Analyse de l'activité de conduite par les chaînes de Markov cachées et les modèles de ruptures multi-phasiques: méthodologie et applications

N. Dapzol

UNIVERSITÉ CLAUDE BERNARD
LYON I

ANALYSE DE L'ACTIVITE DE CONDUITE PAR LES CHAÎNES DE MARKOV CACHÉES ET LES MODÈLES DE RUPTURES MULTI-PHASIQUES : MÉTHODOLOGIE ET APPLICATIONS

THÈSE

Présentée pour obtenir le titre de

DOCTEUR
Discipline : Mathématiques appliquées.

Par

Nicolas DAPZOL

Directeur de thèse : Professeur Christian Mazza

Jury :

Celeux, Gilles	Professeur	*Rapporteur*
Ciuperca, Gabriela	Maître de conférence	*Encadrante Lyon1*
Meizel, Dominique	Professeur	*Rapporteur*
Goldman André	Professeur	
Tattegrain-Veste, Helene	Chargé de Recherche	*Encadrante INRETS*

Remerciements

Je tiens à remercier Christian Mazza pour avoir accepté de diriger cette thèse et Corinne Brusque pour m'avoir accueilli au sein du LESCOT.

Je tiens à exprimer ma reconnaissance aux Professeurs Dominique Meizel et Gilles Celeux de m'avoir fait l'honneur de juger ce mémoire et d'en être les rapporteurs.

Je remercie également le professeur Goldman André pour avoir voulu d'intéresser à ce travail et pour avoir accepté de le juger.

Je remercie Hélène Tattegrain-Veste pour m'avoir offert la liberté et sa rigueur, son enthousiasme comme ses critiques, son inaltérable confiance comme son infatigable soutien.

Je remercie Gabriela Ciuperca pour m'avoir encadré dans la bienveillance et la minutie, pour sa clémence et ses précisions, et pour sa présence comme son inlassable indulgence.

Je remercie l'ensemble de l'équipe du LESCOT pour leur chaleureux accueil, en particulier Arnaud, infatigable et fidèle compagnon pour m'avoir appris la patience, Céline, amicale camarade, pour m'avoir appris l'impatience et le compromis, Olivier apôtre des dialectiques mondaines, pour m'avoir fait entrevoir les bienfaits de la tempérance, et Béa & Fred lumineuses soeurettes qui m'ont guidé dans cet obscur parcours, et l'ensemble innombrable de personnes qui se reconnaîtront et sans qui rien ne serait.

M. & M. votre courage et votre patience furent sans égales.

Bien sur, je n'oublie pas tous ceux avec qui j'ai eu l'honneur et le plaisir de traverser ces quelques années.

Table des matières

Introduction

Le développement de systèmes d'aide à la conduite est aujourd'hui l'une des voies privilégiées pour améliorer la sécurité routière. Dans cette optique, un des enjeux primordiaux de la recherche actuelle est de veiller à ce que l'introduction de ces systèmes dans les véhicules apporte un réel accroissement de la sécurité.

En effet, bien que ces technologies visent à assister le conducteur dans sa tâche de conduite, elles peuvent induire des effets antagonistes. Assister ou informer le conducteur, c'est toujours le faire dans une situation routière particulière. Cette situation peut dépendre tant du contexte routier que du comportement du conducteur dans ce contexte. Ne pas tenir compte de ce contexte et de ce comportement risque de perturber le conducteur dans des moments critiques.

Cet état de fait oriente la recherche dans ce domaine vers une approche centrée sur l'homme (Human Centred Design) plutôt que centrée sur la technologie.

Un système d'assistance sécuritaire doit pouvoir adapter l'aide procurée aux vrais besoins du conducteur [Bellet & Tattegrain-Veste, 2004]. Pour cela il doit:

1. Comprendre le contexte dans lequel se trouve le conducteur,
2. Analyser son comportement,
3. Juger de l'adéquation du comportement à la situation,
4. Adapter son assistance en fonction de ce jugement.

Ainsi, ce type de système ne peut être élaboré qu'en disposant d'une analyse temps réel du comportement du conducteur, analyse basée sur la compréhension et l'interprétation des actions de conduite. Ceci est maintenant techniquement possible sur les véhicules récents. En effet, leur informatisation et leur architecture, basées notamment sur le bus CAN[1], permettent d'accéder facilement à un grand nombre de données en temps réel, aussi bien sur la dynamique du véhicule, que sur les actions de commandes du conducteur.

Cependant, la quantité et l'hétérogénéité de ces données rendent difficiles leur interprétation et leur utilisation dans des systèmes d'assistance.

Aussi une des voies possibles pour analyser ces données est la mise en correspondance des connaissances en sciences humaines sur le conducteur et des connaissances sur l'analyse des paramètres recueillis au sein des véhicules.

En effet, d'une part, un grand nombre de recherches en sciences humaines a permis de mettre en exergue les facteurs permettant de décrire l'activité de conduite. Ces recherches ont, de plus, permis d'élaborer des modèles cognitifs du conducteur facilitant l'analyse de l'activité de conduite.

D'autre part, des recherches plus récentes sur la modélisation de signaux ont permis de définir des modèles mathématiques efficaces pour analyser les données recueillies sur les véhicules.

De notre point de vue, les limites rencontrées par ces dernières recherches peuvent être

1 Controller Area Network, réseau intra-véhicule permettant l'accès aux données du véhicules de façon aisée.

repoussées par une meilleure prise en compte des connaissances des sciences humaines dans les modélisations effectuées.

L'objectif de cette thèse est donc d'établir un cadre d'analyse des signaux numériques issus des véhicules, et de modéliser leur évolution dans différentes situations de conduite rencontrées par le conducteur. Pour cela, nous nous appuierons:

- Sur les connaissances en sciences humaines pour structurer l'analyse de l'activité de conduite,
- Sur une modélisation probabiliste de l'évolution des signaux via le modèle de Markov caché.
- Sur l'utilisation des modèles de ruptures multi-phasiques afin de segmenter l'activité.

D'un point de vue applicatif, l'objectif est d'établir une méthodologie d'analyse de l'activité de conduite focalisée sur le problème suivant: à quelle situation ou à quel groupe de situations se rapportent un enregistrement de données de conduite inconnu ? Cette méthodologie rendra possible la conception d'un module de catégorisation de l'activité de conduite, par rapport à un jeu de capteurs donné, tant en temps réel, pour être utilisé dans des systèmes d'assistance, qu'a posteriori, pour être utilisé dans les études sur le comportement.

Dans un premier temps, afin d'appuyer notre étude sur les résultats des recherches en sciences humaines sur l'activité de conduite, nous présenterons les modèles cognitifs du comportement du conducteur. Puis l'étude des méthodes numériques généralement utilisées pour analyser les signaux portant sur la conduite nous amènera à nous intéresser plus particulièrement au modèle des chaînes de Markov cachées et aux modèles de ruptures.

Dans une deuxième partie, nous présenterons les algorithmes efficaces pour manipuler les chaînes de Markov cachées et nous en proposerons une extension « le modèle Semi-Markovien caché pondéré » puis nous étudierons les modèles de ruptures multi-phasiques et en particulier les propriétés des estimateurs par maximum de vraisemblance dans les cas linéaire et non-linéaire.

Enfin, la troisième partie sera consacrée à la présentation de la méthodologie utilisée pour modéliser l'activité de conduite, à l'expérimentation effectuée, ainsi qu'aux différents résultats de cette recherche.

1 Analyser le comportement du conducteur dans une situation de conduite.

Afin d'étudier le comportement du conducteur, il est nécessaire de comprendre comment ce dernier organise ses activités cognitives et motrices, en vue d'interagir efficacement avec son environnement. C'est pourquoi, en nous appuyant notamment un modèle cognitif du conducteur, le modèle COSMODRIVE, nous nous intéresserons tout d'abord à la structure de sa cognition lors de l'activité de conduite.

Cette étude nous amènera alors à nous interroger sur les différents moyens expérimentaux, permettant d'investiguer le comportement du conducteur.

La mise en correspondance, par le biais de modélisation mathématique, de ces deux types de données, fera l'objet de notre troisième section. Nous verrons ainsi comment des systèmes, en s'appuyant notamment sur les modèles de Markov cachés, permettent d'appréhender certains aspects de la cognition du conducteur.

1.1 Modélisation cognitive du conducteur

1.1.1. Introduction

L'activité « conduire un véhicule automobile » est une activité d'une haute complexité. En effet, pour réaliser ses objectifs (arriver à destination plus ou moins rapidement, réduire ses risques d'accident...), le conducteur doit gérer un environnement potentiellement dangereux comportant un très grand nombre de caractéristiques (densité du trafic, météorologie, géométrie de la route, position et vitesse par rapport aux autres véhicules...). De plus, ces caractéristiques évoluent en fonction du temps et de l'espace.

Dans le cadre de cette activité, le conducteur est en interaction constante avec les autres véhicules, interprétant et modifiant, par son activité, les actions des autres automobilistes.

Par ailleurs, pour gérer son activité de façon aussi sécuritaire que possible, il doit également tenir compte de ses propres capacités (fatigue, perception…).

Enfin, une autre caractéristique de l'environnement routier est qu'en plus d'être dangereux et complexe, le conducteur doit y agir sous des contraintes temporelles fortes.

Du fait de l'ensemble de ces facteurs, et du nombre considérable de déplacements automobiles, on pourrait s'attendre à un nombre élevé d'accidents. Pourtant, on comptabilise proportionnellement au nombre de kilomètres parcourus, peu de situations accidentogènes et encore moins d'accidents.

Ainsi, l'accident est un phénomène rare. En effet, en 2003, en France, sur les 553 milliards de kilomètres parcourus, on compte 2.341.000 accidents [La documentation Française, 2004] (soit

0.0004 accident au kilomètre) [1].

La représentation pyramidale qualitative de Hydèn, cité par Van der Horst, [Van Der Horst, 2005] symbolise clairement la hiérarchie des fréquences relatives des différents évènements routiers (illustration 1.1). Le plus souvent, le conducteur gère non seulement les situations de conduites rencontrées, mais aussi la plupart des situations accidentogènes.

Il faut en conclure que l'être humain est en général un système efficace. Ainsi, il est capable de comprendre l'environnement dans lequel il se trouve, d'analyser les actions possibles, d'établir des scénarios probables pour le futur, puis de prendre des décisions en temps réel pour les mettre en oeuvre.

Cette efficacité est d'autant plus remarquable que pour le moment aucun système expert n'est

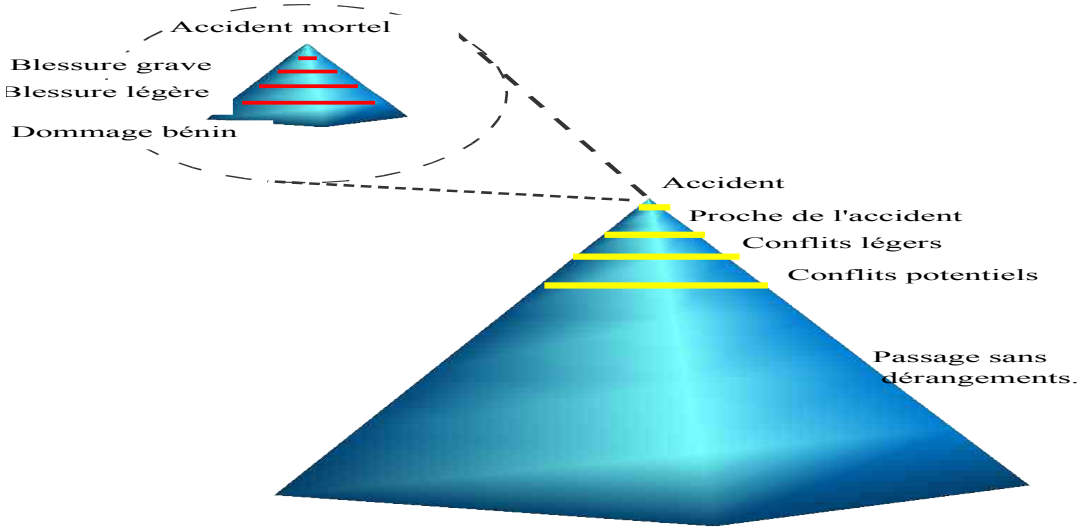


Illustration 1.1 : Hiérarchie des évènements routiers (Hydèn)

parvenu à égaler le conducteur humain dans la totalité des situations de conduite. Ainsi, malgré les différentes études sur les voitures autonomes [Forbes *et al.*, 1995], aucune de celles-ci n'a pu être mise en condition réelle de conduite de façon sécuritaire.

Afin de comprendre comment l'être humain gère aussi efficacement son activité de conduite, mais aussi pourquoi parfois il est défaillant, de nombreux travaux ont porté sur l'analyse de son comportement. Pourtant, au contraire d'une approche globale, on trouve dans la littérature un grand nombre d'études ponctuelles analysant les différents aspects de la conduite automobile. Suivant le niveau d'analyse, le conducteur est appréhendé sous différents aspects :

- Historique, à travers l'étude des différences de conduite entre générations,
- Social, à travers l'étude des interactions entre usagers,
- Individuel, à travers l'étude des bilans pré/post formations, ou des processus d'acquisition de connaissances [Kampfe, 2005],
- Cognitif, à travers l'évaluation des processus mentaux mis en place pour gérer l'activité de conduite,
- Opérationnel, en assimilant, par exemple, la trajectoire de la voiture lors d'un virage, à l'optimisation d'une fonction liée à la vitesse, à la position désirée en sortie de virage, et à la

1 Le nombre d'accident est calculé selon le nombre de dossiers ouverts, en sachant que chaque accident entraîne en moyenne l'ouverture de 1,72 dossiers [La documentation Française, 2004] et que les accidents bénins, n'entraînant pas d'ouverture de dossiers, ne sont pas comptabilisés.

vitesse et la position actuelle [Jurgensohn, 2005].

C'est sur l'aspect cognitif que nous focaliserons notre première partie. Nous présenterons alors divers modèles cognitifs du conducteur puis les différents découpages pertinents de l'activité qui en découlent.

1.1.2. Structure de la cognition du conducteur.

Une étape importante dans la modélisation de la cognition du conducteur fut la modélisation par Michon en 1985 de l'activité de conduite comme une hiérarchie de tâche [Michon, 1985].

a Historique de la modélisation cognitive du conducteur.

Afin d'appréhender comment l'être humain gère son activité de conduite, les recherches se sont tout d'abord focalisées sur l'analyse des tâches de conduite. Cette analyse était faite sans tenir compte des différentes caractéristiques propres au conducteur (âge, compétences, expérience...). A cette époque, McKnight et Adam recensent des comportements de conduite, sans les lier avec ces caractéristiques [McKnight & Adam, 1970].

Puis la nécessité d'améliorer la sécurité routière amena à vouloir mieux comprendre les accidents. C'est ainsi que Näätänen et Sumala, puis Van der Molen et Bötticher étudièrent la notion de risque et de prise de risque chez le conducteur [Naatanen & Summala, 1974],[Van Der Molen & Botticher A.M.T., 1988]. Ceci les amena à décrire mathématiquement les processus de jugement et de prise de décision.

Enfin, dans les années 1990, avec l'accroissement de l'accessibilité et de la puissance des ordinateurs, sont apparus les modèles quantitatifs, pouvant être testés informatiquement.

Ces modèles sont de deux sortes :

1) les modèles tentant d'apporter des connaissances ponctuelles sur le comportement du conducteur (analysant les corrélations entre les caractéristiques des conducteurs et leur comportement général),

2) les modèles se dirigeant vers la construction d'un modèle complet de la cognition du conducteur, c'est à dire simulant « les processus et les états internes du système cognitif en interaction avec son environnement » [Bellet, 1998].

b La modélisation hiérarchique de Michon.

Michon propose de subdiviser l'activité de conduite en trois niveaux :

- Le niveau stratégique (Strategic Level),

  qui prend en compte les tâches de *planification de la conduite* et de la *programmation de l'itinéraire,*

- Le niveau tactique (Manoeuvring Level),

  qui intègre les processus d'*analyse de la situation* routière, de *prise de décision* et de *planification des actions de conduite* (choix d'un but) à engager dans le contexte situationnel du moment (e.g. réaliser ou non un dépassement)

- Le niveau opérationnel (Control Level),

Qui concerne la *planification détaillée* des *actions de conduite* (changer de vitesse, garder un cap, freiner) sélectionnées au niveau tactique.

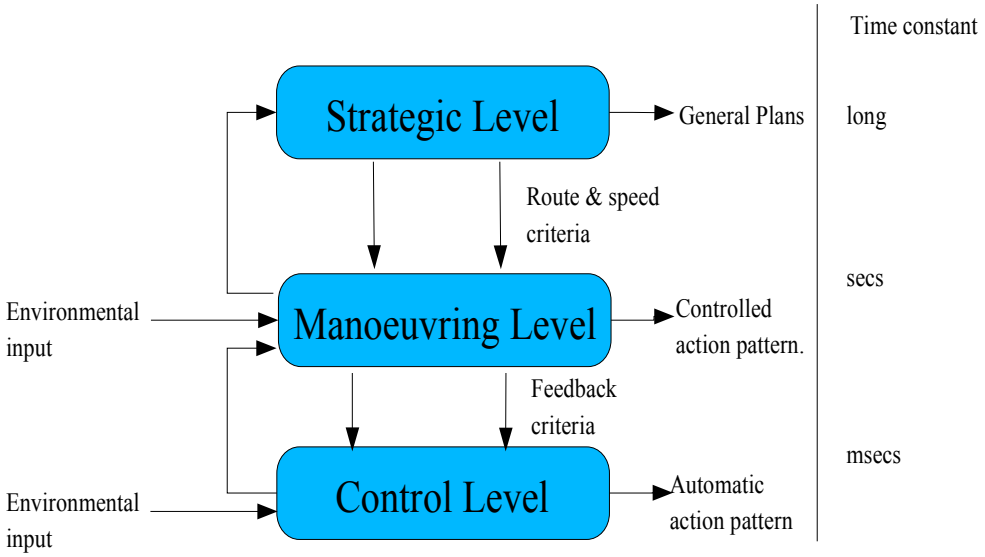


Illustration 1.2 : Représentation des niveaux stratégique, tactique et opérationnel selon Michon (1985)

Cette modélisation offre un découpage généraliste de l'activité de conduite. Ainsi, elle fournit aujourd'hui le cadre d'un grand nombre d'études sur le comportement du conducteur.

Carsten [Carsten, 2005] rappelle que le conducteur peut être considéré, selon les études, comme :

- Un système de choix basé sur le compromis entre risque et utilité des actions de conduite (motivational models). Ce modèle est explicatif et ne peut prédire le comportement du conducteur,
- un système d'input-output-feedback (adaptive control models), basé sur l'idée que le conducteur est un système déterministe adaptant sa conduite en fonction de données perçues dans l'environnement. Ce genre de modèle décrit bien les déviations tolérées par le conducteur dans la position latérale, mais ne rend pas compte des erreurs non délibérées.
- un système de règles formelles (production models), ici ce type de modèle décrit bien comment un conducteur mène une action (changer une vitesse), mais non pourquoi il la mène. Ce type de système s'attache donc à offrir une description du niveau opérationnel.

Le modèle COSMODRIVE, initié par Bellet [Bellet, 1998] reprend les apports de ces différents travaux dans une description « temps réel » des processus cognitifs mis en jeu.

c Le modèle COSMODRIVE.

Le modèle Cosmodrive est un modèle cognitif computationnel[1]. Son but est de simuler informatiquement les processus cognitifs impliqués dans l'élaboration des représentations et des prises de décisions dans le cadre de la conduite automobile.

COSMODRIVE *modélise la cognition humaine* lors de l'activité de conduite via une décomposition en sept modules, qui se comporte comme un système multi-agents. Ces modules correspondent, en partie, à une division fonctionnelle des activités cognitives du conducteur.

Bellet reprend ainsi les concepts liés aux différents *niveaux stratégique, tactique et opérationnel* décrits par Michon et les intègre dans COSMODRIVE sous forme de trois modules.

De plus, il ajoute quatre autres modules :

- Le *Module « Contrôle et Gestion »* rend compte des mécanismes d'allocation des ressources attentionnelles du système cognitif humain,
- Les modules « *Perception »* (traitement des informations sensorielles) et « *Exécution »* (réalisation des actions de conduite, *via* le véhicule) interviennent plus directement dans les interactions entre le modèle et l'environnement routier du moment.

Le dernier module, le *Module « Gestion d'Urgence »* n'intervient qu'en cas de risque manifeste d'accident et simule les raisonnements mis en œuvre en situation critique.

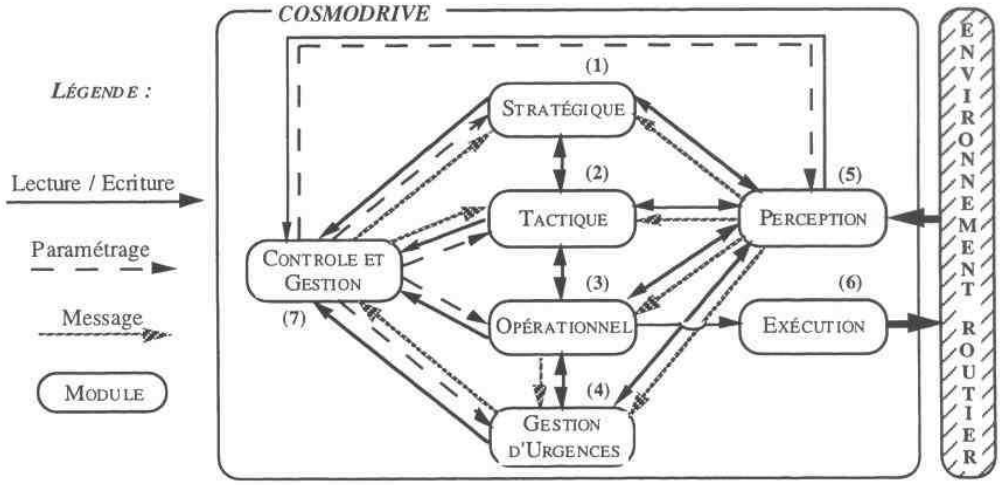


Figure 1.3 : Structure de Cosmodrive, vue d'ensemble[Bellet, 1998]

Les échanges entre les modules peuvent être de différentes natures. Ces dernières peuvent être des procédures soit de paramétrage (allocation des ressources cognitives), soit d'échange de données ou soit d'échange direct de messages entre les processus cognitifs (Figure 1.3).

La modélisation de chacun de ces modules est constituée par un système multi-agents. Chaque module intègre trois types d'objets :

1. des *Agents Cognitifs* (simulant les processus cognitifs),
2. des *Bases de Connaissances* (espace de stockage permanent correspondant à la mémoire

1 Dans le cadre de simulation de processus internes aux systèmes cognitifs, Michon [Michon J. A, 1985] définit la visée des modèles computationels comme la construction d'un « robot conducteur psychologiquement plausible ».

à long terme du conducteur),

3. et des *Tableaux Noirs* (en charge de simuler certaines structures cognitives de stockage temporaire de l'information comme la mémoire de travail, par exemple).

La Figure 1.4 représente les liens entre ces 3 types d'objets dans le module tactique. A titre d'exemple, lorsque le conducteur arrive à une intersection, le « générateur de représentation » envoie une requête à « l'agent catégorisation ». Celui-ci, en fonction des informations perçues et des connaissances préalables du conducteur sur les contextes routiers, engendre une représentation générique du lieu (« une intersection ») et du schéma de conduite à adopter (ex: « Tourner à Gauche, Changer de voie»). Le « générateur de représentation » envoie alors une requête au module de perception pour connaître les spécificités de la situation courante (position des autres véhicules, feux rouges...). Ces données lui permettent alors de construire ce que Bellet nomme une « Représentation Mentale de la Situation Courante ».

