{w}_j^-$ *are defined by* $\begin{cases} \widetilde{w}_j^- = w_2 w_1^{j-1} & \text{if } j < N \\ \widetilde{w}_N^- = w_1^{N-1} \end{cases}$
with (w_1, w_2) *the weights defined in theorem 2.1.*

The equivalence with the recursive expression of the estimator $\widehat{x}_{RT}^{\infty}$ given in definition 2.1 is easily demonstrated by recursion.

Remark 2.1.

1. It is easy to check that the sum of weights $\widetilde{w}_j^-$ is equal to one.
2. For a given sampling time t_i, the estimators $\widehat{x}_j^-(t_i)$, $j = 1, \ldots, N$, are also estimators of the vehicle position at time t_i, each estimator being associated with the j-th GPS measurement obtained before time t_i as shown in Figure 1. The real-time estimator $\widehat{x}_{RT}^{\infty}$ is a weighted sum of these estimators.

[Figure 1 about here.]

The estimator $\widehat{x}_{RT}^{\infty}$ defined in both (2.1) and (2.4) uses all measurements obtained up to time t_i. However, with the non-recursive expression (2.4), it is possible to fix an integer $N < \lfloor \frac{i}{\lambda} \rfloor + 1$ in order to obtain a "truncated" estimator that can be more advantageous to calculate from a computational point of view. In this case, the integer N represents the number of GPS measurements (available before the time t_i) used in the computation of the estimator $\widehat{x}_{RT}^{\infty}$. The choice of the value of N is entirely defined by the user which implies a high flexibility in practice. We then deduce an expression of the variance of the estimator $\widehat{x}_{RT}^{\infty}$ with N fixed, at each sampling time t_i.

Theorem 2.2. *Let* $N \geq 1$ *an integer,* $r = \frac{\sigma_{od}^2}{\sigma_{gps}^2}$ *and* $\lambda = \frac{f_{od}}{f_{gps}} \in \mathbb{N}^*$. *Let* $\widehat{x}_j^-$, $j = 1, \ldots, N$ *and* $\widetilde{w}_j^-$, $j = 1, \ldots, N$ *respectively the estimators and the weights defined in definition 2.2. Then the variance of the real-time estimator* $\widehat{x}_{RT}^{\infty}$ *at a sampling time* t_i, $i = 1, \ldots, n$ *is written in matrix form as follows:*

$$Var[\widehat{x}_{RT}^{\infty}(t_i)] = (\widetilde{\mathbf{w}}^-)^T \, \mathbf{\Sigma}^- \widetilde{\mathbf{w}}^- \quad (2.6)$$

where $\widetilde{\mathbf{w}}^- = (\widetilde{w}_1^-, \ldots, \widetilde{w}_N^-)^T$ *and* $\mathbf{\Sigma}^-$ *is the* $N \times N$ *covariance matrix of the estimators* $\widehat{x}_j^-$. *The covariance matrix* $\mathbf{\Sigma}^-$ *can be decomposed as follows:*

$$\mathbf{\Sigma}^- = \sigma_{gps}^2 (\mathbf{I_N} + r \mathbf{A_N}(d_i)) \quad (2.7)$$

where $\mathbf{I_N}$ *is the identity matrix of size* N, $d_i = i - \lambda \lfloor \frac{i}{\lambda} \rfloor$ *is the number of odometer measurements between* t_i *and the first time of a GPS measurement before* t_i, *and* $\mathbf{A_N}(d_i)$ *is a* $N \times N$ *matrix, function of* d_i, *defined by:*

$$\mathbf{A_N}(d_i) = \begin{bmatrix} d_i & d_i & d_i & \cdots & d_i \\ d_i & d_i + \lambda & d_i + \lambda & \cdots & d_i + \lambda \\ d_i & d_i + \lambda & d_i + 2\lambda & \cdots & d_i + 2\lambda \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d_i & d_i + \lambda & d_i + 2\lambda & \cdots & d_i + (N-1)\lambda \end{bmatrix} \quad (2.8)$$

Then, we can also deduce a linear form of the variance of the real-time estimator $\widehat{x}_{RT}^{\infty}$ *at a sampling time* t_i, $i = 1, \ldots, n$ *as follows:*

$$Var[\widehat{x}_{RT}^{\infty}(t_i)] = \sigma_{gps}^2 [\frac{(1-w_1)^2 + 2w_1^{2N-1}(1-w_1)}{1-w_1^2} + r(d_i + \lambda \frac{w_1^2 - w_1^{2N}}{1-w_1^2})] \quad (2.9)$$

where $w_1 = \frac{\lambda r + 2 - \sqrt{\lambda r (\lambda r + 4)}}{2}$ *is the asymptotically optimal weight defined in theorem 2.1.*

Note that for a fixed N, the variance function defined in (2.9) is periodic with period λ. Moreover, since $\widehat{x}_{RT}^{\infty}$ is an asymptotically minimum variance estimator, the weights $\widetilde{w}_j^-$, $j = 1, \ldots, N$, defined in definition 2.2 are optimal when N tends to infinity, i.e. when we have an infinite number of measurements. Thus, if we assume $d_i = 0$ and if N tends to infinity in the variance expression (2.9), we find the

expression of the asymptotically variance given in (2.3). The speed of convergence of the variance of $\widehat{x}_{RT}^{\infty}$ defined in (2.9) to the asymptotically variance defined in (2.3) will be examined in the section 3. However, the estimator $\widehat{x}_{RT}^{\infty}$ is an asymptotically minimum variance estimator and it is not optimal for estimating the vehicle position at a sampling time t_i close to the initial time t_0. Therefore we have also constructed an optimal real-time estimator with minimum variance at each sampling time t_i.

2.2.2 Minimum variance unbiased estimator for a fixed N

In the definition 2.2, we have written a real-time estimator of the vehicle position as a weighted average of the estimators $\widehat{x}_j^-$ $(j = 1, \ldots, N)$ and we have determined the optimal weights $\widetilde{w}_j^-$ $(j = 1, \ldots, N)$ that minimize the asymptotic variance. In this section, we consider the same real-time estimator but we search the optimal weights w_j^- $(j = 1, \ldots, N)$ that minimize the variance at each sampling time t_i.

Theorem 2.3. *Let $N \geq 1$ an integer and $\mathbf{b} = (1, \ldots, 1)^T$ a vector of length N. Assume that $\widehat{x}_j^-$, $j = 1, \ldots, N$ are the estimators defined in definition 2.2 and $\mathbf{\Sigma}^-$ is the $N \times N$ covariance matrix of these estimators defined in theorem 2.2. The estimator $\widehat{x}_{RT}^{opt}(t_i)$ is a real-time estimator of the vehicle position at the given sampling time t_i, $i = 1, \ldots, n$, defined as follows:*

$$\widehat{x}_{RT}^{opt}(t_i) = \sum_{j=1}^{N} \widehat{w}_j^- \widehat{x}_j^-(t_i) \tag{2.10}$$

where the weight vector $\widehat{\mathbf{w}}^- = (\widehat{w}_1^-, \ldots, \widehat{w}_N^-)^T$ satisfies:

$$\begin{aligned} &\widehat{\mathbf{w}}^- = \frac{1}{c_{rt}} (\mathbf{\Sigma}^-)^{-1} \mathbf{b} \\ &\text{with} \quad c_{rt} = \mathbf{b}^T (\mathbf{\Sigma}^-)^{-1} \mathbf{b} \text{ a constant.} \end{aligned} \tag{2.11}$$

Then, $\widehat{x}_{RT}^{opt}(t_i)$ is a minimum variance unbiased estimator of the vehicle position at the sampling time t_i, and its variance at time t_i is the following:

$$Var[\widehat{x}_{RT}^{opt}(t_i)] = \frac{1}{c_{rt}} \tag{2.12}$$

The variance function defined in (2.12) depends on d_i and is periodic with period λ. The estimator $\widehat{x}_{RT}^{opt}$ is a minimum variance unbiased estimator at each sampling time t_i for a fixed N. However, determining the optimal weights $\widehat{w}_j^-$ requires the inversion of the covariance matrix $\mathbf{\Sigma}^-$ which is inconvenient in practice. Furthermore, the expressions of the optimal weights and the variance of $\widehat{x}_{RT}^{opt}$ are not given explicitly in terms of the integer N, which makes it difficult to study the properties of this estimator depending on N. Thus, in some cases, it may be more advantageous to use the "truncated" estimator $\widehat{x}_{RT}^{\infty}$ with a fixed N that have a simpler expression of weights and variance.

2.3 Post-processing estimator

In this section, we assume that data are post-processed and we can use all the GPS and odometer measurements available on the studied time interval $[0, T]$. Thus, unlike the previous section where we were restricted to use only measurements obtained up to time t_i to estimate the vehicle position at the sampling time t_i, the objective of this section is to use all available information to construct a more accurate estimator than the real-time estimators defined in the section 2.2.

2.3.1 Minimum variance unbiased estimator for a fixed N

The general idea is to extend the real-time estimator defined in definition 2.2 in case we also have measurements obtained after the sampling time t_i.

Definition 2.3. *Let $\lambda = \frac{f_{od}}{f_{gps}} \in \mathbb{N}^*$ and $\lfloor x \rfloor$ the floor function. Assume that $N \geq 1$ is a fixed integer. The estimator $\widehat{x}_{PP}(t_i)$ is a post-processing estimator of the vehicle position at the given sampling time t_i, $i = 1, \ldots, n$, defined as follows:*

$$\begin{aligned} &\widehat{x}_{PP}(t_i) = \sum_{j=1}^{N} (w_j^- \widehat{x}_j^-(t_i) + w_j^+ \widehat{x}_j^+(t_i)) \\ &\text{with} \quad \sum_{j=1}^{N} (w_j^- + w_j^+) = 1 \end{aligned} \tag{2.13}$$

where for $j = 1, ..., N$,

$$\begin{aligned}\widehat{x}_j^-(t_i) &= y_{gps}(t_{g_i^-(j)}) \\ &\quad + \sum_{k=g_i^-(j)+1}^{i} (y_{od}(t_k) - y_{od}(t_{k-1})) \\ &\text{with } \; g_i^-(j) = \lambda \lfloor \frac{i}{\lambda} \rfloor - \lambda(j-1)\end{aligned} \tag{2.14}$$

and

$$\begin{aligned}\widehat{x}_j^+(t_i) &= y_{gps}(t_{g_i^+(j)}) \\ &\quad - \sum_{k=i+1}^{g_i^+(j)} (y_{od}(t_k) - y_{od}(t_{k-1})) \\ &\text{with } \; g_i^+(j) = \lambda \lfloor \frac{i}{\lambda} \rfloor + \lambda\, j\end{aligned} \tag{2.15}$$

A graph of the estimators $\widehat{x}_j^-$ *and* $\widehat{x}_j^+$ *is represented in Figure 2.*

In this case, the integer N represents the number of GPS measurements (available before and after the time t_i) used in the computation of the estimator (i.e. a total of $2N$ GPS measurements around t_i).

[Figure 2 about here.]

The following lemma gives a general expression of the variance of the post-processing estimator defined in (2.13).

Lemma 2.1. *Let* $N \geq 1$ *an integer and* $\widehat{x}_j^-$ *and* $\widehat{x}_j^+$, $j = 1, \ldots, N$, *the estimators defined in definition 2.3. Let* $\widehat{\mathbf{w}} = (\widehat{w}_1^-, \ldots, \widehat{w}_N^-, \widehat{w}_1^+, \ldots, \widehat{w}_N^+)^T$ *the weight vector of length* $2N$. *The variance of the post-processing estimator* $\widehat{x}_{PP}$ *defined in definition 2.3 can be written as follows:*

$$Var[\widehat{x}_{PP}(t_i)] = \mathbf{w}^T \mathbf{\Sigma}\, \mathbf{w} \tag{2.16}$$

where $\mathbf{\Sigma}$ *is the* $2N \times 2N$ *covariance matrix of the estimators* $\widehat{x}_j^-$ *and* $\widehat{x}_j^+$ *defined by* $\Sigma = \begin{bmatrix} \Sigma^- & 0 \\ 0 & \Sigma^+ \end{bmatrix}$ *with* $\mathbf{\Sigma}^-$ *and* $\mathbf{\Sigma}^+$ *respectively the* $N \times N$ *covariance matrix of the estimators* $\widehat{x}_j^-$ *and* $\widehat{x}_j^+$. *Furthermore, we can decomposed* $\mathbf{\Sigma}^-$ *and* $\mathbf{\Sigma}^+$ *as follows:*

$$\begin{aligned}\mathbf{\Sigma}^- &= \sigma_{gps}^2(\mathbf{I_N} + r\mathbf{A_N}(d_i)) \;\; \textit{and} \\ \mathbf{\Sigma}^+ &= \sigma_{gps}^2(\mathbf{I_N} + r\mathbf{A_N}(\lambda - d_i))\end{aligned} \tag{2.17}$$

where $\mathbf{I_N}$ *is the identity matrix of size* N, $d_i = i - \lambda \lfloor \frac{i}{\lambda} \rfloor$ *is the number of odometer measurements between* t_i *and the first time of a GPS measurement before* t_i, *and* $\mathbf{A_N}(d_i)$ *is a* $N \times N$ *matrix, function of* d_i, *defined in (2.8).*

Then, we search the optimal weights (w_j^-, w_j^+) $(j = 1, \ldots, N)$ that minimize the variance of the post-processing estimator $\widehat{x}_{PP}$ at each sampling time t_i. Intuitively, we give more weight to the estimators $\widehat{x}_j^-(t_i)$ and $\widehat{x}_j^+(t_i)$ associated with GPS measurements obtained at times close to t_i. The minimum variance unbiased estimator $\widehat{x}_{PP}^{opt}(t_i)$ of the vehicle position at the sampling time t_i is similar to the minimum variance unbiased estimator $\widehat{x}_{RT}^{opt}(t_i)$ defined in theorem 2.3 by taking $\widehat{\mathbf{w}} = (\widehat{w}_1^-, \ldots, \widehat{w}_N^-, \widehat{w}_1^+, \ldots, \widehat{w}_N^+)^T$ as weight vector, $\mathbf{\Sigma}$ defined in lemma 2.1 as covariance matrix, and $\mathbf{b} = (1, \ldots, 1)^T$ a vector of length $2N$. Thus the constant c_{rt} becomes the constant $c_{pp} = \mathbf{b}^T(\mathbf{\Sigma})^{-1}\mathbf{b}$. However, as in the case of the real-time estimator, the computation of the optimal weights $\widehat{w}_j^-$ and $\widehat{w}_j^+$ requires the inversion of the covariance matrix $\mathbf{\Sigma}$ which is inconvenient in practice. Thus, we study the post-processing estimator with asymptotically minimum variance.

2.3.2 Asymptotically minimum variance unbiased estimator

Theorem 2.4. *Let* $N \geq 1$ *an integer,* $r = \frac{\sigma_{od}^2}{\sigma_{gps}^2}$ *and* $\lambda = \frac{f_{od}}{f_{gps}} \in \mathbb{N}^*$. *Let* $\widehat{x}_j^-$ *and* $\widehat{x}_j^+$, $j = 1, \ldots, N$, *the estimators defined in definition 2.3. Assume that* (w_1, w_2) *are the weights defined in theorem 2.1. Then* $\widehat{x}_{PP}^{\infty}(t_i)$ *is an asymptotically minimum variance unbiased estimator of the position of the vehicle at the sampling time* t_i, *defined as follows:*

$$\widehat{x}_{PP}^{\infty}(t_i) = \widetilde{w}_1 \widetilde{x}_{PP}^-(t_i) + \widetilde{w}_2 \widetilde{x}_{PP}^+(t_i) \tag{2.18}$$

where

$$\widetilde{x}^{-}_{PP}(t_i) = \sum_{j=1}^{N} \widetilde{w}^{-}_{j} \widehat{x}^{-}_{j}(t_i) \quad \textit{and} \quad \widetilde{x}^{+}_{PP}(t_i) = \sum_{j=1}^{N} \widetilde{w}^{+}_{j} \widehat{x}^{+}_{j}(t_i) \tag{2.19}$$

$$\textit{with} \begin{cases} \widetilde{w}^{-}_{j} = \widetilde{w}^{+}_{j} = w_2 w_1^{j-1} & \textit{for } j < N \\ \widetilde{w}^{-}_{N} = \widetilde{w}^{+}_{N} = w_1^{N-1} \end{cases} .$$

The weights $(\widetilde{w}_1, \widetilde{w}_2)$ whose sum is equal to one, can be written as follows:

$$\widetilde{w}_1 = \frac{Var[\widetilde{x}^{+}_{PP}(t_i)]}{Var[\widetilde{x}^{-}_{PP}(t_i)] + Var[\widetilde{x}^{+}_{PP}(t_i)]} \quad \textit{and} \quad \widetilde{w}_2 = \frac{Var[\widetilde{x}^{-}_{PP}(t_i)]}{Var[\widetilde{x}^{-}_{PP}(t_i)] + Var[\widetilde{x}^{+}_{PP}(t_i)]}$$

where

$$Var[\widetilde{x}^{-}_{PP}(t_i)] = \sigma^2_{gps}[\frac{(1-w_1)^2 + 2w_1^{2N-1}(1-w_1)}{1-w_1^2} + r(d_i + \lambda \frac{w_1^2 - w_1^{2N}}{1-w_1^2})] \quad \textit{and}$$

$$Var[\widetilde{x}^{+}_{PP}(t_i)] = \sigma^2_{gps}[\frac{(1-w_1)^2 + 2w_1^{2N-1}(1-w_1)}{1-w_1^2} + r(\lambda \frac{1-w_1^{2N}}{1-w_1^2} - d_i)].$$

Then, we deduce the following expression for the variance of the estimator $\widehat{x}^{\infty}_{PP}$:

$$Var[\widehat{x}^{\infty}_{PP}(t_i)] = \frac{Var[\widetilde{x}^{-}_{PP}(t_i)] Var[\widetilde{x}^{+}_{PP}(t_i)]}{Var[\widetilde{x}^{-}_{PP}(t_i)] + Var[\widetilde{x}^{+}_{PP}(t_i)]} \tag{2.20}$$

As for the asymptotically minimum variance real-time estimator, when N is fixed, we obtain a truncated estimator. The asymptotic variance is obtained when N tends to infinity in the expression (2.20).