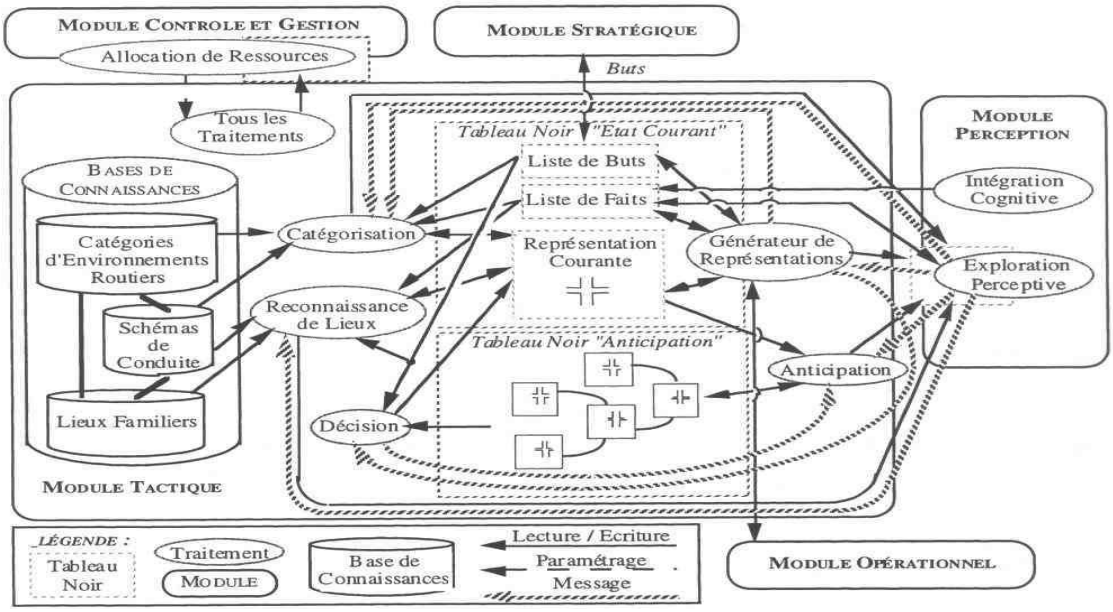


Figure 1.4 : Architecture de la représentation tactique courante [Tattegrain-Veste et al., 2001]

Il faut noter que l'ensemble de ces tâches a un «coût cognitif » et que l'être humain a une capacité cognitive limitée. Aussi l'agent « contrôleur de ressources » est chargé de décider à chaque instant à quel « agent » allouer les ressources disponibles.[Tattegrain-Veste *et al*, 2001]

A ce jour, l'ensemble de la modélisation n'est pas encore implémenté. Cependant « l'architecture du modèle a été proposée et les principaux traitements cognitifs impliqués au niveau tactique de l'activité ont fait l'objet d'une modélisation computationnelle... »[Bellet, 1998]. Aussi, cette modélisation est efficace pour décrire et comprendre le fonctionnement cognitif du conducteur.

L'avantage de cette modélisation est double. Non seulement elle offre une description structurée des différents niveaux cognitifs de l'activité de conduite, mais elle inclut également la description de l'aspect dynamique de cette activité. Ce dernier aspect est essentiel pour rendre compte des contraintes temporelles inhérentes à la conduite (liées d'une part au rapport entre la charge mentale et les capacités cognitives du conducteur, et d'autre part à l'environnement

potentiellement dangereux).

Le concept central du modèle est la notion de Représentation Mentale. En effet, le conducteur n'agit pas en fonction de l'environnement réel mais en fonction de la représentation qu'il se fait de cet environnement. Pour élaborer cette représentation, il utilise les connaissances qu'il a acquises sur les contextes routiers (et en particulier sur celui qu'il a identifié) et les informations qu'il prélève dans l'environnement.

Ceci signifie que le conducteur possède une base de connaissances des différents contextes routiers via des *Représentations Génériques* de ceux-ci (comprenant ce qu'il faut y faire, quels sont les événements possibles, comment identifier la situation, où chercher l'information pour élaborer une représentation efficace du réel), sur laquelle il s'appuie. On dit alors que le conducteur *instantie* la représentation générique par les informations décelées dans l'environnement, pour constituer la Représentation Mentale Courante.

L'ensemble de cette modélisation des connaissances est basé sur une approche spécifique de la structuration des connaissances du conducteur, centrée autour du concept de frame.

1.1.3. Les frames

a Le concept de frame

Ce concept est apparu avec la théorie des schémas dans les années 1970 par un rapprochement entre des conceptions psychologiques et des conceptions d'intelligence artificielle [Minsky, 1975]. Le but de cette notion a été de rendre compte « de la capacité de l'esprit humain à manipuler des unités mentales complexes et structurées à un niveau élevé de granularité »[Bellet, 1998].

Selon Bellet [Bellet, 1998], « Un frame (ou cadre) correspond à une structure de données mémorisées (et de traitement à ces données) qui représentent un objet typique ou une situation stéréotypée. ». Il correspond ainsi à l'ensemble des actions prévues, en fonction des évènements potentiels dans une situation donnée. Il vise à décrire l'organisation des connaissances en mémoire humaine.

L'instanciation d'un frame se fait selon un processus en trois étapes :

1) le choix d'un frame sur la base d'heuristiques et de prévisions (à partir d'autres frames). *Par exemple, le conducteur connaît l'itinéraire et sait qu'il va arriver à un rond- point. Il sait aussi qu'à cette heure le trafic est important. Dès lors, il envisage déjà la série d'actions à effectuer : freiner, regarder attentivement, tourner.*

2) l'appariement qui tente de mettre en correspondance l'information contenue dans le frame avec celle observée dans l'environnement, les valeurs par défaut étant alors éventuellement remplacées par les valeurs prélevées.

Par exemple, le trafic est moins important que prévu, la taille du rond-point est plus importante que celle imaginée. Le conducteur adapte alors son schéma de conduite en fonction de ces nouvelles données.

3) le remplacement du frame choisi si celui ne correspond pas à la réalité. Dans ce cas, il

fournit éventuellement un frame de remplacement en transférant les informations prélevées. *Par exemple, des travaux ont lieu et le conducteur doit adopter une conduite différente de celle initialement prévue.*

b Les frames dans COSMODRIVE

Au sein de COSMODRIVE, les frames sont à la base de la représentation mentale du conducteur. Ainsi, pour chaque situation reconnue, l'agent cognitif « génération de la représentation » *instantie* le frame de la situation reconnue en fonction :

- des paramètres prélevés dans l'environnement,
- de ses connaissances préalables (tant sur son parcours en particulier, que sur l'environnement routier).

Cette instanciation donne lieu à la représentation tactique courante de la situation.

Cette représentation, parce qu'elle précède et explique toute action du conducteur est au centre même de l'activité de conduite.

1.1.4. La représentation tactique courante.

a Présentation

La représentation tactique courante est instanciée par le module tactique via la base des frames (ou schémas) de conduite. Nous reprenons ici les résultats de Bellet [Bellet, 1998].

Ainsi, dans le cadre de la modélisation cognitive, cette base contient l'ensemble du savoir opératif du conducteur, c'est à dire l'ensemble des séquences d'action à accomplir pour remplir un but particulier dans une infrastructure particulière (dans ce cadre, une infrastructure peut aussi bien être un environnement routier qu'un des lieux familiers du conducteur.).

Le triptyque (but, infrastructure, et séquence d'actions) constitue un schéma de conduite. Pour mener à bien les différentes actions du conducteur, le schéma de conduite repose sur un découpage de l'environnement, pour le conducteur, en différentes zones fonctionnelles. Ces zones peuvent être:

- *des zones perceptives* où le conducteur peut chercher de l'information utile pour gérer la situation (paramétrage de la représentation, détection d'évènements susceptibles de se produire dans certaines sections de l'espace routier),
- *des zones de déplacements* où le conducteur peut faire évoluer le véhicule suivant des séquences d'actions déterminées.

  Chaque zone de déplacement est ainsi caractérisée par un état initial (position et vitesse du véhicule en entrée de zone), un état but (position et vitesse en fin de zone) et par une série d'actions élémentaires à réaliser. A ces actions sont aussi associées des conditions portant sur des événements probables (exemple : présence d'objet).

L'ensemble des zones de déplacement constitue alors la trajectoire du véhicule permettant de passer de l'état initial de la situation à l'état final souhaité.

La Figure 1.5 illustre cette représentation lors d'un Tourner à gauche.

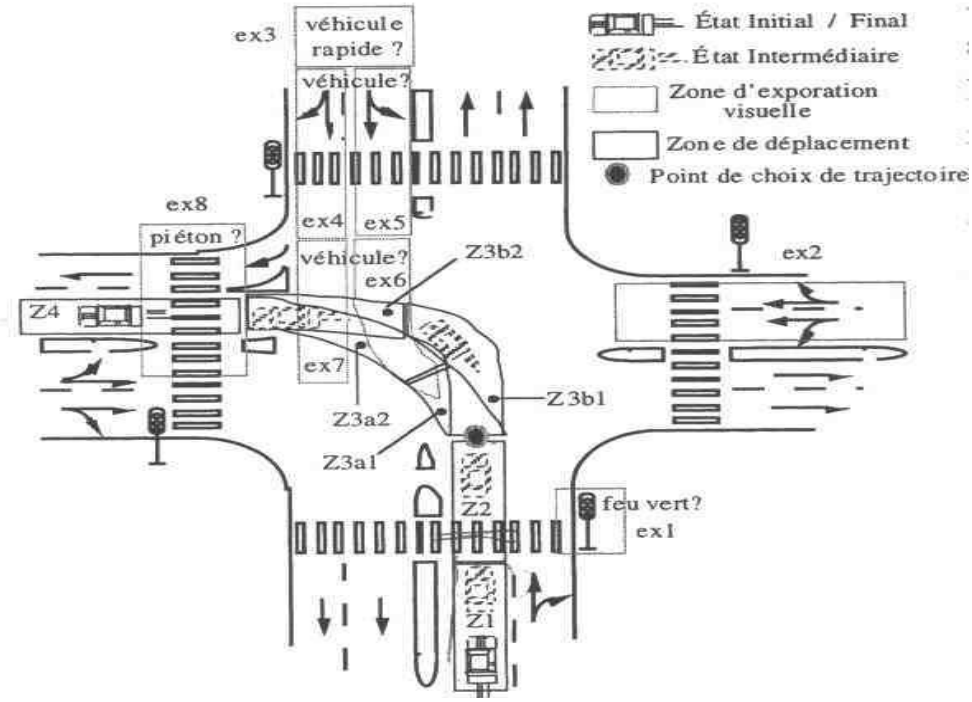


Figure 1.5 : Frame du tourne à gauche dans carrefour à feux [Bellet, 1998].

Ainsi, en arrivant dans un lieu, le conducteur a une pré-connaissance de l'objectif qu'il doit achever et des actions qu'il doit entreprendre pour réaliser son objectif. Les diverses informations qu'il pourra recueillir, modifieront sa perception du lieu et donc les actions qu'il accomplira pour réaliser au mieux son objectif principal.

Chaque action réalise ainsi un sous-objectif comme atteindre telle zone à telle vitesse.

b Les catégories d'environnement.

Ainsi, le frame instantié par le conducteur dépend de la catégorie d'environnement identifié. On peut donc légitimement s'interroger sur les différentes catégories d'environnement identifiables par le conducteur.

Il existe plusieurs manières d'envisager ce problème :

- soit on relie les différentes situations de conduite par une ressemblance perceptive pour le conducteur (même aspect),
- soit on les relie en fonction des similarités de comportement dans la situation.

Fleury et al [Fleury, *et al* 1993], cités par Bellet, montrent que ces deux types de catégorisation peuvent se recouper lorsqu'elles reposent sur « des configurations d'indices

signifiants univoques qui contribuent à une définition « claire » des représentations tant du point de vue des caractéristiques génériques (habitat, nature du paysage...) qu'en relation avec l'usage [qui est fait] de la voie qui traverse cet espace» ou en d'autres termes, lorsque dans l'environnement il y a des signes, non ambigus, définissant où le conducteur se trouve et ce qu'il doit faire.

Ces catégorisations se disjoignent lorsque dans le site il y a des contraintes importantes pour l'activité de conduite (présence d'autres véhicules, de piétons, feux...) [Fleury *et al.*, 1993] p 139].

Mazet [Mazet, 1991] note ainsi que « les intersections en rase campagne » ainsi que les « entrées et sorties de périphériques ou d'autoroutes » sont clairement identifiées dans les deux types de catégorisations. Par contre, les situations « carrefour en ville » et « carrefour en rase campagne » ne sont pas nettement différenciées. En effet, ces dernières possèdent d'autres propriétés décisives pour l'activité de conduite (présence de véhicule, feux, obstacles...). Ces diverses propriétés ont pour conséquence de séparer ces deux situations en de multiples sous catégories (avec obstacles, avec feux...). Ces sous-catégories, mêlant des situations en ville et en campagne, sont alors plus homogènes au niveau comportemental.

Dans ce cadre, Bellet a eu une approche plus systémique de cette catégorisation. Huit sujets ont eu à classifier 400 photographies représentant un large panel de scènes routières. Les sujets devaient opérer cette classification en s'imaginant au volant d'un véhicule. Puis, ces sujets devaient donner les raisons de leur choix.

Une hiérarchie de partition est apparue commune à tous les sujets.

Cette partition utilise « des critères d'ordre descriptif et concerne surtout les abords du site ainsi que certaines caractéristiques de l'infrastructure (dimensions, aménagements divers comme les glissières de sécurité, certains marquage au sol). »

D'autres critères peuvent aussi intervenir comme la présence « d'événements potentiels ayant une forte ou faible probabilité d'occurrence dans chacun de ces univers », la prise en compte des « règles du code de la route propres à chacun de ces univers », et des « risques potentiels qui leur sont plus spécifiques »[Fleury *et al.*, 1993].

Cette hiérarchie se découpe en quatre niveaux : catégories d'univers routiers, catégories de routes, catégories d'infrastructure et catégories de situations. Ces catégories tiennent d'abord compte de l'environnement puis du type de route, puis de l'infrastructure routière, et enfin de la familiarité du lieu (voir Figure 1.6).

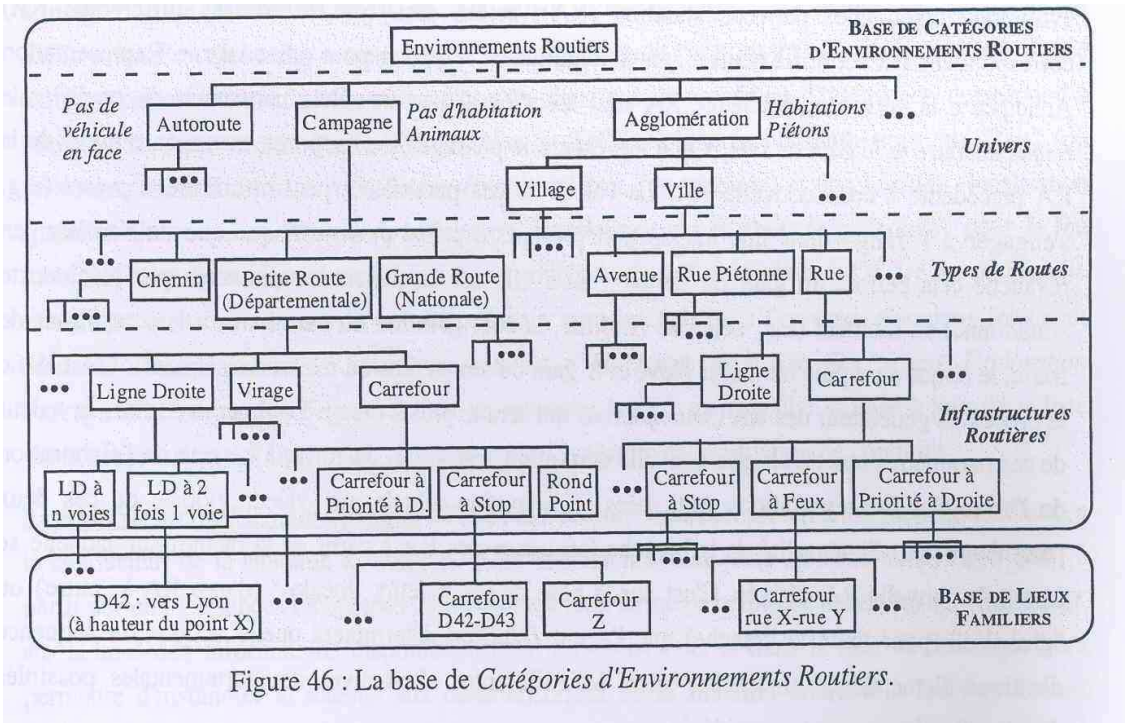


Figure 46 : La base de *Catégories d'Environnements Routiers*.

Figure 1.6 : Catégorisation des évènements routiers. [Bellet, 1998]

Remarque sur les lieux familiers : On remarque que même si collectivement les catégories d'environnements routiers identifiées sont homogènes, les connaissances spécifiques sur certains lieux familiers mettent en exergue les connaissances individuelles de chacun des sujets.

Dans ces lieux, le conducteur a un savoir-faire hautement spécialisé. Pour gérer au mieux la situation de conduite, il n'utilise pas les connaissances génériques sur l'infrastructure et la situation mais ses connaissances plus précises sur le lieu, ce qu'il faut y observer et les différents évènements qui peuvent survenir.

L'image opérative qu'aura le conducteur (et donc les actions qu'il mettra en oeuvre) dans un lieu qui lui est familier ne pourra qu'être différente de celle des autres conducteurs non familiers des lieux.

Ainsi, les résultats apportés par les sciences cognitives et décris dans ce chapitre (infrastructure perçue, division de l'activité suivant l'objectif tactique, zones de déplacements et actions dépendantes d'évènements potentiels) permettent de hiérarchiser l'analyse de l'activité de conduite. Comme nous le verrons dans le chapitre 3, ces concepts serviront à structurer notre modélisation de l'activité.

1.2 Les moyens d'investigation du comportement

Cette thèse a pour but d'établir une correspondance entre l'activité de conduite et les données pouvant être recueillies à bord de véhicules instrumentés. Aussi, il est primordial de nous intéresser d'une part, à la nature des données pouvant être enregistrées et d'autre part, à leur indication sur l'activité.

1.2.1. Les différents niveaux d'enregistrement.

Les nombreuses études dans le domaine de l'analyse de la conduite ont révélé l'importance des variabilités inter-individuelles et inter-situationnelles.

Ces études montrent qu'il est primordial d'enregistrer l'activité de conduite à 3 niveaux distincts : au niveau du conducteur, du véhicule et au niveau de l'environnement.

a Le conducteur

Au niveau du conducteur, l'activité de conduite se manifeste tout d'abord par *ses actions sur les organes de contrôle et de commande.* Ces actions, même si elles peuvent être une réaction directe à l'environnement immédiat, sont le plus souvent témoins d'une anticipation par rapport à une situation future prévue par le conducteur.

En même temps qu'il agit, le conducteur est aussi en train de chercher des informations sur son environnement. Il importe alors de comprendre ce qu'il perçoit de celui-ci. L'élaboration d'une image mentale de l'environnement s'effectue en utilisant des informations visuelles (et dans une moindre mesure des informations sonores). Aussi, l'étude des regards (par oculométrie ou par exploitation des enregistrements vidéos) et l'étude des réactions corporelles (gestuelles...) peuvent être essentielles (illustration 1.7).

Illustration 1.7 : L'étude des regards peut être extrêmement importante : dans cet exemple le conducteur arrive face à un véhicule lent. Son regard témoigne d'une attention soutenue. Elle peut être interprété comme une prise de conscience du danger potentiel.

Ce type de données a permis par exemple d'étudier :

- l'utilisation de certains systèmes sur l'activité de conduite (voir notamment Bruyas et al sur l'impact du téléphone au volant [Bruyas et al, 2005])
- la possibilité de l'introduction de systèmes d'assistances adaptatifs [Tattegrain-Veste *et al.*, 2004].
- les différences de comportements entre plusieurs types de populations.

  Par exemple en 2005, Etienne et al étudient l'effet du vieillissement et l'effet de la maladie d'Alzheimer sur le comportement de conduite [Etienne & Marin-Lamellet, 2005].

Enfin, le conducteur n'a pas toujours son attention dirigée vers la conduite, aussi l'étude des données physiologiques permet de diagnostiquer le niveau de vigilance du conducteur par rapport à sa tâche de conduite. Morel et al étudient ainsi l'évolution des réponses électrodermales lorsque des tâches supplémentaires sont effectuées en conduisant [Morel *et al.*, 2005].

b Le véhicule

En complément des données recueillies sur le comportement du conducteur, les données sur la dynamique du véhicule jouent un rôle essentiel dans l'analyse de l'activité.

Elles nous renseignent sur la vitesse et l'accélération du véhicule suivant les axes longitudinaux et latéraux.

Il faut noter qu'il y a une redondance d'information entre les mesures sur les actions de contrôle et celles sur la dynamique du véhicule. En effet, Peltier [Peltier, 1993] signale que, « la fonction de transfert liant actions de contrôle et réponse du véhicule agit comme un filtre passe-bas » et que « le signal de sortie est donc une version lissée de la commande ». Ainsi, les micro-corrections apportées par le conducteur tant sur la trajectoire que sur la vitesse sont souvent trop faibles pour se refléter sur la dynamique du véhicule. De plus, du fait de l'inertie du véhicule, sa dynamique est influencée par les actions du conducteur avec un retard non négligeable

Pourtant, bien qu'il y ait une perte d'information brute entre les deux types de données, celles sur le véhicule nous permettent d'interpréter a posteriori, plus facilement qu'avec les actions de contrôle, l'objectif poursuivi par le conducteur (accélération, changement de cap...). En effet, une augmentation de la vitesse est directement déchiffrée par l'analyse de la dynamique du véhicule, alors que, du fait de sa plus forte variation, l'analyse de la pédale d'accélérateur est plus difficile.

Aussi, dans la visée d'un système de diagnostic temps réel de l'activité de conduite, si l'étude des actions de contrôle permet d'avoir un diagnostic précoce, l'analyse des changements de dynamique du véhicule rendra le diagnostic plus robuste.

c L'environnement

Le dernier point dans l'étude de l'analyse de l'activité est l'enregistrement du contexte dans lequel a lieu l'activité de conduite.

Ce contexte peut être caractérisé par trois types de critères :

- à long terme, par toutes les variables ne changeant pas pendant l'expérimentation (météorologie, conduite diurne/nocturne...),

- à moyen terme, par les variables relativement stables mais pouvant changer pendant l'expérimentation (trafic, contexte routier (ville, campagne, autoroute)...),
- à court terme, par les variables définissant l'environnement immédiat du conducteur (nombre et position des voitures à ses cotés, infrastructure).

Par le passé, l'ensemble de ces données était codé « manuellement » par l'expérimentateur en temps réel ou a posteriori à l'aide des enregistrements vidéo. Dans un futur proche, la fiabilisation du GPS associé aux bases cartographiques permettra la saisie automatique de nombre de ces facteurs (contexte urbain, infrastructure...). De même, les récentes innovations technologiques (laser à balayage, stéréovision) permettront d'appréhender l'environnement proche du conducteur de manière plus automatisée (détection d'obstacles). La catégorisation des environnements routiers deviendra plus simple et offrira alors la possibilité d'études à grande échelle sur le comportement du conducteur.

Remarque : Nous nous sommes focalisés ici uniquement sur les études en environnement réel. Les études menées en simulateur de conduite sont différentes. Elles ont l'avantage d'avoir non seulement un environnement déterminé, mais aussi de pouvoir, de par leur nature, enregistrer l'ensemble des interactions entre le conducteur et l' environnement. Cependant, il n'est pas certain que les actions de conduite soient les mêmes dans un simulateur qu'en situation réelle.

d Relation entre les niveaux d'enregistrement

Le rôle du capteur sur les organes de commande est de nous montrer quelles sont les actions physiques effectuées par le conducteur. Pourtant, celles-ci ne peuvent prendre sens qu'en comprenant d'une part dans quel contexte elles sont effectuées (comment est l'environnement) et d'autre part comment elles influencent l'état du véhicule (vitesse et position) dans cet environnement.

Ce triptyque nous permet alors de comprendre la relation entre les différents niveaux

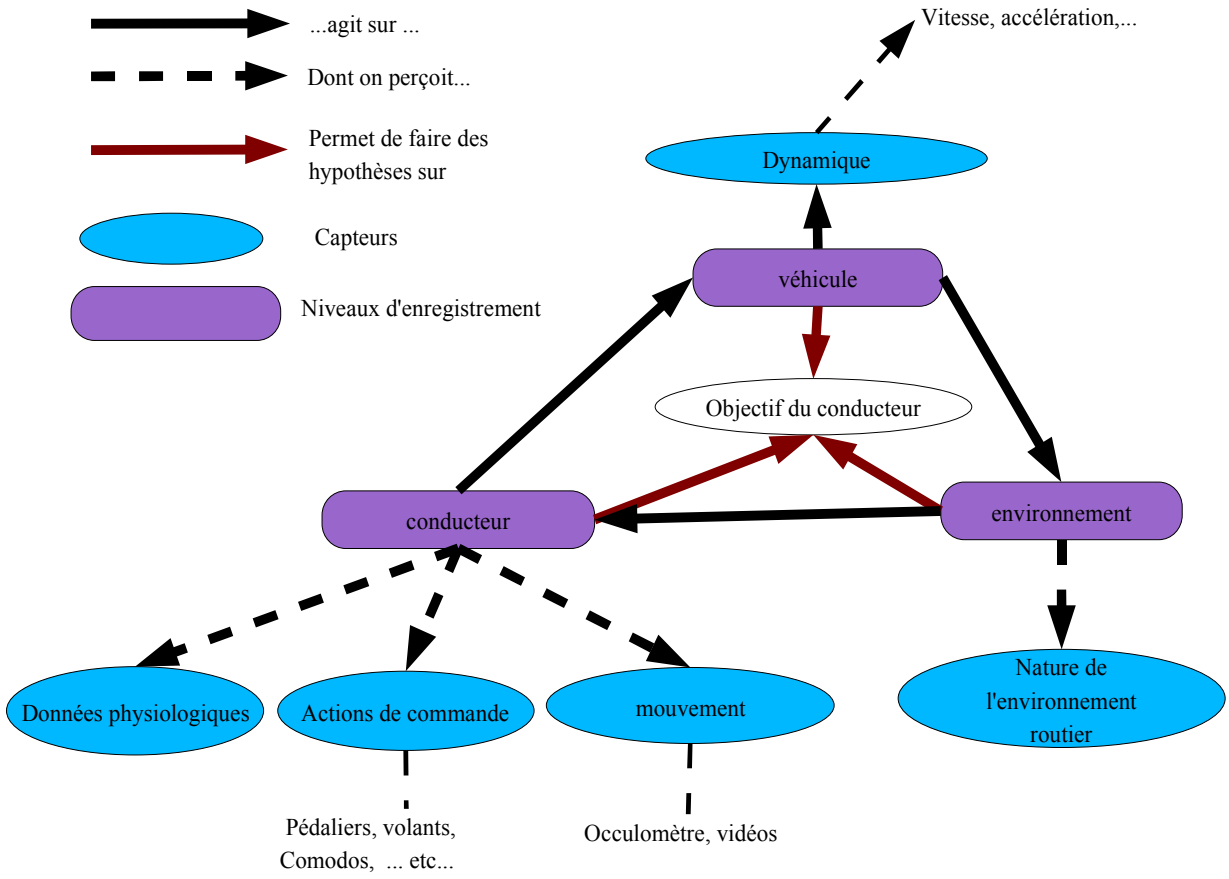


Figure 1.8 : Relation entre les différentes catégories de capteurs et les hypothèses pouvant être faites sur la cognition du conducteur.

d'enregistrement et de faire des hypothèses sur l'objectif du conducteur (Figure 1.8). Lors de notre étude sur les données de conduite, c'est donc sur ce schéma que nous nous appuierons pour interpréter l'objectif poursuivi par le conducteur.

Cette vision, conçue pour l'étude de signal, est fondée sur la conception du véhicule en tant qu'objet ayant des caractéristiques propres (vitesse, accélération), non nécessairement lié avec les actions de commande du conducteur.

Elle est donc différente d'un modèle classique, le DVE (Driver-Vehicule-Environment), s'intéressant aux rapports entre le conducteur et le véhicule et considérant le signal sur les pédales comme une caractéristique du véhicule [Cacciabue & Hollnagel, 2005].

1.2.2. Accessibilité des mesures.

Au delà des trois catégories préalablement établies, il semble aussi important de déterminer une autre classification basée sur l'utilisation des données issues capteurs dans des systèmes embarqués ou à titre expérimental.

En effet, bien qu'un grand nombre de capteurs nous permet de mieux comprendre le comportement humain, certains sont inutilisables dans un véhicule autre qu'expérimental. Des

caméras rendant compte de l'expression du visage du conducteur sont un exemple de cette dichotomie entre une utilisation expérimentale et au sein d'un système embarqué. En effet, l'analyse par leur biais est difficilement automatisable, et leur mise en place à bord d'un véhicule particulier paraît difficile.

Cette distinction permet alors de comprendre la portée scientifique et/ou industrielle des systèmes développés.

On peut diviser ces capteurs en trois catégories :

1. Ceux dont les mesures sont disponibles, ou seront disponibles, à peu de frais dans un avenir proche. C'est le cas des capteurs utilisés pour l'ensemble des actions de commande et sur la dynamique véhicule dont les mesures se font déjà dans les véhicules récents et dont l'accessibilité est aisée via le bus CAN[1]. C'est aussi le cas des systèmes GPS et des bases de données géographiques dont la généralisation croissante permet d'envisager une utilisation pour les systèmes d'assistance dans un avenir proche.
2. Ceux dont les mesures sont utiles en laboratoire pour l'analyse de l'environnement du conducteur, mais dont l'exploitation par des systèmes experts n'aura pas d'application à court terme dans le parc automobile. Caméras en stéréovision, télémètre à balayage, radar sont des exemples de ce type de capteurs.
3. Ceux qui nous renseignent sur le comportement humain, mais qui sont inexploitables pour l'implémentation de systèmes d'assistance : caméra, occulomètre mais aussi système d'enregistrement des données physiologiques.

Plus les systèmes conçus utiliseront des capteurs complexes, c'est-à-dire appartenant aux dernières catégories, plus les résultats seront performants, mais plus leur diffusion sera lente.

1.3 Modélisation de l'évolution des capteurs lors d'une situation de conduite.

1.3.1. Points de vues sur le découpage de l'activité de conduite : quelques définitions.

Analyser l'activité du conducteur, c'est avant tout segmenter l'activité de conduite en différentes séquences de conduite.

Suivant les études et le point de vue choisi, ces divisions peuvent être différentes.

Elles ont toujours comme spécificité d'être homogènes suivant des critères particuliers. Généralement, ces critères sont définis suivant :

- l'objectif tactique du conducteur (situation de conduite subjective)(Bellet),
- les caractéristiques objectives de la situation (situation de conduite objective)(Kumagai).

Nous introduisons un troisième type de critère, qui nous apparaît utile lors de l'utilisation de données véhicule :

1 Controller Area Network : ce bus de données, développé par Bosh dans les années 80 pour les communication série dans les véhicules automobiles, a fait l'objet d'une normalisation Iso et est maintenant couramment répandu. Il permet l'accès aux données du véhicule de façon aisée.

- Les caractéristiques mesurées par le biais de capteurs numériques[2] (situation de conduite mesurée).

Chaque type de segmentation de l'activité associe alors à une séquence de conduite, une série de caractéristiques. Ces caractéristiques permettent de regrouper les séquences en différentes catégories. Nous appellerons alors « situation de conduite » un regroupement de séquences de conduite, fonction de ces caractéristiques.

a Séquence et Situation de Conduite Vécue

La thèse de Bailly [Bailly, 2004] est révélatrice de la distance qu'il peut y avoir entre la représentation mentale que le conducteur a de la scène routière et la situation routière objective (informations non importantes pour la conduite omises, ou informations primordiales non prises en compte...).

Cette image mentale doit donc être considérée à part entière et différenciée de la vraie situation de conduite.

Nous définissons une Séquence de Conduite Vécue comme une période continue de l'activité de conduite où le conducteur a un objectif tactique homogène et caractérisé par l'objectif du conducteur et l'environnement tel que le conducteur se le représente. Son début correspond dans COSMODRIVE, à l'instanciation d'un frame tactique, et sa fin, à l'instanciation d'un frame tactique différent.

Cette séquence est donc caractérisée par:

- les caractéristiques physiques de l'environnement *assimilées par le conducteur*. Elles peuvent être liées à l'infrastructure, au trafic, à la météorologie, à la luminosité, aux obstacles...
- l'objectif tactique du conducteur à ce moment. (exemple : tourner à droite, changer de voie...).

Une Séquence de Conduite Vécue est donc un temps de conduite délimité par l'instanciation d'un même frame tactique et caractérisé par l'ensemble des éléments ci-dessus. Une Situation de Conduite Vécue est alors le regroupement de Séquences de Conduite Vécues suivant ces éléments.

b Séquence et Situation Réelle de Conduite

Il nous faut distinguer cette «Séquence de Conduite Vécue» de la « Séquence Réelle de Conduite ». Cette dernière est définie par l'ensemble des éléments caractérisant objectivement l'environnement du conducteur, ainsi que l' objectif tactique du conducteur.

Ces deux types de situation pourraient se confondre dans le cas où le conducteur assimilerait l'ensemble des données de l'environnement pour agir, ou se disjoindre lorsque, comme souvent, il ne prend en compte qu'une partie des informations contenues dans l'environnement pour gérer son activité de conduite (exemple : passer une intersection, comme s'il s'agissait d'une ligne droite).

2 On entend par « capteur numérique », tout dispositif, à l'intérieur du véhicule, rendant compte d'une activité par un signal numérique.

Ainsi, une Séquence Réelle de Conduite est un temps de conduite délimité temporellement par un objectif tactique homogène et caractérisé par l'ensemble des éléments caractérisant objectivement l'environnement.

Une Situation Réelle de Conduite est alors le regroupement de Séquences Réelles Conduite suivant ces éléments.

c Séquence et Situation de Conduite Mesurée

En parallèle de ces 2 premières catégorisations de l'activité de conduite, une troisième catégorisation, basée sur la mesure de l'activité de conduite grâce aux capteurs, s'avère primordiale. En effet, les capteurs disposés sur les véhicules instrumentés sont limités et bruités. Ils n'offrent ainsi qu'une vue réduite de la situation réelle comme de la situation vécue. Aussi nous définissons une Séquence de Conduite Mesurée comme un temps de conduite délimité temporellement par un objectif tactique homogène et caractérisé par l'ensemble des éléments numériques accessibles par les capteurs du véhicule.

Une Situation de Conduite Mesurée sera alors un regroupement de Séquences de Conduite Mesurées de manière à ce que la distance[1] entre l'évolution des capteurs dans chaque séquence soit faible.

Ainsi, si on ne prend en compte que le volant et la vitesse, les 2 situations « arrêté au feux » et « arrêté derrière un obstacle » sont différentes « réellement » et « cognitivement » mais peuvent être les mêmes au niveau de leur mesure via les capteurs.

Cette distinction s'avérera essentielle dans l'élaboration de systèmes d'analyse numérique de l'activité humaine : une forme particulière de l'évolution des capteurs peut être reconnue. Cette forme peut être associée à un panel de situations possibles. Par contre, le diagnostic de la situation précise, avec les moyens limités du système, s'avère difficile.

Par exemple, Oliver et al rappellent que « certaines manoeuvres comme doubler et changer de voies ne peuvent être distinguées clairement en utilisant seulement les informations du véhicule. » [Oliver & Pentland, 2000].

Pourtant, le fait de déterminer que le conducteur est dans l'une ou l'autre de ces situations, même sans connaître précisément laquelle, peut être important lors de l'élaboration de systèmes d'assistance.

d Comprendre la relation entre les 3 catégorisations.

Bien que ces 3 types de catégorisation soient distincts, pour certains auteurs, le type de catégorie étudiée n'est pas toujours explicité.

Ainsi, Pentland et Liu [Pentland & Liu,1999] demande aux sujets de verbaliser leurs actions lors de la conduite. Ce protocole lui permet de connaître en partie la situation vécue du conducteur (qu'il assimile avec la Situation Réelle). Il met ensuite en regard cette situation avec des enregistrements de données sur le comportement du conducteur (c'est-à-dire la Situation Mesurée). Le principal inconvénient de ne pas expliciter clairement les différentes situations étudiées est de ne

1 Cette distance pouvant être définie de plusieurs manières (décomposition par ondelettes, Fourier, distance euclidienne...). Nous verrons dans la partie 3 une distance particulière basé sur les chaînes de Markov cachées.

pas comprendre pourquoi le diagnostic d'une situation est faux. En effet, celui-ci peut provenir d'une confusion

- dûe à l'expertise de la situation réelle (éléments importants non pris en compte par l'analyste),
- dûe à l'expertise de la situation vécue (éléments non pris en compte par le conducteur, objectif différent...),
- dûe à un déficit technologique.

Ainsi, les diagnostics de ces différentes visions de l'activité sont, par nature, entremêlé. En effet, le diagnostic de la Situation de Conduite Mesurée peut nous informer (sans que le lien soit univoque), sur la situation réelle ou vécue de conduite.

Par exemple, si avec des capteurs simples, on peut diagnostiquer la Situation Mesurée qu'un conducteur est arrêté, on ne peut pas affirmer quelle est la Situation Réelle (« arrêté à un feu », « arrêté à un carrefour en T », « arrêté devant un obstacle »...). Dans ce cas, il est alors difficile d'inférer si le conducteur a pris en compte telle ou telle caractéristique de l'environnement (piéton traversant...). Pour faire ces distinctions, l'ajout de capteurs est nécessaire (oculométrie, GPS).

La connaissance et la catégorisation d'une situation mesurée ne peuvent que restreindre le champ de recherche de la situation vécue. De la même manière, des informations sur la Situation de Conduite Réelle nous aident à comprendre la Situation de Conduite Vécue par le conducteur.

Aussi, afin d'éviter toute confusion, il est important de distinguer les 3 classes de catégorisation.

Ainsi, l'ensemble des capteurs concourt donc à la définition de la Situation de Conduite Mesurée.

Les capteurs sur l'environnement et sur la dynamique du véhicule sont primordiaux pour connaître la Situation de Conduite Réelle. Néanmoins, l'avis d'un expert par l'étude notamment des données vidéo est souvent nécessaire.

Ce dernier peut alors s'appuyer d'une part sur les diagnostics des *situations de conduite réelles* et *mesurées* et d'autre part sur les différents supports à l'expertise pour inférer celle *vécue* par le conducteur.

Ces supports peuvent être de nature différente [Wiell Janssen, 2004] : tests sur les caractéristiques cognitives des conducteurs, étude qualitative du comportement (par données vidéo), verbalisation du conducteur [Bailly, 2004].

Le schéma 1.9 résume ainsi le lien les unissant et les reliant aux différents types de capteurs.

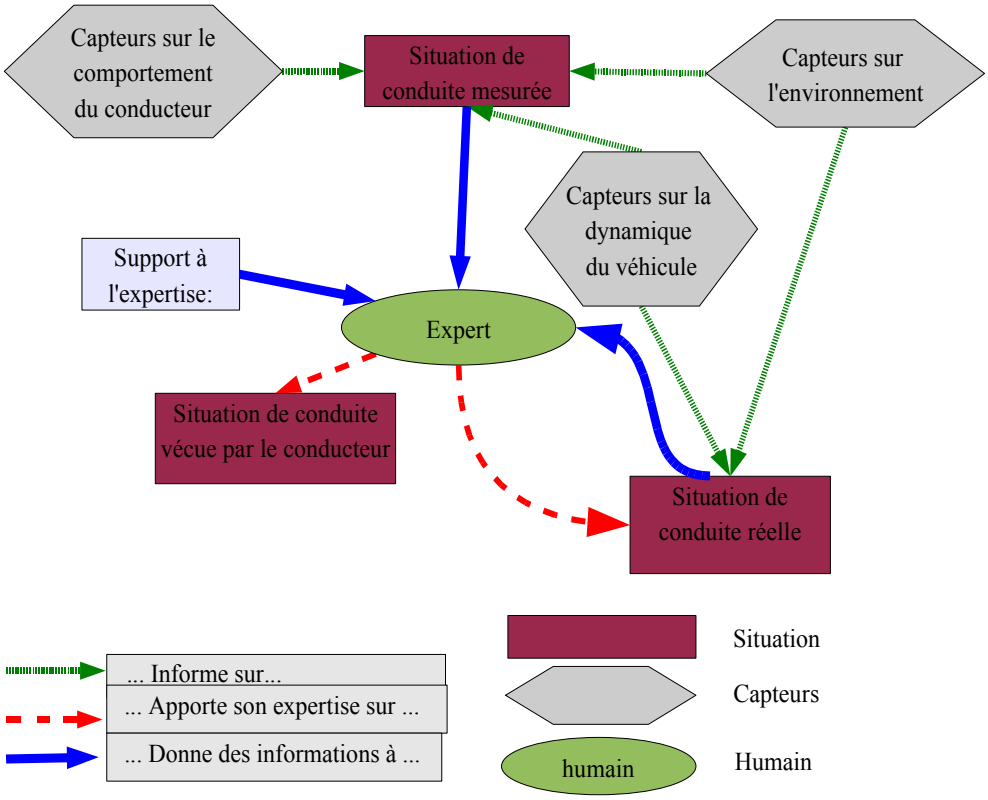


Illustration 1.9 : Relation entre les catégories de situation de conduite et les capteurs

La distinction que nous venons de proposer entre ces différentes types de situations, nous sera utile pour structurer et préciser notre analyse de l'activité de conduite.

Dans le paragraphe suivant, nous examinerons comment des études ont pu mettre en correspondance les Situations Vécues ou Réelles avec celles Mesurées.

1.3.2. Différentes approches pour la classification des situations de conduite

Différentes méthodes ont été utilisées pour modéliser l'évolution des signaux portant sur la conduite et les mettre en regard avec l'activité du conducteur.

Nous en ferons ici un rapide résumé puis nous nous focaliserons sur l'utilisation des chaînes de Markov cachées, qui présentent en terme d'interprétabilité et de performance un grand nombre d'avantages.

a Les modèles contrôle/ commande.

Ces modèles tentent de décrire la réalisation de manoeuvres ponctuelles comme la conduite en ligne droite, le suivi de trajectoire sur route sinueuse, ou les manoeuvres d'évitement d'obstacle. Ils

considèrent le conducteur comme un système physique caractérisé par une fonction de transfert liant le recueil d'informations aux actions sur les organes de contrôle du véhicule [Peltier, 1993].

Toute manoeuvre est alors conditionnée par la prise d'information. Les informations utilisées par le conducteur sont supposées essentiellement d'origine visuelle. Celui-ci échantillonne à intervalle de temps régulier sa position sur la voie, son angle de cap et les distances qui le séparent des différents obstacles et ces différentes informations sont alors traitées de manière à anticiper les événements routiers par une action adéquate sur les organes de contrôle.

Pratiquement, la modélisation se fait souvent via des modèles linéaires et non linéaires ayant pour caractéristique l'introduction d'un retard pur dans la fonction de transfert [Weir & Chao, 2005].

Ces modèles expliquent bien la trajectoire globale, mais peinent à expliquer les micro-corrections sans cesse effectuées par le conducteur. Or, ces micro-corrections peuvent révéler un changement de stratégie de contrôle.

Ces modèles sont des modèles mécaniques. Ils sont adéquats pour décrire deux aspects de la conduite : la collecte d'information et les actions de contrôle. Par contre, ils sont inaptes à intégrer l'influence des capacités d'apprentissage et d'anticipation.

b Les méthodes Bayésiennes.

Face à la complexité de l'analyse des situations de conduite, l'exploitation de méthodes « souples » de type « réseaux bayésiens » s'est avérée pertinente et efficace dans certains cas. Elles ont servi tout d'abord à classifier, en terme de dangerosité, un ensemble limité de situations (passage d'intersection, conduite dangereuse sur autoroute), sans mesurer l'impact de ces critères sur l'ensemble des situations routières (voir. Peltier [Peltier, 1993] pour l'utilisation des formes floues et Pribe & Rogers [Pribe & Rogers, 1999] pour l'utilisation des réseaux de neurones). Par contre, elles ne se sont pas avérées assez robustes pour une analyse de l'ensemble des situations routières.

Ce problème a déjà été rencontré au LESCOT (Laboratoire d'Ergonomie et de Science Cognitive pour les Transports) lors d'expérimentations antérieures. Le projet CEMVOCAS [Tattegrain-Veste *et al.*, 2004] visait à l'élaboration d'un diagnostic en temps réel de disponibilité du conducteur à recevoir des messages vocaux. Ce système, fonctionnant en conduite réelle, devait être capable de fournir un diagnostic en continu, donc de traiter l'ensemble des situations rencontrées par le conducteur. L'approche utilisée dans le projet était l'utilisation des réseaux de neurones. Elle a permis une classification de 85% des situations en terme de disponibilité. Mais, l'amélioration des taux de reconnaissance était difficile du fait du caractère « boîte noire » des réseaux de neurones. En effet, à cause de la non interprétabilité du système, l'ajout de nouveaux facteurs explicatifs était hasardeux. Bien que ces ajouts amélioraient la reconnaissance de certaines situations, ils la réduisaient pour d'autres.

c Les méthodes à base de règles

Par la simplicité de leur mise en oeuvre, les méthodes à base de règles sont souvent utilisées. Ces dernières sont construites autour de l'analyse visuelle de l'évolution des données et de la définition de critères pertinents.

Bien que cette approche permet d'obtenir un système interprétable, elle a deux inconvénients

majeurs. D'une part, l'élaboration des règles est fastidieuse (l'analyse visuelle des fichiers de données est longue, et le processus d'élaboration des règles se fait par essais/erreurs), d'autre part l'absence d'analyse automatique des données ne permet pas de faire émerger des informations sur le comportement.

Cette approche a permis, lors du projet CEMVOCAS d'obtenir les mêmes taux de reconnaissance que ceux obtenus par réseaux de neurones [Tattegrain-Veste *et al.*, 2004]. Son principal avantage était alors de permettre une amélioration des taux de reconnaissance des situations mal reconnues, sans changer le taux des autres.

d Les méthodes statistiques

Les méthodes statistiques issues de l'analyse de données (segmentation, classification, analyse discriminante) permettent de faire émerger des différences entre plusieurs groupes caractérisés par un ensemble important de données. Ces méthodes ont l'avantage d'être simples à mettre en oeuvre, d'avoir des résultats interprétables et de traiter des ensembles multidimentionnels importants.

Une étude menée au LESCOT avait pour objectif de trouver des différences entre des comportements dans des situations différentes en utilisant des méthodes statistiques classiques (segmentation, classification) [Dapzol, 2003]. L'activité d'un conducteur sur un circuit comprenant 6 situations différentes fût enregistrée et analysée (le conducteur effectuait 5 fois un parcours urbain).

L'objectif était, à partir des données numériques, de catégoriser les situations rencontrées par le conducteur. L'utilisation des méthodes statistiques classiques sur les valeurs moyennes ne furent d'aucune utilité. De plus, du fait de la présence de dilatations temporelles fortes entre plusieurs séquences d'une même situation, le prétraitement des données par filtre ne fut pas efficace.

La recherche de ruptures dans les signaux et l'utilisation des techniques simples d'analyse de données (classification hiérarchique, analyse discriminante)sur les moyennes des données sur les segments formés donna de meilleurs résultats dans des situations simples. Par contre, la reconnaissance devenait bien plus aléatoire pour des situations complexes (présence d'un autre véhicule, ou autre perturbation).

Néanmoins, les résultats obtenus ont permis de voir les limites de ces approches et nous ont incité à rechercher des modèles plus complexes.

1.3.3. Approche utilisant les chaînes de Markov cachées pour la classification des situations de conduite

Nous avons vu dans la partie 1.1.4 que les travaux en modélisation cognitive ont montré qu'une situation de conduite pouvait se découper en différentes phases, de durée variable, s'enchaînant de manière logique.

Par ailleurs, les résultats d'un domaine de recherche comportant les mêmes spécificités (bruitage, dilatation temporelle, phase s'enchaînant logiquement), la reconnaissance de la parole, nous ont orienté vers un champ de recherche efficace, les modèles de Markov cachés (MMC). En effet, le formalisme MMC convient, aussi bien pour la prise en compte des transitions entre phases, que pour sa possibilité structurelle à prendre en compte les dilatations temporelles.

Nous ne présenterons ici que les études sur ce formalisme qui nous ont paru être les plus pertinentes et les plus révélatrices de son adéquation pour étudier le comportement du conducteur.

Lors de leurs travaux sur la conduite automatisé, Forbes et al. [Forbes *et al.*, 1995] ont développé une méthode basée sur un MMC à entrée-sortie dont l'état dépend non seulement de l'état précédent, mais aussi de la décision en cours. Ils ont montré que l'architecture utilisée apportait *une solution pertinente au problème de bruit et d'incertitude au niveau des capteurs et des interactions avec les autres véhicules* et qu'il était intéressant d'associer *les MMCs avec des règles de connaissances expertes.* Ce modèle de conducteur a été testé en pilotant un véhicule dans un simulateur de trafic, et en vérifiant que le comportement de conduite était cohérent.

Un deuxième type d'application des MMCs a été réalisé par Nechyba et Xu [Nechyba & Xu, 1998] pour l'identification de stratégies de contrôle sur un simulateur. À chaque personne est associé un MMC. Les modèles sont construis sur des indicateurs de comportements, obtenus par quantification vectorielle des actions de contrôle du sujet (modélisées par leurs spectres et leurs transformées en ondelettes). La classification d'une nouvelle série de données se fait grâce à un indice de similarité entre les MMCs préalablement construits et celui de la série à classifier. Cet indice *« capable de comparer des trajectoires stochastiques, dynamiques, et multidimensionnelles »* surpasse un classifieur bayésien classique.

Un troisième type d'application s'inspire des résultats en psychologie cognitive décrivant l'activité de conduite en états successifs. Par exemple, Kuge et al ont cherché à caractériser et à détecter des manoeuvres routières. Ils montrent comment modéliser le positionnement sur la voie (même voie, changement de voie urgent, ordinaire) par une série de MMCs [Kuge et al,, 2000]. Les résultats (98,3% de reconnaissance) sur simulateur de conduite dynamique montrent l'intérêt de cette approche *pour gérer la variabilité des comportements de conduite*. Par ailleurs, Kumagaï et al [Kumagaï *et al.*, 2003] se sont intéressés aux situations potentiellement accidentogènes aux arrivées sur carrefour, en prédisant les actions futures du conducteur. Pour cela, ils ont utilisés un modèle de détection des manoeuvres d'arrêt à un carrefour. Un intérêt important de l'article de Kumagai et al. est « *l'interprétation de la topologie du modèle, en terme de phase de conduite, aussi bien au niveau des états que des transitions »*. Bien qu'elle soit liée à la simplicité des variables utilisées (la vitesse et l'enfoncement de la pédale de frein), cette interprétabilité nous conforte dans l'intérêt des MMCs pour la compréhension de l'activité de conduite.