3 Data processing and discussion

In this section, we present simulation and real data results and a comparison of the different estimators defined in the previous section with a classical Kalman filter. They have been obtained on a DELL T3400 workstation equipped with a Intel E8400 core 2 duo processor. The Kalman filter is constructed using only the odometer and GPS measurements in order to fairly compare this Kalman filter with the real-time estimators defined in the section 2.2. Thus, the state vector is only composed with one component x_i where x_i is the distance traveled by the vehicle at time t_i from the initial position x_0. The dynamic equation is given by:

$$x_{i+1} = x_i + (y_{od,i+1} - y_{od,i}) + \varepsilon_{od,i} \tag{3.1}$$

where $y_{od,i}$ is the odometer measurement at time t_i and $\varepsilon_{od,i} \sim N(0, \sigma^2_{od})$. The measurement equation using only GPS measurement is given by:

$$y_{gps,i+1} = x_{i+1} + \varepsilon_{gps,i} \tag{3.2}$$

where $y_{gps,i}$ is the GPS measurement at time t_i and $\varepsilon_{gps,i} \sim N(0, \sigma^2_{gps})$.
Then the step prediction is performed as follows:

$$\begin{cases} \widehat{x}_{i+1|i} = \widehat{x}_{i|i} + (y_{od,i+1} - y_{od,i}) \\ P_{i+1|i} = P_{i|i} + \sigma^2_{od} \end{cases} \tag{3.3}$$

where $\widehat{x}_{i|i}$ is the state estimate at time t_i knowing the measures until t_i, and $P_{i|i}$ is the related covariance matrix of the estimation error (here, $P_{i|i}$ is a real number). The update step is performed as follows:

$$\begin{cases} K_{i+1} = P_{i+1|i} \ (P_{i+1|i} + \sigma^2_{gps})^{-1} \\ \widehat{x}_{i+1|i+1} = \widehat{x}_{i+1|i} + K_{i+1}(y_{gps,i+1} - \widehat{x}_{i+1|i}) \\ P_{i+1|i+1} = (1 - K_{i+1}) P_{i+1|i} \end{cases} \tag{3.4}$$

The filter is initialized as follows:

$$\begin{cases} \widehat{x}_{0|0} = y_{gps}(t_0) \\ P_{0|0} = \sigma^2_{gps} \end{cases} \tag{3.5}$$

Later in the document, the Kalman filter will be denoted $\widehat{x}^{KF}_{RT}$. More details on the Kalman filter can be found in [24] and [25].

3.1 Simulation results

Given a reference vehicle trajectory length of 4000m and traveled in about 300s, the odometer and map-matched GPS data were simulated from the model 2.1 with an odometer error standard deviation σ_{od} equal to 0.05m and a GPS error standard deviation σ_{gps} equal to 3m. These sensor simulated data represent the measurement data $y_{od}(t_i)$ and $y_{gps}(t_i)$ at each time t_i used in the calculation of each estimators. We suppose that the odometer and GPS frequencies are respectively equal to 10Hz and 1HZ, so that the ratio λ is equal to 10. 100 simulations of GPS and odometer measurements were generated, each simulation involved generating a new set of sensor data of the reference distance traveled and computing the estimated location at each time t_i with each estimators defined in the previous section (real-time estimators and post-processing estimators) and with a Kalman filter. For each estimator, the Root Mean Square Error (RMSE) for all 100 simulations was computed every second (i.e. at each time t_i for which a GPS measurement is available).
The RMSE is a good measure of the accuracy of an estimator and has the advantage of being expressed in the same units as the quantity being estimated (i.e. in meters). The RMSE of an estimator $\widehat{X}$ of a vector X is defined as follows:

$$\begin{aligned} RMSE(\widehat{X}) &= \sqrt{MSE(\widehat{X})} \\ &= \sqrt{E((\widehat{X}-X)^T(\widehat{X}-X))} \\ &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{X}_i - X_i)^2} \end{aligned} \tag{3.6}$$

where $\widehat{X}_i$ (resp. X_i) is the i-th component of the vector $\widehat{X}$ (resp. X). There are other types of errors (e.g. mean absolute error, geometric average error), but the choice of the MSE is justified by its interpretation in terms of bias and variance:

$$MSE(\widehat{X}) = [Bias(\widehat{X})]^2 + Var(\widehat{X}) \tag{3.7}$$

where $Bias(\widehat{X}) = E[\widehat{X}] - X$. Thus, the best estimator between two unbiased estimators is the one that has the smallest variance, and an unbiased estimator of minimum variance is generally regarded as the best estimator possible. It is moreover well as real-time estimator $\widehat{x}_{RT}^{opt}$ and post-processing estimator $\widehat{x}_{PP}^{opt}$ were built.

Figures 3 and 4 contain the RMSE of the vehicle location obtained with each estimator. In these two figures, the RMSE of the simulated sensors are represented by green circles for GPS and blue dots for the odometer. Each estimator (real-time and post-processing) is compared with the Kalman filter $\widehat{x}_{RT}^{KF}$ defined in the beginning of the section 3 (denoted x_hat_RT_KF in Figure 3 and 4) and represented by orange dashed line. Figure 3 presents the comparaison between the RMSE of the vehicle location obtained with the real-time estimators defined in section 2.2 and the Kalman filter $\widehat{x}_{RT}^{KF}$. The chocolate dashed line represent the asymptotically minimum variance estimator $\widehat{x}_{RT}^{\infty}$ (denoted x_hat_RT_inf in Figure 3) defined recursively in theorem 2.1 and initialized with $\widehat{x}_{RT}^{\infty}(t_0) = y_{gps}(t_0)$. The turquoise line represented the truncated estimator $\widehat{x}_{RT}^{\infty}$ with a fixed N defined in theorem 2.2 (and denoted x_hat_RT_inf_Nfix in Figure 3) and the red line represent the minimum variance estimator $\widehat{x}_{RT}^{opt}$ for a fixed N defined in theorem 2.3 (and denoted x_hat_RT_opt_Nfix in Figure 3). These two estimators depending on N were computed for three different values of N:

- N=4 is a small value chosen at random;
- N=20 is the threshold above which the difference between the standard deviation of the minimum variance estimator $\sqrt{Var[\widehat{x}_{RT}^{opt}(t_i)]}$ defined in theorem 2.3 (with $d_i = 0$ since the time step is 1s and the GPS frequency is 1Hz) and the square root of the asymptotic variance of $\widehat{x}_{RT}^{\infty}$ defined in (2.3) is less than 0.1m;
- N=40 is the threshold above which the difference between the standard deviation of the truncated asymptotically minimum variance estimator $\sqrt{Var[\widehat{x}_{RT}^{\infty}(t_i)]}$ with N fixed defined in theorem 2.2 (with $d_i = 0$) and the square root of the asymptotic variance of $\widehat{x}_{RT}^{\infty}$ defined in (2.3) is less than 0.1m.

Figure 3 shows that the Kalman filter is the best estimator of the vehicle location, mainly at each time t_i of the beginning of the path, but after around 50s the RMSE curve of the Kalman filter and that

of the asymptotically minimum variance estimator $\widehat{x}_{RT}^{\infty}$ are merged. Similarly, when N=20, the RMSE curve of the minimum variance estimator $\sqrt{Var[\widehat{x}_{RT}^{opt}(t_i)]}$ is approximately merged with the RMSE curve of the Kalman filter, and it is the same for the truncated version of the estimator $\widehat{x}_{RT}^{\infty}$ when N=40. The average and maximum RMSE of each estimator represented in Figure 3 are given in Table 1. These values confirm the results described in Figure 3. The asymptotic standard deviation achieved by all estimators and equal to $\sqrt{Var[\widehat{x}_{RT}^{\infty}(t_i)]}$ when $t_i \to \infty$ can be calculated with the formula given in (2.3). We then obtained an optimal standard deviation equal to 0.68m which corresponds approximatively to the RMSE obtained with each estimator after 50s when N is sufficiently large.

[Figure 3 about here.]

[Table 1 about here.]

Figure 4 and Table 2 are similar to Figure 3 and Table 1 but with a comparison between the Kalman filter and the post-processing estimators defined in section 2.3. Thus, the turquoise line represents the truncated estimator $\widehat{x}_{PP}^{\infty}$ with a fixed N defined in theorem 2.4 (and denoted x_hat_PP_inf_Nfix in Figure 4) and the red line represents the minimum variance estimator $\widehat{x}_{PP}^{opt}$ for a fixed N defined after the lemma 2.1 (and denoted x_hat_PP_opt_Nfix in Figure 4). These two estimators depending on N were computed for N=4 as in the real-time case and also for the two following values:

- N=17 is the threshold above which the difference between the standard deviation of the minimum variance estimator $\sqrt{Var[\widehat{x}_{PP}^{opt}(t_i)]}$ (with $d_i = 0$) and the square root of the asymptotic variance of $\widehat{x}_{PP}^{\infty}$ obtained in (2.20) when N tend to infinity, is less than 0.1m;
- N=36 is the threshold above which the difference between the standard deviation of the truncated asymptotically minimum variance estimator $\sqrt{Var[\widehat{x}_{PP}^{\infty}(t_i)]}$ with N fixed defined in theorem 2.4 (with $d_i = 0$) and the square root of the asymptotic variance of $\widehat{x}_{PP}^{\infty}$ obtained in (2.20) when N tend to infinity, is less than 0.1m.

Figure 4 and Table 2 show that when we use measurements obtained after time t_i (post-processing case), the accuracy of the estimate of the position of the vehicle is improved and is better than using the Kalman filter except at the end of the path where a side effect appears. The asymptotic standard deviation achieved by all estimators when N is sufficiently large, except on the boundaries, and corresponding to $\sqrt{Var[\widehat{x}_{PP}^{\infty}(t_i)]}$ when $N \to \infty$ in (2.20), is equal to 0.49m.

[Figure 4 about here.]

[Table 2 about here.]

3.2 Real data results

In this section, real data collected from a trip provided on test tracks at Versailles-Satory (France) were used. The trip length was around 4000m and a travel time of 300s. The odometer measurements have been collected on CAN (Controller Area Network) bus of the vehicle and have been provided at a 10Hz sampling frequency. Two GPS were also located on the roof of the vehicle: A GlobalSat BR-355 GPS receiver (with SIRF Star III) and a Thales Sagitta RTK-GPS receiver. The BR-355 GPS provides position measurements at a 1Hz sampling frequency with a 10m accuracy and the RTK-GPS (Real-Time Kinematic Global Positioning System) provides position measurements at a 10Hz sampling frequency with a centimeter accuracy. Thus, the RTK-GPS measurements were used as the "true" locations of the vehicle and were considered as the reference trajectory. A simple map-matching algorithm was used in order to project the GPS measurements on the road, and the position measurements from the odometer and the two GPS were synchronized in time. According to the accuracy of the sensors, we assume that the standard deviations of the errors of the odometer σ_{od} and the GPS σ_{gps} are respectively equal to 0.03m and 3m. The ratio λ between the odometer and GPS sampling frequencies is equal to 10 as in the previous section. Tables 3.a and 3.b contain the RMSE of the vehicle location of each estimator on the complete trip, and the total computation time of the estimated positions with each estimator at each time t_i with a sampling frequency of 10Hz. The Kalman filter is compared with the real-time estimators in Table 3.a. For comparaison, the values of N are the same as in the previous section with simulated data. Contrary to previous results obtained

with simulated data, Table 3.a shows that the minimum variance estimator $\widehat{x}_{RT}^{opt}$ and the truncated version of the estimator $\widehat{x}_{RT}^{\infty}$ with a fixed N are better than the Kalman filter and the asymptotically minimum variance estimator $\widehat{x}_{RT}^{\infty}$. However the results show that the GPS is more accuracy (RMSE=3.07m) than all the real-time estimators except the minimum variance estimator $\widehat{x}_{RT}^{opt}$ with N=4 (RMSE=2.59m). Furthermore, increasing the interval smoothing (i.e. increase the value of N) does not improve the accuracy of the real-time estimators that depend on N. Table 3.b shows that all the post-processing estimators are better than the real-time estimators and are more accuracy than the GPS. However the computational time of the post-processing estimators are bigger than the real-time estimators mainly when N is large.

[Table 3 about here.]

3.3 Discussion

The asymptotically minimum variance estimator $\widehat{x}_{RT}^{\infty}$ is similar to the Kalman filter $\widehat{x}_{RT}^{KF}$ with a fixed gain K which is optimal when the estimation time t_i tends to infinity. Indeed, the weights (w_1, w_2) defined in definition 2.1 and theorem 2.1 are fixed which saves computation time (twice as fast as the Kalman filter with the real data). Furthermore, the construction of $\widehat{x}_{RT}^{\infty}$ provides a simple expression of the asymptotic variance (equation (2.3)). However, the optimality of these weights only at infinity implies a poor accuracy of the estimator $\widehat{x}_{RT}^{\infty}$ at the beginning of the trip, even if its convergence to the optimal estimator is relatively fast (50s with the simulated data). Thus, when the vehicle distance traveled to be estimated is quite long in time, the estimator $\widehat{x}_{RT}^{\infty}$ can be more effective.
By definition, the estimator $\widehat{x}_{RT}^{opt}$ is the optimal estimator at each time t_i for a N fixed and the Kalman filter $\widehat{x}_{RT}^{KF}$ is the optimal estimator using all measurements available up to t_i. Therefore, at a given time t_i, the Kalman filter $\widehat{x}_{RT}^{KF}(t_i)$ is similar to the estimator $\widehat{x}_{RT}^{opt}(t_i)$ with $N = \lfloor \frac{i}{\lambda} \rfloor + 1$ (corresponding to use all measurements up to time t_i). The Kalman filter $\widehat{x}_{RT}^{KF}(t_i)$ then corresponds to $\widehat{x}_{RT}^{opt}$ with a non-fixed interval filtering which increases over time. However, contrary to the simulation results, the real data results have shown that increasing the size of the interval filtering, by increasing N, does not improve the accuracy of the estimator $\widehat{x}_{RT}^{opt}$. Indeed, adding too much information to estimate the position at a sampling time t_i can bias the estimation. Measurements obtained at times close to t_i are supposed to contain the most accurate information to estimate the position at time t_i unless such measures are very noisy. In the case of noisy measurements around t_i, it is better to increase the interval filtering (or the interval smoothing in the post-processing case) even if the computational time increase. Thus, the best estimator could be an estimator with a variable interval filtering that would be optimal at each time t_i even if the computation time of such an estimator would certainly be large. However, it is important to note that Tables 3.a and 3.b give the computational time to estimate the whole trip, i.e. a vector of size n containing the estimated distance traveled at each time t_i, which benefits recursive estimators as the Kalman filter. But if we want only one estimated distance traveled at a given time t_i, a recursive expression requires to compute all the estimated distance up to t_i, which can significantly increase the computational time when t_i is large. For example, with the real data used in section 3.2, the computational time of the estimated position at time $t_i = 300s$ is respectively 6.30ms for the Kalman filter and 2.45ms for the estimator $\widehat{x}_{RT}^{opt}$ with $N = 20$. Thus, in some cases, the estimator $\widehat{x}_{RT}^{opt}(t_i)$ is faster to compute, even if the computational time depends on the size of the interval filtering or smoothing.
Finally, the estimator $\widehat{x}_{RT}^{opt}$ (resp. $\widehat{x}_{PP}^{opt}$) is accurate but requires the inversion of a matrix of size $N \times N$ (resp. $2N \times 2N$) which can be a disadvantage when N is large. In such cases, the truncated version of $\widehat{x}_{RT}^{\infty}$ (resp. $\widehat{x}_{PP}^{\infty}$) with a fixed N can be a good alternative. Indeed, this estimator is faster to compute than $\widehat{x}_{RT}^{opt}$ (resp. $\widehat{x}_{PP}^{opt}$) for the same N and converge to the optimal estimator when N tends to infinity, even if it is on average less accuracy.

4 Conclusion

Some real-time and post-processing estimators of the distance traveled by a vehicle on a road segment was developed and compared to a classic Kalman filter. These estimators were based on a simple statistical error model of the odometer and the GPS which makes the model generalizable to other ap-

plications. Firstly, a recursive asymptotically minimum variance filter, similar to the Kalman filter with a fixed gain K which is optimal when time tends to infinity, was developed. This estimator is two times faster to compute than the Kalman filter and converges quickly to the optimal estimator (error less than 1m after 50s with simulated data). Then, two more flexible filters was developed using only measurements included in a moving fixed-interval: an optimal filter that requires the inversion of the covariance matrix, and a truncated version of the asymptotically minimum variance filter that is less accurate but have a simpler expression of weights and variance. Real-data results have shown the interest of using moving fixed-interval filter instead of recursive filter such as the Kalman filter: the error can be averaged less than 3m with a good choice of filtering window size. Finally, two moving fixed-interval smoothers derived from the two previous filters was also developed for post-processing cases. These smoothers estimate the vehicle position at a time t by using the measurement over a specified window around t. Taking into account both past and future observations allows to achieve better estimation accuracy (error less than 2m with real data).
In future work, the robustness of the estimators presented in this paper will be tested. Furthermore, this work have shown the interest of using moving fixed-interval filter/smoother but the choice of the best filtering/smoothing windows size for the whole trip is difficult. The development of a nonfixed-interval filter/smoother with an optimal windows size at each time t could be a good alternative even if it would certainly be at the expense of a higher computation time.