Pentland et Liu [Pentland & Liu, 1999], quant à eux, se sont focalisés sur l'analyse d'un ensemble de séquences de conduite plus diversifié (s'arrêter à la prochaine intersection, tourner à gauche à la prochaine intersection, tourner à droite à la prochaine intersection, changer de voie, doubler une voiture, aller tout droit) sur simulateur de conduite. Ils ont supposé que le contrôle de l'humain sur le véhicule est différent suivant les phases de l'activité de conduite. Par exemple, ils divisent le changement de voie en 6 étapes successives : (1) centrer la voiture sur la voie initiale, (2) regarder si la voie est libre, (3) initier le changement de direction, (4) le changement de voie, (5) la fin du changement, et (6) recentrer la voiture sur la nouvelle voie.

À chaque étape est associé un filtre de Kalman, et chaque séquence est modélisée par un MMC dont les paramètres sont les indices d'adéquation à chacun des filtres. Ils ont montré que ce type d'approche permettait d'avoir des résultats significatifs. En effet, 1,5 seconde après le début de l'action, le taux de reconnaissance de la situation de conduite est de 95%. Cette recherche montre donc *qu'il est possible de catégoriser rapidement l'objectif du conducteur dans une infrastructure donnée*.

Dans la même optique, Oliver et al [Brand, Oliver & Pentland, 1996] enregistrent l'activité de conduite de 70 personnes pendant 1h15 chacune. Les enregistrements effectuées comprennent des données sur le véhicule et des données oculométriques. Ces enregistrements sont segmentés en situations de conduite et pour chaque situation les auteurs modélisent l'évolution conjointe de ces deux types de données par des C-MMC[1]. Leurs résultats sont encourageants. En moyenne ils peuvent prévoir la manoeuvre en cours, une seconde après son commencement.

Ces résultats nous ont incités à poursuivre dans cette voie en tenant compte des trois limites suivantes :

- Certains capteurs utilisés sont pour le moment indisponibles sur la plupart des véhicules récents (capteurs de positionnement latéral, oculomètre), et ne permettent pas une utilisation aisée de la modélisation. De plus il est difficile de savoir si leurs résultats provient de l'utilisation de capteurs performants ou d'une modélisation adéquate.
- Les auteurs ont sélectionné quelques situations à comparer parmi l'ensemble des situations rencontrées par le conducteur. Leurs résultats sont biaisés par cette vue réduite de l'activité. Il importe de modéliser le maximum de situation de conduite afin de pouvoir comprendre les différences entre chacune et construire des modules de diagnostic efficaces.
- Le taux et la rapidité de reconnaissance, bien qu'importants, seront-ils suffisants pour une utilisation dans les systèmes d'assistance ? Ils pourraient s'avérer dévastateurs si les 5% de confusion comprennent des situations critiques.

1.3.4. Brève vue critique des MMCs pour l'analyse de la conduite

Ainsi, l'utilisation des MMCs semble efficiente dans le cadre de l'étude de l'activité de conduite. En effet, les modèles de Markov cachés offrent des possibilités de modélisation intéressantes:

1 : un apprentissage automatique vers un optimum local.

Pour une topologie fixée (nombre d'états et graphe associé), l'algorithme itératif de Baum-Welch permet d'ajuster les autres paramètres (fonction de densité, matrice de transition) vers un optimum local de la vraisemblance du modèle. La convergence de cet algorithme est assurée. Grâce à une initialisation adéquate, la rapidité de la convergence peut être améliorée.

2 : une possibilité d'interpréter par la correspondance entre état et phase de conduite

Comme les travaux précédents l'ont suggéré, il est possible d'établir une correspondance entre les états des chaînes de Markov et les phases de conduite ou les états du conducteur. Grâce à l'apprentissage automatique, nous pouvons découvrir des découpages pertinents de l'activité de conduite. Au contraire d'autres modèles à apprentissage automatique (réseaux neuronaux), le modèle des chaînes de Markov cachées peut être explicité et faire sens. L'analyste a alors la possibilité d'introduire sa connaissance du domaine dans le modèle. Par cette action, il peut combler des lacunes et/ou rectifier des erreurs dans les données.

1 Modèle de Markov Cachée Couplée : extension apporté par Oliver et Pentland qui proposent de relier l'état de différents MMCs par des probabilités conditionnelles.

3 : un indice en temps réel d'adéquation des observations aux modèles

L'algorithme Forward permet, en temps réel, d'obtenir la probabilité que les observations aient été générées par le modèle. Grâce à cette probabilité, la comparaison entre différents modèles est aisée. Il est donc possible, au cours du temps, de décider à quel modèle il est probable que les observations appartiennent. Cette possibilité s'avérera extrêmement importante pour l'utilisation des MMCs dans des systèmes d'assistance.

L'apprentissage de la topologie est autrement plus délicat. Il constitue un des points faibles de ce type de modèle. Au problème, il n'existe pas de solution exacte. Dans la majorité des applications, les experts construisent des MMCs selon leur conception *a priori* du modèle ou procèdent par essais / erreurs pour obtenir de meilleurs modèles. Pourtant, des solutions existent pour se rapprocher d'une topologie optimale (algorithme génétique, test d'adéquation des états, agglomération/division d'état...).

Un autre inconvénient des MMCs est la nécessité d'un ensemble de données d'apprentissage important. Mais une initialisation correcte avec l'introduction de connaissances expertes peut pallier ce problème.

En conséquence, les qualités intrinsèques de ce type de modèle ainsi que les résultats de l'ensemble des recherches précédentes, ont confirmé notre choix en terme de modélisation de l'activité de conduite.

Cependant, si ce modèle est performant pour décrire une situation de conduite, il est inapte à prendre en compte les changements de situations. Aussi, la nécessité d'analyser ces changements nous a amenés à étudier les modèles de ruptures.

1.4 Conclusion

Nous avons donc vu dans cette partie que les recherches en psychologie cognitive nous ont amenés à considérer le conducteur comme un ensemble interconnecté de processus cognitifs. Ces derniers sont centrés autour de la Représentation Mentale, représentation que se fait le conducteur de la situation de conduite et qui est différente de la situation réelle de conduite. Pour élaborer cette représentation, il utilise les informations qu'il prélève dans l'environnement et les connaissances qu'il a préalablement acquises sur les contextes routiers.

Bellet [Bellet, 1998] nomme alors frame, la connaissance que le conducteur a d'une situation de conduite générale (comprenant ce qu'il faut y faire, quels sont les événements possibles, comment identifier la situation, où chercher l'information pour élaborer une représentation efficace du réel). Par exemple, le conducteur sait qu'en général lorsqu'il arrive à un rond-point, il faut freiner puis regarder à gauche pour savoir si une voiture arrive.

Cet auteur nomme alors *instanciation* le mécanisme par lequel le conducteur adapte ce frame, en utilisant les informations prélevées dans l'environnement pour générer la Représentation Mentale Courante.

Cette représentation repose sur un découpage de l'environnement en différentes zones de déplacements associées à des actions potentielles et des zones des perceptions, associées à des événements potentiels. La Représentation Mentale Courante dépend alors du conducteur, de ses objectifs, de sa situation initiale et de l'environnement perçu.

Cette notion de frame nous a permis de dégager certains concepts fondamentaux qui nous seront utiles pour structurer notre modélisation de l'activité de conduite : infrastructure perçue, division de l'activité suivant l'objectif tactique, zones de déplacements associées à des actions potentielles et dépendantes d'évènements extérieurs.

Par la suite, nous avons vu que, pour accéder à cette représentation mentale, trois niveaux d'enregistrement de l'activité de conduite s'avèrent pertinents, au niveau du conducteur, du véhicule, et de l'environnement.

Ces trois types de données ne sont pas indépendants et chacun nous renseigne partiellement sur l'activité de conduite et sur la cognition du conducteur.

Par ailleurs, analyser l'activité du conducteur, c'est avant tout segmenter l'activité de conduite en différentes séquences de conduite. Aussi, suivant les auteurs et les buts poursuivis, ces séquences peuvent être définies soit suivant l'objectif du conducteur, soit suivant les caractéristiques objectives de ces séquences.

Pour éviter toute confusion, nous avons donc défini 3 types de séquences de conduite : Séquences de Conduite Vécues, Séquences de Conduite Réelles et Séquences de Conduite Mesurées. Nous avons nommé Situation un regroupement de ces séquences suivant des critères particuliers.

Enfin, nous avons vu que pour modéliser l'évolution des données enregistrées lors de ces

séquences, plusieurs approches ont été utilisées. L'une d'elle, basée sur les chaînes de Markov cachées, s'est révélée particulièrement pertinente pour modéliser l'activité de conduite. En effet, ses qualités intrinsèques (intégration des dilatations temporelles, possibilité d'apprentissage automatique, indice temps réel d'adéquation de nouvelles données à un modèle) et la possibilité d'interpréter la structure de ce modèle (rendant accessible l'intégration de connaissances sur le comportement) ont montré leur efficacité lors de précédentes études sur le comportement du conducteur.

Pour l'ensemble de ces raisons, nous nous sommes donc focalisés sur ce type d'approche. Ceci nous a conduit notamment à développer certains aspects pratiques et théoriques de ces modèles pour accroître leur efficacité en vue de leur application à l'analyse de la conduite. Le chapitre suivant sera consacré à leurs descriptions.

Par ailleurs, si le modèle de Markov caché est efficace pour modéliser l'évolution des signaux numériques dans des situations de conduite, il est inadéquat pour prendre en compte les changements de situations.

Aussi nous avons recherché un autre type de modèle permettant de reconnaître une situation dans un flot de données inconnues. Pour cela, l'étude des ruptures dans l'évolution de la probabilité d'appartenance de données inconnues à une situation, nous a semblé une approche intéressante. Ceci nous a conduit à étudier les modèles dit de ruptures multiples dont nous montrerons quelques résultats théoriques sur les propriétés des estimateurs.

2 Chaînes de Markov Cachées et modèles multi-phasiques.

Cette partie est consacrée à la présentation des développements théoriques nécessaires pour atteindre les objectifs de cette thèse. Ces développements sont de deux ordres. D'une part, ils concernent le modèle de Markov caché dont nous expliciterons les méthodes classiques permettant son utilisation et donnerons une structure adaptée à l'étude des données de conduite. D'autre part, l'utilisation d'un modèle multi-phasiques pour analyser les données nous amènera à déterminer les propriétés de convergence des estimateurs par maximum de vraisemblance.

2.1 Chaînes de Markov cachées : aspects théoriques et pratiques

Cette partie est consacrée à la description du modèle des chaînes de Markov cachées (MMC) et d'une extension de ce modèle « le modèle Semi-Markovien caché pondéré ». Ce dernier a été conçu et mis en place pendant la thèse et sera à la base de notre modélisation de l'activité de conduite.

Dans un premier temps, nous rappellerons les définitions et les propriétés du modèle général. Puis, nous en présenterons les principaux algorithmes nécessaires à sa mise en oeuvre et leurs propriétés. La deuxième section sera consacrée à l'étude de quelques développements utilisés pour approcher la topologie optimale d'un modèle.

Enfin, c'est dans la dernière section que nous introduirons le modèle Semi-Markovien caché pondéré.

2.1.1. Définitions

Le modèle des chaînes de Markov cachées a été introduit par Baum et al dans les années 1960-1970. Ce modèle a tout d'abord été utilisé dans le domaine de la reconnaissance vocale [Rabinner, 1989]. Puis son domaine d'application s'est étendu tant à la reconnaissance de texte manuscrit que récemment à l'analyse des séquences biologiques (ADN ou protéine).

Dans la suite, nous considérons $(S_t)_{t\in\mathbb{N}}$ une suite de variables aléatoires définies sur un espace de probabilité $(\Omega, B, \mathbb{P})$ et à valeurs dans un ensemble fini $E=\{E_1, E_2, ..., E_K\}$ dit espace d'état. Pour plus de lisibilité, on notera $E=\{1,2,...K\}$.

Par la suite, on considère T fixé, un entier naturel, supérieur à 1.

Chaîne de Markov finie

Définition : Une suite aléatoire $(S_t)_{t\in[1:T]}$ est une chaîne de Markov finie si

$$\begin{gathered}\mathrm{p}(S_{t+1}\in A/S_t=i,\{S_1\in B_1,....S_{t-1}\in B_{t-1}\})=\mathrm{p}(S_{t+1}\in A/S_t=i)\\ \forall A,B_t\in E\ ;\ \forall i\in E;\forall t\in[1:T-1]\end{gathered} \tag{1}$$

Chaîne de Markov cachée

Définition : Une chaîne de Markov cachée est un processus aléatoire (S_t,Y_t) tel que :

i. S_t est une chaîne de Markov finie.

ii. Y_t est une suite de variables aléatoires et Y_t | S_t est une suite de variables aléatoires indépendantes.

iii La loi conditionnelle de $Y_t|\{S_t, t\in[1:T]\}$ est la loi de $Y_t|S_t$.

Le nom de chaîne de Markov cachée est motivé par la supposition que la suite S_t n'est pas observable. Ainsi, toute estimation n'est basée que sur la suite $(Y_t)_{t\in[1:T]}$.

Soit $g(y_t/S_t)$ la densité de Y_t.

Pour simplifier l'écriture, on note $\forall t\in[1:T],\ j\in[1:K]$

- $\pi_j := \mathrm{p}(S_1=j)$,
- $\pi=(\pi_1,\ldots\pi_K)$ est dit le vecteur de probabilité initiale,
- $p_{i,j} := \mathrm{p}(S_{t+1}=j/S_t=i)$
- $p=\{p_{i,j}\}\quad 1\le i,j\le K$ est dite la matrice de transition,
- $g_j(y_t) := g(y_t/S_t=j)$,
- $\varphi=\{\varphi_1\ldots\varphi_K\}$ avec φ_j les paramètres qui caractérisent g_j,
- $\theta\in\Theta\in\mathbb{R}^d$, $d\ge 1$ tel que $\theta=[\pi,p,\varphi]$ les paramètres d'un modèle.

La Figure 2.1 est l'illustration usuelle des chaînes de Markov cachées.

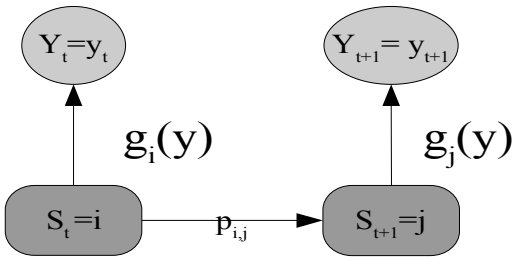


Figure 2.1 : Représentation des modèles de Markov cachés.

Lors des paragraphes suivant, nous nous intéresserons aux solutions algorithmiques apportées aux trois problèmes classiques posées par ce genre de modèle :

a) le calcul de la probabilité d'une observation $y_{1\ldots T}$,

b) le calcul de la séquence d'états cachés $s_{1\ldots T}$ ayant le plus probablement générée $y_{1\ldots T}$,

c) et l'estimation des paramètres d'un modèle.

2.1.2. Algorithmes pour les Chaînes de Markov cachées

a Probabilité d'observation d'une séquence $y_{1\ldots T}$: algorithme Forward-Backward

On cherche la probabilité d'une réalisation $(Y_1=y_1)\cap(Y_2=y_2)\ldots\cap(Y_T=y_T)$ Pour cela, on note cet évènement $Y_{1\ldots T}=y_{1\ldots T}$. De la même manière, on note $(S_1=s_1)\cap(S_2=s_2)\ldots\cap(S_T=s_T)$, par $S_{1\ldots T}=s_{1\ldots T}$, où $s_t\in[1:K]$.

Théoriquement, la probabilité de $Y_{1\ldots T}=y_{1\ldots T}$ est la somme sur tous les chemins possibles de la suite S, des probabilités conjointes de Y et de S. Ainsi,

$$\mathrm{p}(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}})=\sum_{s_1\in E}\ldots\sum_{s_T\in E}\mathrm{p}(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}},S_{1\ldots T}=s_{1\ldots T}) \quad (2)$$

Par ailleurs, la probabilité de $(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}},S_{1\ldots T}=s_{1\ldots T})$ est égale à

$$\mathrm{p}(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}},S_{1\ldots T}=s_{1\ldots T})=\pi_{s_1}\prod_{t=1}^{(T-1)}p_{s_t,s_{t+1}}\prod_{t=1}^{T}g_{s_t}(y_t) \quad (3)$$

Aussi pour connaître la probabilité de $Y_{1\ldots T}=y_{1\ldots T}$ la formule des probabilités totales, $\mathrm{p}(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}})=\sum_{s_{1\ldots\mathrm{T}}\in(1\ldots\mathrm{K})^T}\mathrm{p}(Y_{1\ldots\mathrm{T}}=y_{1\ldots\mathrm{T}}|S_{1\ldots T}=s_{1\ldots T})\mathrm{p}(S_{1\ldots T}=s_{1\ldots T})$ pourrait être utilisable. Cependant, ce calcul requiert T^K opérations correspondant au nombre de terme de la somme. Une méthode plus efficace, l'algorithme Forward-Backward, est souvent utilisée. Selon Murphy [Murphy K, 2002], cet algorithme permet alors de résoudre ce problème en $K*T^2$ opérations.

Il est basé sur la définition de deux fonctions, la fonction Forward et la fonction Backward.

- La première, la fonction Forward est définie comme la probabilité à l'instant *t* d'être dans l'état i connaissant les t premières réalisations de Y, $\alpha(t,i):=\mathrm{p}(S_t=i/Y_{1\ldots\mathrm{t}}=y_{1\ldots\mathrm{t}})$:

$\alpha(1,i)=\pi_i g_i(y_1)$ et pour t>1, $\alpha(t,i)=g_i(y_t)\sum_{j=1}^{K} p_{j,i}\alpha(t-1,j)$ (4)

- Inversement la fonction Backward est définie comme la probabilité à l'instant t d'être à l'état i connaissant les T-t+1 dernières réalisations de Y, $\beta(t,i):=\mathrm{p}(S_t=i \mid Y_{t+1\ldots T}=y_{t+1\ldots T})$:

$$\beta(T,i)=1 \;;1\le i\le K \text{ et pour t<T,}$$
$$\beta(t,i)=\sum_{j=1}^{K} p_{i,j}.g_j(y_{t+1})\beta(t+1,j)\ . \qquad (5)$$

Zhong et al [Zhong & Ghosh, 2001] illustre ainsi ces deux fonctions :

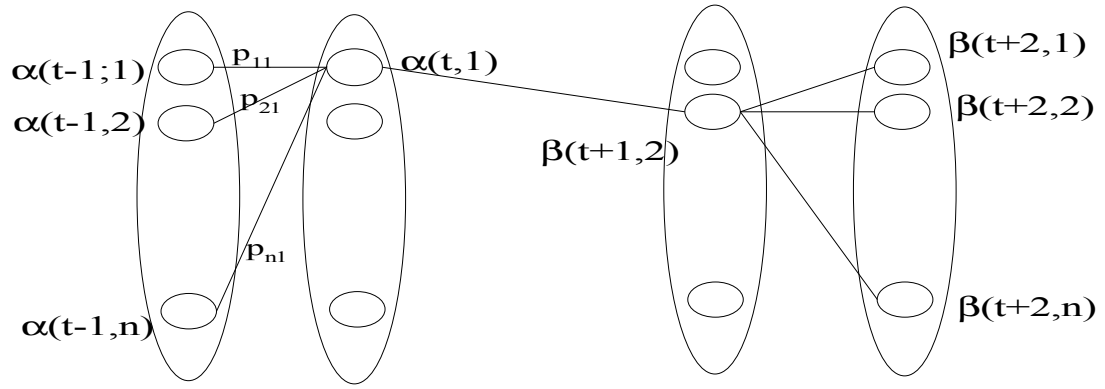


Illustration 2.2 : Illustration de la procédure Forward-Backward par Zhong & Ghosh [Zhong & Ghosh, 2001]

Pour tout t (1<t<T), la probabilité de $(Y_{1\ldots T}=y_{1\ldots T})$ est alors égale à la somme des α et des β sur l'ensemble des états possibles.

$$\begin{aligned}\mathrm{p}(Y_{1\ldots T}=y_{1\ldots T}) &= \sum_{i=1}^{K} \mathrm{p}(Y_{1\ldots T}=y_{1\ldots T}, S_t=i)\\ &= \sum_{i=1}^{K} \mathrm{p}(Y_{1\ldots t}=y_{1\ldots t}, S_t=i)\,\mathrm{p}(Y_{t+1\ldots T}=y_{t+1\ldots T}, S_t=i)\\ &= \sum_{i=1}^{K} \alpha(t,i)\beta(t,i)\end{aligned}$$

Le calcul itératif de α et de β et ces égalités résolvent donc le problème du calcul de la probabilité d'une séquence $y_{1\ldots T}$.

b Séquence des états cachées la plus probable : l'algorithme de Viterbi.

Un des problèmes importants avec ce type de modèle à données incomplètes est d'inférer les observations manquantes.

Dans cette optique, l'algorithme de Viterbi cherche à retrouver la séquence $s_{1...T}$ la plus probablement réalisée par la suite de variables aléatoires $S_{1...T}$, en connaissant $y_{1...T}$.

Ainsi la suite $s_1^* \dots s_T^*$ recherchée est celle maximisant $\mathrm{p}(S_{1...T}=s_{1...T} / Y_{1...T}=y_{1...T})$.

Pour parvenir à cet objectif, Viterbi [Viterbi, 1967], définit la fonction $r_t(i)$ comme la réalisation la plus probable de S à l'instant t, connaissant les réalisations passées de S et de Y.

Il pose ainsi pour tout $i,j \in [1:K], t \in [1:T]$

$$r_t(j) = max_{s_1,..s_{t-1}} \mathrm{p}(Y_{1..t}=y_{1..t}, S_{1..(t-1)}=s_{1..(t-1)}, S_t=j) \quad (6)$$

On a alors :

$$r_1(j) = \pi_j . g_j(y_t)$$

$$r_t(j) = max_{i \in 1:K} r_{t-1}(i) . p_{i,j} g_j(y_t) .$$

Pour t=T, $\max_{i \in [1:K]} r_T(i) = s_T^*$ correspond à l'état final le plus probable.

Par la suite, en définissant la fonction $d_t(i)$ comme l'état le plus probable précédant l'état i au temps t, $d_t : E \to E$

$$d_t : i \to \underset{j \in 1:K}{argmax}[p_{j,i} . r_t(j)] \quad (7)$$

Viterbi reconstruit alors le chemin menant à l'état final le plus probable, en posant $s_{t-1}^* = d_t(s_t^*)$.

La suite $s_{1...T}^*$ trouvée maximise alors $\mathrm{p}(S_{1...T}=s_{1...T} | y_{1...T})$.

L'algorithme consiste alors à calculer itérativement $r_t(j)$ puis $d_t(j)$ pour $j \in [1:K], t \in [1:T]$ et à déduire $s_1^* \dots s_T^*$ qui est la suite recherchée.

c Problème de l'estimation de modèle : algorithme EM

Le nombre d'état K est fixé. Dans ce cas, on veut trouver un estimateur de $\theta^0 = \{\pi_i^0, p_{i,j}^0, \varphi_i^0, i,j \in [1:K]^2\} \in \Theta$, le vrai paramètre.

Nous nous intéressons uniquement ici à la procédure numérique d'apprentissage basée sur l'algorithme Expectation-Maximisation (E-M). Cet algorithme est basé sur une procédure itérative permettant l'augmentation de la vraisemblance des données par rapport au modèle.

Bien que d'autres procédures ont l'avantage d'intégrer le concept de discrimination entre classes,

en optimisant un critère dit d'information mutuelle (sur ce sujet, voir par exemple McDonnough & Waibel [Mcdonough & Waibel, 2003] et Kamal & Hasegawa-Johnson [Kamal & Hasegawa-Johnson, 2003]), leurs inconvénients majeurs est l'importance de leurs coûts en calcul. Dès que le nombre de classe devient élevé, il s'avère difficile de les mettre en place. Aussi nous nous focaliserons ici uniquement sur l'algorithme E-M, dont nous rappellerons le principe dans le cas général et ses propriétés dans le cas des chaînes de Markov cachées.

c.1 Algorithme EM : principe général

Cet algorithme utilisé pour l'identification des chaînes de Markov cachées a été introduit par Dempster et al [Dempster et al, 1977] dans le cadre général de modèles à données incomplètes. Il s'agit d'une procédure itérative permettant d'approcher le maximum de vraisemblance lorsque sa maximisation analytique est difficile.