5 Appendix A: Nomenclature

$x(t_i)$	vehicle position at time t_i, $t_i \in [0, T]$
$y_{od}(t_i), y_{gps}(t_i)$	odometer and map-matched GPS measurements at time t_i, $t_i \in [0, T]$
$\sigma^2_{od}, \sigma^2_{gps}$	variances of the odometer and the GPS
r	ratio between the variance of the odometer and the variance of the GPS
λ	ratio between the sampling frequency of the odometer and the sampling frequency of the GPS
d_i	Number of odometer measurements between the time t_i and the first time of a GPS measurement before t_i
N	Number of GPS measurements used in the computation of the fixed-interval filter/smoother
$\widehat{x}^{KF}_{RT}(t_i)$	Kalman filter estimator of the vehicle position at time t_i
$\widehat{x}^{\infty}_{RT}(t_i)$	Real-time unbiased estimator of the vehicle position at time t_i with asymptotically minimum variance
$\widehat{x}^{opt}_{RT}(t_i)$	Real-time unbiased estimator of the vehicle position at time t_i with minimum variance for a fixed N
$\widehat{x}^{\infty}_{PP}(t_i)$	Post-processing unbiased estimator of the vehicle position at time t_i with asymptotically minimum variance
$\widehat{x}^{opt}_{PP}(t_i)$	Post-processing unbiased estimator of the vehicle position at time t_i with minimum variance for a fixed N

6 Appendix B: Proofs of Theorems

Proof of Theorem 2.1:
Unbiased: Since $\varepsilon_{od,i}$ and $\varepsilon_{gps,j}$ are centered errors, it is easy to prove that $E[\widehat{x}^{\infty}_{RT}(t_i) - x(t_i)] = 0$, i.e. the estimator is unbiased.
Convergence of variance: Considering only the sampling time where we have both odometer and GPS measurements, the two equations defined in (2.2) can be written as follows: $\widehat{x}^{\infty}_{RT}(t_{\lambda j}) = w_1\,[\widehat{x}^{\infty}_{RT}(t_{\lambda(j-1)}) + \sum_{k=\lambda(j-1)}^{\lambda j-1}(y_{od}(t_{k+1}) - y_{od}(t_k))] + w_2\; y_{gps}(t_{\lambda j}), \quad j = 1, \ldots, m$. Thus, the variance

of the estimator can be represented as an arithmetico-geometric sequence: $Var[\widehat{x}^{\infty}_{RT}(t_{\lambda j})] = w_1^2\ Var[\widehat{x}^{\infty}_{RT}(t_{\lambda(j-1)})] + w_1^2\ \lambda\ \sigma^2_{od} + w_2^2\ \sigma^2_{gps},\quad j = 1,...,m$. Since $|w_1^2| < 1$, the sequence converges and its limit is $\sigma^2_{od}\ \frac{\lambda\ w_1^2 + \frac{1}{r} w_2^2}{1-w_1^2}$ where $r = \frac{\sigma^2_{od}}{\sigma^2_{gps}}$.

Calculation of asymptotically optimal weights: The sum of weights is equal to one. Thus, the asymptotic variance can be written as a function of one variable defined on $[0,1]$ as follows: $\varphi(w_1) = \sigma^2_{od}\ \frac{\lambda\, w_1^2 + \frac{1}{r}(1-w_1)^2}{1-w_1^2}$. Since φ is convex on $[0,1]$, φ has a global minimum which satisfies the quadratic equation $\varphi'(w_1) = 0$. This equation has only one solution on $[0,1]$: $w_1 = \frac{\lambda\, r + 2 - \sqrt{\lambda\, r(\lambda\, r+4)}}{2}$. □

Proof of Theorem 2.2:

The matrix form 2.6 of the variance of $\widehat{x}^{\infty}_{RT}$ is derived from the definition 2.4. Furthermore, for a given sampling time t_i and for $j = 1,\ldots,N$, $Var[\widehat{x}^{-}_{j}(t_i)] = \sigma^2_{gps} + (d_i + (j-1)\lambda)\,\sigma^2_{od}$ according to the definition of the estimators $\widehat{x}^{-}_{j}$ given in 2.14, and $Cov(\widehat{x}^{-}_{j}(t_i), \widehat{x}^{-}_{j'}(t_i)) = Var[\sum_{k=g_i^-(j)+1}^{i} \varepsilon_{od,k}] = (d_i + (j-1)\lambda)\,\sigma^2_{od}$ by independence of $\varepsilon_{od,i}$, which proves the expression 2.17 of the covariance matrix $\mathbf{\Sigma}^{-}$. Now, it remains to show the linear expression 2.9 of the variance of $\widehat{x}^{\infty}_{RT}$. We have shown that: $Var[\widehat{x}^{\infty}_{RT}(t_i)] = (\widetilde{\mathbf{w}}^{-})^T\,\mathbf{\Sigma}^{-}\,\widetilde{\mathbf{w}}^{-} = \sigma^2_{gps}[(\widetilde{\mathbf{w}}^{-})^T\,\widetilde{\mathbf{w}}^{-} + r\,(\widetilde{\mathbf{w}}^{-})^T\,\mathbf{A_N}(d_i)\,\widetilde{\mathbf{w}}^{-}]$ where $(\widetilde{\mathbf{w}}^{-})^T\,\widetilde{\mathbf{w}}^{-} = \sum_{j=1}^{N}(\widetilde{w}_j^{-})^2 = w_2^2\sum_{j=1}^{N-1}(w_1^2)^{j-1} + (w_1^2)^{N-1} = (1-w_1)^2\frac{1-(w_1^2)^{N-1}}{1-w_1^2} + (w_1^2)^{N-1} = \frac{(1-w_1)^2 + 2w_1^{2N-1}(1-w_1)}{1-w_1^2}$ and $(\widetilde{\mathbf{w}}^{-})^T\,\mathbf{A_N}(d_i)\,\widetilde{\mathbf{w}}^{-} = (\widetilde{\mathbf{w}}^{-})^T\,(d_i\mathbf{H_N} + \lambda\mathbf{\Delta_N})\,\widetilde{\mathbf{w}}^{-}$ with $\mathbf{H_N} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$ and

$$\mathbf{\Delta_N} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & 1 & \cdots & \cdots & 1 \\ \vdots & \vdots & 2 & \cdots & 2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2 & \cdots & N-1 \end{bmatrix}.$$

We prove easily that $(\widetilde{\mathbf{w}}^{-})^T\,\mathbf{H_N}\,\widetilde{\mathbf{w}}^{-} = (\sum_{j=1}^{N}(\widetilde{w}_j^{-}))^2 = 1$ and if we decompose $\mathbf{\Delta_N}$ as follows:

$$\begin{aligned}\mathbf{\Delta_N} &= \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \vdots & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & 1 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 1 & \cdots & 1 \end{bmatrix} \\ &\quad + \cdots + \begin{bmatrix} 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix} \\ &= \mathbf{C_1} + \mathbf{C_2} + \ldots + \mathbf{C_{N-1}}\end{aligned}$$

we obtain that: $(\widetilde{\mathbf{w}}^{-})^T\,\mathbf{\Delta_N}\,\widetilde{\mathbf{w}}^{-} = \sum_{k=1}^{N-1}((\widetilde{\mathbf{w}}^{-})^T\,\mathbf{C_k}\,\widetilde{\mathbf{w}}^{-}) = \sum_{k=1}^{N-1}(\sum_{j=k+1}^{N}(\widetilde{w}_j^{-}))^2 = \sum_{k=2}^{N}(\sum_{j=k}^{N}(\widetilde{w}_j^{-}))^2$ where $\sum_{j=k}^{N}\widetilde{w}_j^{-} = \sum_{j=k}^{N-1} w_2 w_1^{k-1}\frac{1-w_1^{N-k}}{1-w_1} + w_1^{N-1} = w_1^{k-1}(1-w_1^{N-k}) + w_1^{N-1} = w_1^{k-1}$. Then we deduce that $(\widetilde{\mathbf{w}}^{-})^T\,\mathbf{\Delta_N}\,\widetilde{\mathbf{w}}^{-} = \sum_{k=2}^{N} w_1^{2(k-1)} = \frac{w_1^2 - w_1^{2N}}{1-w_1^2}$ which completes the proof of the expression 2.9. □

Proof of Theorem 2.3:

We search the optimal weights $\widehat{\mathbf{w}}^{-} = (\widehat{w}_1^{-},\ldots,\widehat{w}_N^{-})^T$ that minimize the variance of the estimator under the condition that the sum of weights is equal to 1, i.e. the following constraint optimization problem:

$$\begin{cases} \min\limits_{\mathbf{w}} \mathbf{w}^T\mathbf{\Sigma}^{-}\,\mathbf{w} \\ \text{subject to}\ \ \mathbf{w}^T\mathbf{b} = 1 \end{cases}$$

This is an optimization problem of a quadratic function with an equality constraint. The Lagrange multiplier method is used to solve this optimization problem. The Lagrangian function is $L(\mathbf{w},\lambda) = \mathbf{w}^T\mathbf{\Sigma}^{-}\,\mathbf{w} + \lambda(1 - \mathbf{w}^T\mathbf{b})$ where λ is the Lagrange multiplier. The first optimality condition is:

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \ \Leftrightarrow\ 2\mathbf{\Sigma}^{-}\mathbf{w} - \lambda\mathbf{b} = 0 \ \Leftrightarrow\ \mathbf{w} = \frac{1}{2}\lambda(\mathbf{\Sigma}^{-})^{-1}\mathbf{b} \tag{6.1}$$

Note that the covariance matrix $\mathbf{\Sigma}^{-}$ is symmetric positive definite and so invertible. Adding the con-

straint, we obtain:

$$\mathbf{w}^T\mathbf{b} = 1 \;\Leftrightarrow\; \mathbf{b}^T\mathbf{w} = 1 \;\Leftrightarrow\; \lambda\mathbf{b}^T(\mathbf{\Sigma}^-)^{-1}\mathbf{b} = 2$$
$$\Leftrightarrow\; \lambda = \frac{2}{c_{rt}}$$
$$\text{where}\;\; c_{rt} = \mathbf{b}^T(\mathbf{\Sigma}^-)^{-1}\mathbf{b} \;\; \text{is a constant.}$$

Finally, substituting in (6.1), we obtain the optimal weights $\widehat{\mathbf{w}}^- = \frac{1}{c_{rt}}(\mathbf{\Sigma}^-)^{-1}\mathbf{b}$. Then, since $Var[\widehat{x}^{opt}_{RT}(t_i)] = (\widehat{\mathbf{w}}^-)^T\,\mathbf{\Sigma}^-\widehat{\mathbf{w}}^-$, we deduce the expression of the variance, which completes the proof. □

Proof of Lemme 2.1:

The estimators $\widehat{x}_j^-$ and $\widehat{x}_j^+$ are independent, so $Cov(\widehat{x}_j^-(t_i), \widehat{x}_j^+(t_i)) = 0$, and then:

$$Var[\widehat{x}_{PP}(t_i)] = \sum_{j=1}^{N}\{(w_j^-)^2\,Var[\widehat{x}_j^-(t_i)] + (w_j^+)^2\,Var[\widehat{x}_j^+(t_i)]\}$$
$$+\,2\sum_{1\le j<j'\le N}\{w_j^-w_{j'}^-\,Cov(\widehat{x}_j^-(t_i), \widehat{x}_{j'}^-(t_i)) + w_j^+w_{j'}^+\,Cov(\widehat{x}_j^+(t_i), \widehat{x}_{j'}^+(t_i))\}$$

with $\forall j = 1,\ldots,N$ and $j < j'$, the expression of $Var[\widehat{x}_j^-(t_i)]$ and $Cov(\widehat{x}_j^-(t_i), \widehat{x}_{j'}^-(t_i))$ have been proven in theorem 2.2, $Var[\widehat{x}_j^+(t_i)] = \sigma^2_{gps} + (j\lambda - d_i)\,\sigma^2_{od}$ according to (2.15), and $Cov(\widehat{x}_j^+(t_i), \widehat{x}_{j'}^+(t_i)) = Var[\sum_{k=i+1}^{g_i^+(j)}\varepsilon_{od,k}] = (j\lambda - d_i)\,\sigma^2_{od}$ by independence of $\varepsilon_{od,i}$. □

Proof of Theorem 2.4:

The estimator $\widetilde{x}^-_{PP}(t_i)$ is equivalent to the estimator $\widehat{x}^{\infty}_{RT}(t_i)$ defined in definition 2.2. Thus, by symmetry, we deduce the expression of $\widetilde{x}^+_{PP}(t_i)$ and then that the estimator $\widehat{x}^{\infty}_{PP}(t_i)$ defined in (2.18) is an asymptotically minimum variance unbiased estimator of the position of the vehicle at the sampling time t_i. Since $\sum_{j=1}^N \widetilde{w}_j^- = \sum_{j=1}^N \widetilde{w}_j^+ = 1$, we normalize $\widehat{x}^{\infty}_{PP}(t_i)$ weighting by $(\widetilde{w}_1, \widetilde{w}_2)$ such that their sum is equal to one. The expression of weights $(\widetilde{w}_1, \widetilde{w}_2)$ are then deduce by the weighted least squares method. Finally, the expression of the variance of $\widetilde{x}^-_{PP}(t_i)$ is derived from the equivalence between this estimator and $\widehat{x}^{\infty}_{RT}(t_i)$. Thus, $Var[\widetilde{x}^-_{PP}(t_i)] = Var[\widehat{x}^{\infty}_{RT}(t_i)]$ whose expression is given in (2.9), and the expression of $Var[\widetilde{x}^+_{PP}(t_i)]$ is deduced using lemma 2.1: $Var[\widetilde{x}^+_{PP}(t_i)] = (\widetilde{w}^-)^t\,\Sigma^+\widetilde{w}^- = (\widetilde{w}^-)^t\,\sigma^2_{gps}(I_N + r\,A_N(\lambda - d_i))\,\widetilde{w}^-$ which completes the proof. □

References

[1] SHRP 2 Naturalisting Driving Study. http://www.trb.org/SHRP2, 2010.

[2] EuroFOT project. http://www.eurofot-ip.eu, 2009.

[3] A. Laureshyn. Automated video analysis and behavioural studies based on individual speed profiles. In *Proceedings of 18th ICTCT, Helsinki*, October 2005.

[4] H.M. Barbosa, M.R. Tight, and A.D. May. A model of speed profiles for traffic calmed roads. *Transportation Research Part A: Policy and Practice*, 34(2):103–123, 2000.

[5] A. Lahrech, C. Boucher, and J.C. Noyer. Fusion of gps and odometer measurements for map-based vehicle navigation. In *Industrial Technology, 2004. IEEE ICIT'04. 2004 IEEE International Conference on*, volume 2, pages 944–948. IEEE, 2004.

[6] A.N. Kealy, M. Tsakiri, and M. Stewart. Land vehicle navigation in the urban canyon-a kalman filter solution usingintegrated gps, glonass and dead reckoning. In *Proceedings of the 12th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1999)*, pages 509–518, 1999.

[7] L. Zhao, W.Y. Ochieng, M.A. Quddus, and R.B. Noland. An extended kalman filter algorithm for integrating gps and low cost dead reckoning system data for vehicle performance and emissions monitoring. *Journal of Navigation*, 56(2):257–275, 2003.

[8] K.W. Chiang and N. El-Sheimy. Ins/gps integration using neural networks for land vehicle navigation applications. In *Proceedings of the 15th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 2002)*, pages 535–544, 2001.

[9] D. Loebis, R. Sutton, J. Chudley, and W. Naeem. Adaptive tuning of a kalman filter via fuzzy logic for an intelligent auv navigation system. *Control engineering practice*, 12(12):1531–1539, 2004.

[10] C. Boucher, A. Lahrech, and J.C. Noyer. Nonlinear filtering for land vehicle navigation with gps outage. In *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, volume 2, pages 1321–1325. IEEE, 2004.

[11] A.N. Ramjattan and P.A. Cross. A kalman filter model for an integrated land vehicle navigation system. *Journal of Navigation*, 48(02):293–302, 1995.

[12] R. Da and G. Dedes. Nonlinear smoothing of dead reckoning data with gps measurements. In *Proceedings of the 8th International Technical Meeting of the Satellite Division of The Institute of Navigation (ION GPS 1995)*, pages 1285–1294, 1995.

[13] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[14] H.E. Rauch, C.T. Striebel, and F. Tung. Maximum likelihood estimates of linear dynamic systems. *Journal of the American Institute of Aeronautics and Astronautics*, 3(8):1445–1450, 1965.

[15] DQ Mayne. A solution of the smoothing problem for linear dynamic systems. *Automatica*, 4(2):73–92, 1966.

[16] D.C. Fraser. *A new technique for the optimal smoothing of data*. PhD thesis, Massachusetts Institute of Technology, 1967.

[17] J.E. Wall Jr, A.S. Willsky, and N.R. Sandell Jr. On the fixed-interval smoothing problem. *Stochastics: An International Journal of Probability and Stochastic Processes*, 5(1-2):1–41, 1981.

[18] M. A. Quddus, W.Y. Ochieng, and R.B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C*, 15:312–328, 2007.

[19] G. Taylor, C. Brunsdon, J. Li, A. Olden, D. Steup, and M. Winter. Gps accuracy estimation using map-matching techniques: Applied to vehicle positioning and odometer calibration. *Computers, Environments, and Urban Systems*, 30:757–772, 2006.

[20] M.S. Grewal, L.R. Weill, and A.P. Andrews. *Global Positioning Systems, Inertial Navigation, and Integration*. John Wiley & Sons, 2007.

[21] C.A. Scott and C.R. Drane. An optimal map-aided position estimator for tracking motor vehicles. In *IEEE Proceedings of the 6th International Conference on Vehicle Navigation and Information Systems. p. 360-367*, Jul. 1995.

[22] N. Viandier, A Rabaoui, J. Marais, and E. Duflos. Gnss pseudorange error density tracking using dirichlet process mixture. In *Proceedings of 13th International Conference on Information Fusion, EICC Edinburgh, UK*, July 2010.

[23] J. Borenstein and L. Feng. Measurement and correction of systematic odometry errors in mobile robots. In *IEEE Transactions on Robotics and Automation*, volume 12, pages 869–880, 1996.

[24] C.K. Chui and G. Chen. *Kalman filtering with real-time applications*. Springer series in information sciences. Springer-Verlag, 1987.

[25] M.S. Grewal and A.P. Andrews. *Kalman Filtering: Theory and Practice Using MATLAB*. Wiley, 2008.

List of Figures

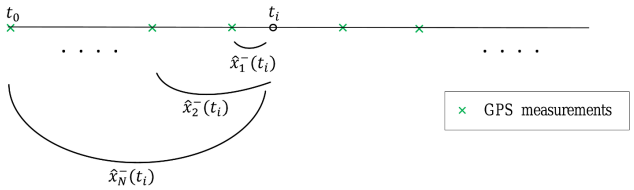


Figure 1: Graph of estimators $\widehat{x_j^-},\ j = 1, \ldots, N$.

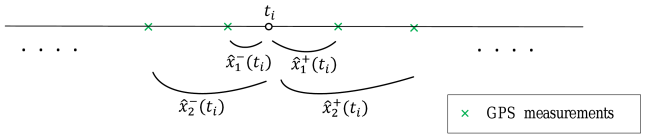


Figure 2: Graph of estimators $\widehat{x}_j^-$ and $\widehat{x}_j^+$, $j = 1, \ldots, N$.