Etant donnée une séquence observée, l'estimateur du maximum de vraisemblance s'obtient en maximisant la log-vraisemblance $\log(\mathsf{P}(y|\theta))$ pour $\theta \in \Theta$. Cependant l'existence des régimes cachés $s = s_{1..T}$ rend difficile cette maximisation. La log-vraisemblance des données complètes $\log(\mathsf{P}(y, s|\theta))$ est plus facilement manipulable. On rappelle que la probabilité conjointe de $(y_1, \ldots y_T, s_1, s_2, \ldots . s_T)$ est donnée par (3).

Ainsi, parce que , $\mathsf{P}(y) = \sum_{s_1=1}^{K} \ldots . \sum_{s_T=1}^{K} \mathsf{P}(y, s|\theta)$ chaque itération de l'algorithme va consister dans un premier temps à remplacer $\log(\mathsf{P}(y, s|\theta))$ par son espérance conditionnelle connaissant la séquence observée $y_{1..T}$ et une valeur du paramètre θ', candidate au maximum.

Cette espérance conditionnelle est alors maximisée en θ dans un second temps, pour obtenir une nouvelle valeur du paramètre.

L'algorithme se base alors sur la définition de la fonction Q :

$$Q(\theta|\theta') = \mathsf{e}[\log \mathsf{P}(y, s|\theta)|\, y, \theta'] = \sum_{s \in E^T} \log \mathsf{P}(y|\theta)\mathsf{P}(s|y, \theta') \tag{8}$$

Commençant par une valeur initiale du paramètre $\theta^{(0)}$ et sachant la valeur courante $\theta^{(m)}$, l'itération m de l'algorithme est alors décrite par les deux étapes suivantes

- ➔ étape *Expectation* : calculer $Q(\theta|\theta^{(m)})$
- ➔ étape Maximisation : choisir $\theta^{(m+1)} = \underset{\theta \in \Theta}{argmax}\, Q(\theta|\theta^{(m)})$

L'algorithme repose sur le fait que pour toute suite $(\theta^{(m)})_{m \geq 0}$ générée par l'algorithme EM, $\mathsf{P}(y|\theta^{(m+1)}) \geq \mathsf{P}(y|\theta^{(m)})$ [Dempster *et al.*, 1977]. Ainsi à chaque itération l'estimation de θ est améliorée du point de vue de la log-vraisemblance.

c.2 Algorithme de Baum-Welch

Rabinner [Rabinner, 1989] note que l'algorithme EM correspond, pour les chaînes de Markov cachées, à l'algorithme de réestimation de Baum-Welch. Ainsi, pour estimer θ , Baum et al [Baum & al, 1970] utilisent une procédure itérative consistant à maximiser une fonction auxiliaire qui est exactement la fonction $Q(\theta|\theta')$ et pour laquelle ils démontrent que cette technique de maximisation fait croître la log-vraisemblance $\mathsf{p}(y)$ à chaque itération avec p définit comme en (3). De la même manière, Zhong et Gosh [Zhong & Ghosh, 2001] montrent aussi que cette procédure peut être aussi vu comme un problème d'optimisation.

Dans les trois cas, les formules de réestimation restent les mêmes. Nous présentons en annexe (4.1) le détail des formules de réestimations données Rabinner [Rabinner, 1989].

Le calcul itératif de $\theta^{(m)}$ apportent ainsi une solution au problème de l'estimation.

2.1.3. Déterminer la topologie : méthodes usuelles

Graphe d'état:

Définition: On définit par graphe d'état, la matrice M définissant les arcs entre les différents états.

$$M(i,j)=1 \ si \ p_{i,j}>0$$
$$0 \ si \ p_{i,j}=0 \ .$$

On définit la topologie par le nombre d'état et le graphe associé. Dans la partie précédente, nous avons travaillé avec une topologie fixée. Et nous avons vu comment, dans ce cas précis, déterminer les paramètres $[\pi^*, p^*, \varphi^*]$ maximisant la vraisemblance des données. Aussi, le choix de ces deux premiers attributs est déterminant pour le modèle.

Pourtant, au problème de déterminer la topologie la plus adéquate pour modéliser un ensemble de données, il n'existe pas de réponse simple. En effet, pour comparer deux modèles aux dimensions différentes, la seule comparaison des vraisemblances est inadéquate.

Aussi, pour faire face à ce type de problème, la vraisemblance est généralement pénalisée en fonction de la complexité du modèle. Par la suite, nous présenterons quelques-unes des fonctions de pénalisation couramment utilisées dans le cas des modèles de Markov cachées (MMC). Puis nous aborderons différentes méthodes permettant de construire un ensemble restreint de modèles à comparer.

a Critère de sélection de modèle

Pour mesurer l'adéquation d'un modèle aux données, la seule vraisemblance ($\mathsf{P}(y/M)$) s'est avérée insuffisante [Akaike, 1973]. En effet, plus un modèle est complexe (plus les paramètres seront nombreux), plus il sera probable qu'il ait généré les données observées, mais moins il aura de facultés généralisantes (capacité à être adapté à de nouvelles données.). Cet équilibrage entre complexité nécessaire et faculté généralisante est un problème connu en statistique.

Aussi, lorsque les connaissances a priori sur les données ne permettent pas de définir un cadre

unique, l'une des réponses possibles est la minimisation d'un critère pénalisé [Peltier, 1993].

Ce critère est ainsi, une fonction de la vraisemblance, pénalisée par la complexité du modèle *(Critère(modèle) =Log(P(Observation / modèle))- pénalité)* et résout le problème suivant :

On dispose d'une collection de modèle $\{M_1, M_2, M_{3...,} M_m\}$. A chaque modèle M_i correspond une densité g_{Mi} de paramètres θ_i de dimension $\mathbb{R}^{q_i}$.

Par ailleurs, on dispose d'une série d'observations $y=\{y_{1,}y_{2,}\cdots,y_T\}$ de variables indépendantes de densité inconnue « f ».

Selon Lebarbier & Mary-Huard [Lebarbier & Mary-Huard, 2004], le critère Cr recherché est tel que si le modèle M* est solution de $M^* = \underset{i\in[1:m]}{argmax}\, Cr(M_i)$ alors M* est aussi solution de

$$M^* = \underset{M_i, i\in[1:m]}{argmin} \mathbb{E}\left(\int \log\left(\frac{f(x)}{g_{Mi}(x,\widehat{\theta_i})} f(x)dx\right)\right) \quad . \tag{9}$$

Ces critères Cr peuvent être divisés en deux catégories basées

1. Soit sur une validation croisée.
2. Soit sur la robustesse propre du modèle.

L'utilisation des critères basés sur une validation croisée nécessite un nombre de calcul important, ce qui les rends difficilement utilisable dans le cadre des MMC. Pourtant, certains auteurs ont adapté spécifiquement ces méthodes pour les chaînes de Markov cachées en s'intéressant à des familles particulières de modèles (par exemple Durand [Durand, 2003] adaptent ces méthodes pour les Chaînes de Markov cachées au graphe « moral »).

Pour notre part, ces contraintes (calcul important ou spécification d'une famille particulière de modèle) ont fait que nous n'avons pu nous appuyer sur ce type de méthodes, et que nous nous sommes appuyé sur les critères du second type.

Ces derniers pénalisent la vraisemblance d'un modèle par un terme positif fonction de sa complexité.

Ainsi, pour résoudre l'équation (9), on approxime la dissimilarité entre la fonction f et la fonction $g_{Mi..}$ Selon le type d'approximation utilisé, plusieurs critères sont apparus (AIC, BIC, MDL, et C_p de Mallow). Nous présenterons ici les deux premiers qui sont aussi les plus couramment utilisés.

D'un point de vue théorique, le critère d'information d'Akaike (AIC) est basé sur un développement limité de la vraisemblance du modèle autour du vrai modèle. Sa valeur est de :

$$C_{AIC}(M_i) = -2*\log(\mathrm{p}(y/M_i)) + 2W_{i.} \tag{10}$$

avec W_i le nombre de paramètre du modèle M_i.

Le critère d'information Bayesien (BIC) est lui basé sur une approximation du calcul de la vraisemblance conditionnellement aux données et au modèle fixé. Sa valeur est de :

$$C_{BIC}(M_i) = -2*\log(\mathrm{p}(y/M_i)) + W_i*\log(T) \tag{11}$$

Ces différences se traduisent en pratique par le fait que lors des tests par simulation de donnée sur des modèles simples, BIC sélectionne le vrai modèle et AIC le vrai modèle ou un modèle plus grand. Cependant, sur des modèles plus complexes, on constate que BIC devient moins performant qu'AIC car même pour des grandes tailles d'échantillons, BIC sélectionne des modèles sous-ajustés [Lebarbier & Mary-Huard, 2004].

En outre, les résultats obtenus sur des données simulées montrent que la complexité du modèle, celle des modèles candidats, ainsi que la taille des échantillons, influencent beaucoup les performances pratiques de ces critères.

Pourtant, les critères BIC et AIC sont souvent utilisés indistinctement quel que soit le problème posé [Reschenhoffer, 1996]. Or ces deux critères diffèrent dans leurs objectifs : l'un opte pour un modèle prédictif et l'autre pour un modèle explicatif.

Ainsi, comme le font remarquer Lebarbier et Mary-Huard [Lebarbier & Mary-Huard, 2004], l'objectif de l'analyse et la connaissance des données conditionnent le choix d'un critère et peuvent donner un sens à la notion de supériorité d'un critère sur l'autre.

C'est donc la nature de notre problématique basée sur un objectif de prédiction (voir chapitre 3.3.1.c.2), qui nous orientera vers l'utilisation de BIC.

Pourtant comme le rappelle Bouchard et Celeux [Bouchard & Celeux, 2004], le critère BIC mesure de toute façon l'adéquation du modèle aux données et ne détermine pas forcément un classifieur performant. A cette fin, Bouchard et Celeux [Bouchard & Celeux, 2004] ont développé le critère BEC (Bayesian Entropic Criterion). Cependant, l'aspect incrémental de notre procédure de modélisation (voir partie 3.3) et le nombre de catégories étudiées rendent difficile son application dans notre cas.

Remarque : De plus, dans le cadre des chaînes de Markov cachées, Durand ([Durand, 2003]) rappelle que, la justification théorique d'AIC n'est pour le moment pas prouvée (et en particulier la normalité asymptotique de l'estimateur de maximum de vraisemblance)

b Construire un ensemble de modèles à comparer.

Nous venons de définir un critère de comparaison de modèles. Pourtant, le nombre de modèles possibles est, a priori, illimité. Aussi de nombreux auteurs ont décrit des procédures itératives de construction de modèles. Elles fonctionnent toutes selon le principe suivant : elles partent d'un modèle initial «M(0) », puis une procédure de modification a lieu donnant naissance à un nouveau modèle « M(1) ». Si ce modèle augmente un critère Cr(M) de validité du modèle (vraisemblance, critère d'information...), il est remodifié pour donner naissance à un nouveau modèle « M(2) », et ainsi de suite jusqu'à ce que la vraisemblance n'augmente plus (Figure 2.3).

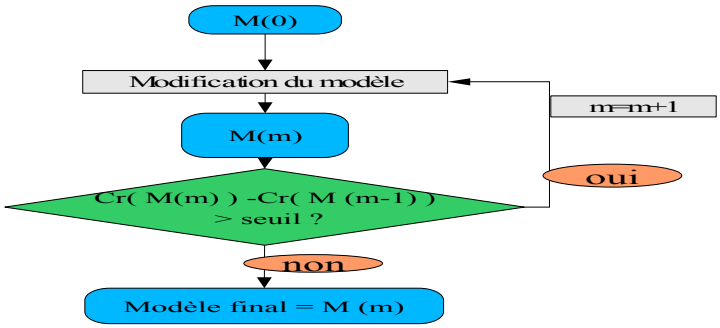


Figure 2.3 : Procédure permettant de sélectionner itérativement le modèle adéquat pour représenter des données.

Selon les applications et les auteurs, les procédures de modification et le choix du modèle initial sont différents. *Dans le cadre des chaînes de Markov cachées,* les trois méthodes suivantes sont généralement utilisées :

1. Algorithmes génétiques.
2. Agglomération d'états
3. Division d'états

Bien que Thomsen [Thomsen, 2002], montrent que les algorithmes génétiques peuvent être utilisé pour augmenter les performances prédictives de modèles de Markov cachés, Kwong et al [Kwong *et al.*, 2001] rappellent que l'absence de convergence de ces algorithmes et les nombreux paramètres à définir a priori, rendent leurs utilisations périlleuses et réservées à des domaines où *aucune connaissance « a priori » ne peut être utilisée*. Aussi nous nous intéresserons ici uniquement aux deux autres procédures. C'est sur ces dernières que nous baserons notre algorithme d'apprentissage de la topologie, algorithme que nous présenterons dans la partie 3.3.1.c.2.

b.1 Agglomération d'états

Cette méthode est basée sur la définition d'un modèle initial avec un nombre d'états important, lié suivant un graphe complet (tous les états sont reliés entre eux). Puis, chaque modification du modèle consiste à enlever un état. Mais pour choisir celui-ci, Stenger et al [Stenger *et al.*, 2001] rappellent que deux méthodes sont utilisées:

a) l'état le moins instructif :

A l'étape « m+1 », l'état C ($C \in [1:K]$ choisi est celui dont le retrait fait le moins baisser la vraisemblance du modèle :

$$\{\mathrm{P}(y_{1..T}/M_C(m+1))\} = \max_{i\in[1:K]} \{\mathrm{P}(y_{1..T}/M_i(m+1))\} \qquad (12)$$

Avec $M_i(m+1)$ le modèle construit en enlevant l'état « i » au modèle $M(m)$.

b) l'état le moins utilisé :

L'état C choisi est celui où le processus S passe le moins souvent.

$$Card\{s_t = C \mid t \in [1:T]\} = min\{Card\{s_t = i \mid t \in [1:T]\}\}. \quad (13)$$

Stenger et al donnent alors les différentes façon de définir la nouvelle matrice de transition.

b.2 Division d'état

De façon symétrique à la méthode présentée ci-dessus, une autre méthode est utilisée.

Un premier modèle très général (un seul état) peut être choisi. Puis chaque modification du modèle consiste à choisir un état C qui sera « divisé » en deux. Cette division complexifiera le modèle, augmentant ainsi sa vraisemblance. Stenger et al [Stenger *et al.*, 2001], utilisent cette technique pour catégoriser des images vidéo

Le choix de l'état à diviser peut se faire aussi ici selon plusieurs critères.

a) Si c'est la robustesse globale du modèle qui importe, le critère utilisé est celui de l'augmentation de la vraisemblance du modèle. Dans ce cas, l'état C choisi est

$$\{P(y_{1..T} / M_C(m+1))\} = \max_{i \in [1:K]} \{P(y_{1..T} / M_i(m+1))\} \quad (14)$$

Avec $M_i(m+1)$ le modèle construit en divisant l'état « i » au modèle $M(m)$.

b) Si au contraire c'est la pertinence de chaque état qui importe, l'état choisi est celui dont les observations (provenant de ce dernier) sont les moins cohérentes. C'est à dire en notant $A_i = \{y_t \mid S_t = i, \quad t \in [1:T]\}$ l'ensemble des observations[1] issues de l'état i, l'état à diviser est alors C_i tel que

$$P(A_C) = \min_{i \in [1:K]} P(A_i) \quad (15)$$

Cet état correspond alors à celui dont les observations sont le moins en accord avec la fonction de densité associée g_i.

Dans les deux cas, la division d'un état donne lieu à la création de deux nouveaux états dont les paramètres peuvent être :

- soit une variation aléatoire des paramètres de l'état « père »,
- soit issus des résultats d'une classification par k-means[2] sur l'ensemble A_C , en prenant la variance et la moyenne des groupes trouvés comme paramètres des nouveaux états.

Parallèlement à ces techniques générales, de multiples méthodes, chacune spécifique aux phénomènes étudiés sont apparues (voir notamment Stolcke et Omohundro [Stolcke & Omohundro,

1 L'ensemble A_i est approché par l'algorithme de Viterbi.

2 La technique des k-means est une méthode statistique classique permettant de partitionner une population, caractérisée par un ensemble de données continu, en groupes d'individu homogènes.

1994] qui déterminent une procédure permettant une représentation plus compacte de modèles grammaticaux).

Lors de notre étude sur les données de conduite, nous avons conçu, élaboré et développé un algorithme spécifique d'apprentissage de la topologie, basé sur le critère BIC et l'association des deux méthodes précédentes. Celui-ci sera présenté dans la partie 3.1.4.c.c.2.

Le paragraphe suivant propose d'analyser une autre solution souvent utilisée pour définir une topologie efficace. Cette méthode consiste à étendre le modèle Markov caché en y intégrant des contraintes spécifiques. Nous présenterons des exemples de modèles ainsi construits puis définirons un nouveau modèle, le Modèle Semi-Markovien Caché Pondéré.

2.1.4. Extension des modèles de Markov cachés

Une des solutions pour modéliser au mieux des données par un MMC est d'insérer dans sa définition des contraintes spécifiques. Ces contraintes sont généralement fixées suivant la connaissance a priori du phénomène à modéliser [Pal & Hu, 2001].

Les études sur la reconnaissance vocale ont permis ainsi l'édification de nombreux modèles spécialisés. La structure fortement hiérarchisée de la parole, et les riches études sur le domaine ont rendu cette édification possible et performante.

Dès lors de nombreuses extensions aux modèles de Markov cachés sont apparus. Chacune permet de modéliser des contraintes particulières. Par exemple, le modèle Semi-Markovien caché modélise ainsi spécifiquement le temps passé dans chaque état.

Pourtant à notre connaissance, il n'existe pour l'instant aucun modèle intégrant le concept de pondération. C'est pourquoi après avoir présenté un état de l'art général sur ce type d'extension, nous présenterons le modèle de Markov caché pondéré développé lors de cette thèse.

a État de l'art

a.1 Présentation générale

Une des premières extensions développées fut le FHMM (Factorial Hidden Markov Model) Celui-ci avait pour but de représenter les différents niveaux du langage (phonèmes, syllabes, mots, phrases). Pourtant, bien que Logan et Moreno [Logan & Moreno, 1998] aient indiqué que cette modélisation n'augmentait pas les performances générales en matière de reconnaissance vocale, elle montre son efficacité quand le signal à traiter est issu de plusieurs processus indépendants (voir à ce sujet, Logan et al [Logan & Moreno, 1998], et Reyes et al [Reyes-Gomez et al, 2003]).

Fine et al proposent une structure plus générale : les MMCs hiérarchiques [Fine et al, 1998]. Ceux-ci se présentent comme une structure à plusieurs niveaux. Chaque état de la chaîne de Markov représente lui-même un MMC. Ce modèle est particulièrement adapté pour modéliser les structures à plusieurs niveaux (parole, écriture…). Ainsi, les résultats de Murphy et Paskin [Murphy & Paskin, 2001] sur la reconnaissance d'écriture manuscrite se sont avérés prometteurs. Selon ces auteurs, les modèles construits capturent des corrélations entre événements proches dans le temps ou éloignés.

Pourtant, l'intérêt de ce genre de modèle (intégration de fortes contraintes données par

l'expert) fait aussi leur faiblesse. En effet, dans les domaines où les différentes études n'ont pu donner une structure hiérarchique valide, l'implémentation de ces modèles est périlleuse. De plus, l'absence de possibilité de prendre en compte des variables manquantes à un instant donné rend caduc ce type d'approche pour le couplage de données hétérogènes (à une cadence d'enregistrement différente). Or, ce dernier point est essentiel, dans de nombreux domaines (paroles, automobiles...), pour coupler des sources vidéo, interprétées automatiquement via des algorithmes de reconnaissance de formes et des données de type signaux numériques.

C'est dans ce cadre que Brand et al [Brand *et al.*, 1996] ont construit des MMCs dits couplées permettant l'intégration de différentes sources d'information. Ces derniers modèle supposent qu'on dispose de deux ou plusieurs MMCs en parallèle dont les états à un instant donné dépendent de l'ensemble des états des autres MMCs. Ainsi, cette modélisation permet de façon simple de fusionner des données hétérogènes (voir à ce sujet, Zhong et Gosh, [Zhong & Ghosh, 2002], et Rezek et al [Rezek et al, 2000]).

Le dernier cas qui nous intéresse est l'intégration dans la chaîne de Markov d'une dépendance à l'égard de facteurs extérieurs connus. En effet, l'ajout d'une relation entre un événement extérieur connu, l'état de la chaîne et les observations permettent de modéliser des systèmes de réponses complexes. Bengio et al ont présenté l'architecture de ces modèles [Bengio & Frasconi, 1995] et leur efficacité pour résoudre le problème de reconnaissance de grammaire [Bengio & Frasconi 1996]. Par ailleurs, Kim et al [Kim et al, 2002]en montrent l'intérêt pour modéliser les flux boursiers. Ceux-ci sont alors considéré comme une conséquence de « l'état » de la bourse et des ordres d'action passés.

Un cas particulier des MMCs à entrées-sorties sont les Modèle Semi Markovien Cachés. Ceux-ci constituant le socle de notre travail, nous les présenterons en détail dans la section suivante.

a.2 **Modèle Semi-Markovien Caché** (MSMC) :

Murphy [Murphy, 2002] note que dans la littérature, ce type de réseau se trouve sous deux noms différents : « Hidden semi-Markov models », ou Variable-Length MMCs.

Il rappelle alors que leur principe commun de ce type de modèle est de modéliser explicitement la durée durant laquelle le processus reste dans chaque état. En effet, dans un MMC «habituel », selon Rabinner [Rabinner, 1989], la durée « d », durant laquelle le processus reste dans un état « i », suit une loi exponentielle $(p_{i,j})^{d-1}(1-p_{i,j})$.

Pourtant, dans nombre d'applications cette hypothèse n'est pas justifiée. Aussi, dans un SMMC, cette durée est modélisée par une probabilité spécifique : $h_i(d)$, avec D une variable aléatoire, représentant la durée pour le i^{ème} état.

Dans ce cas, Rabinner [Rabinner, 1989] note que les formules de réestimations des paramètres ainsi que le calcul de $\mathrm{P}(y/M)$ change. Zen et al [Heiga *et al.*, 2004] donnent les formules de réestimations dans le cas où les paramètres d'émission suivent une loi normale et Ge et Smyth [Ge & Smyth, 2001] adapte l'algorthime de Viterbi pour ce type de MMCs.

Remarque : Pour résoudre le problème de l'augmentation du nombre de paramètres dû à $h_i(d)$, certains auteurs comme Levinson [Levinson, 1985], ont modélisé la durée durant laquelle le processus reste dans un état par des lois paramétriques (gaussienne ou gamma).

b Développement du concept de pondération pour les modèles de Markov cachées.

Dans un grand nombre d'application, comme celle de l'analyse de la conduite, le nombre de variable enregistrée est important. Certaines sont inutiles et leurs présences perturbent, soit globalement soit localement, l'analyse des données. Aussi, il nous a semblé important d'introduire le concept de pondération appliqué aux variables et aux différents états d'un modèle de Markov caché.

b.1 Modèles de Markov Cachées Pondéré

La fonction de densité $g_i(y)$ caractérisant $P(Y_t = y_t / S = i)$ est souvent défini soit comme une loi multinomiale, soit comme un mélange de lois normales.

Dans ces deux cas, toutes les dimensions de Y ont la même « influence » sur l'état de S.

Pourtant, cette hypothèse est souvent injustifiée dans de nombreux modèles. A notre connaissance, mise à part Heiga et al [Heiga *et al.*, 2004], aucune étude n'a introduit le concept de pondération sur les dimensions de Y.

Ainsi, ces auteurs se sont intéressés à *l'union des évènements* $(Y_t^r = y_t^r)$ pondérés chacun par un facteur $c_{j,r}$. avec r la r^ième^ dimension de y. Ils définissent la fonction de densité associé à l'état i, ainsi $g(y_t | S_t = j) = g_j(y) = \sum_{r=1}^{R} c_{j,r} g_{r,j}(y^{(r)})$, $avec \sum_{r=1}^{R} c_{j,r} = 1$ et R=dim(y). Ils dérivent alors les formules de réestimations pour cette fonction de densité particulière.

Nous souhaitons nous appliquer le concept de pondération à *l'intersection des évènements* $(Y_t^r = y_t^r)$ $pour$ $r = [1:R]$.

Pour cela, on introduit $c_{j,r}$ $j = [1:K], r = [1:R]$ avec R= dim(y) et on pose :

$$\tilde{g}_j(y) = \prod_{r=1}^{R} [g_{j,r}(y^{(r)})]^{c_{j,r}} \quad avec \prod_{r=1}^{R} c_{j,r} = 1 \ . \tag{16}$$

$y^{(r)}$ est la *r^ième^ dimension de y*. $y = [y^{(1)}, \ldots y^{(r)}, \ldots y^{(R)}]$

$c_{j,r}$ est la pondération de la r^ième^ dimension pour le j^ième^ état.

et $g_{j,r}$ la densité d'une loi normale de paramètre $[m_{j,r} \ , \ \sigma_{j,r}]$.

Un tel modèle a donc comme paramètre $\theta = \{\pi, p, m, \sigma, c\}$.

Dans ce cas les algorithmes pour le calcul de la probabilité d'observer une séquence (2.1.2.a) et pour estimer la séquence d'état le plus probable (2.1.2.b) vus dans le cas des MMCs sont encore valables. La fonction de densité, associée à chaque état, est alors $\tilde{g}_i(y)$ spécifiée comme en (16).

Pour calculer la formule de réestimation de c, on utilise l'algorithme E-M, sur

$$\begin{aligned}\mathrm{p}(y,s|M') &= \pi_{s_1}\tilde{g}_{s_1}(y_1)\prod_{t=2}^{T} p_{s_{t-1},s_t}\tilde{g}_{s_t}(y_t) \\ &= \pi_{s_1}\tilde{g}_{s_1}(y_1)\prod_{t=2}^{T} p_{s_{t-1},s_t}\prod_{r=1}^{R}[g_{j,r}(y_t)]^{c_{r,s_t}}\end{aligned}$$

En reprenant la définition de la fonction Q, spécifié en (8), on a :

$$\begin{aligned}Q(\theta^{(m+1)},\theta^{(m)}) &= \sum_{s\in[1:K]^T}\mathrm{p}(y,s,\theta^{(m)})\log(\mathrm{p}(y,s|\theta^{(m+1)})) \\ &= \sum_{s\in[1:K]^T}\log(\pi_{s_1}^{(m+1)})\,\mathrm{p}(y,s,\theta^{(m)}) + \sum_{s\in[1:K]^T}\sum_{t=2}^{T}\log(p_{s_{t-1},s_t}^{(m+1)})\,\mathrm{p}(y,s|\theta^{(m)}) \\ &\quad + \sum_{s\in[1:K]^T}\sum_{t=1}^{T}\sum_{r=1}^{R} c_{s_t,r}^{(m+1)}\log(g_{j,r}(y_t^{(r)}))\,\mathrm{p}(y,s|\theta^{(m)})\end{aligned}$$

De plus, pour intégrer la contrainte $\prod_{r=1}^{R} c_{j,r}=1$, on maximise Q par rapport $c_{j,r}$ en utilisant la méthode du multiplicateur de Lagrange. Finalement on obtient :

$$c_{j,r}^{(m+1)} = \frac{\eta_j}{\sum_{t=1}^{T}\alpha(t,j)\beta(t,j)\log(g_{j,r}(y_t^{(r)}))} \tag{17}$$

avec $\eta_j=[\prod_{r=1}^{R}\sum_{t=1}^{T}\alpha(t,j)\beta(t,j)\log(g_{j,r}(y_t^{(r)}))]^{1/R}$ $\forall r\in[1:R],\ j\in[1:K]$.

b.2 Modèle Semi-Markovien Caché Pondéré (MSMCP)

Pour les besoins de notre étude, nous avons dû développer un MMC mixte alliant

1. Les avantages des semi-MMCs qui incluent la possibilité de modéliser explicitement la durée des phases.
2. Les avantages des MMCs pondérés qui incluent la notion d'importance relative d'une variable dans la définition d'un état.

C'est à dire :

1. la durée de passage D dans un état j est modélisée par une loi de probabilité spécifique : h_j de paramètre ς_j
2. La fonction de densité $\tilde{g}_j(y)$ associée à chaque état est de la forme (16).

Un tel modèle a donc pour paramètre $\theta=\{\pi, p, m_{j,r}, \sigma_{j,r}, c, \varsigma_j, j\in[1:K], r\in[1:R]\}$.

Inférence et décodage :

Dans ce cas, en spécifiant la fonction de densité $\tilde{g}_j(y)$ comme en (16), les algorithmes vus dans le cas des Semi-MMCs sont encore valables.

De la même manière que précédemment, on détermine la formule de réestimation du paramètre c. On a alors la formule suivante:

$$c_{j,r}^{(m+1)}=\frac{\eta_j}{\sum_{t=1}^{T}\sum_{d=1}^{D}\alpha^*(t-d,j)\,h_j(d)\,\beta^*(t-d,j)\log(z_{j,r}(y_{t\ldots t+d}))} \tag{18}$$

avec $\forall r\in[1:R], j\in[1:K]$,

$$\eta_j=[\prod_{r=1}^{R}\sum_{t=1}^{T}\sum_{d=1}^{D}\alpha^*(t-d,j)\,h_j(d)\beta^*(t-d,j)\,z_j(y_{t\ldots t+d})\log(\tilde{z}_{j,r}(y_{t\ldots t+d}))]^{1/R} ,$$

$$\tilde{z}_j(y_{t\ldots t+d})=\prod_{s=t}^{t+d}g_j(y_s) ,$$

$$z_{j,r}(y_{t\ldots t+d})=\prod_{s=t}^{t+d}g_{j,r}(y_s) ,$$

et $\mathrm{p}(y_t, s_t=j/\theta^{(m)})=\sum_{d=1}^{D}\alpha^*(t-d,j)\,h_j(d)\,\beta^*(t-d,j)$.

La définition et le calcul des fonctions α^* et β^* sont donnés dans l'annexe 1 .

Dans cette partie, nous avons donc vu les principales méthodes, ayant trait aux chaînes de Markov cachées, dont l'application fera l'objet de la troisième partie de ce mémoire.

La section suivante est consacrée à la description et à l'étude des propriétés d'un autre type de modèle, le modèle multi-phasique.

2.2 Estimation paramétrique dans un modèle multi-phasique par maximum de vraisemblance.

Dans cette partie nous nous intéressons à l'estimation des paramètres de modèles multi-phasiques phases linéaires et non linéaires. Pour cela nous nous baserons sur les travaux de Koul & Qian [Koul & Qian, 2002] ainsi que sur les travaux de Ciuperca [Ciuperca, 2004].

Le problème d'estimation des points de ruptures a été largement étudié dans la littérature en utilisant différentes approches (voir Csorgo et Horva [Csorgo & Horvath, 1997] pour une revue complète sur le sujet). Carlstein [Carlstein, 1988], Facer et Muller [Facer & Muller, 2003], Loader [Loader, 1996], Eubank et Speckman [Eubank & Speckman, 1994], Wu et Chu [Wu & Chu, 1993] ont développé l'approche non paramétrique. Pour l'approche paramétrique, Feder [Feder, 1975], [Feder, 1975] obtient la distribution asympotique de l'estimateur des moinfadres carrés pour des modèles biphasique. Rukhin et Vajda [Rukhin A & Vajda I, 1997] considèrent le M-estimateur pour une régression non-linéaire. Bhattacharya [Bhattachary, 1994] examine le comportement limite de l'estimateur du maximum de vraisemblance (MLE) (voir aussi Schulze [Schulze, 1987], Van der Geer [Van der Geer, 1988], Gill et Baron [Gill & Baron, 2004], et Gill [Gill, 2004]).

Pour des mesures aléatoires, à notre connaissance, les seules références sont Koul et Qian [Koul & Qian, 2002] pour le cas linéaire et Ciuperca [Ciuperca, 2004] pour le cas non linéaire.

Koul et Qian [Koul & Qian, 2002] ont étudié le cas d'un modèle de régression linéaire biphasique discontinue, avec un saut fixe, et une distribution des erreurs générales. Dans ce cas, ils ont montré que l'estimateur du MLE est n-constistant et que le processus de vraisemblance converge faiblement vers un processus de Poisson composé. Puis en 2004, Ciuperca [Ciuperca, 2004] a étendu ces résultats au cas non linéaire.

Ces auteurs ont étudié le cas d'un unique point de rupture. Pourtant, dans de nombreuses applications, il y a plusieurs point de ruptures. Aussi par la suite, nous étudierons les propriétés des estimateurs pour une régression paramétrique avec $K \geq 1$ point de ruptures inconnues. Nous verrons qu' à cause de la multiplicité des phases et des points de ruptures les démonstrations sont légèrement différentes du cas à rupture unique.

Dans une première partie, nous ferons des hypothèses sur le modèle étudié et nous introduirons des notations. Puis nous étudierons, pour le cas linéaire, la convergence de l'estimateur du maximum de vraisemblance et sa vitesse de convergence. Dans une troisième partie, nous analyserons la distribution limite du maximum de vraisemblance et la limite faible du processus de vraisemblance en fonction des paramètres de rupture. Enfin, nous étendrons, dans une dernière partie, ces résultats au cas non-linéaire.

2.2.1. Notations et premiers résultats

Soit le modèle multi-phasique linéaire à K ruptures :

$$Y_i=(a_0+b_0X_i)\mathbb{1}_{(0<X_i\le r_1)}+(a_1+b_1X_i)\mathbb{1}_{(r_1<X_i\le r_2)}+(a_2+b_2X_i)\mathbb{1}_{(r_2<X_i\le r_3)} \\ +\ldots+(a_k+b_kX_i)\mathbb{1}_{(r_{r_{k-1}}<X_i\le r_k)}+\ldots+(a_K+b_KX_i)\mathbb{1}_{(r_K<X_i<1)}+\epsilon_i,\quad i=1....n \tag{19}$$

$0\le X_i\le 1,\ a_i,b_i,r_i\in\mathbb{R},\quad r_{k-1}<r_k\quad \forall i,k>1$

Les suites $(X_i,\epsilon_i)_{1<i<n}$ sont des suites de variables aléatoires et indépendantes avec la même distribution que $(X_{1,}\epsilon_1)$ et

1. $\theta=(\theta_1,\tau)$, avec $\theta_1=(a_{0\,,}\ b_{0\,,}\ a_k,b_k)_{k\in[1:K]}$, $\theta_1\in\Gamma^{(K+1)*2}\subseteq\mathbb{R}^{(K+1)*2}$, $\Omega=\Gamma^{(K+1)*2}*(0,1)^K$ Γ un compact et $\tau\in(0,1)^K$.

2. la loi de X a une densité de Lebesgue g absolument continue et positive sur $(0,1)$.

3. $X_i et\epsilon_i$ sont indépendants, $\forall\, i\in[1:n]$

4. La variable aléatoire ϵ_i a une fonction de densité φ et $E\epsilon_i=0$, $E\,\epsilon_i^2<\infty\quad i=1....n$.

De plus, nous supposons que les conditions suivantes sont vérifiées :

(C1) φ *est absolument continue et positive sur* $\mathbb{R}$ *avec* φ' *'sa dérivée et* $u=\dfrac{\varphi'}{\varphi}$

(C2) la fonction u *est Lip(1) et* $I(\varphi):=\int u^2(x)\varphi(x)\,dx<\infty$.

(C3) u *est différentiable et sa dérivée* u' *est Lip(1).*

(C4) $\mathbb{E}\,(|X|^3)<\infty$

En outre, on suppose que la taille du k^{ème} saut à l'endroit du point de rupture r_k est différente de 0, pour tout $k\in[1...K]$.

$$d_k=a_k-a_{k-1}+r_k(b_k-b_{k-1})\neq 0\quad \forall k\in[1...K] \tag{20}$$

Le paramètre r=[r_1,$r_{2,\ldots}r_k$] est appelé le vecteur de rupture de la fonction de régression.

On suppose que $\theta=(a_{0\,,}\ b_{0\,,}\ a_k,b_k,\tau_k)_{k\in[1:K]}$ est un point intérieur de l'espace des paramètres Ω et on note $\tau=(\tau_k)_{k\in[1:K]}$.

On définit la fonction $f:\mathbb{R}*\Omega\to\mathbb{R}$ par

$$f(x,\theta)=(a_0+b_0X_i)1_{(0<X_i\le\tau_1)}+....(a_k+b_kX_i)1_{(\tau_{k-1}<X_i\le\tau_k)}+...+(a_K+b_KX_i)1_{(\tau_K<X_i\le 1)}$$

De plus, pour toute fonction $\phi:\mathbb{R}*\Omega\to\mathbb{R}$, on défini

$\phi'(x,\theta):=\partial\phi/\partial x$, $\dot{\phi}(x,\theta):=\partial\phi/\partial\theta_1$ et $\ddot{\phi}(x,\theta):=\partial^2/\partial\theta_1^2$.

Par commodité, nous écrirons $\varphi_\theta(y,x):=\varphi(y-f_\theta(x))$, $\psi_\theta:=\log(\phi_\theta/\phi_{\theta_0})$

La vraisemblance est alors notée

$$L_n(\theta):=\prod_{i=1}^{n}\varphi(Y_i-f(X_i,\theta))=\prod_{i=1}^{n}\varphi_\theta(Y_i,X_i) \tag{21}$$

Pour tout vecteur, on définit $\|.\|$ pour la norme Euclidienne, et pour toute matrice $A=(a_{i,j})$, $\|A\|=\sum_{i,j}|a_{i,j}|$, et C une constante générique, positive et finie, ne dépendant pas de n.

a Notations

Soit θ^0 le vrai paramètre (inconnu), $\theta^0:=(\theta_1^0,r)$ avec $\theta_1^0=(a_0^0,b_0^0,....a_K^0,b_K^0)$ et $r=(r_1,...,r_K)$.

De plus, on définit : $\theta_\tau=(\theta_1,\tau)$, $\theta_r=(\theta_1,r)$, $\theta_\tau^0=(\theta_1^0,\tau)$

et $\hat{\theta}_{(n),1}(\tau)=\underset{\theta_1\in\Gamma^{(K+1)*2}}{argmax}\ L_n(\theta_\tau)$, pour $\tau\in[0,1]$.

Par la suite, pour faciliter la lecture on pose $\phi_{\_k}(x):(0,1)^K*\mathbb{R}\to\mathbb{R}$:

$$\phi_{\_k}(x,\tau):=\phi(\tau_1,...\tau_k,r_{k+1},...r_K)(x)\quad\forall k\in[1:K]\ . \tag{22}$$

Pour tout $\forall\eta>0$, on définit un $\eta-$ voisinage de θ par

$$\Omega_\eta(\theta):=\{\theta^*=(\theta_1^*,\tau^*)\in\bar{\Omega}:\|\theta_1^*-\theta_1\|\le\eta,\|\tau^*-\tau\|\le\eta\}\ .$$

Pour $B\in(0,\infty)$ on définit D tel que $D(r,B/n):=\{\tau/\|\tau-r\|\le B/n\}$.

On pose $\psi(x,y,\theta)=\log\dfrac{\varphi(y-f(x,\theta))}{\varphi(y-f(x,\theta^0))}$ $(x,y)\in\mathbb{R}^2$.

Enfin, pour tout $k=1...K$, $i=1...n$, on définit $\xi_{k,i}:\Omega\to\mathbb{R}$ et $v_k:\mathbb{R}*\mathbb{R}\to\mathbb{R}$ par

$$\xi_{k,i}(\theta_1,\tau) := \psi_{\_k}(X_i;(\theta_1,\tau)) - \psi_{\_(k-1)}(X_i;(\theta_1,\tau)) = \log\frac{f_{\_k}(X_i;\theta_1,\tau)}{f_{\_(k-1)}(X_i;\theta_1,\tau)} \quad (23)$$

et

$$\nu_k(x,z) := \log\frac{\varphi(z+(a_k^0-a_{k-1}^0)+(b_k^0-b_{k-1}^0)x)}{\varphi(x)} \quad k=1...K \quad (24)$$

b Calcul du MLE

Pour calculer le MLE, nous procédons comme suit :

Etape 1) Pour chaque $\tau=(\tau_k)_{k\in 1:K}\, avec\, \tau_k<\tau_{k+1}$ fixé, nous calculons $\hat{\theta}_{(n),1}(\tau)$ maximisant $L(\hat{\theta}_{(n),1}(\tau),\tau)$ avec $\hat{\theta}_{(n),1}(\tau)$ dans K.

Remarque : i) $\hat{\theta}_{(n),1}(\tau)$ est constant pour chaque intervalle entre 2 valeurs consécutives des X_i ordonnés.

ii) la fonction de vraisemblance $L(\hat{\theta}_{(n),1}(\tau),\tau)$ admet un nombre fini de valeurs possibles. Ces valeurs sont prises pour τ tel que $\{\tau_k\in\{X_{1,}...X_n\}\ /k\in[1..K]\}$.

Etape 2) On calcule $\hat{r}_n = \max_{\tau\in\{X_i,\, i=1:n\}^K\}} (L(\hat{\theta}_{(n),1}(\tau),\tau))$.

Etape 3) L'estimateur $\hat{\theta}_n=(\hat{\theta}_{(n),1}(\hat{r}_n),\hat{r}_n)=(\hat{\theta}_{(n),1},\hat{r}_n)$ est l'estimateur du maximum de vraisemblance du paramètre θ^0 .

c Quelques résultats préliminaires

Dans notre cas, on a :

$$\dot{f}_\tau(x)=\partial/\partial\theta_1 f_\tau(x,\theta_1)$$
$$=(1_{(x\le\tau_1)}, x1_{(x\le\tau_1)}, 1_{(\tau_1<x\le\tau_2)}, x1_{(\tau_1<x\le\tau_2)}, ...1_{(\tau_{k-1}x\le\tau_k)}, x1_{(\tau_k<x\le\tau_{k-1})}, ..., 1_{(\tau_K<x\le 1)}, x1_{(\tau_K<x\le 1)})$$

On observe que $||\dot{f}_\tau(x)||=\sqrt{1+x^2}\quad x\in R, \tau\in(0,1)^K$ et que $f'_\tau(x)=\theta_1{}'\dot{f}_\tau(x)$ (25).

Aussi pour $\forall x\in R, \forall\theta^*, \theta\in\Theta$, on a :

$$
\begin{aligned}
&||f(x,\theta)|| \le ||\theta_1|| \sqrt{1+x^2} \\
&||f_\tau(x,\theta_1) - f_\tau(x,\theta_1^*)|| \le ||\theta_1 - \theta_1^*|| \sqrt{1+x^2}
\end{aligned}
\tag{26}
$$

De plus, pour $\forall x \in R, \theta^*, \tau \in (0,1)^K, \tau^* \in (0,1)^K$, on observe que :

$$
||\dot{f}_\tau(x) - \dot{f}_{\tau^*}(x)|| \le \sqrt{2(1+x^2)} \left[\sum_{k=1}^{K} 1_{(min(\tau_k, \tau_k^*) \le x \le max(\tau_k, \tau_k^*))} \right] \tag{27}
$$

et donc que

$$
||\dot{f}_\tau(x) - \dot{f}_{\tau^*}(x)|| \le \sqrt{2(1+x^2)} \left[\sum_{k=1}^{K} 1_{(|x-\tau_k| \le |\tau_k - \tau_k^*|)} \right] \tag{28}
$$

2.2.2. Convergence de l'estimateur

Nous nous intéressons tout d'abord à la convergence forte de l'estimateur $\hat{\theta}_{(n)}$.

Par la suite nous avons besoin du lemme suivant

Lemme 1 : *On suppose que l'hypothèse (C1) est satisfaite, et*

- *soit* u *est bornée et* $E|X| < \infty$ *(29)*
- *soit* $(C2)$ *est vérifié et* $E X^2 < \infty$ *(30)*

alors

$$
E\left[\sup_{\theta^* \in \Omega_\eta(\theta)} |\psi(X,Y,\theta^*) - \psi(X,Y,\theta)| \right] \to 0. \text{ pour } \eta \to 0 \tag{31}
$$

Preuve du Lemme 1 : Fixons $\theta \in \Theta$ et définissons ϵ *et* δ tel que $\epsilon(\theta) = Y - f(X,\theta)$ et $\delta(X,\theta^*) := f(X,\theta) - f(X,\theta^*) = \theta_1' \dot{f}_\tau(X) - \theta_1'^* \dot{f}_{\tau^*}(X), \quad \forall \theta^* \in \Omega$.

Pour tout $\eta > 0$, $\tau \in [0,1]^K$, on définit $\tau_k(\eta)$ par la relation $|\tau_k(\eta) - \tau_k| = \eta$, $k = 1...K$.

Par (26) et (28), on a pour $\theta^* \in \Omega_n(\theta)$ et pour tout $\tau \in [0,1]$,

$$
\begin{aligned}
|(\delta(X,\theta^*))| &\le |\theta_1' [\dot{f}_\tau(X) - \dot{f}_{\tau^*}(X)]| + |[\theta_1 - \theta_1^*]' \dot{f}_{\tau^*}(X)| \\
&\le [\sqrt{2}\, ||\theta_1|| [\sum_{k=1}^{K} 1_{(|X-\tau_k| \le |\tau_k^* - \tau_k|)}] + ||\theta_1 - \theta_1^*||] \sqrt{1+X^2} \\
&\le [\sqrt{2}\, ||\theta_1|| [\sum_{k=1}^{K} 1_{(|X-\tau_k| \le |\tau_k^* - \tau_k|)}] + \eta] \sqrt{1+X^2} \\
&= \Delta(\eta, X)
\end{aligned}
\tag{32}
$$

Puis, avec l'absolue continuité de $\log\varphi$, on a :

$$\left|\psi(X,Y,\theta^*)-\psi(X,Y,\theta)\right| \le \int_{-\Delta(\eta,X)}^{\Delta(\eta,X)} \left|\phi(\epsilon(\theta)+v)\right| dv \quad \text{Avec} \quad \epsilon(\theta)=Y-f_\theta(X) \ . \tag{33}$$

Le reste de la preuve est identique à celle du Lemme 3.1 dans Koul et Qian [Koul & Qian, 2002] pour un seul point de rupture.

■

Théorème 1 : *Sous les hypothèses du Théorème 1, on a pour tout* $\theta\in\bar{\Omega}$

Alors $\hat{\theta}_{(n)} \underset{n\to\infty}{\overset{p.s}{\to}} \theta_0$.

Preuve du Théorème 1 :En utilisant le Lemme 1, on a de la même manière que dans Koul et Qian [Koul & Qian, 2002] la preuve du Théorème 1.

■

Remarque : Si ϵ_i suit une loi normale, logistique, double exponentielle ou de Student à p>2 degrés liberté, les conditions du théorème sont satisfaites.

2.2.3. Vitesse de convergence

Dans ce paragraphe nous étudions la vitesse de convergence de l'EMV. Pour cela, nous avons besoin de quelques résultats préliminaires:

$\forall x\in R$, $\forall k\in[1..K]$, on définit $u=(u_1,\dots u_K)\in(0,1)^K$ tel que

$r_{k+1}>r_k+u_k \quad \forall k=1,\dots,K$ et les fonctions suivantes:

$$p_k(x)=\mathbb{E}_\epsilon[\nu_k(x,\epsilon)],\quad p_{1,k}(x)=\mathbb{E}_\epsilon[\left|\nu_k(x,\epsilon)\right|],\quad p_{2,k}(x)=\mathbb{E}_\epsilon[\nu_k^2(x,\epsilon)]$$

$$R_{k,n}(u_k)=n^{-1}\sum_{i=1}^{n}\nu_k(X_i,\epsilon_i)\,\mathbb{1}_{\min(r_k,r_k+u_k)<X_i\le\max(r_k,r_k+u_k)}$$

$$r_{k,n}(u_k)=n^{-1}\sum_{i=1}^{n}p_k(X_i)\,\mathbb{1}_{\min(r_k,r_k+u_k)<X_i\le\max(r_k,r_k+u_k)}$$

$$G_k(u_k)=E_X[\mathbb{1}_{\min(r_k,r_k+u_k)<X_i\le\max(r_k,r_k+u_k)}],\quad G_{k,n}(u_k)=n^{-1}\sum_{i=1}^{n}\mathbb{1}_{\min(r_k,r_k+u_k)<X_i\le\max(r_k,r_k+u_k)}$$

$$G(u)=\sum_{k=1}^{K}G_k(u_k)\quad G_n(u)=\sum_{k=1}^{K}G_{k,n}(u_k)$$

Définissons aussi les fonctions $D_{k,n}, d_{k,n}:\mathbb{R}\to\mathbb{R}$ par

$$D_{k,n}(u_k)=n^{-1}\sum_{i=1}^{n} \nu_k(X_i,\epsilon_i).\,1_{min(r_k,r_k+u_k)<X_i\le max(r_k,r_k+u_k)}$$

$$d_{k,n}(u_k)=n^{-1}\sum_{i=1}^{n} p_k(X_i,\epsilon_i).\,1_{min(r_k,r_k+u_k)<X_i\le max(r_k,r_k+u_k)}$$

Lemme 2 : *On suppose que les fonctions $p_{1,k}$ et $p_{2,k}$ sont bornées sur des ensembles bornés. Alors pour chaque $\gamma>0,\ \eta>0$ il existe une constante $B<\infty$ tel que $\forall\, 0<\delta<1$, $\forall\, n\ge[B/\delta]+1$, et $\forall\, k\in[1..K]$ on a :*

$$P\Big(\sup_{B/n<u\le\delta}\left|G_n(u)/G(u)-1\right|<\eta\Big)>1-\gamma \tag{34}$$

$$P\Big(\sup_{B/n<u\le\delta}\left|\frac{R_{k,n}(u_k)-R_k(u_k)}{G(u)}\right|<\eta\Big)>1-\gamma \tag{35}$$

$$P\Big(\sup_{B/n<u\le\delta}\sum_{k=1}^{K}\left|\frac{D_{k,n}(u_k)-d_k(u_k)}{G(u)}\right|<\eta\Big)>1-\gamma \tag{36}$$

Preuve du Lemme 2: On démontre seulement la relation (34). La démonstration des relations (35) et (36) est similaire.