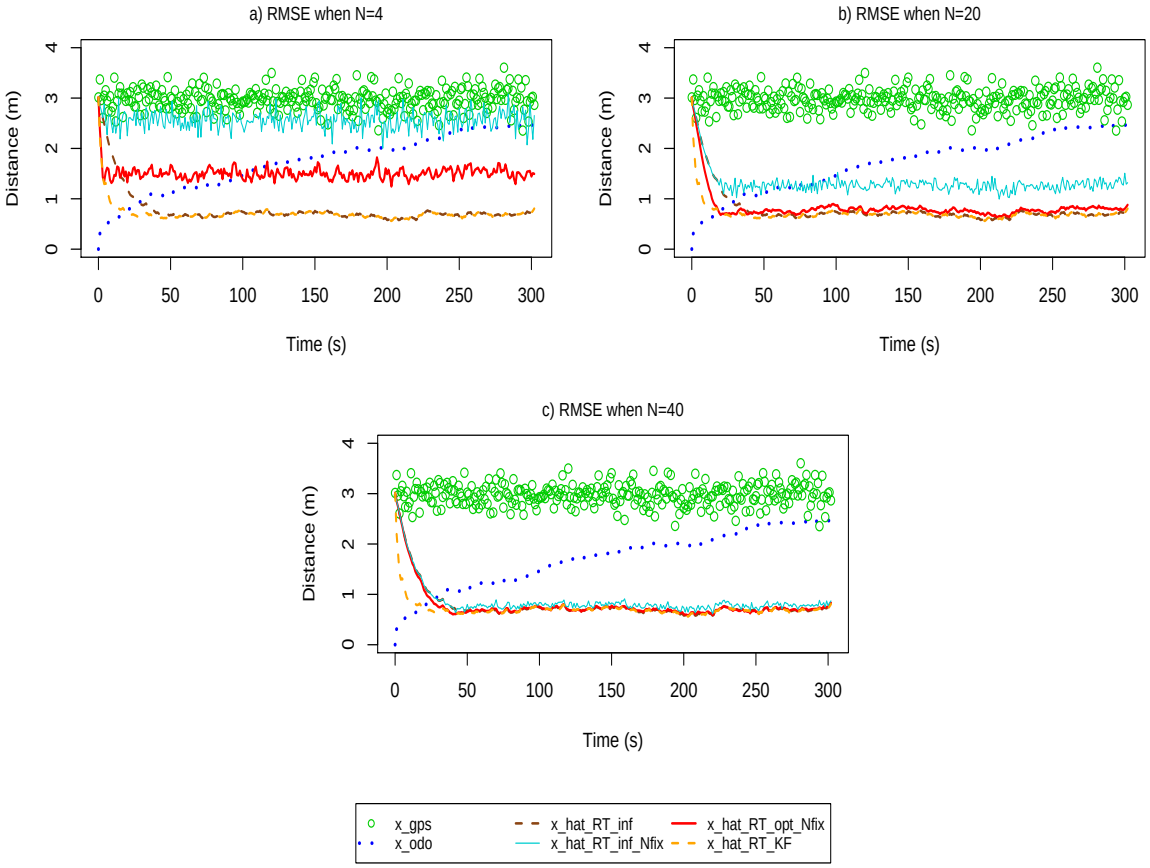


Figure 3: RMSE of real-time estimators ($\widehat{x}_{RT}^{\infty}$ and $\widehat{x}_{RT}^{opt}$) and Kalman filter.

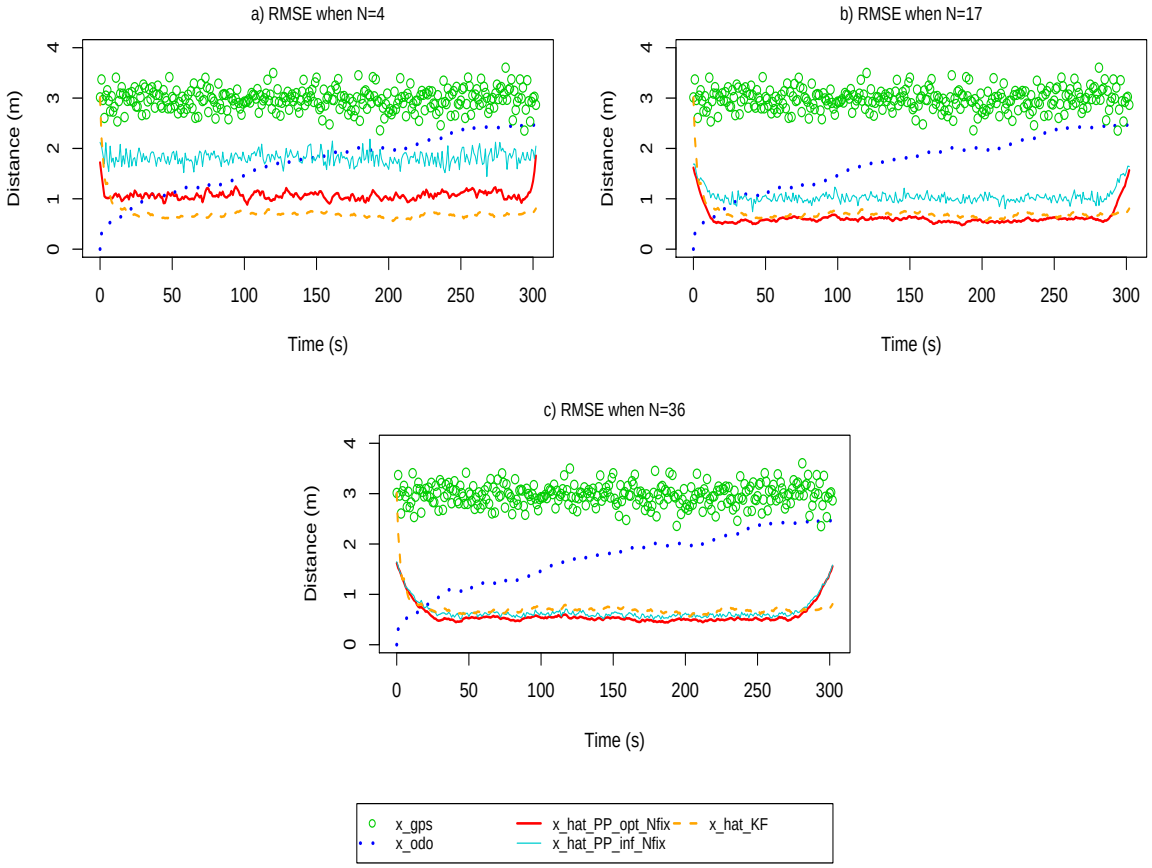


Figure 4: RMSE of real-time estimators ($\widehat{x}_{PP}^{\infty}$ and $\widehat{x}_{PP}^{opt}$) and Kalman filter.

List of Tables

	Mean RMSE (m)	Max RMSE (m)
Odometer	1.72	2.55
GPS	2.97	3.60
$\widehat{x}_{RT}^{\infty}$	0.79	3.01
$\widehat{x}_{RT}^{\infty}$ with N fixed (N=4)	2.57	3.08
$\widehat{x}_{RT}^{\infty}$ with N fixed (N=20)	1.31	3.01
$\widehat{x}_{RT}^{\infty}$ with N fixed (N=40)	0.86	3.01
$\widehat{x}_{RT}^{opt}$ with N fixed (N=4)	1.51	3.01
$\widehat{x}_{RT}^{opt}$ with N fixed (N=20)	0.83	3.01
$\widehat{x}_{RT}^{opt}$ with N fixed (N=40)	0.78	3.01
$\widehat{x}_{RT}^{KF}$	0.71	3.01

Table 1: Mean and maximum RMSE of the real-time estimators compared to the Kalman filter.

	Mean RMSE (m)	Max RMSE (m)
Odometer	1.72	2.55
GPS	2.97	3.60
$\widehat{x}_{PP}^{opt}$ with N fixed (N=4)	1.07	1.86
$\widehat{x}_{PP}^{opt}$ with N fixed (N=17)	0.62	1.62
$\widehat{x}_{PP}^{opt}$ with N fixed (N=36)	0.59	1.62
$\widehat{x}_{PP}^{\infty}$ with N fixed (N=4)	1.82	2.18
$\widehat{x}_{PP}^{\infty}$ with N fixed (N=17)	1.04	1.69
$\widehat{x}_{PP}^{\infty}$ with N fixed (N=36)	0.66	1.65
$\widehat{x}_{RT}^{KF}$	0.71	3.01

Table 2: Mean and maximum RMSE of the post-processing estimators compared to the Kalman filter.

a) Real-time estimators

	RMSE (m)	Computing time (s)
Odometer	20.59	-
GPS	3.07	-
$\widehat{x}_{RT}^{KF}$	4.37	0.06
$\widehat{x}_{RT}^{\infty}$	4.50	0.03
$\widehat{x}_{RT}^{\infty}$ (N=4)	3.21	0.26
$\widehat{x}_{RT}^{\infty}$ (N=20)	3.91	1.01
$\widehat{x}_{RT}^{\infty}$ (N=40)	3.93	1.94
$\widehat{x}_{RT}^{opt}$ (N=4)	2.59	0.28
$\widehat{x}_{RT}^{opt}$ (N=20)	3.08	1.10
$\widehat{x}_{RT}^{opt}$ (N=40)	3.52	2.09

b) Post-processing estimators

	RMSE (m)	Computing time (s)
Odometer	20.59	-
GPS	3.07	-
$\widehat{x}_{PP}^{opt}$ (N=4)	2.28	0.54
$\widehat{x}_{PP}^{opt}$ (N=17)	1.54	1.94
$\widehat{x}_{PP}^{opt}$ (N=36)	1.54	3.97
$\widehat{x}_{PP}^{\infty}$ (N=4)	2.42	0.49
$\widehat{x}_{PP}^{\infty}$ (N=17)	1.67	1.78
$\widehat{x}_{PP}^{\infty}$ (N=36)	2.08	3.72

Table 3: RMSE and computing time of the real-time and post-processing estimators compared to the Kalman filter.

Annexe C

Caractérisation de l'éco-conduite et travaux associés

C.1 L'éco-conduite : définition et règles de base

Si la gestion de la vitesse est une préoccupation majeure des autorités publiques en terme de sécurité, la prise de conscience des problèmes liés à l'augmentation des émissions de gaz à effet de serre a conduit à prendre également en compte les aspects d'économies d'énergie. La recherche d'une conduite optimale en termes à la fois de sécurité, d'efficacité économique (ex : temps de trajet) et d'économies d'énergie, est ainsi devenue un enjeu majeur de notre mobilité moderne. L'éco-conduite est connu pour être un moyen rapide et efficace de réduire la consommation de carburant, cette baisse pouvant aller jusqu'à 20 %. Elle peut être définie comme la recherche, à tout moment, de la meilleure conduite (en termes de consommation d'énergie, de sécurité, de temps, de confort...) lors de différentes tâches de conduite (navigation, choix des manoeuvres, contrôle) (voir figure C.1).

Figure C.1 - Définition de l'éco-conduite.

Le concept d'éco-conduite répond à la mise en oeuvre de certains principes de base (http://www.ecodrive.org) tels que :

1. Passer à la vitesse supérieure dès que possible.
   Passer à la vitesse supérieure entre 2000 et 2500 tr/min.
2. Maintenir une allure constante.
   Enclencher la plus haute vitesse possible et conduire avec un régime moteur faible.
3. Anticiper le trafic.
   Regarder le plus loin possible et anticiper le trafic environnant.
4. Décélérer progressivement.
   S'il faut ralentir ou s'arrêter, décélérer progressivement en relâchant l'accélérateur à temps et en laissant la voiture en prise.

C.2 Travaux de recherche : caractérisation de l'éco-conduite et conception d'un eco-index

Plusieurs travaux de recherche sur le thème de l'éco-conduite ont été réalisés au LIVIC. L'objet de cette section concerne les travaux auxquels j'ai contribué en parallèle de ma thèse et qui s'inscrivent dans la continuité de mon stage de master 2 recherche effectué au LIVIC sur cette thématique.

L'objet de la première étude est une analyse statistique de deux modes d'apprentissage de l'éco-conduite : le premier en suivant de simples conseils sur l'éco-conduite tels que les règles de base énoncées à la section précédente, et le second après avoir suivi une formation à l'éco-conduite avec un moniteur professionnel. L'effet de chacun de ces deux modes d'apprentissage dans la mise en oeuvre des consignes d'éco-conduite a été étudié à partir d'une expérimentation de type "Naturalistic driving" (conduite en situation naturelle) en associant chacune de ces consignes à un indicateur quantitatif, puis en modélisant la probabilité d'être en situation d'éco-conduite à l'aide de différents modèles de régression logistique. Les détails de cette étude sont donnés dans l'article suivant joint dans cette annexe :

C. Andrieu and G. Saint Pierre (2012), Comparing effects of ecodriving training and simple advices on driving behavior, *Proceedings of the 15th Euro Working Group on Transportation (EWGT 2012)*, Paris, France, septembre 2012.

L'objet de la seconde étude s'inscrit dans la continuité de la première et consiste à utiliser la méthodologie développée pour construire un indicateur global d'éco-conduite, appelé eco-index, tenant compte de divers paramètres liés au comportement du conducteur. Un modèle de régression logistique a ainsi été élaboré avec pour variables explicatives les quatre indicateurs de performance associés à chacune des principales règles d'éco-conduite citées à la section précédente. Les détails de

cette étude sont donnés dans l'article suivant joint également dans cette annexe :

C. Andrieu and G. Saint Pierre (2012), Using statistical models to characterize eco-driving style with an aggregated indicator, *Proceedings of the IEEE Intelligent Vehicle Symposium (IV 2012)*, Alcala de Henares, Espagne, juin 2012.

Un tel indicateur agrégé peut être utilisé pour concevoir un système d'aide à l'éco-conduite (EDAS, "*Ecological Driving Assistance System*") puisqu'il permet de détecter une conduite souple ou consommatrice, et d'afficher des informations ciblées destinées à permettre au conducteur de corriger et d'évaluer l'efficacité énergétique de sa conduite. Un système d'aide à l'éco-conduite pour smartphone basé sur cet eco-index est notamment en cours de développement dans le cadre du projet européen "ecoDriver" (http://www.ecodriver-project.eu).

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 54 (2012) 211 – 220

Procedia Social and Behavioral Sciences

EWGT 2012

15th meeting of the EURO Working Group on Transportation

Comparing effects of eco-driving training and simple advices on driving behavior

Cindie Andrieu, Guillaume Saint Pierre*

IFSTTAR, IM, LIVIC, 14, route de la minière, 78000 Versailles-Satory, France

Abstract

Eco-driving style is widely known to induce up to 20% fuel consumption reduction, but little is known on the effects of different learning methods. In order to evaluate the potential impacts of future ecological driving assistance system (EDAS), two kinds of experiments are analyzed in this paper: In the first one, simple advices are given to the participants, while in the second one, full courses with eco-driving experts were used. Different kind of statistical models are discussed, among which we choose to apply the ordinary logistic regression to assess the effects of each driving advice separately.

Keywords: eco-driving; logistic regression; driving behavior.

1. Introduction

Driving more efficiently is part of the solution to reduce the surface transportation greenhouse gas emissions but it is a highly complex task, comprising over hundreds of separate tasks (Walker et al., 2001). Drivers need to simultaneously control the vehicle, adjust their speed and trajectory according to driving environment, deal with hazards, and make strategic decisions such as navigation to progress toward their goal (Young et al., 2010). Since climate change and humanity responsibility has been widely accepted, many drivers have a new goal in mind: fuel efficiency. Eco-driving style is therefore often referred as smart driving because of the necessary complex

* Corresponding author. Tel.: +33 (0) 1 40 43 29 33.
Email address: guillaume.saintpierre@ifsttar.fr

1877-0428 © 2012 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of the Program Committee
doi:10.1016/j.sbspro.2012.09.740

trade off between the multiple goals the driver has to manage with. Studies usually simplifies the green way to drive using simple advices easily understood by drivers (CIECA, 2007), but sometimes leading to a misunderstanding of the fuel efficient driving strategy. Other studies used trial experiments before and after a training program to assess the eco-driving impact (Symmons et al., 2009). Effects of eco-driving on fuel consumption are well described in the literature, but results are often optimistic: CO2 emissions reduction can be up to 30% according to many studies. The key question for policy makers is "how big" of an emission reduction we can get by encouraging an eco-driving style, taking into account the diversity in the way to learn eco-driving: just reading a few driving tips, taking a course with a professional, or doing practical exercises with equipped vehicles?… Moreover, there is a need to understand the best way to teach and learn eco-driving style, especially for young drivers.

This work present the statistical analysis of two different data sets, one with subjects following simple eco-driving advices, the other with subjects driving the way they learned in a course with professional eco-drivers. For the analysis needs, eco-driving style is summarized into four different simple advices, each one of them being associated to a quantitative indicator build to reflect the associated driving behavior. Different kind of statistical models are discussed, among which we choose to apply the ordinary logistic regression to assess the effects of each driving advice separately. The significance of the differences for each indicator between normal and eco-driving trips allow us to evaluate which advice is practically used by the drivers, according to the way they learned eco-driving. The same analysis is done for each different speed limit zone to take into account the effects of the driving environment.

2. The experiments

2.1. Experiment 1: simple advices on eco-driving

The experiment goal was to clearly identify two classes of driving behavior on the same test track: "normal" and fuel efficient way to drive commonly known as "eco-driving". Twenty drivers participated in this experiment that took place in June and July 2009 in Ponchartrain (Yvelines) in France. Four of these drivers were eco-driving instructors while others were recruited among one thousand persons working in two different research institutes. In order to minimize traffic influence, the chosen route is of inter-urban type and a length of 14km. The trips were all performed under free flow conditions and with dry weather. The vehicle used was a petrol-driven Renault Clio III with manual gearshift. First of all, the journey is discovered by the subjects while seeing the experimenter driving and giving safety and direction instructions. Then, the trip was driven twice by each driver: once while driving normally, and secondly while following the "Golden Rules" of eco-driving extracted from the Ecodrive project (Ecodrive, 2009) and summarized in Table 1. These rules were given just before the ecological trip. To eliminate a learning effect of the journey, trip's order has been counter-balanced. An on-board logging device was used to monitor key driving parameters. The device is connected to the controller area network (CAN) of the vehicle, logging most of the relevant parameters related to engine state, vehicle dynamic, and driver actions on pedals. The vehicle has been also equipped with a GPS, a camera in front of the vehicle and a fuel flow meter. We used a fuel flow meter DFL1x-5bar to validate the fuel consumption logged with the CAN. Additional variables were post-processed such speed limits, gear ratio, and many indicators inspired from Ericsson (2001).