Le Lemme 3.2 dans Koul et Qian [Koul & Qian, 2002] implique que, pour chaque $\gamma>0,\ \eta>0$, il existe une constante $B<\infty$ tel que $\forall\, 0<\delta<1$, $\forall\, k\in[1..K]$, $\forall\, n\ge[B/\delta]+1$,

$$P\Big(\sup_{B/n<u\le\delta}\left|G_{k,n}(u_k)/G_k(u)-1\right|<\eta\Big)>1-\gamma \tag{37}$$

$$P\Big(\sup_{B/n<u\le\delta}\left|\frac{R_{k,n}(u_k)-R_k(u_k)}{G_k(u_k)}\right|<\eta\Big)>1-\gamma \tag{38}$$

$\forall\, k=1\ldots K$, on a alors

$$P\Big(\sup_{B/n<u\le\delta}\left|\frac{\sum_{k=1}^{K}G_{k,n}(u_k)}{\sum_{k=1}^{K}G_k(u)}-1\right|<\eta\Big)>P\Big(\sup_{B/n<u\le\delta}\left|\frac{\sum_{k=1}^{K}\left|G_{k,n}(u_k)-G_k(u_k)\right|}{\sum_{k=1}^{K}G_k(u)}\right|<\eta\Big)>$$

$$P\Big(\sup_{B/n<u\le\delta}\left|\frac{\sum_{k=1}^{K}\left|G_{k,n}(u_k)-G_j(u_k)\right|}{G_k(u)}\right|<\eta\Big)>1-\gamma$$

ce qui conclu la démonstration.

■

Nous pouvons maintenant démontrer le théorème suivant:

Théorème 2 : *Si (C1), et les relations (20) et (35) sont vérifiées*
alors $|n(\hat{r}_{(n),k}-r_k)|=O_p(1) \quad \forall k\in[1,,K]$.

Preuve du Théorème 2: Pour démontrer le théorème 2, nous avons besoin du résultat suivant:
$\forall \gamma>0$, $\exists aB<\infty, \gamma_1>0$, $0<\delta<1,$ et $n_0\in\mathbb{N}$, et $\forall n>n_0$,

$$\mathrm{P}\left[\sup_{B/n<u\le\delta,\theta\in\Omega(\delta)}\frac{L_n(\theta_1,\tau)-L_n(\theta_1,r)}{G|\tau-r|}<-\gamma_1\right]>1-2\gamma \quad \forall n>n_0 \tag{39}$$

Or, l'égalité suivante est vérifiée:

$$L_n(\theta_1,\tau)-L_n(\theta_1,r)=[\ (L_n(\theta_1,\tau)-L_n(\theta_1,r))\ -\ (L_n(\theta_1^0,\tau)-L_n(\theta_1^0,r))]+(L_n(\theta_1^0,\tau)-L_n(\theta_1^0,r))$$
$$:=L_n^1(\theta)-L_n^2(\theta)$$

Pour démontrer (39), on prouve que pour $\forall \gamma>0$, $\exists B<\infty, \gamma_1>0$, $0<\delta<1,$ et $n_0\in\mathbb{N}$, et $\forall n>n_0$, on a

$$\mathrm{P}\left[\sup_{B/n<u\le\delta,\theta\in\Omega(\delta)}\frac{|L_n^1(\theta)|}{G|\tau-r|}>\gamma_1\right]<\gamma \tag{40}$$

$$\text{et}\quad \mathrm{P}\left[\sup_{B/n<u\le\delta}\frac{L_n^2(\theta)}{G|\tau-r|}<-2\ \gamma_1\right]<1-\gamma \tag{41}$$

On a

$$L_n^1(\theta)=n^{-1}\sum_{k=1}^{K}\sum_{i=1}^{n}\int_0^1\dot{\xi}_{k,i}{}'(\theta_1^0+v(\theta_1-\theta_1^0),\tau)\,dv(\theta_1-\theta_1^0)^t \tag{42}$$

On définit alors $\theta_1(v)=\theta_1^0+v(\theta_1-\theta_1^0)$ et $\theta(v):=(\theta_1^0(v),\tau)$ avec $v\in[0,1]$.

En notant $\dot{f}_{\_k}$, la fonction $\dot{f}_{\_k}(X_i;\theta_1(v),\tau)$, l'égalité suivante est vérifiée.

$$\dot{\xi}_{k,i}(\theta_1(v),\tau) = -u(Y_i-\theta_1'\dot{f}_{\_k})\dot{f}_{\_k}+u(y_i-\theta_1'\dot{f}_{\_k\text{-}1})\dot{f}_{\_k\text{-}1}(X_i,\tau)$$
$$= -[u(Y_i-\theta_1'\dot{f}_{\_k}(X_i,\tau))-u(Y_i-\theta_1'\dot{f}_{\_k-1}(X_i,\tau))]\dot{f}_{\_k\text{-}1}(X_i,\tau)$$
$$-u(Y_i-\theta_1'\dot{f}_{\_k\text{-}1}(X_i,\tau))[\dot{f}_{\_k}(X_i,\tau)-\dot{f}_{\_k\text{-}1}(X_i,\tau)]$$

Alors en utilisant le fait que $u\in Lip(1)$, on obtient les inégalités suivantes :

$$\dot{\xi}_{k,i}(\theta_1,\tau) \le C\Big[\|\theta_1\|.\|\dot{f}_{\_k}-\dot{f}_{\_k\text{-}1}\|(1+X_i)^{1/2}+|u(y_i-\theta_1'\dot{f}_{\_k\text{-}1})|.\|\dot{f}_{\_k}-\dot{f}_{\_k\text{-}1}\|\Big]$$
$$\le C\Big[\|\theta_1\|.\|\dot{f}_{\_k}-\dot{f}_{\_k\text{-}1}\|(1+X_i)^{1/2}$$
$$+C\Big[\|\theta_1\|\sum_{j=1}^{k-1}.\|\dot{f}_{\_j}-\dot{f}_{\_j\text{-}1}\|\Big]+\|\theta_1-\theta_1^0(1+X_i)^{1/2}\|+u(\epsilon_i)\Big]\|\dot{f}_{\_k}-\dot{f}_{\_k\text{-}1}\|$$

En utilisant $||\dot{f}_\tau(x)-\dot{f}_{\tau^*}(x)|| \le \sqrt{2(1+x^2)}\Big[\sum_{k=1}^{K} 1_{(|x-\tau_k|\le|\tau_k-\tau_k^*|)}\Big]$, et le fait que $\|\theta_1\|$ soit borné, on obtient alors :

$$\dot{\xi}_{k,i}(\theta_1,\tau) \le C\Big[(\sqrt{(1+X_i^2)}\,1_{|X_i-\tau_k|\le|\tau_k-r_k|})+$$
$$\sqrt{(1+X_i^2)}\sum_{j=1}^{k-1}1_{|X_i-\tau_j|\le|\tau_j-r_j|}+\delta\sqrt{(1+X_i^2)}+u(\epsilon_i)\Big]\sqrt{(1+X_i^2)}\,1_{|X_i-\tau_k|\le|\tau_k-r_k|} \qquad (43)$$

Par conséquent, l'inégalité suivante est vérifiée:

$$\dot{\xi}_{k,i}(\theta_1,\tau) \le C\Big[\sqrt{(1+X_i^2)}\sum_{j=1}^{k-1}1_{|X_i-\tau_j|\le|\tau_j-r_j|}+\delta\sqrt{(1+X_i^2)}+u(\epsilon_i)\Big]\Big|\sqrt{(1+X_i^2)}\,1_{|X_i-\tau_k|\le|\tau_k-r_k|} \qquad (44)$$

En utilisant les relations (42) et (44), on obtient :

$$|L_n^1(\theta)| \le C\delta\, n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\Big[\sum_{j=1}^{k}1_{|X_i-\tau_k|\le|\tau_k-r_k|}+\delta+|u(\epsilon_i)|\Big]1_{r_k<X_i\le r_k+u_k}$$
$$\le C\delta\, n^{-1}\Big[\sum_{i=1}^{n}\sum_{k=1}^{K}1_{r_k<X_i\le r_k+u_k}+\sum_{i=1}^{n}\sum_{k=1}^{K}|u(\epsilon_i)|\,1_{|X_i-\tau_k|\le|\tau_k-r_k|}\Big]$$
$$\le C\delta\Big[G_n(|\tau-r|)+n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}|u(\epsilon_i)|\,1_{|X_i-\tau_k|\le|\tau_k-r_k|}\Big]$$

Or on a $u=\tau-r$ et on suppose maintenant que $\tau\ge r$. Donc

$$|L_n^1(\theta)| \le C\delta\sum_{k=1}^{K}\Big[|G_{k,n}(u_k)|+|R_{k,n}(u_k)-r_{k,n}(u_k)|+|r_{k,n}(u_k)|\Big]$$

Or $r_{k,n}(u_k) < C.\,G_{k,n}(u_k)$ donc, on obtient l'inégalité suivante :

$$\left|L_n^1(\theta)\right| \le C\delta \sum_{k=1}^{K} \left[\left|R_{k,n}(u_k) - r_{k,n}(u_k)\right| + \left|G_{k,n}(u_k)\right|\right]$$

D'où :

$$\frac{\left|L_n^1(\theta)\right|}{G(u)} \le C\delta \left[\left| \frac{\sum_{k=1}^{K} \left[R_{k,n}(u_k) + r_{j,n}(u_k)\right]}{\sum_{k=1}^{K} G_k(u_k)} - 1 \right| + \frac{\sum_{k=1}^{K} G_{k,n}(u_k)}{\sum_{k=1}^{K} G_k(u_k)}\right].$$

Cette relation avec le Lemme 2 implique la relation (40).

Pour démontrer la relation (41), on utilise l'égalité suivante :

$$L_n^2(r+u) = L_n(\theta_1^0, r+u) - L_n(\theta_1^0, r) = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \nu_k(X_i, \epsilon_i)\, \mathbb{1}_{r_k < X_i \le r_k + u_k}$$

On a alors

$$\begin{aligned} \frac{L_n^2(r+u)}{G(u)} \le & \frac{\left| \sum_{k=1}^{K} \frac{1}{n} \sum_{i=1}^{n} \left[\nu_k(X_i, \epsilon_i) - p_k(X_i)\right] \mathbb{1}_{r_k < X_i \le r_k + u_k} \right|}{G(u)} \\ & + \sum_{j=1}^{K} \left| \frac{1}{n} \frac{\sum_{i=1}^{n} \left[p_k(X_i) - p_k(r_k)\right] \mathbb{1}_{r_k < X_i \le r_k + u_k}}{G(u)} \right| \\ & + \sum_{k=1}^{K} \left|p_k(r_k)\right| \frac{\left|G_{k,n}(u_k) - G_k(u_k)\right|}{G(u)} \\ & + \sum_{k=1}^{K} p_k(r_k) \frac{G_k(u_k)}{G(u)} \end{aligned} \qquad (45)$$

Nous allons majorer les quatre termes de (45). Pour cela, rappelons tout d'abord que $G(u) = \sum_{k=1}^{K} G_k(u_k)$.

Pour le premier terme de (45), on a la relation (36).

Pour le deuxième terme de (45), on a l'inégalité suivante :

$$\sup_{\|u\|} \sum_{k=1}^{K} \left| n^{-1} \frac{\sum_{i=1}^{n} \left[p_k(X_i) - p_k(r_k)\right] \mathbb{1}_{r_k < X_i \le r_k + u_k}}{G(u)} \right| \le \sup_{0 \le x \le \delta} \left|p_k(r_k + x) - p_k(r_k)\right| . \sup_{\|u\|} \frac{\sum_{k=1}^{K} G_{k,n}(u_k)}{G(u)} .$$

Donc parce que la fonction p_k est continue, on a pour $\delta \to 0$ $\sup_{0 \le x \le \delta} \left|p_k(r_k + x) - p_k(r_k)\right| = O_P(1)$ et

en conséquence du Lemme 2 :

$$\sup_{\|u\|} \frac{\sum_{k=1}^{K} G_{k,n}(u_j)}{G(u)} = 0_P(1) .$$

Donc le supremum du deuxième terme de (45) est $0_P(1)$

Pour la troisième terme de (45), on a d'abord le fait que $|p_k(r_k)| < \infty$, $\forall k = 1...K$, puis pour $\delta \to 0$, on a $\frac{\sum_{k=1}^{K} |G_{k,n}(u_k) - G_k(u_k)|}{G(u)} = \sum_{k=1}^{K} \left| \frac{G_{k,n}(u_k)}{G_k(u_k)} - 1 \right| . \frac{G_k(u_k)}{G(u)} = O_P(1) \frac{\sum_{k=1}^{K} G_k(u_k)}{G(u)} = O_P(1)$.

Puis, $0_P(1)$ étant uniforme en $u \in (\frac{B}{n}, \delta]^K$, le troisième terme de (45) est $0_P(1)$, uniformément en $u \in (\frac{B}{n}, \delta]^K$, pour $\delta \to 0$.

Enfin pour le quatrième terme de (45) on a tout d'abord le fait que $p_k(r_k) < 0$. La démonstration de cette inégalité est similaire à la preuve du Théorème 3.2 dans l'article de Koul et Qian [Koul & Qian, 2002].

Puis en utilisant le Lemme 2, on a:

$$\sup_{B/n < \|u\| < \delta} n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K} p_k(r_k) \frac{G_k(u_k)}{G(u)} < 0 .$$

Avec l'ensemble de ces arguments, la relation (41) est donc prouvée et le théorème aussi.

■

2.2.4. Distributions asymptotiques

Cette section donne la distribution limite du maximum de vraisemblance et la limite faible du « calcul de la vraisemblance » en fonction des paramètres de rupture.

Tout d'abord, on s'intéresse à la distribution limite de $\hat{\theta}_{(n),1}$

De la même manière que le Lemme 1, on a le résultat d'une « consistance uniforme » :

Lemme 3 : *si on suppose (C1), (C2) vérifiées, et* $E\ X^2 < \infty$

alors $\forall\ 0 < B < \infty$, $\sup_{\|\tau - r\| \le B/n} \left| \hat{\theta}_{(n),1}(\tau) - \theta_1^0(r) \right| = Op(1)$

La preuve de ce lemme est facilitée par le fait que sous les conditions précédentes

$$\mathsf{E}\Big(\sup_{\theta^*\in\Omega_\eta(\theta),\tau\in[0,1]^K}\Big|\psi(X,\epsilon,\theta_1^*,\tau)-\psi(X,\epsilon,\theta_1,\tau)\Big|\Big)\to 0 \quad quand\ \eta\to 0 \quad \forall\,\theta_1\in\Gamma^{(K+1)*2}$$

En utilisant ce fait, la démonstration du Lemme 3 est alors la même que celle du Théorème 1 lors de l'utilisation du Lemme 1.

■

On introduit les notations suivantes : $a_i(\theta_1,\tau):=-u(Y_i-\theta_1{}'\dot f_{\theta_1,\tau}(X_i))\dot f_{\theta_1,\tau}(X_i)$, $A_\tau(x):=\dot f_{\theta_\tau}(x)\dot f_{\theta_\tau}(x)$ $A_0(x):=\dot f_{\theta^0}(x)\dot f_{\theta^0}(x)$ $V:=I(\varphi)E_X(A_0(X))$.

Théorème 3 : *On suppose que les conditions (C1)-(C4) sont vérifiées alors* $\forall\,0<B<\infty$

$$\sup_{\|\tau-r\|\le B/n}\Big\|n^{1/2}(\hat\theta_{(n),1}(\tau)-\theta_1^0(r))+V^{-1}n^{-1/2}\sum_{i=1}^n a_i(\theta_1^0(r))\Big\|=Op(1)$$

$$(\hat\theta_{(n),1}-\theta_1^0)\underset{n\to\infty}{\overset{L}{\to}}N(0,\ V^{-1})$$

Preuve Théorème 3 : La démonstration est la même que dans le cas à une seule rupture.

■

On a $L_n(\theta_1,\tau)=\dfrac{1}{n}\sum_{i=1}^n\varphi(Y_i,X_i,\theta_1,\tau)$.

Pour plus de lisibilité on pose, pour $t=(t_1,\dots t_K)\in\mathbb{R}^K$:

$$\hat L_n(t)=n[L_n(\hat\theta_{(n),1}(r+t/n),r+t/n)-L_n(\hat\theta_{(n),1}(r),r)]$$

$$\tilde L_n(t)=n[L_n(\theta_1^0,r+t/n)-L_n(\theta_1^0,r)]\ .$$

Théorème 4 : *Si on suppose que (C1)-(C5) sont vérifiées, alors*

$$\forall\,0<B<\infty\ ,\ \sup_{|t|\le B}\big|\hat L_n(t)-\tilde L_n(t)\big|=O_{\mathsf{P}}(1)$$

Preuve duThéorème 4 L'égalité suivante est vérifiée :

$$\begin{aligned}\hat L_n(t)-\tilde L_n(t)=\sum_{i=1}^n\Big[\log\frac{\varphi(Y_i-\hat\theta_{(n),1}(r+t/n).\dot f_{\theta_1,r+t/n}(X_i))}{\varphi(Y_i-\hat\theta_{(n),1}(r+t/n).\dot f_{\theta_1,r}(X_i))}-\log\frac{\varphi(Y_i-\theta_1^0.\dot f_{\theta_1,r+t/n}(X_i)}{\varphi(Y_i-\theta_1^0.\dot f_{\theta_1,r}(X_i))}\\ +\sum_{i=1}^n\log\frac{\varphi(Y_i-\hat\theta_{(n),1}(r+t/n).\dot f_{\theta_1,r}(X_i))}{\varphi(Y_i-\theta^0_{(n),1}.\dot f_{\theta_1,r}(X_i))}:=\hat L_{1,n}(t)+\hat L_{2,n}(t)\end{aligned}\qquad(46)$$

On a donc,

$$\hat{L}_{1,n}(t)=\sum_{i=1}^{n}\sum_{k=1}^{K}[\xi_k(\hat{\theta}_{(n),1}(r+t/n),r+t/n)-\xi_k(\theta_1^0,r+t/n)] \tag{47}$$

Pour montrer le théorème on montrera que $\forall B\in(0,\infty)$

$$\sup_{\|t\|\le B}\left|\hat{L}_{1,n}(t)\right|=O_P(1) \tag{48}$$

et $$\sup_{\|t\|\le B}\left|\hat{L}_{2,n}(t)\right|=O_P(1) \tag{49}$$

Du fait du Théorème 4, pour montrer (48), il est suffisant de montrer que $\forall b\in(0,\infty)$,et pour $\sqrt{n}\|\theta_1-\theta_1^0\|\le b$, pour tout pour tout $k=1\ldots\mathrm{K}$, on a

$$\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\sum_{i=1}^{n}\xi_{k,i}(\theta_1(r+t/n),r+t/n)-\xi_{k,i}(\theta_1^0,r+t/n)]=0(n^{-1/2}) \tag{50}$$

Par commodité, et comme $f_{\_\mathrm{k}}(X_i;\theta_1,r+t/n)$ n'est pas dépendant de θ_1 , nous utiliserons par la suite, $f_{\_\mathrm{k}}(r+t/n)$ comme abréviation.

Aussi, pour tout $k=1\ldots\mathrm{K}$,

$$\begin{aligned}\dot{\xi}_{k,i}(\theta_1,r+t/n)&=-u(Y_i-\theta_1'\dot{f}_{_k}(r+t/n))\dot{f}_{_k}(r+t/n)+u(Y_i-\theta_1'\dot{f}_{_k\text{-}1}(X_i,r+t/n))\dot{f}_{_k\text{-}1}(r+t/n)\\
&=[-u(Y_i-\theta_1'\dot{f}_{_k}(r+t/n))+u(Y_i-\theta_1'\dot{f}_{_k\text{-}1}(r+t/n))].\dot{f}_{_k}(r+t/n)\\
&\quad+[-u(Y_i-\theta_1'\dot{f}_{_k\text{-}1}(r+t/n))-u(Y_i-\theta_1^{0\prime}\dot{f}_{_k\text{-}1}(r+t/n))]\\
&\qquad.[\dot{f}_{_k}(r+t/n)-\dot{f}_{_k\text{-}1}(r+t/n)]\\
&\quad+[-u(Y_i-\theta_1^{0\prime}\dot{f}_{_k\text{-}1}(r+t/n))+u(Y_i-\theta_1^{0\prime}\dot{f}_{_k\text{-}1}(r))]\\
&\qquad.[\dot{f}_{_k}(r+t/n)-\dot{f}_{_k\text{-}1}(r+t/n)]\\
&\quad-u(Y_i-\theta_1^{0\prime}\dot{f}_{_k\text{-}1}(r))[\dot{f}_{_k}(r+t/n)-\dot{f}_{_k\text{-}1}(r+t/n)]\end{aligned}$$

De part l'absolue continuité de $\log\varphi$, on a :

$$\begin{aligned}\xi_{k,i}(\theta_1,r+t/n)-\xi_{k,i}(\theta_1^0,r+t/n)&=\int_0^1\xi_{k,i}(\theta_{1,\gamma},t)\,d\gamma\,(\theta_1-\theta_1^0)\\
&\equiv U_{1,i,k}(\theta_1,t)+U_{2,i,k}(\theta_1,t)+U_{3,i,k}(\theta_1^0,t)+U_{4,i,k}(\theta_1^0,t)\end{aligned} \tag{51}$$

où $\theta_{1,\gamma}=\theta_1^0+\gamma(\theta_1-\theta_1^0)$ et

$$U_{1,i,k}(\theta_1,t)=\int_0^1 [-u(Y_i-\theta_{1,\gamma}'\dot{f}_{\_k}(r+t/n)) + u(Y_i-\theta_{1,\gamma}'\dot{f}_{\_k\text{-}1}(r+t/n))]\dot{f}_{\_k}(r+t/n)\,d\gamma(\theta_1-\theta_1^0) \tag{52}$$

$$U_{2,i,k}(\theta_1,t)=\int_0^1 [-u(Y_i-\theta_1'\dot{f}_{\_k\text{-}1}(r+t/n)) + u(Y_i-\theta_1^0'\dot{f}_{\_k\text{-}1}(X_i,r+t/n))][\dot{f}_{\_k}(r+t/n)-\dot{f}_{\_k\text{-}1}(r+t/n)]\,d\gamma(\theta_1-\theta_1^0) \tag{53}$$

$$U_{3,i,k}(\theta_1,t)=[-u(Y_i-\theta_1^0'\dot{f}_{\_k\text{-}1}(r+t/n))+u(Y_i-\theta_1^0'\dot{f}_{\_k\text{-}1}(r))] .[\dot{f}_{\_k}(r+t/n)-\dot{f}_{\_k\text{-}1}(r+t/n)](\theta_1-\theta_1^0) \tag{54}$$

$$U_{4,i,k}(\theta_1,t)=-u(Y_i-\theta_1^0'\dot{f}_{\_k\text{-}1}(r))[\dot{f}_{\_k}(r+t/n)-\dot{f}_{\_k\text{-}1}(r+t/n)](\theta_1-\theta_1^0) \tag{55}$$

Pour $U_{1,i,k}$, en utilisant le fait que $u\in Lip(1)$ et les relations (27) et (28), nous avons,

$$n^{1/2}\|\theta_1-\theta_1^0\|\le b\ ,\ \sup_{\|t\|\le B}|U_{1,i,k}(\theta_1,t)|\le C\,n^{-1/2}[\|\theta_1^0\|+b\,n^{-1/2}](1+X_i^2).1_{r_k<X_i\le r_k+t_k/n}$$

et en utilisant (C5)

$$\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}]\left|\sum_{i=1}^n\sum_{k=1}^K U_{1,i,k}(\theta_1,t)\right|\le C\,n^{-1/2}\sum_{k=1}^K \mathsf{E}_X\sum_{i=1}^n 1_{r_k<X_i\le r_k+t_k/n}=C\,n^{-1/2}\sum_{k=1}^K G_{k,n}(r_k+t_k/n) \tag{56}$$

Or, en conséquence du Lemme 2 et parce que la densité g de X est continue :

$$G_{k,n}(r_k+t_k/n)=O_P(G_k(r_k+t_k/n))=O_{\mathsf{P}}(n^{-1}) \tag{57}$$

donc

$$E_{(\epsilon,X)}[\sup_{\|t\|\le B}\left|\sum_{i=1}^n\sum_{k=1}^K U_{1,i,k}(\theta_1,t)\right|]=O_{\mathsf{P}}(n^{-1}) \tag{58}$$