2.2. Experiment 2: eco-driving training

Nineteen drivers (who have not participated in the experiment 1) participated in this experiment that took

place near Toulouse in 2004. The trials goal was to evaluate the effect of an embedded EDAS produced by the GERICO project funded by the French program of research, experimentation and innovation in land transport (Barbé et al., 2008). The original design was to compare a control group, a group applying eco-driving, and another group using the system without any advices. For the purpose of this study, only data for the eco-driving group was used. The chosen route contains various network categories (urban, rural, motorway) and has a length of 70km. The vehicle used was a Renault Megane Scenic with a four-speed sequential gearbox. The trip was driven twice by each driver: once while driving normally and secondly after an eco-driving training with professional eco-drivers. In this case, trips are not counter balanced and effects of the eco-driving teaching may be over estimated because of a learning effect.

3. Methodology

3.1. Selection of indicators associated with each of the main rules of eco-driving

Driving style to reduce fuel consumption is related to the implementation of the four main eco-driving rules set out in Table 1. Due to this link, each of these instructions was associated with an indicator. The proposed indicators are summarized in Table 1. So the first rule state to shift up early. Therefore, it is natural to associate the indicator *AvgRPMShiftUp* which is the average engine speed (in rpm) at the shift into a higher gear. The second rule is related both to the gear and the engine speed.

Table 1. Main rules of eco-driving and indicators associated

Instruction	Indicator	Abbreviation
1. Shift up as soon as possible: Shift up between 2.000 and 2.500 revolutions per minute.	Average engine speed at the shift into a higher gear.	Avg_RPM_Shift_Up
2. Maintain a steady speed: Use the highest gear possible and drive with low engine RPM.	Index of gear ratio distribution and engine speed associated.	Index_Gear_RPM
3. Anticipate traffic flow: Look ahead as far as possible and anticipate the surrounding traffic.	Positive Kinetic Energy.	PKE
4. Decelerate Smoothly: When you have to slow down or to stop, decelerate smoothly by releasing the accelerator in time, leaving the car in gear.	Percentage of time in engine brake.	Time_Engine_Brake

So we created an indicator, called *IndexGearRPM*, summarizing these two variables and calculated as follows:

$$IndexGearRPM = \frac{1}{3500}(TimeNeut \times AvgRPMNeut + Gear1 \times AvgRPMGear1 + \cdots + Gear5 \times AvgRPMGear5) \quad (1)$$

where *Time_Neutral* is the percentage time in neutral gear, *AvgRPMNeut* is the average engine speed in neutral gear, *Gear1* is the percentage time in gear 1 (with pressing the accelerator pedal), etc. Note that the condition of pressing the accelerator pedal ensure to ignore the time in engine brake which is associated to the fourth rule. Note also that the division by 3500, representing the maximum engine speed, is just a normalization factor. Then the third rule related to the anticipation of traffic is associated to the parameter PKE (Positive Kinetic Energy) calculated as follows:

$$PKE = \frac{\sum (v_f^{\,2} - v_i^{\,2})}{x} \quad \text{when} \quad \frac{dv}{dt} > 0 \quad (2)$$

where v_f and v_i are respectively the final and the initial speed (in m/s) at each time interval for which $dv/dt > 0$, and x is the total distance travelled (in m). This indicator represents the ability to keep the vehicle's kinetic energy as low as possible. So a nervous driving will be associated with a high PKE, and conversely a smoothly driving will be associated with a PKE close to zero. Finally, the fourth rule is naturally associated with the percentage of time in engine brake characterized by the following conditions: non zero speed, no neutral, no pressure the brake pedal and the accelerator pedal.

3.2. Statistical models

The objective of this study is to compare effects of simple advices (experiment 1) and eco-driving training (experiment 2) on driving behavior. Our approach relies on developing a predictive model of economic driving behavior based on easily interpretable variables. Assuming trips are clustered according to the two driving conditions, it is worth trying a statistically based approach to predict the driving style.

Such models are well suited in estimating the relationship between an outcome variable and a set of explanatory variables. In this paper, the outcome variable is from a binary distribution with two possible values:

$$Y_{ij} = \begin{cases} 1 \; if \; ecodriving \\ 0 \; if \; not \end{cases} \quad i = 1, \dots, I \; ; \; j = 1, \dots, T_i \tag{3}$$

where I is the number of drivers and T_i is the number of observations for the driver *i*. Logistic regression is a form of statistical modeling that is often appropriate for binary outcome variables. Assume Y_{ij} follows a Bernoulli distribution with parameter $p_{ij} = P(Y_{ij} = 1)$ where p_{ij} represent the probability that the event occurred for the observation Y_{ij}. The relationship between the event probability p_{ij} and the set of factors is modeled through a logit link function with the following form:

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = X'_{ij}\beta \tag{4}$$

where X_{ij} is the vector of explanatory variables and β is the vector of regression parameters (Agresti, 2002). The ordinary logistic regression assumes independent observation and the vector β is estimated by the method of maximum likelihood. However, the assumption of data independence does not suit our data very well, as it will contain unavoidable driver-specific correlations (i.e. observations from the same driver are assumed to be correlated) that should be treated as random effects. The standard errors from the ordinary logistic regression are then biased because the independence assumption is violated.

To account for these driver specific correlations as random effects, more sophisticated statistical models need to be applied. These models are particularly useful for naturalistic driving study (Guo and Hankey, 2010; Benminoun et al., 2011) and specially event based approach (EBA) which basic principle is to identify time segments that can be predictive of an event (e.g. crash, near-crash, …). Indeed, these models include additional parameters to deal with correlations, and confounding factors are viewed as explicative variables that can be used to predict event probability. One such model is the "Generalized Estimated Equations" (GEE) model or marginal models, originally developed to model longitudinal data by Liang and Zeger (1986), which assumes that observations are marginally correlated. Another approach for modeling correlated data is "Generalized Linear Mixed Models" (GLMM). The GLMM model introduces a random effect specific to each subject whereas the GEE approach models the marginal distributions by treating correlation as a nuisance parameter. Therefore the inference is individual (subject-specific approach) in contrast to marginal models that model the average population (population-averaged approach). However, in our study, we didn't use these two sophisticated

statistical models because of the small sample size (see Section 4.2 for more details). So we used only ordinary logistic regression models.

4. Results

4.1. Overall effects of eco-driving rules and eco-driving training

Numeric results are summarized in Table 2. A paired *t*-test was performed to assess whether the mean of each parameter differ significantly according to the driving style. Table 2 indicates the p-values of these tests. Among the most interesting ones, the average fuel consumption across drivers decreased by 12.5% between normal driving and eco-driving for the experiment 1 and decreased by 11.3% for the experiment 2. These similar results between the two experiment show that it seems quite simple to reduce fuel consumption by applying some basic rules of eco-driving. The average speed decreased by 5.8% for the experiment 1 and 10.1% for the experiment 2, and the percentage of time beyond the legal speed limit decreased by 30.1% for the experiment 1 and 36.1% for the experiment 2.

Table 2. Effects of eco-driving rules on different parameters

Parameter	Description	Experiment 1			Experiment 2		
		Mean "Normal"	Mean "Eco"	Variation (%)	Mean "Normal"	Mean "Eco"	Variation (%)
AvgFuelConsum	Average fuel consumption (l/100km).	6.86	6.00	-12.5^{***}	9.01	7.99	-11.3^{***}
AvgRPMShiftUp	Average engine speed at the shift into a higher gear (associated with rule 1).	2737.5	2232.8	-18.4^{***}	3177.3	2465.6	-22.4^{***}
IndexGearRPM	Index of gear ratio distribution and engine speed associated (associated with rule 2).	61.0	52.9	-13.3^{***}	70.8	60	-15.3^{***}
PKE	Positive Kinetic Energy (associated with rule 3).	0.343	0.243	-29.2^{***}	0.293	0.197	-32.8^{***}
TimeEngineBrake	Percentage of time in engine brake (associated with rule 4).	20.3	26.3	$+29.6^{**}$	16.2	16.8	+ 0.04
AvgSpeed	Average speed (km/h)	50.85	47.89	-5.8^{**}	61.45	55.22	-10.1^{***}
AvgAccel	Average acceleration (ms^{-2})	0.498	0.387	-22.3^{***}	0.596	0.473	-20.6^{***}
AvgDecel	Average deceleration (ms^{-2})	-0.619	-0.523	-15.5^{***}	-0.672	-0.599	-10.9^{***}
AvgRPM	Average engine speed (rpm)	2097.4	1835.5	-12.5^{***}	2379.6	2009.6	-15.5^{***}
TimeNonLegalSpeed	Percentage of time beyond the legal speed limit	37.9	26.5	-30.1^{***}	28.5	18.2	-36.1^{***}

$^{*}\ p < 0.05;\ ^{**}\ p < 0.01;\ ^{***}\ p < 0.001.$

These reductions reflect a better compliance with speed limits with economical driving regardless of the learning mode. As regards the application of eco-driving rules, the four associated indicators are significantly different among the two driving conditions, indicating that the instructions were applied with the two learning mode. However, in the experiment 2, the engine brake (associated with the fourth rule of eco-driving) does not seem to have been used correctly. Furthermore, the average acceleration and deceleration both decrease significantly in the two experiments which is in agreement with the second and the third rules of eco-driving.

4.2. Separated effects of the main eco-driving rules

The aim of this study is to assess the effects of each driving advice after two learning mode: one with subjects following simple eco-driving advices (experiment 1), and the other with eco-driving training (experiment 2). Our approach is to construct, for each experiment, a predictive model of the probability of being in an eco-driving situation using a binomial logistic regression model with the four indicators in Table 1 as explanatory variables. According to our experiment, we predict the binary variable named "Trip" which takes the value 0 in normal driving (noted "normal") and 1 in eco-driving (noted "eco"). Thus, the significance of the differences of each indicator between normal and eco-driving trips allow us to evaluate which advice is practically used by the drivers, according to the way they learned eco-driving.

However, in our two experiments, both the number of clusters (20 in the experiment 1 and 19 in the experiment 2) and the cluster size (2 in the two experiments) are small, which implies various constraints. In a first part, the smallness of our sample size limits the number of predictors for which effects can be estimated precisely. Peduzzi et al. (1996) suggests there should ideally be at least ten outcomes of each type for every predictor. This result constrains us to assess the effects of each driving advice separately and consequently to construct one logistic regression model with each of the four indicators as predictor. In a second part, the smallness of our sample size does not allow us to use the appropriate statistical models taking into account driver specific correlations. Indeed, we tested the GEE method using the PROC GENMOD of the SAS software, but the parameters estimates were closed to zero. Ziegler et al. (1998) recommend an application of the GEE only, if the number of clusters is at least 30 for a cluster size of about 4 for a low to moderate correlation. We also tested the generalized linear mixed models using the PROC GLIMMIX of SAS but a statement indicates that one of the estimated variance parameters was negative. This result is an underestimate of the true variance component that occurs when the number of observations per random effect category is small or when the ratio of the true variance component to the residual is small. Moreover, several studies (Moineddin (2007), Theall (2011)) have shown that parameters estimates are unbiased with either fixed or random effects logistic models when the number of clusters and the cluster size are small. However these studies show that the estimates of the random intercept and random slope have larger biases compared to the fixed effect parameters. Thus, later in this paper, we use an ordinary logistic regression.

The logistic model can be written as:

$$Logit[P(Trip = Eco)] = \alpha + \beta X \quad (5)$$

where α is the intercept, X is one of the four indicators associated with the main rules of eco-driving (Table 1) and β is the parameter estimate of the predictor X. The results from each logistic model are listed in Table 3 for the experiment 1 and Table 4 for the experiment 2. For each logistic model, we indicate the explanatory variable X, the estimated parameter β, its standard error SE and the p-value of the Wald test. We also indicate the odds ratios (OR) and their 95% Wald confidence limits. The usefulness of each model is measure by the Nagelkerke R^2, denoted R_N^2, which is an adjusted version of the Cox & Snell R^2 and which is similar to the coefficient of determination R^2 in linear regression. This parameter does not measure the goodness of fit of the model but indicate how useful the explanatory variable is in predicting the response variable. Finally, the predictive power of each model is measure by the area under the ROC curve (AUC). This parameter, ranges from zero to one and identical to the concordance index, assess the discrimination power of the model. In our study, it measures the model's ability to discriminate between eco-driving trips versus normal trips. More details on these various parameters are given in Agresti (2002) or Hosmer and Lemeshow (2000).

In Table 3 and Table 4, the four logistic models, assessing the implementation of each rules of eco-driving, are ranked in descending order of both parameters R_N^2 and AUC and thus represents the order of implementation of each driving advice. Table 3 shows that all the indicators are significant (p-value lower than 0.01 and 95% confidence interval including one) in the experiment 1 but the indicators associated with the first three rules are most significant: relatively high R_N^2 reflecting that the three indicators *AvgRPMShiftUp*, *IndexGearRPM* and *PKE* are useful in predicting eco-driving trip, and AUC greater than 0.8 reflecting a high discriminatory power

of this three models. On the contrary, the indicator *TimeEngineBrake* is not very useful in predicting eco-driving trip (R_N^2=0.289) even if the discriminatory power of this model is acceptable (0.7 ≤ AUC ≤ 0.8). Table 4 shows the results obtained in the experiment 2. The results are globally similar to those obtained in the experiment 1 except that the indicator *TimeEngineBrake* is no longer significant (one is excluded of the 95% confidence interval) and the model associated is not very useful in predicting eco-driving behavior (R_N^2 close to zero and AUC close to 0.5 indicating poor discrimination of the model).

Table 3. Experiment 1: logistic regression models with each of the four indicators associated with the main rules of eco-driving and ranked in descending order of implementation of each driving advice.

Models	β	SE	OR	95% CI	R_N^2	AUC
X= AvgRPMShiftUp (Rule 1)	-0.0068^{**}	0.002	0.993	0.989 - 0.997	0.608	0.908
X=PKE (Rule 3)	-34.0893^{**}	10.622	< 0.001	< 0.001 - < 0.001	0.594	0.898
X= IndexGearRPM (Rule 2)	-0.3068^{**}	0.103	0.736	0.601 - 0.901	0.491	0.866
X= TimeEngineBrake (Rule 4)	0.1849^{**}	0.071	1.203	1.047 - 1.383	0.289	0.780

$^{*}\ p < 0.05;\ ^{**}\ p < 0.01;\ ^{***}\ p < 0.001.$

SE: standard error; OR: odds ratio; CI: confidence interval; R_N^2: Nagelkerke R^2; AUC: area under the ROC curve.

Table 4. Experiment 2: logistic regression models with each of the four indicators associated with the main rules of eco-driving and ranked in descending order of implementation of each driving advice.

Models	β	SE	OR	95% CI	R_N^2	AUC
X= IndexGearRPM (Rule 2)	-1.4262^{*}	0.677	0.240	0.064 - 0.906	0.922	0.989
X= AvgRPMShiftUp (Rule 1)	-0.0126^{**}	0.004	0.987	0.979 - 0.996	0.878	0.976
X=PKE (Rule 3)	-63.7126^{**}	21.715	< 0.001	< 0.001 - < 0.001	0.744	0.952
X= TimeEngineBrake (Rule 4)	0.0254	0.065	1.026	0.902 - 1.166	0.005	0.568

$^{*}\ p < 0.05;\ ^{**}\ p < 0.01;\ ^{***}\ p < 0.001.$

SE: standard error; OR: odds ratio; CI: confidence interval; R_N^2: Nagelkerke R^2; AUC: area under the ROC curve.

4.3. Eco-driving effects for different speed limits

Assuming that eco-driving behavior depends on the road conditions, previous logistic models were extended to more complex models taking into account the speed limits. The variable "Speed limit" is used as a stratification variable in order to derive specific models. Thus, for each trip of the two experiments, sections corresponding to a specific speed limit were merged for analysis. The calculation of the four indicators defined in Table 1 was then adapted on these new trip to take into account the grouping of sections not necessarily continuous.

Table 5,

Table 6 and Table 7 contain the estimated parameter, its standard error, the Nagelkerke R^2 and the AUC for the three main speed limits: 50km/h, 70km/h and 90km/h.

Table 5 shows similar results for the two experiments when the speed limit is 50km/h: the three indicators *AvgRPMShiftUp*, *IndexGearRPM* and *PKE* are most significant while the indicator *TimeEngineBrake* is not very useful in predicting eco-driving behavior.

Table 6, corresponding to the speed limit 70km/h, shows that in the experiment 1, the four driving advices have been applied while in the experiment 2, only the first three advices have been applied. Finally, Table 7 shows that when the speed limit is 90km/h, the indicators *AvgRPMShiftUp* and *IndexGearRPM* are most significant in the two experiments whereas the indicator PKE is less significant than with the previous speed limitations. As for areas limited to 50km/h, the indicator *TimeEngineBrake* is not useful in predicting eco-driving behavior and in the experiment 1, the estimated parameter is negative (but no significant) which means that engine brake seems to have been less used during eco-driving trips than during normal trips.

Table 5. Logistic regression models for 50km/h speed limit.

Models	Experiment 1				Experiment 2			
	β	SE	R_N^2	AUC	β	SE	R_N^2	AUC
X= AvgRPMShiftUp (Rule 1)	-0.007^{***}	0.002	0.62	0.909	-0.012^{**}	0.004	0.83	0.964
X= IndexGearRPM (Rule 2)	-0.371^{**}	0.124	0.56	0.903	-0.908^{*}	0.362	0.85	0.978
X=PKE (Rule 3)	-36.022^{**}	11.969	0.59	0.896	-34.859^{**}	11.301	0.64	0.922
X= TimeEngineBrake (Rule 4)	0.108^{*}	0.045	0.24	0.745	0.074	0.074	0.07	0.676

$^{*}\ p < 0.05;\ ^{**}\ p < 0.01;\ ^{***}\ p < 0.001.$

SE: standard error; R_N^2: Nagelkerke R^2; AUC: area under the ROC curve.

Table 6. Logistic regression models for 70km/h speed limit.

Models	Experiment 1				Experiment 2			
	β	SE	R_N^2	AUC	β	SE	R_N^2	AUC
X= AvgRPMShiftUp (Rule 1)	-0.005^{**}	0.002	0.48	0.871	-0.013^{**}	0.005	0.88	0.986
X= IndexGearRPM	-0.293^{**}	0.105	0.43	0.851	-0.459^{***}	0.139	0.76	0.938

(Rule 2)								
X=PKE	-25.350^{***}	7.705	0.46	0.863	-31.830^{**}	10.967	0.61	0.922
(Rule 3)								
X= TimeEngineBrake	0.178^{**}	0.063	0.35	0.795	0.034	0.050	0.02	0.562
(Rule 4)								

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

SE: standard error; R_N^2: Nagelkerke R^2; AUC: area under the ROC curve.

Table 7. Logistic regression models for 90km/h speed limit.

Models	Experiment 1				Experiment 2			
	β	SE	R_N^2	AUC	β	SE	R_N^2	AUC
X= AvgRPMShiftUp	-0.005^{**}	0.002	0.47	0.868	-0.019^{*}	0.009	0.90	0.989
(Rule 1)								
X= IndexGearRPM	-0.225^{**}	0.077	0.43	0.850	-0.494^{**}	0.159	0.78	0.956
(Rule 2)								
X=PKE	-14.054^{*}	5.463	0.27	0.745	-21.121^{*}	8.280	0.33	0.758
(Rule 3)								
X= TimeEngineBrake	-0.015	0.080	0.001	0.521	0.038	0.051	0.02	0.651
(Rule 4)								

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

SE: standard error; R_N^2: Nagelkerke R^2; AUC: area under the ROC curve.

5. Conclusion

This study provides the statistical analyses of two learning mode of eco-driving: one with simple eco-driving advices and the other with eco-driving training. The study of different parameters like average fuel consumption, average speed, or average acceleration shows a real positive impact of eco-driving style regardless of the learning mode.

The association of each of the main eco-driving rules with a quantitative indicator allows us to assess the effect of each driving advice separately using logistic regression models. It is shown that drivers succeed efficiently in applying advices related to constant speed or gearshift strategy regardless of the learning mode of eco-driving, while they are less efficient in using engine brake (small parameter influence for experiment 1 and insignificant for experiment 2). The same analysis is done for each different speed limit zone in order to take into account the effects of the driving environment. Results are all together in line although significant differences are found for the engine brake related rule. On 70km/h limited areas the engine brake was not correctly used in experiment 2 (with eco-driving training) while all the four driving advices were correctly implemented in experiment 1 (with simple advices). On the contrary, on 90km/h limited areas, the 4th rule effect is insignificant for both experiments although the engine brake seems to have been less used during eco-driving trips than during normal trips.

Golden rules indicators show that fuel efficient driving is better implemented after a course than just applying

eco-driving tips (greater R_N^2 and AUC). Differences are small due to the bias introduced by the presence of an experimenter in the car in both experiment. Suitable experimental designs and specific studies are needed to quantify precisely the size of the differences between the two leaning modes.

Data sets used in this paper are small and lack of consistency between controlled factors for each experiment (different drivers, cars, driving conditions, etc.) but it is worth trying a meta-analysis to improve veracity of the results. Effects sizes are in line all together showing the ability of our indicators to represent eco-driving capacities. Our work show that just reading simple eco-driving advices allows drivers to reduce significantly their fuel consumption and to adopt an eco-driving behavior although performances are better after a course. The important question now is to find how long a fuel efficient driving behavior last depending on the way drivers learned it. This issue will be the scope of our future research.

References

Agresti, A. (2002). *Categorical data analysis*. 2nd edition. Wiley Series in Probability and Mathematical Statistics, Chichester.

Barbé, J., Boy, G.A. & Sans, M. (2008). GERICO: A human centered eco-driving system. IFAC analysis, Design, and Evaluation of Human-Machine Systems, Volume # 10 / Part # 1 (http://www.ifac-papersonline.net/Detailed/39408.html).

Benminoun, M., et al. (2011). Safety analysis method for assessing the impacts of advanced driver assistance systems within the European large scale filed test "EuroFOT". *8th ITS European Congress 2011, Lyon.*

CIECA (International commission for drivertesting authorities), internal project on'Eco-driving' in category Bdriver training& the driving test2007. [Online].Available: http://www.cieca.be/

Ecodrive project (2009). found at: http//www.ecodrive.org.

Ericsson, E. (2001). Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. *Transportation Research Part D: Transport and Environment*, 6, 325-345.

Guo, F. and Hankey, J. (2010). *Modeling 100-Car Safety Events: A Case-Based Approach for Analyzing Naturalistic Driving Data*. The National Surface Transportation Safety Center for Excellence.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression.* Second edition. John Wiley and Sons, New York, NY.

Liang, K.-Y., and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

Moineddin R, Matheson FI, Glazier RH. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Med Res Methodol 2007*;7:34.

Peduzzi, P., et al. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, **49:** 1372-1379.

Symmons, M., Rose, G., VanDoorn, G., Ecodrive as a road safety tool for Australian conditions, Monash University, 2009.

Theall, K.P., et al. (2011). Impact of small group size on neighbourhood influences in multilevel models. *J Epidemiol Community Health*, 65 (8), 688-

Walker, G.H., Stanton, N.A. and Young, M.S. (2001). Hierarchical Task Analysis of Driving: A New Research Tool, In M.A.Hanson (Ed), Contemporary Ergonomics, Taylor & Francis Ltd., London, 435-440.

Young, M.S., et al., Safe driving in a green world: A review of driver performance benchmarks and technologies..., Applied Ergonomics (2010).

Ziegler, A., Kastner, C. and Blettner, M. (1998). The generalised estimating equations: an annotated bibliography, *Biometrical Journal*, 40 (2), 115-139.

2012 Intelligent Vehicles Symposium
Alcalá de Henares, Spain, June 3-7, 2012

Using statistical models to characterize eco-driving style with an aggregated indicator

Cindie Andrieu and Guillaume Saint Pierre
IFSTTAR
Vehicle-Infrastructure-Driver Interactions Research Unit
14, route de la minière
78000 Versailles-Satory, France
Email: cindie.andrieu@ifsttar.fr and guillaume.saintpierre@ifsttar.fr

***Abstract*— This paper presents the construction of an aggregated indicator of a fuel-efficient driving style, in order to construct an efficient Ecological Driving Assistance System (EDAS). Such an eco-index can be used to detect eco-driving behaviour, but also to give to the driver useful advices to help him improving his driving efficiency without deteriorating safety. The logistic regression is used to model our experimental dataset of twenty subjects driving twice the same route: normally or following the golden rules of eco-driving. Depending on some driving indicators, the estimated probability of being an eco-driver is used as an eco-index to characterize that driving pattern. This work show how such a simple aggregated indicator, related to driving dynamics rather than fuel consumption, can be useful for driver monitoring and information. Two models, from the simplest to the most complicated, are compared, and their performances analysed.**

Keywords: Eco-driving, EDAS, Driving behaviour, Logistic regression.

I. INTRODUCTION

Speed management is a preoccupation for public authority as speed is considered as the main cause of traffic injury accidents. Since the climate change evidence, many attempts are made to merge traditional speed management methods and greenhouse gas emission reduction techniques, optimizing both safety and fuel consumption. Among the most promising solutions to solve such a challenge, the ecological way to drive usually referred as the eco-driving style appears to be one of the best.

The characteristics of eco-driving are generally well defined and easily characterized (see for example [1] and [2]) even if eco-driving rules are slightly different among countries [3]. The advantages of eco-driving, of course, go beyond CO2 reductions [4]. They include reducing the cost of driving to the individual and producing tangible and well-known safety benefits (with fewer accidents and traffic fatalities) [1]. Disadvantages, are little public understanding of the nature of eco-driving, and seemingly ingrained driving habits. According to [3], some potential sources of driving conflicts exist if the eco-driving rules are misinterpreted.

Helping the driver to choose the best compromise between safety and CO2 reduction driving techniques is the goal of a new type of advanced systems called ecological driving assistance systems (EDAS). Even GPS devices or smartphones applications are sometimes providing a dynamic fuel efficiency indicator, while some software are specifically dedicated to eco-driving and are able to deliver tips to help decrease fuel consumption. Most of these devices (embedded or not) are using miles per gallon (or liters for 100km in Europe) as displayed parameters, while some of them use more sophisticated approach and compute a global indicator. According to expert's knowledge (interview with eco-driving professionals), displaying instant fuel use, or battery gauge, is not sufficient to help the drivers in understanding the dynamic relationship between driving actions and fuel efficiency. As most of the people want to keep ecological driving assistance systems (EDAS) simple (see for example [5]), we believe that a global indicator, merging different driving parameters can be more efficient than fuel consumption.

978-1-4673-2118-1/$31.00 ©2012 IEEE

The aim of this study is to provide a methodology suitable to compute an aggregated eco-driving indicator based on statistical models estimated using naturalistic driving data, and evaluated with an experimental study. Four indicators were chosen, each associated with one of the main rules of eco-driving. After a discussion about various suitable statistical models, we demonstrate the interest of using logistic regression for model based estimation of driving fuel efficiency. A second model has been developed in order to allow its implementation on smartphones not connected to the vehicle.

II. Methodology

A. Experimental design

The experiment goal was to clearly identify two classes of driving behavior on the same test track: "normal" and fuel efficient way to drive commonly known as "eco-driving". Twenty drivers participated in this experiment that took place in June and July 2009 in Ponchartrain (Yvelines) in France. In order to minimize traffic influence, the chosen route is of inter-urban type and a length of 14km. The trips were all performed under free flow conditions and with dry weather. The vehicle used was a petrol-driven Renault Clio III with manual gearshift. First of all, the journey is discovered by the subjects while seeing the experimenter driving and giving safety and direction instructions. Then, the trip was driven twice by each driver: once while driving normally, and secondly while following the "Golden Rules" of eco-driving extracted from the Ecodrive project [2] and summarized in Table I. These rules were given just before the ecological trip. To eliminate a learning effect of the journey, trip's order has been counter-balanced. An on-board logging device was used to monitor key driving parameters. The device is connected to the controller area network (CAN) of the vehicle, logging most of the relevant parameters related to engine state, vehicle dynamic, and driver actions on pedals. The vehicle has been also equipped with a GPS, a camera in front of the vehicle and a fuel flow meter. We used a fuel flow meter DFL1x-5bar to validate the fuel consumption logged with the CAN. Additional variables were post-processed such speed limits, gear ratio, and many indicators inspired from [6].

B. Selection of indicators associated with each of the main rules of eco-driving

Driving style to reduce fuel consumption is related to the implementation of the four main eco-driving rules set out in Table I. Due to this link, each of these instructions was associated with an indicator. The proposed indicators are summarized in Table I. So the first rule state to shift up early. Therefore, it is natural to associate the indicator *AvgRPMShiftUp* which is the average engine speed (in rpm) at the shift into a higher gear. The second rule is related both to the gear and the engine speed. So we created an indicator, called *IndexGearRPM*, summarizing these two variables and calculated as follows:

$$\begin{aligned} &IndexGearRPM = \\ &\tfrac{1}{3500}(TimeNeut \times AvgRPMNeut \\ &\quad + TimeGear1 \times AvgRPMGear1 + \dots \\ &\quad + TimeGear5 \times AvgRPMGear5) \end{aligned}$$

where *TimeNeut* is the percentage time in neutral gear, *AvgRPMNeut* is the average engine speed (in rpm) in neutral gear, *TimeGear1* is the percentage time in gear 1 (with pressing the accelerator pedal), etc. Note that the condition of pressing the accelerator pedal ensure to ignore the time in engine brake which is associated to the fourth rule. Note also that the division by 3500, representing the maximum engine speed, is just a normalization factor. Then the third rule related to the anticipation of traffic is associated to the parameter *PKE* (Positive Kinetic Energy) calculated as follows:

$$PKE = \frac{\sum(v_f^2 - v_i^2)}{x} \text{ when } \frac{dv}{dt} > 0$$

where v_f and v_i are respectively the final and the initial speed (in m/s) at each time interval for which $\frac{dv}{dt} > 0$, and x is the total distance traveled (in m). This indicator represents the ability to keep the vehicle's kinetic energy as low as possible. So a nervous driving will be associated with a high *PKE*, and conversely a smoothly driving will be associated with a *PKE* close to zero. Finally, the fourth rule is

naturally associated with the percentage of time in engine brake characterized by the following conditions: non zero speed, no neutral, no pressure the brake pedal and the accelerator pedal.

C. Statistical models

The objective of this study is to construct an aggregated indicator of an eco-driving style. Our approach relies on developing a predictive model of economic driving behavior based on easily interpretable variables. Assuming trips are clustered according to the two driving conditions, it is worth trying a statistically based approach to predict the driving style. Such models are well suited in estimating the relationship between an outcome variable and a set of explanatory variables. In this paper, the outcome variable is from a binary distribution with two possible values:
For $i = 1, \ldots, I$ and $j = 1, \ldots, T_i$,

$$Y_{ij} = \begin{cases} 1 \ if \ ecodriving \\ 0 \ if \ not \end{cases}$$

where I is the number of drivers and T_i is the number of observations for the driver *i*. Logistic regression is a form of statistical modeling that is often appropriate for independant binary outcome variables. Assume Y_{ij} follows a Bernoulli distribution with parameter $p_{ij} = P(Y_{ij} = 1)$ where p_{ij} represent the probability that the event occurred for the observation Y_{ij}. The relationship between the event probability p_{ij} and the set of factors is modeled through a logit link function with the following form:

$$logit(p_{ij}) = \log\left(\frac{p_{ij}}{1 - p_{ij}}\right) = X'_{ij}\beta$$

where X_{ij} is the vector of explanatory variables and β is the vector of regression parameters [7]. The ordinary logistic regression assumes independent observation and the vector β is estimated by the method of maximum likelihood. However, the assumption of data independence does not suit our data very well, as it will contain unavoidable driver-specific correlations (i.e. observations from the same driver are assumed to be correlated) that should be treated as random effects. The standard errors from the ordinary logistic regression are then biased because the independence assumption is violated.

To account for these driver specific correlations as random effects, more sophisticated statistical models need to be applied. These models are particularly useful for naturalistic driving study [8] and specially event based approach (EBA) which basic principle is to identify time segments that can be predictive of an event (e.g. crash, near-crash, ...). Indeed, these models include additional parameters to deal with correlations, and confounding factors are viewed as explicative variables that can be used to predict event probability. One such model is the "Generalized Estimated Equations" (GEE) model or marginal models, originally developed to model longitudinal data by Liang and Zeger [9], which assumes that observations are marginally correlated. Another approach for modeling correlated data is "Generalized Linear Mixed Models" (GLMM). The GLMM model introduces a random effect specific to each subject whereas the GEE approach models the marginal distributions by treating correlation as a nuisance parameter. Therefore the inference is individual (subject-specific approach) in contrast to marginal models that model the average population (population-averaged approach). However, in our study, we didn't use these two sophisticated statistical models because of the small sample size (see Section III-B for more details). So we used only ordinary logistic regression models.

III. RESULTS

A. Overall effects of eco-driving rules

Numeric results are summarized in Table II. A paired *t*-test was performed to assess whether the mean of each parameter differ significantly according to the driving style. Table II indicates the p-values of these tests. Among the most interesting ones, the average fuel consumption across drivers decreased by 12.5% between normal driving and eco-driving; this fall being of 26% for some drivers. These results show that it seems quite simple to reduce fuel consumption by applying some basic rules of eco-driving. The average speed decreased by 5.8% and the percentage of time beyond the legal speed limit decreased by 30.1%. These reductions reflect a better

Instruction	Indicator	Abbreviation
1. Shift up as soon as possible: Shift up between 2.000 and 2.500 revolutions per minute.	Average engine speed at the shift into a higher gear.	AvgRPMShiftUp
2. Maintain a steady speed: Use the highest gear possible and drive with low engine RPM.	Index of gear ratio distribution and engine speed associated.	IndexGearRPM
3. Anticipate traffic flow: Look ahead as far as possible and anticipate the surrounding traffic.	Positive Kinetic Energy.	PKE
4. Decelerate Smoothly: When you have to slow down or to stop, decelerate smoothly by releasing the accelerator in time, leaving the car in gear.	Percentage of time in engine brake.	TimeEngineBrake

TABLE I
MAIN RULES OF ECO-DRIVING AND INDICATORS ASSOCIATED

compliance with speed limits with economical driving. As regards the application of eco-driving rules, the four associated indicators are significantly different among the two driving conditions, indicating that the instructions were applied. Furthermore, the average acceleration and deceleration both decrease significantly which is in agreement with the second and the third rules of eco-driving.

B. Construction of an eco-index based on the main rules of eco-driving

The aim of this work is the development of a predictive model of economic driving behavior based on easily interpretable variables excluding the variable on fuel consumption. Indeed, fuel consumption is closely related to road type (urban, interurban, motorway) and traffic, but it cannot be considered itself as an eco-driving indicator. It is obvious as even a very efficient driver will have a high fuel consumption when driving under congestion or on hilly roads. An efficient indicator of eco-driving should not depend too much on such external conditions and rely more on driver actions. Thus, we constructed a predictive model of the probability of being in an eco-driving situation using a binomial logistic regression model with the four indicators in Table II as explanatory variables. According to our experiment, we predict the binary variable named "Trip" which takes the value 0 in normal driving (noted "normal") and 1 in eco-driving (noted "eco").

In our experiment, both the number of clusters (20) and the cluster size (2) are small. These constraints do not allow us to use the appropriate statistical models taking into account driver specific correlations. Thus, Ziegler et al. [10] recommend an application of the GEE only, if the number of clusters is at least 30 for a cluster size of about 4 for a low to moderate correlation. Moreover, several studies (e.g. [11]) have shown that parameters estimates are biased with both fixed or random effects logistic models when the number of clusters and the cluster size are small. However these studies show that the estimates of the random intercept and random slope have larger biases compared to the fixed effect parameters. Thus, later in this paper, we use an ordinary logistic regression. The estimated logistic model is the following:

$$\begin{aligned} &logit\left[P\left(Trip = Eco\right)\right] = 8.967 \\ &- 0.007 \times X_1 + 0.242 \times X_2 \\ &- 31.684 \times X_3 + 0.148 \times X_4 \end{aligned} \quad (1)$$

where $X_1, \ldots, X_4$ are the four indicators associated respectively with the four instructions of eco-driving (Table I). The usefulness of the model is measured by the Nagelkerke R^2, denoted R^2_N, which is an adjusted version of the Cox & Snell R^2 and which is similar to the coefficient of determination R^2 in linear regression. This parameter does not measure the goodness of fit of the model but indicate how useful the explanatory variables are in predicting the response variable. The model 1 reached a total $R^2_N = 0.74$, with a strong influence of the variable *PKE*, leading to increase the probability of being in a situation of eco-driving. Using a decision rule's cutoff value of 0.5, the model correctly classified 85% of true positives ("normal") and 80% of true negatives ("eco") even though this

Parameter	Description	Mean "Normal"	Mean "Eco"	Variation (%)
Avg_Fuel_Consum	Average fuel consumption (l/100km).	6.86	6.00	-12.5^{***}
Avg_RPM_Shift_Up	Average engine speed at the shift into a higher gear (associated with rule 1).	2737.5	2232.8	-18.4^{***}
Index_Gear_RPM	Index of gear ratio distribution and engine speed associated (associated with rule 2).	61.0	52.9	-13.3^{***}
PKE	Positive Kinetic Energy (associated with rule 3).	0.343	0.243	-29.2^{***}
Time_Engine_Brake	Percentage of time in engine brake (associated with rule 4).	20.3	26.3	$+29.6^{**}$
Avg_Speed	Average speed	50.85	47.89	-5.8^{**}
Avg_Accel	Average acceleration	0.498	0.387	-22.3^{***}
Avg_Decel	Average deceleration	-0.619	-0.523	-15.5^{***}
Avg_RPM	Average engine speed	2097.4	1835.5	-12.5^{***}
Time_NonLegal_Speed	Percentage of time beyond the legal speed limit	37.9	26.5	-30.1^{***}

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE II
EFFECTS OF ECO-DRIVING RULES ON DIFFERENT PARAMETERS

classification results from using all observations to fit the model, which can bias the results. For pedagogical purposes, we call "eco-index" of the observed trip, the model output probability $P\,(Trip = Eco)$ multiplied by one hundred. So we obtain an index of eco-driving which varies between 0 and 100 for easier interpretation. One of the main objectives of eco-driving is to reduce fuel consumption; evaluating the performance of such an eco-index can be done by studying the relationship strength between our eco-index and the average fuel consumption. We conducted a linear regression between these two parameters for all 40 trips from our experiment. This model reached a total coefficient of determination $R^2 = 0.70$, which shows that our eco-index is closely related to the average fuel consumption.

C. Construction of an eco-index based on a simple indicator: eco-index for smartphone

In this section, we use the same method as in the previous section to build a new model of eco-index based only on the parameter *PKE*. Indeed, on the one hand, we observed that this parameter had a strong influence with the probability of being in an eco-driving situation. On the other hand, the advantage of this parameter is its easiness to be calculated since it depends only on speed. Thus, a model of eco-index based solely on *PKE* can be implemented easily on a smartphone. The logistic model is the following:

$$\begin{aligned} &logit\left[P\left(Trip = Eco\right)\right] = 9.773 \\ &- 34.089 \times PKE \end{aligned} \tag{2}$$

This model reached a total Nagelkerke $R^2_N = 0.59$ and correctly classified 80% of true positives ("normal") and 85% of true negatives ("eco"). The linear regression between the eco-index derived from the model 2 and the average fuel consumption has a coefficient of determination $R^2 = 0.69$. This simple model has good features, similar to the complete model.

D. Factorial analysis

A principal component analysis (PCA) was performed with the forty original trips using the four indicators based on the main rules of eco-driving. The first factorial plan with the value of the eco-index related to the model 1, distinguishing "normal" and "eco" trips, is represented in Fig. 1.

The first axis is correlated with the three first indicators defined in Table 1, and the second axis is correlated with the fourth indicator *TimeEngineBrake*. We observe that the two driving styles are well discriminated by the four indicators. Moreover, these results confirm that the eco-index is a well eco-driving indicator since "eco" trips

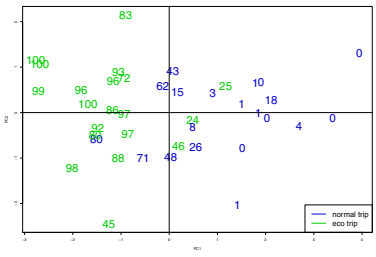


Fig. 1. Principal component analysis and eco-index (blue for "normal" trips and green for "eco" trips).

are associated with high eco-index whereas "normal" trips are associated with lower eco-index.

IV. Conclusion

This study provides a methodology suitable to compute a global eco-driving indicator based on statistical models, taking into account various behavior related parameters. Two logistic regression models of this eco-index, from the simplest to the most complicated, have been developed: the first one is based on the four performance indicators associated with each of the main rules of eco-driving, the second one is based only on the variable PKE. The first model provides the most appropriate information to be displayed in a future ecological driving assistance system (EDAS). Indeed, each performance indicator being associated with a rule of eco-driving, it is possible to display quantitative feedback to the driver, specifically for each one of the four main rules of eco-driving. The second model based on PKE has the advantage of being easily calculated and therefore suitable for nomadic devices implementation.

Assuming that eco-driving behavior depends on the road conditions, we have extended the full model 1 to a more complex model taking into account the speed limit as a stratification variable. This third model, not introduced in this paper, improves the properties of the full model and allows to inform the driver on the network categories (urban, rural, ...) on which he can improve his efficiently driving. However, this model needs the knowledge of speed limits for the traveled route.

Other statistical models taking into account driver specific correlations, namely GEE and mixed models, have been mentioned but we could not implement them because of the small sample size of our experiment. However, it might be interesting to test bootstrap methods suitable if the number of clusters is small, as discussed in [12].

Future works will focus on the validation of the two logistic models presented in this paper, and on the development of a dynamic eco-index providing information to the driver during the trip and allowing self-evaluation throughout the journey.

References

[1] J. Barkenbus, "Eco-driving: An overlooked climate change initiative," *Energy Policy*, vol. 38, no. 2, pp. 762–769, 2010.

[2] "Ecodrive project," 2009. [Online]. Available: http://www.ecodrive.org

[3] CIECA (International commission for driver testing authorities), *CIECA internal project on 'Eco-driving' in category B driver training and the driving test*. CIECA, 2007. [Online]. Available: http://www.cieca.be/

[4] R. Vermeulen, "The effects of a range of measures to reduce the tail pipe emissions and/or the fuel consumption of modern passenger cars on petrol and diesel." TNO science and industry, Tech. Rep., 2006.

[5] M. S. Young, S. A. Birrell, , and N. A. Stanton, *Lecture Notes in Computer Science: Engineering Psychology and Cognitive Ergonomics*. Springer Berlin / Heidelberg, 2009, ch. Design for Smart Driving: A Tale of Two Interfaces, pp. 477–485.

[6] E. Ericsson, "Independent driving pattern factors and their influence on fuel-use and exhaust emission factors," *Transportation Research Part D: Transport and Environment*, vol. 6, pp. 325–345, 2001.

[7] A. Agresti, *Categorical data analysis. 2nd ed.* Chichester: Wiley Series in Probability and Mathematical Statistics., 2002.

[8] F. Guo and J. Hankey, "Modeling 100-car safety events: A case-based approach for analyzing naturalistic driving data." Report No. 09-UT-006, Virginia Tech Transportation Institute, Tech. Rep., 2009.

[9] K. Y. Liang and S. L. Zeger, "Longitudinal data analysis using generalized linear models," *Biometrika*, vol. 73, pp. 13–22, 1986.

[10] A. Ziegler, C. Kastner, and M. Blettner, "The generalised estimating equations : an annotated bibliography," *Biometrical Journal*, vol. 40 (2), pp. 115–139, 1998.

[11] R. Moineddin, F. I. Matheson, and R. H. Glazier, "A simulation study of sample size for multilevel logistic regression models," *BMC Medical Research Methodology*, p. 7:34, 2007.

[12] L. H. Moulton and S. L. Zeger, "Analyzing repeated measures on generalized linear models via the bootstrap," *Biometrics*, vol. 45, pp. 381–394, 1989.

Bibliographie

[1] Aarts, L. et Van Schagen, I. (2006). Driving speed and the risk of road crashes : A review. *Accident Analysis and Prevention.* **38**(2), p. 215–224.

[2] Abramowitz, M. et Stegun, I. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* U.S. National Bureau of Standards.

[3] Adams, R. A. et Fournier, J. J. F. (2003). *Sobolev Spaces.* Elsevier Science.

[4] Allain, G. (2008). *Prévision et analyse du traffic routier par des méthodes statistiques.* Thèse doct. Université Paul Sabatier, Toulouse III.

[5] Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics,* Taylor & Francis, **16**(1), p. 125–127.

[6] Andersson, J. (2000). *Image processing for analysis of road user behaviour - a trajectory based solution.* Lund Institute of Technology, Department of Technology et Society, Traffic Engineering, Bulletin 212.

[7] Andrieu, C., Saint Pierre, G. et Bressaud, X. (2013). Estimation of space-speed profiles: A functional approach using smoothing splines. *Proceedings of the IEEE Intelligent Vehicle Symposium (IV 2013), Gold Coast, Australie, juin 2013*.

[8] Andrieu, C., Saint Pierre, G. et Bressaud, X. (2012). Modélisation fonctionnelle et lissage sous contraintes d'un profil spatial de vitesse. *44ème Journées de Statistique, Bruxelles, Belgique*.

[9] Antoniadis, A., Bigot, J. et Lambert-Lacroix, S. (2010). Peaks detection and alignment for mass spectrometry data. *Journal de la Société Française de Statistique.* **151**(1), p. 17–37.

[10] Aronszajn, N. (1950). Theory of Reproducing Kernels. *Transactions of the American Mathematical Society.* **68**(3), p. 337–404.

[11] Atteia, M. (1965). Fonctions-splines généralisées. *Compte Rendu Académie des Sciences Paris.* **261**, p. 2149–2152.

[12] Atteia, M. (1965). Généralisation de la définition et des propriétés des "splines-fonctions". *Compte Rendu Académie des Sciences Paris.* **260**, p. 3550–3553.

[13] Bagdadi, O. et Várhelyi, A. (2011). Jerky driving - An indicator of accident proneness? *Accident Analysis & Prevention.* **43**(4), p. 1359–1363.

[14] Barbosa, H. M., Tight, M. R. et May, A. D. (2000). A model of speed profiles for traffic calmed roads. *Transportation Research Part A: Policy and Practice.* **34**(2), p. 103–123.

[15] Berlinet, A. et Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics.* Springer.

[16] Bertero, M. (1986). Regularization methods for linear inverse problems. *Inverse Problems*. In: éd. Talenti, G., Springer, p. 52–112.

[17] Besse, P. (1979). *Etude descriptive des processus : Approximation et interpolation.* Thèse doct. Université Paul Sabatier, Toulouse, France.

[18] Besse, P. et Thomas-Agnan, C. (1989). Le lissage par fonctions splines en statistique: revue bibliographique. *Statistique et Analyse des données.* **14**(1), p. 55–84.

[19] Besse, P. et Cardot, H. (1996). Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Canadian Journal of Statistics.* **24**(4), p. 467–487.

[20] Bigot, J. (2003). *Recalage de signaux et analyse de variance fonctionnelle par ondelettes : application au domaine biomédical.* Thèse doct. Universite Joseph Fourier, Grenoble I.

[21] Bigot, J. et Gadat, S. (2010). Smoothing under diffeomorphic constraints with homeomorphic splines. *SIAM Journal on Numerical Analysis.* **48(1)**, p. 224–243.

[22] Boonsiripant, S. (2009). *Speed profile variation as a surrogate measure of road safety based on GPS-equipped vehicle data.* Thèse doct. School of civil et environmental engineering, Georgia Institute of Technology, Atlanta.

[23] Bosq, D. (2000). *Linear processes in function spaces : theory and applications.* Springer-Verlag, New York.

[24] Bratt, H. et Ericsson, E. (1999). Estimating speed and acceleration profiles from measured data. *Conference Proceedings of the 8th International Symposium "Transport and Air Pollution", Graz, Austria*.

[25] Brunk, H. D. (1955). Maximum likelihood estimates of monotone parameters. *Annals of Mathematical Statistics.* **26**, p. 607–616.

[26] Burden, R. et Faires, J. (2011). *Numerical analysis, Ninth Edition.* Brooks Cole Publishing Company.

[27] Calderon, C., Martinez, J., Carroll, R. et Sorensen, D. (2010). P-splines using derivative information. *Multiscale Modeling & Simulation.* **8**(4), p. 1562–1580.

[28] Cleveland, W. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association.* **74**, p. 829–836.

[29] Cleveland, W. et Devlin, S. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association.* **83**, p. 596–610.

[30] Comte, S. L. (1998). Evaluation of in-car speed limiters: Simulator study. *Managing Speeds of Traffic on European Roads: Working Paper.* **3**,

[31] Conan-Guez, B. (2002). *Modélisation supervisée de données fonctionnelles par perceptron multi-couches.* Thèse doct. Université Paris-IX Dauphine, France.

[32] Cox, D. (1988). Approximation of method of regularization estimators. *The Annals of Statistics.* p. 694–712.

[33] Craven, P. et Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik,* Springer, **31**(4), p. 377–403.

[34] Cuevas, A., Febrero, M. et Fraiman, R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational statistics & data analysis.* **51**(2), p. 1063–1074.

[35] Cuevas, A., Febrero, M. et Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics.* **22**(3), p. 481–496.

[36] Dauxois, J. et Pousse, A. (1976). *Les analyses factorielles en calcul des probabilités et en statistique : essai d'étude synthétique.* Thèse doct. Université Paul Sabatier, Toulouse, France.

[37] De Boor, C. (2001). *A Practical Guide to Splines.* (éd. Edition, R.) Springer.

[38] Delecroix, M. et Thomas-Agnan, C. (2000). Spline and kernel regression under shape restrictions. *Smoothing and Regression: Approaches, Computation, and Application*. In: Wiley Online Library,

[39] Delecroix, M., Simioni, M. et Thomas-Agnan, C. (1996). Functional estimation under shape constraints. *Journaltitle of Nonparametric Statistics.* **6**(1), p. 69–89.

[40] Delsol, L. (2008). *Régression sur variable fonctionnelle : Estimation, Tests de structure et Applications.* Thèse doct. Université Paul Sabatier, Toulouse, France.

[41] Demailly, J. P. (2006). *Analyse numérique et équations différentielles.* EDP Sciences.

[42] Dette, H. et Pilz, K. (2006). A comparative study of monotone nonparametric kernel estimates. *Journal of Statistical Computation and Simulation.* **76, no. 1**, p. 41–56.

[43] Dette, H., Neumeyer, N. et Pilz, K. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli.* **12, no. 3**, p. 469?490.

[44] Deville, J. C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Annales de l'INSEE.* **15**, p. 7–97.

[45] Dieudonné, J. (1975). *Eléments d'analyse.* tomes 1, 2, 6. Gauthier-Villars.

[46] Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. *Constructive theory of functions of several variables*. In: Springer, p. 85–100.

[47] Dudley, R. (1989). *Real Analysis and Probability.* Chapman et Hall.

[48] Edgar, A. (2006). A Consistent Method for Rural Speed Zoning : 85th Percentile Speed Profile vs Risk Based Calculation. *Research into Practice: 22nd ARRB Conference, Canberra, Australia*.

[49] Ehrlich, J. (2009). Towards ISA deployment in Europe : state of the art, main obstacles and initiatives to go forward. *Proceedings from the 2009 Intelligent Speed Adaptation Conference, Sidney*.

[50] Eilers, P. et Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical science.* **11**, p. 89–102.

[51] Ericsson, E. (2001). Independent driving pattern factors and their influence on fuel-use and exhaust emission factors. *Transportation Research Part D: Transport and Environment.* **6**(5), p. 325–345.

[52] Ericsson, E. (2000). Variability in urban driving patterns. *Transportation Research Part D.* **5**, p. 337–354.

[53] Eubank, R. (1999). *Nonparametric regression and spline smoothing.* Marcel Dekker.

[54] Evgeniou, T., Pontil, M. et Poggio, T. (2000). Regularization networks and support vector machines. *Advances in Computational Mathematics.* **13 (1)**, p. 1–50.

[55] Fan, J. et Gijbels, I. (1996). *Local polynomial modelling and its applications.* Chapman & Hall.

[56] Febrero, M., Galeano, P. et González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOx levels. *Environmetrics.* **19**(4), p. 331–345.

[57] Ferraty, F. et Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice.* Springer-Verlag, New York.

[58] Ferraty, F. et Vieu, P. (2011). Richesse et complexité des données fonctionnelles. *Revue Modulad.* **43**, p. 25–43.

[59] Fitzpatrick, K. *et al.* (2003). Design Speed, Operating Speed, and Posted Speed Practices. *Publication NCHRP Report 504, Transportation Research Board.*

[60] Fox, J. (2000). *Nonparametric Simple Regression: Smoothing Scatterplots.* SAGE Publications, Incorporated.

[61] Fraiman, R. et Muniz, G. (2001). Trimmed means for functional data. *Test.* **10**(2), p. 419–440.

[62] Friedman, J. H. et Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics,* Taylor & Francis, **31**(1), p. 3–21.

[63] Friedman, J. H. et Tibshirani, R. (1984). The Monotone Smoothing of Scatterplots. *Technometrics.* **26**, p. 243–250.

[64] Fritsch, F. et Carlson, R. (1980). Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis.* **17**(2), p. 238–246.

[65] Gallen, R. (2010). *Assistance à la conduite en conditions atmosphériques dégradées par la prise en compte du risque routier.* Thèse doct. Université Pierre et Marie Curie, Paris.

[66] Gallen, R., Hautière, N. et Glaser, S. (2010). Advisory Speed for Intelligent Speed Adaptation in Adverse Conditions. *IEEE Intelligent Vehicles Symposium (IV'10), San Diego, California, USA.* p. 107–114.

[67] Gasser, T. et Kneip, A. (1995). Searching for structure in curve samples. *Journal of the American Statistical Association.* **90**(432), p. 1179–1188.

[68] Gijbels, I. (2004). *Monotone regression.* Discussion paper 0334, Institute de Statistique, Université Catholique de Louvain. http://www.stat.ucl.ac.be.

[69] Green, P. et Silverman, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach.* Chapman & Hall.

[70] Grenander, U. (1981). *Abstract Inference.* John Wiley & Sons.

[71] Gu, C. (1989). RKPACK and its applications: Fitting smoothing spline models. *Proceedings of the Statistical Computing Section, ASA.* p. 42–51.

[72] Gu, C. (2002). *Smoothing spline ANOVA models.* Springer Verlag.

[73] Hall, P. et Huang, L.-S. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics.* **29**, p. 624–647.

[74] Hall, P. et Yatchew, A. (2007). Nonparametric estimation when data on derivatives are available. *The Annals of Statistics.* **35**(1), p. 300–323.

[75] Hastie, T. et Tibshirani, R. (1990). *Generalized additive models.* Chapman & Hall.

[76] Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction. Second Edition.* Springer.

[77] He, X. et Shi, P. (1998). Monotone B-Spline Smoothing. *Journal of the American Statistical Association.* **14**, p. 315–337.

[78] Hellstrom, E. (2010). *Look-ahead control of heavy vehicles.* Thèse doct. Linkoping University, Institute of Technology, Sweden.

[79] Härdle, W. (1992). *Applied Nonparametric Regression.* Cambridge University Press.

[80] Hyden, C. et Varhelyi, A. (2000). The effects on safety, time consumption and environment of large scale use of roundabouts in an urban area: a case study. *Accident Analysis & Prevention.* **32**(1), p. 11–23.

[81] Hyndman, R. et Shang, H. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics.* **19**(1),

[82] Jun, J., Guensler, R. et Ogle, J. (2006). Smoothing methods to minimize impact of Global Positioning System random error on travel distance, speed, and acceleration profile estimates. *Transportation Research Record: Journal of the Transportation Research Board,* Trans Res Board, **1972**, p. 141–150.

[83] Kelly, C. et Rice, J. (1990). Monotone smoothing with application to dose response curves and the assessment of synergism. *Biometrics.* **46**, p. 1071–1085.

[84] Kerper, M., Wewetzer, C. et Mauve, M. (2012). Analyzing vehicle traces to find and exploit correlated traffic lights for efficient driving. *Intelligent Vehicles Symposium (IV), 2012 IEEE*. p. 310–315.

[85] Kimeldorf, G. et Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics,* JSTOR, p. 495–502.

[86] Kimeldorf, G. et Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications,* Elsevier, **33**(1), p. 82–95.

[87] Kimeldorf, G. et Wahba, G. (1970). Spline functions and stochastic processes. *Sankhya: The Indian Journal of Statistics, Series A,* JSTOR, p. 173–180.

[88] Klein, L. (2001). *Sensor technologies and data requirements for ITS.* Artech House, Incorporated.

[89] Kloeden, C., Ponte, G. et McLean, A. (2001). *Travelling speed and the risk of crash involvement on rural roads.* Rapport technique, CR-204, Australian Transport Safety Bureau.

[90] Kneip, A. et Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *The Annals of Statistics.* p. 1266–1305.

[91] Ko, J., Hunter, M. et Guensler, R. (2007). Measuring control delay using second-by-second GPS speed data. *Transportation Research Board 86th Annual Meeting, TRB, National Research Council, Washington, D.C.*.

[92] Laurent, P. J. (1972). *Approximation et optimisation.* Université Scientifique et Médicale de Grenoble.

[93] Laureshyn, A. (2005). Automated video analysis and behavioural studies based on individual speed profiles. *Proceedings of 18th ICTCT, Helsinki, October 2005*.

[94] Laureshyn, A. et Ardö, H. (2006). Automated video analysis as a tool for analysing road user behaviour. *ITS World Congress, London, UK*. **1** p. 8.

[95] Laureshyn, A., Åström, K. et Brundell-Freij, K. (2009). From Speed Profile Data to Analysis of Behaviour. *IATSS research.* **33**(2), p. 89.

[96] Levitin, D. J., Nuzzo, R. L., Vines, B. W. et Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology,* Canadian Psychological Association, **48**(3), p. 135–155.

[97] López-Pintado, S. et Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association.* **104**(486), p. 718–734.

[98] Louah, G. (2005). The accuracy of a speed profile estimation method combining continuous and spot speed measurements. *Road Safety on Four Continents: 13th International Conference, Warsaw, Poland*.

[99] Louah, G. (2002). *Vitesse ponctuelle des véhicules en fonction des caractéristiques de la route - Bibliographie commentée.* Rapport d'étude, Setra, CETE de l'Ouest.

[100] Luo, Z. et Wahba, G. (1997). Hybrid adaptive splines. *Journal of the American Statistical Association,* Taylor & Francis, **92**(437), p. 107–116.

[101] Luu, H. (2011). *Développement de méthodes de réduction de la consommation en carburant d'un véhicule dans un contexte de sécurité et de confort : un compromis entre économie et écologie.* Thèse doct. Université d'Evry Val d'Essonne.

[102] Mammen, E. (1991). Estimating a smooth monotone regression function. *Annals of Statistics.* **19(2)**, p. 724–740.

[103] Mammen, E. et Thomas-Agnan, C. (1999). Smoothing splines and shape restrictions. *Scandinavian Journal of Statistics.* **26**, p. 239–252.

[104] Mammen, E., Marron, J., Turlach, B. et Wand, M. (2001). A general projection framework for constrained smoothing. *Statistical Science,* JSTOR, p. 232–248.

[105] Mandava, S., Boriboonsomsin, K. et Barth, M. (2009). Arterial velocity planning based on traffic signal information under light traffic conditions. *12th International IEEE Conference on Intelligent Transportation Systems, 2009. ITSC'09*. p. 1–6.

[106] Mardia, K., Kent, J., Goodall, C. et Little, J. (1996). Kriging and splines with derivative information. *Biometrika.* **83**(1), p. 207–221.

[107] Matzkin, R. (1991). Semiparametric estimation of monotone and concave utility functions for polychotomous choice models. *Econometrica: Journal of the Econometric Society.* p. 1315–1327.

[108] Maza, E. (2004). *Prévision de trafic routier par des méthodes statistiques. Espérance structurelle d'une fonction aléatoire.* Thèse doct. Université Paul Sabatier, Toulouse III.

[109] Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple. *Zeitschrift für angewandte Mathematik und Physik ZAMP.* **30**(2), p. 292–304.

[110] Meyer, M. (2012). Constrained Penalized Splines. *Canadian Journal of Statistics.* **40**(1), p. 190–206.

[111] Meyer, M. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics.* p. 1013–1033.

[112] Moreno, A. T. et García, A. (2013). Use of speed profile as surrogate measure: Effect of traffic calming devices on crosstown road safety performance. *Accident Analysis & Prevention.* (In press, Available online),

[113] Mukerjee, H. (1988). Monotone nonparametric regression. *Annals of Statistics.* **16**, p. 741–750.

[114] Nilsson, G. (1982). The effects of speed limits on traffic crashes in Sweden. *Proceedings of the international symposium on the effects of speed limits on traffic crashes and fuel consumption, Dublin*.

[115] Nilsson, G. (2004). *Traffic safety dimensions and the Power Model to describe the effect of speed on safety, Lund Institute of Technology and Society.* Bulletin 221, Lund Institute of Technology, Sweden.

[116] Nygard, M. (1999). *A method for analysing traffic safety with help of speed profiles.* Thèse doct. Tampere University of Technology, Department of Civil Engineering, Finlande.

[117] OCDE (2007). *La gestion de la vitesse.* Editions OCDE.

[118] Ogle, J. H. (2005). *Quantitative assessment of driver speeding behavior using instrumented vehicles.* Thèse doct. School of civil et environmental engineering, Georgia Institute of Technology, Atlanta.

[119] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statistical science,* JSTOR, p. 502–518.

[120] Pearce, N. D. et Wand, M. P. (2006). Penalized splines and reproducing kernel methods. *The american statistician,* American Statistical Association, **60**(3), p. 233–240.

[121] Quddus, M. A., Ochieng, W. et Noland, R. (2007). Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C.* **15**, p. 312–328.

[122] Quiroga, C. A. et Bullock, D. (1998). Travel time studies with global positioning and geographic information systems: an integrated methodology. *Transportation Research Part C: Emerging Technologies.* **6**(1), p. 101–127.

[123] R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing (Vienna, Austria, 2008). <http://www.R-project.org>.

[124] Rakha, H. et Ding, Y. (2002). Impact of stops on vehicle fuel consumption and emissions. *Journal of Transportation Engineering,* American Society of Civil Engineers, **129**(1), p. 23–32.

[125] Rakha, H., Dion, F. et Sin, H. (2001). Using Global Positioning System Data for Field Evaluation of Energy and Emission Impact of Traffic Flow Improvement Projects: Issues and Proposed Solutions. *Transportation Research Record: Journal of the Transportation Research Board.* **1768**, p. 210–223.

[126] Ramsay, J. O. (1982). When the data are functions. *Psychometrika,* Springer, **47**(4), p. 379–396.

[127] Ramsay, J. O. et Dalzell, C. J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological),* JSTOR, **53**, p. 539–572.

[128] Ramsay, J. O. et Silverman, B. W. (2002). *Applied Functional Data Analysis: Methods and Case Studies.* Springer-Verlag, New York.

[129] Ramsay, J. O. et Silverman, B. W. (1997). *Functional Data Analysis.* Springer-Verlag, New York.

[130] Ramsay, J. O. et Silverman, B. W. (2005). *Functional Data Analysis, Second Edition.* (éd. in Statistics, S. S.) Springer-Verlag, New York.

[131] Ramsay, J. O., Hooker, G. et Graves, S. (2009). *Functional Data Analysis with R and MATLAB.* Springer-Verlag, New York.

[132] Ramsay, J. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **60(2)**, p. 365–375.

[133] Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science,* JSTOR, p. 425–441.

[134] Ramsay, J. et Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* Wiley Online Library, **60**(2), p. 351–363.

[135] Reinsch, C. H. (1967). Smoothing by spline functions. *Numerische Mathematik,* Springer, **10**(3), p. 177–183.

[136] Rousseeuw, P., Ruts, I. et Tukey, J. (1999). The bagplot: a bivariate boxplot. *The American Statistician.* **53**(4), p. 382–387.

[137] Rudin, W. (1975). *Analyse reelle et complexe: trad. de l'americain par N. Dhombres et F. Hoffman.* Masson.

[138] Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of computational and graphical statistics,* Taylor & Francis, **11**(4), p. 735–757.

[139] Ruppert, D., Wand, M. P., Holst, U. et Hösjer, O. (1997). Local polynomial variance-function estimation. *Technometrics.* **39**(3), p. 262–273.

[140] Ruppert, D., Wand, M. et Carroll, R. (2003). *Semiparametric regression.* Cambridge University Press.

[141] Saint Pierre, G. et Ehrlich, J. (2008). Impact of Intelligent Speed Adaptation systems on fuel consumption and driver behaviour. *Proceedings of the 15th ITS World Congress, New York, Novembre 2008*.

[142] Sakoe, H. et Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on.* **26**(1), p. 43–49.

[143] Schimek, M. (2012). *Smoothing and Regression: Approaches, Computation, and Application.* Wiley.

[144] Schölkopf, B. et Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* MIT press.

[145] Schölkopf, B., Herbrich, R. et Smola, A. (2001). A generalized representer theorem. *Computational learning theory*. p. 416–426.

[146] Schölkopf, B., C., B. et Smola, A. (1998). *Advances in Kernel Methods: Support Vector Learning.* MIT Press.

[147] Schoenberg, I. J. (1946). Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math.* **4**(Parts A and B), p. 45–99.

[148] Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America,* National Academy of Sciences, **52**(4), p. 947.

[149] Schumaker, L. (2007). *Spline Functions: Basic Theory, Third Edition.* Cambridge University Press.

[150] Schwetlick, H. et Kunert, V. (1993). Spline smoothing under constraints on derivatives. *BIT Numerical Mathematics.* **33**(3), p. 512–528.

[151] SETRA (2006). *Comprendre les principaux paramètres de conception géométrique des routes.* SETRA, Note technique, janvier 2006.

[152] SETRA (2008). *La vitesse pratiquée ou V85 - Formules de calcul.* SETRA : Note d'information, série Conception Sécurité Équipement Exploitation, n° 127.

[153] SETRA (1986). *Vitesses pratiquées et géométrie de la route.* SETRA : Note d'information n° 10, série circulation sécurité exploitation, avril 1986.

[154] Shawe-Taylor, J. et Cristianini, N. (2004). *Kernel methods for pattern analysis.* Cambridge university press.

[155] Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Methodological).* p. 1–52.

[156] Silverman, B. (1984). Spline smoothing: the equivalent variable kernel method. *The Annals of Statistics.* p. 898–916.

[157] Simonoff, J. (1996). *Smoothing methods in statistics.* Springer.

[158] Solomon, D. (1964). *Crashes on main rural highways related to speed, driver and vehicle.* Bureau of Public Roads. U.S. Department of Commerce. United States Government Printing Office, Washington D.C.

[159] Stone, C. J., Hansen, M. H., Kooperberg, C. et Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics,* Institute of Mathematical Statistics, **25**(4), p. 1371–1470.

[160] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. Series B (Methodological).* p. 111–147.

[161] Sun, Y. et Genton, M. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics.* **20**(2),

[162] Tastambekov, K., Puechmorel, S., Delahaye, D. et Rabut, C. (2010). Trajectory prediction by functional regression in Sobolev space. *42èmes Journées de Statistique*.

[163] Tikhonov, A. et Arsenin, V. (1977). *Solutions of ill-posed problems.* Winston.

[164] Tukey, J. (1977). Exploratory data analysis. *Reading, MA.* **231**,

[165] Tukey, J. (1975). Mathematics and the picturing of data. *Proceedings of the international congress of mathematicians*. **2** p. 523–531.

[166] Turlach, B. (2005). Shape constrained smoothing using smoothing splines. *Computational Statistics.* **20**(1), p. 81–104.

[167] Utreras, F. I. (1985). Smoothing noisy data under monotonicity constraints: Existence, characterization and convergence rates. *Numerishe Mathematik.* **47**, p. 611–625.

[168] Várhelyi, A., Hjälmdahl, M., Hyden, C. et Draskóczy, M. (2004). Effects of an active accelerator pedal on driver behaviour and traffic safety after long-term use in urban areas. *Accident Analysis & Prevention.* **36**(5), p. 729–737.

[169] Villalobos, M. et Wahba, G. (1987). Inequality-constrained multivariate smoothing splines with application to the estimation of posterior probabilities. *Journal of the American Statistical Association.* **82**(397), p. 239–248.

[170] Violette, E., Hublart, A., Subirats, P. et Louah, G. (2010). *Estimation de la vitesse pratiquée sur un itinéraire - Méthode et application.* CETE Normandie Centre, DITM, ESM.

[171] Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics.* p. 1378–1402.

[172] Wahba, G. (1975). Smoothing noisy data with spline functions. *Numerische Mathematik,* Springer, **24**(5), p. 383–393.

[173] Wahba, G. (1990). *Spline models for observational data.* SIAM.

[174] Wahba, G. et Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates. *Statistics & probability letters.* **25**(2), p. 105–111.

[175] Wahba, G. et Wendelberger, J. (1980). Some new mathematical methods for variational objective analysis using splines and cross-validation. *Monthly weather review.* **108**, p. 1122–1145.

[176] Wahba, G. et Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Communications in Statistics: Theory and Methods,* Taylor & Francis, **4**(1), p. 1–17.

[177] Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Computational Statistics.* **15**(4), p. 443–462.

[178] Wand, M. et Jones, M. (1995). *Kernel smoothing.* Chapman & Hall.

[179] Wang, K. et Gasser, T. (1997). Alignment of curves by dynamic time warping. *The Annals of Statistics.* **25**(3), p. 1251–1276.

[180] Wang, Y. (1997). GRKPACK: Fitting smoothing spline ANOVA models for exponential families. *Communications in Statistics-Simulation and Computation.* **26**(2), p. 765–782.

[181] Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association.* **93**(441), p. 341–348.

[182] Wang, Y. (2011). *Smoothing Splines: Methods and Applications.* Chapman & Hall/CRC Press.

[183] Wang, Y. et Ke, C. (2004). ASSIST: A suite of S functions implementing spline smoothing techniques. *University of California, Santa Barbara.*

[184] Wegman, E. J. et Wright, I. W. (1983). Splines in statistics. *Journal of the American Statistical Association.* **78**(382), p. 351–365.

[185] Whittaker, E. T. (1922). On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society,* Cambridge Univ Press, **41**(1), p. 63–75.

[186] Zuo, Y. et Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics.* p. 461–482.

Auteur : Cindie ANDRIEU

Directeurs de thèse : Xavier BRESSAUD et Guillaume SAINT PIERRE

Date et lieu de soutenance : le 24 septembre 2013 à l'Université Toulouse III Paul Sabatier

---

Modélisation fonctionnelle de profils de vitesse en lien avec l'infrastructure et méthodologie de construction d'un profil agrégé

La connaissance des vitesses pratiquées est une caractéristique essentielle du comportement des conducteurs et de leur usage du réseau routier. Cette information est rendue disponible grâce à la généralisation des véhicules connectés, mais aussi des smartphones, qui permettent d'accroître le nombre de "traceurs" susceptibles de renvoyer leur position et leur vitesse en temps réel. Dans cette thèse, nous proposons d'utiliser ces traces numériques et de développer une méthodologie, fondée sur une approche fonctionnelle, permettant d'extraire divers profils de vitesse caractéristiques. Dans une première partie, nous proposons une modélisation fonctionnelle des profils spatiaux de vitesse (i.e. vitesse vs distance parcourue) et nous étudions leurs propriétés (continuité, dérivabilité).

Dans une seconde partie, nous proposons une méthodologie permettant de construire un estimateur d'un profil spatial de vitesse à partir de mesures bruitées de position et de vitesse, fondée sur les splines de lissage et la théorie des espaces de Hilbert à noyau reproduisant (RKHS).

Enfin, la troisième partie est consacrée à la construction de divers profils agrégés (moyen, médian). Nous proposons notamment un alignement des profils par landmarks au niveau des arrêts, puis nous proposons la construction d'enveloppes de vitesse reflétant la dispersion des vitesses pratiquées.

Mots clés : Profils de vitesse ; Analyse de données fonctionnelles ; Régression non paramétrique ; Splines de lissage ; Espaces de Hilbert à noyau reproduisant.

Functional modeling of speed profiles adapted to the infrastructure and methodology of construction of an aggregated speed profile

The knowledge of the actual vehicle speeds is an essential characteristic of drivers behavior and their road usage. This information become available with the generalization of connected vehicles, but also smartphones, which increase the number of "tracers" likely to refer their position and speed in real time. In this thesis, we propose to use these digital traces and to develop a methodology, based on a functional approach, to produce several reference speed profiles. In a first part, we propose a functional modeling of space-speed profiles (i.e. speed vs position) and we study their properties (continuity, differentiability).

In a second part, we propose a methodology to construct an estimator of a space speed profile from noisy measurements of position and speed, based on smoothing splines and the theory of reproducing kernel Hilbert spaces (RKHS).

The third part is devoted to the construction of several aggregated profiles (average, median). In particular, we propose a landmark-based registration of profiles at stops, and we propose the construction of speed corridors reflecting the dispersion of actual speeds.

Keywords : Speed profiles ; Functional data analysis ; Nonparametric regression ; Smoothing splines ; Reproducing kernel Hilbert spaces.