Pour $U_{2,i,k}$ en utilisant les mêmes arguments que pour $U_{1,i,k}$, on obtient :

$$\mathsf{E}_{(\epsilon,X)}\Big[\sup_{|t|\le B,\,n^{-1/2}}\Big|\sum_{i=1}^{n}U_{2,i,k}(\theta_1,t)\Big|\Big]\le C\,n^{-1/2}\,\mathsf{E}_X\Big\|\dot f_{\_k}(r+t/n)-\dot f_{\_k\text{-}1}(r+t/n)\Big\|.\int_0^1\Big\|\theta_{1,\gamma}-\theta_1^0\Big\|d\gamma \tag{59}$$
$$\le Cn^{-1}\,\mathsf{E}_X\Big[\sum_{i=1}^{n}1_{r_k<X_i<r_k+t_k/n}\Big]=C\,n^{-1}\,O_{\mathsf{P}}(1)=O_{\mathsf{P}}(n^{-1})$$

Pour $U_{3,i,k}$ En utilisant le fait que $u\in Lip(1)$, on a :

$$\mathsf{E}_{(\epsilon,X)}\Big[\sup_{|t|\le B}\Big|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{3,i,k}(\theta_1,t)\Big|\Big]\le C\Big\|\theta_1-\theta_1^0\Big\|\sum_{k=1}^{K}\mathsf{E}_{(\epsilon,X)}\Big[\Big\|\dot f_{\_k\text{-}1}(r+t/n)-\dot f_{\_k\text{-}1}(r)\Big\|.$$
$$\Big\|\dot f_{\_k}(r+t/n)-\dot f_{\_k\text{-}1}(r+t/n)\Big\|\Big]$$
$$\le n^{-1/2}\sum_{k=1}^{K}\mathsf{E}_X\Big[(1+X^2)\sum_{j=1}^{k-1}1_{r_j<X_i<r_j+t_j/n}\,1_{r_k<X_i<r_k+t_k/n}\Big]$$
$$\le n^{-1/2}\sum_{k=1}^{K}\mathsf{E}_X^{1/2}\Big[(1+X^2)\sum_{j=1}^{k-1}1_{r_j<X_i<r_j+t_j/n}\Big]\mathsf{E}_X^{1/2}\Big[1_{r_k<X_i<r_k+t_k/n}\Big]$$
$$\le C\,n^{-1/2}n^{-1/2}$$

Donc

$$\mathsf{E}_{(\epsilon,X)}\Big[\sup_{|t|\le B}\Big|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{3,i,k}(\theta_1,t)\Big|\Big]=O_{\mathsf{P}}(n^{-1}) \tag{60}$$

- Par ailleurs, pour $U_{4,i,k}$, en utilisant les mêmes arguments que pour $U_{1,i,k}$ et en appliquant l'inégalité de Cauchy-Schwartz, on obtient:

$$\mathsf{E}_{(\epsilon,X)}\Big[\sup_{|t|\le B}\Big|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{4,i,k}(\theta_1,t)\Big|\Big]\le C\,n^{-1/2}\,\mathsf{E}_{(\epsilon,X)}^{1/2}\sup_{|t|\le B}\sum_{i=1}^{n}\sum_{k=1}^{K}u^2(Y_i-\theta_1^0\,\dot f_{\_k\text{-}1}(r)).\,\mathsf{E}_{(X)}^{1/2}\Big[\sum_{i=1}^{n}1_{r_k<X_i\le r_k+t_k/n}\Big]$$

Or, on a l'égalité suivante $\mathsf{E}_{(X)}^{1/2}\Big[\sum_{i=1}^{n}1_{r_k<X_i\le r_k+t_k/n}\Big]=O_{\mathsf{P}}(1)$. Aussi en utilisant le fait que $u^2(Y_i-\theta_1^0\,\dot f(r)_{k\text{-}1})=u^2(Y_i-\theta_1^0\,f_{\theta^0}(r))$, on obtient : $\mathsf{E}_{(\epsilon,X)}^{1/2}u^2(Y_i-\dot f_{\theta^0}(r))=\mathsf{E}_{(\epsilon)}^{1/2}(\epsilon_i)<\infty$.

En conclusion, on a pour $U_{4,i,k}$ pour $|t|\le B$,

$$\mathsf{E}_{\epsilon,X}\Big[\sup\Big|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{4,i,k}(\theta_1,t)\Big|\Big]=O_{\mathsf{P}}(n^{-1/2}) \tag{61}$$

Les relations (58), (59), (60) et (61) impliquent (50) donc (48).

Montrons maintenant (49).

On pose pour $t\in\mathbb{R}$, $\theta_{(n),1,\gamma}(r+t/n)=\hat{\theta}_{(n),1}(r)+\gamma[\hat{\theta}_{(n),1}(r+t/n)-\hat{\theta}_{(n),1}(r)]$, avec $\gamma\in[0,1]$

Alors par l'absolue continuité du $\log\varphi$ on a

$$\hat{L}_{2,n}(t)=\sum_{i=1}^{n}[-u(Y_i-\theta_{(n),1,\gamma}(r+t/n).\dot{f}_{\theta^0}(X_i))+u(Y_i-\theta_1^0.\dot{f}_{\theta^0}(X_i))].d\gamma[\hat{\theta}_{(n),1}(r+t/n)-\hat{\theta}_{(n),1}(r)]$$
$$-\sum_{i=1}^{n}[u(Y_i-\theta_1^0(r+t/n)\dot{f}_{\theta^0}(X_i))f_{\theta^0}(X_i)[\hat{\theta}_{(n),1}(r+t/n)-\hat{\theta}_{(n),1}(r)]]$$

Le reste de la démonstration est comme la fin de la preuve du Théorème 4.2 de Koul et Qian [Koul & Qian, 2002].

■

Pour étudier les distributions limites de $\tilde{L}_n$ il nous faut définir pour k=1...K $(\{\tilde{L_k^{(1)}}(-t),t\geq 0\},\{\tilde{L_k^{(2)}}(t),t\geq 0\})$, deux processus de Poisson composés indépendants avec le même taux $g(r_k)$ et tel que $\tilde{L_k^{(1)}}(0)=\tilde{L_k^{(2)}}(0)=0$.

La distribution des sauts est donnée respectivement par la distribution conditionnelle de

$$\log\frac{\varphi(\epsilon+(a_k^0-a_{k-1}^0)+(b_k^0-b_{k-1}^0)X)}{\varphi(\epsilon)}|X=r_k^+$$

$$\text{et } \log\frac{\varphi(\epsilon+(a_k^0-a_{k-1}^0)+(b_k^0-b_{k-1}^0)X)}{\varphi(\epsilon)}|X=r_k^-$$

ce qui est équivalent à

$$\log\frac{\varphi(\epsilon+(a_k^0-a_{k-1}^0)+(b_k^0-b_{k-1}^0)r_k)}{\varphi(\epsilon)}$$

$$\text{et } \log\frac{\varphi(\epsilon-(a_k^0-a_{k-1}^0)-(b_k^0-b_{k-1}^0)r_k)}{\varphi(\epsilon)} .$$

Théorème 5 : *Si on suppose que (C1) est vérifiée et que* $g(r_k)>0 \quad \forall k \quad 1\leq k\leq K$ *, alors*

$\tilde{L}_n(t)$ *converge faiblement dans* $D[0,\infty)^K$ *vers la somme de K processus de Poissons composé indépendant :*

$$\{\tilde{L}(v)=\sum_{k=1}^{K}\tilde{L}_k(v_k);v\in\mathbb{R}_+^K\} \text{ avec } \tilde{L}_k(v_k)=\begin{cases}\tilde{L}_k^{(1)}(v_k) & si \quad t_k\leq 0\\ \tilde{L}_k^{(2)}(v_k) & si \quad t_k\geq 0\end{cases}$$

Pour démontrer le Théorème 5, il suffit de poser $\forall k \quad 1\leq k\leq K$

$$\tilde{L}_n(t)=\sum_{k=1}^{K}\sum_{j=1}^{n} v_k(X_j,\epsilon_j)\,\mathbb{1}_{(min(r_k,r_k+t_k/n)<X_j\le max(r_k,r_k+t_k/n))}=\sum_{k=1}^{K}\tilde{L}_{k,n}(t_k)\ ,$$

avec $v_k(X_j,\epsilon_j)$ la fonction définit en (24) et $\tilde{L}_{k,n}(t_k)=\tilde{L}^{(1)}_{k,n}(-t_k)\,\mathbb{1}_{(-t_k\ge 0)}+\tilde{L}^{(2)}_{k,n}(t_k)\,\mathbb{1}_{(t_k\ge 0)}$.

Puis on utilise sur chaque processus $\tilde{l}_{k,n}(t_k)$ les résultats de Koul et Qian [Koul & Qian, 2002] pour un seul point de rupture.

■

Par la suite, on pose, pour k=1...K, $\tilde{l}_k(t_k)=\tilde{l}^{(1)}_k(-t_k)\,\mathbb{1}_{(-t_k\ge 0)}+\tilde{l}^{(2)}_k(t_k)\,\mathbb{1}_{(t_k\ge 0)}$, $t_k\in\mathbb{R}$. Par Koul et Qian [Koul & Qian, 2002] on pose $[M_{k-},M_{k+})$, l'intervalle aléatoire sur lequel le processus atteint son maximum. En appelant M_ le vecteur aléatoire $M\_=(M_{1\_},....M_{K\_})$, on a le résultat suivant:

Théorème 6 : *Si on suppose que (C1) -(C4) sont vérifiés et que les sauts de ruptures sont non nuls*
Alors $n(\hat{r}_{(n)}-r)$ *converge faiblement vers* $M\_$.

2.2.5. le cas non linéaire

Nous nous sommes intéressés jusqu'à présent au cas où f est linéaire. Nous allons élargir nos premiers résultats au cas non linéaire.

Dans ce cas, on pose comme précédemment :

$$Y_i=f_\theta(X_i)+\epsilon_i \quad 0<X_i<1, \qquad i=1....n$$

où (ϵ_i,X_i) sont des suites de variables indépendantes et aléatoires avec la même distribution jointe (ϵ,X) , mais on suppose maintenant que

$$\begin{aligned} f_\theta(X_i)&=h_{\alpha_0}(X_i)\,\mathbb{1}_{(X_i\le\tau_1)}+...+h_{\alpha_k}(X_i)\,\mathbb{1}_{(\tau_k<X_i\le\tau_{k+1})}+...+h_{\alpha_K}(X_i)\,\mathbb{1}_{(\tau_K<X_i\le 1)}\\ &=\sum_{k=0}^{K}h_{\alpha_k}(X_i)\,\mathbb{1}_{(\tau_k<X_i\le\tau_{k+1})} \end{aligned} \qquad (62)$$

$\tau_k\in[0,1],\tau_{k-1}<\tau_k\ \forall\, k\ge 1$.

avec $\theta=(\alpha_0,\alpha_1,...\alpha_K,\tau_1,...\tau_k,...\tau_K)\in\Theta$ et $\alpha_j\in\Gamma\subset\mathbb{R}^d,\Gamma$ *un compact et* $\tau_k\in\mathbb{R}$.

Le vrai paramètre θ^0 est alors égale à $\theta^0=(\alpha_0^0,\alpha_1^0,...\alpha_K^0,r_1,...r_k,...r_K)\in\Theta$.

On suppose que $\forall\, x\in(0,1)$, les dérivées $\partial^3 h_\alpha(x)/\partial\alpha^3$, $\partial^2 h_\alpha(x)/\partial\alpha^2$, $\partial h_\alpha(x)/\partial\alpha$ existent

et qu'il existe les fonctions $M_0, M_1, M_2 \in L^2(0,1)$ tel que :

$$\begin{aligned} &\sup_{\alpha\in\Gamma} |h_\alpha|(x) \le M_0(x) \\ &\sup_{\alpha\in\Gamma} \left\| \partial^k h_\alpha(x)/\partial\alpha^k \right\| \le M_k \quad k=1,2 \end{aligned} \tag{63}$$

ϵ_i a une fonction de densité f de Lebesgue et $E\epsilon_i = 0$, $E\epsilon_i^2 < \infty \quad i = 1....n$.

En outre, on suppose que f satisfait les conditions (C1), (C2) et (C3) du cas linéaire et

(C4bis) $\sup_{\theta,\theta^*} E_{\epsilon,X}[\phi^2(\epsilon + f_\theta(X) - f_{\theta^*}(X))] < \infty$

dans ce cas on pose $I(f) := E_\epsilon[\phi^2(\epsilon)]$.

On fait toujours la supposition d'identifiabilité

$$h_{\alpha_{k-1}(r_k)} \neq h_{\alpha_k(r_k)} \quad \forall \alpha_{k-1}, \alpha_k \in \Gamma, \quad \forall k = 1...K \tag{64}$$

Et on note dans ce cas

$$v_k(x,z) := \log \frac{\varphi(z + sgn(\tau_k - r_k).(h_{\alpha_k^0}(x) - h_{\alpha_{k-1}^0}(x)))}{\varphi(z)} \tag{65}$$

Nous allons généraliser les résultats obtenus dans la partie 1, au cas non linéaire. Les démonstrations restent en partie les mêmes. Seule, une difficulté supplémentaire est introduite du fait de la non linéarité de h. Ceci entraîne que les égalités (27) et (28) ne sont plus vérifiées.

a Convergence de l'estimateur

Le lemme suivant est une généralisation du Lemme 1.

Lemme 4 : *On suppose que l'hypothèse (C1) est satisfaite, et*

$$soit \qquad u \ est \ bornée \tag{66}$$

$$soit \ les \ conditions \ (C2) \ et \ (C4\,bis) \ sont \ vérifiées \tag{67}$$

Alors, $\forall \theta \in \Omega$

$$E\left(\sup_{\theta^* \in \Omega_\eta(\theta)} \left| \psi(X,Y,\theta^*) - \psi(X,Y,\theta) \right| \right) \to 0 \quad quand \ \eta \to 0 \tag{68}$$

Preuve du Lemme 4: Notons que pour $\tau \in (0,1)^K$ l'inégalité suivante est vérifiée:

$$\left|\left(\delta(X,\theta^*)\right)\right| \le C\left[\eta \left\| \sup_{\alpha\in\Gamma} \frac{\partial h_\alpha(X)}{\partial \alpha} \right\| + 2 \sup_{\alpha\in\Gamma} \left|h_\alpha(X)\right| \sum_{k=1}^{K} 1_{(|X_i-\tau_k|<\eta)}\right] .$$

La démonstration est alors la même que dans le cas à une seule rupture [Ciuperca, 2004].

■

De la même manière, on peut démontrer que le Théorème 7, généralisation du Théorème 3, est vérifié.

Théorème 7: Sous les mêmes hypothèses que celle du Lemme 4, alors

$$\hat{\theta}_{(n)} \underset{n\to\infty}{\overset{p.s}{\to}} \theta^0 .$$

Le théorème suivant est une généralisation du Théorème 2, portant sur la vitesse de convergence.

b Vitesse de convergence

Théorème 8 : *si les conditions (64), (C1),(C2) et (C4bis) sont vérifiées*

alors $\left|n(\hat{r}_{(n)} - r)\right| = O_P(1)$.

Preuve du Théorème 8 : Nous allons montrer que les inégalités (40) et (42) sont aussi vérifiées pour le cas non-linéaire. On définit donc $\theta_1(v) := \theta_1^0 + v(\theta_1 - \theta_1^0)$ et $\theta(v) := (\theta_1(v), \tau)$ avec $\tau \in [0,1]$.

La relation (42) du cas linéaire devient alors :

$$L_n^1(\theta) = n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \int_0^1 -[u_{(\theta_1(v))}(\tau_k) - u_{(\theta_1(v))}(\tau_{k-1})] \dot{f}_{\_k} dv - \int_0^1 u_{(\theta_1(v))}(\tau_{k-1})[\dot{f}_{\_k} - \dot{f}_{\_k-1}] dv$$

En notant I_1 et I_2 tel que

$$I_1 := n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \int_0^1 [u_{(\theta_1(v))}(\tau_k) - u_{(\theta_1(v))}(\tau_{k-1})] \left\| \dot{f}_{\_k} \right\| dv$$

$$I_2 := n^{-1} \sum_{k=1}^{K} \sum_{i=1}^{n} \int_0^1 u_{(\theta_1(v))}(\tau_{k-1}) \left\| \dot{f}_{\_k} - \dot{f}_{\_k-1} \right\| dv$$

On en déduit que $\left|L_{n,k}^1(\theta)\right| \le C(I_1 + I_2)$.

Sans perte de généralité, on peut prendre $\tau_k = r_k + \rho \quad \forall k = 1 \ldots K \quad \text{et} \quad \rho > 0$.

Pour I_1, par C2, (63), et l'inégalité de Cauchy-Schwartz, on a

$$I_1 \le C\delta \sum_{k=1}^{K} G_{k,n}(\rho) \tag{69}$$

Par ailleurs, l'inégalité suivante est vérifiée :

$$I_2 \le C n^{-1} \delta \sum_{k=1}^{K} \sum_{i=1}^{n} \int_0^1 u_{(\theta_1(v))}(\tau_{k-1}) \left\| \frac{\partial h_{\alpha_k^0 + v(\alpha_k - \alpha_k^0)}}{\partial \alpha} - \frac{\partial h_{\alpha_{k-1}^0 + v(\alpha_{k-1} - \alpha_{k-1}^0)}}{\partial \alpha} \right\| dv \, 1_{r_k < X_i \le \tau_k}$$
$$\le C n^{-1} \delta \sum_{k=1}^{K} R_{k,n}(\rho)$$

Ces deux inégalités, avec le Lemme 4, montrent l'exactitude de la relation (42) dans le cas non linéaire. La relation (40) se montre comme dans le cas linéaire en prenant v_k comme en (65).

■

c Distributions limites

En conséquence du cas avec une seule rupture non linéaire étudié par Ciuperca [Ciuperca, 2004], on a :

$$\forall\, k \in [1:K],\ n^{-1} \sum_{i=1}^{n} 1_{|X_i - r_k| \le B/n} = O_{\mathrm{P}}(n^{-1}) \tag{70}$$

soit $a_i(\theta_1, \tau) := -u(Y_i - \theta_1' \dot{f}_{\theta_1, \tau}(X_i)) \dot{f}_{\theta_1, \tau}(X_i)$, $A_\tau(x) := \dot{f}_{\theta_\tau}(x) \dot{f}_{\theta_\tau}(x)$ $A_0(x) := \dot{f}_{\theta^0}(x) \dot{f}_{\theta^0}(x)$ et $V := I(\varphi) E_X(A_0(X))$.

Avec ces notations, le Théorème 9 est vérifié.

Théorème 9 : *On suppose que les conditions (C1)-(C4bis) sont vérifiées alors* $\forall\, 0 < B < \infty$

$$\sup_{\|\tau - r\| \le B/n} \left\| n^{1/2}(\hat{\theta}_{(n),1}(\tau) - \theta_1^0) + V^{-1} n^{-1/2} \sum_{i=1}^{n} a_i(\theta_1^0, r) \right\| = O_{\mathrm{P}}(1)$$

et pour toute séquence $\tau_n = r + t_n / n$ *avec* t_n *borné,*

$$(\hat{\theta}_{(n),1}(\tau_n) - \theta_1^0) \underset{n \to \infty}{\overset{L}{\to}} N(0,\ V^{-1})$$

En conséquence :

$$\sup_{\|\tau - r\| \le B/n} n^{1/2}(\hat{\theta}_{(n),1}(\tau) - \theta_1^0) \to N(0,\ \Gamma^{-1}),\ \ quand\ \ n \to \infty \tag{71}$$

Preuve du Théorème 9 : La démonstration repose sur le fait que pour r fixé, le développement de Taylor de $\dot{L}_n(\hat{\theta}_{(n),1}(r), r)^t = 0$ en $\theta_1 = \theta_1^0$ est :

$$\begin{aligned} 0 = n^{-1} \sum_{i=1}^{n} a_i(\theta_1^0, r) + n^{-1} \sum_{i=1}^{n} [& -u(Y_i - f_{\theta^*_{(n),1}(r)}(X_i)) \ddot{f}_{\theta^*_{(n),1}(r)}(X_i)(\hat{\theta}_{(n),1}(r) - \theta_1^0)^t + \\ & u'(Y_i - f_{\theta^*_{(n),1}(r)}(X_i)) A_r(X_i)] [\hat{\theta}_{(n),1}(r) - \theta_1^0]^t \end{aligned} \tag{72}$$

avec $\theta^*_{(n),1}(r)$ dans une boule de rayon $\eta_{r,n}$ et $\left\| \theta^*_{(n),1}(r) - \theta_1^0 \right\| \le \left\| \hat{\theta}_{(n),1}(r) - \theta_1^0 \right\|$.

On considère séparément les 3 termes de (72) :

Si on note $S_n(\theta) := n^{-1}\sum_{i=1}^{n} u'(Y_i - f_{\theta^*_{(n),1}(r)}(X_i))\,A_\tau(X_i)]$, on a la décomposition $S_n(\theta) := S_1 + S_2 + S_3$, avec

$$S_1 = n^{-1}\sum_{i=1}^{n} [u'(X_i - f_{\theta_{(n)},1^*,\tau}(X_i)) - u'(X_i - f_{\theta^0_\tau,\tau}(X_i))]A_\tau(X_i)$$

$$S_2 = n^{-1}\sum_{i=1}^{n} [u'(X_i - f_{\theta_{(n),1},\tau}(X_i)) - u'(\epsilon_i)]\,A_\tau(X_i)\ ,$$

et $S_3 = n^{-1}\sum_{i=1}^{n} u'(\epsilon_i)\,A_\tau(X_i)$.

On a d'abord $\|S_1\| = O_{\mathrm{P}}(1)$ (voir Ciuperca [Ciuperca, 2004])

Pour S_2 en utilisant la relation (70), on obtient :

$$\|S_2\| \le C\,n^{-1}\sum_{i=1}^{n} \left|f_{\theta^0_1,\tau} - f_{\theta^0_1,r}\right| \le C\,n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K} 1_{|X_i - r_k| \le B/n} = O_{\mathrm{P}}(n^{-1})\ .$$

On a $\|S_3 - V\| = O_{\mathrm{P}}(1)$ (voir Ciuperca [Ciuperca, 2004]).

On en déduit que $\sup_{|\tau - r| \le B/n} \|S_n(\theta) - V\| = O_{\mathrm{P}}(1)\ \ \forall\, B > 0$

Le reste de la preuve est similaire à celle de Ciuperca [Ciuperca, 2004] en utilisant (70).

■

Théorème 10 : *Si les conditions (C1)-(C4bis) sont vérifiées, alors on a*

$$\forall\, B > 0\ ,\ \sup_{|\xi| < B} \left|\hat{L}_n(t) - \tilde{L}_n(t)\right| = O_{\mathrm{P}}(1)\ .$$

Preuve du Théorème 10 : Comme dans le cas linéaire (46), on peut écrire

$$\hat{L}_n(t) - \tilde{L}_n(t) = \hat{L}_{1,n}(t) + \hat{L}_{2,n}(t)\ .$$

Théorème 11 : *Si les conditions (C1)-(C4bis) sont vérifiées, alors on a*

$$\forall\, B > 0\ ,\ \sup_{|\xi| < B} \left|\hat{L}_n(t) - \tilde{L}_n(t)\right| = O_{\mathrm{P}}(1)\ .$$

Preuve du Théorème 11 : Comme dans le cas linéaire (46), on peut écrire

$$\hat{L}_n(t) - \tilde{L}_n(t) = \hat{L}_{1,n}(t) + \hat{L}_{2,n}(t)\ .$$

On montre que:

$$\sup_{\|t\| \le B} \left|\hat{L}_{1,n}(t)\right| = O_P(1) \tag{73}$$

et

$$\sup_{\|t\| \le B} \left|\hat{L}_{2,n}(t)\right| = O_P(1) \tag{74}$$

On a une décomposition similaire à (51) avec:

$$U_{1,i,k}(\theta_1,t)=\int_0^1 [-u(Y_i-\theta_{1,\gamma}'\dot{f}_{k}(r+t/n)) + u(Y_i-\theta_{1,\gamma}'\dot{f}_{k-1}(r+t/n))]\dot{f}_{k}(r+t/n)\,d\gamma(\theta_1-\theta_1^0) \tag{75}$$

$$U_{2,i,k}(\theta_1,t)=\int_0^1 [-u(Y_i-\theta_1'\dot{f}_{k-1}(r+t/n)) + u(Y_i-\theta_1^0'\dot{f}_{k-1}(X_i,r+t/n))][\dot{f}_{k}(r+t/n)-\dot{f}_{k-1}(r+t \tag{76}$$

$$U_{3,i,k}(\theta_1,t)=[-u(Y_i-\theta_1^0'\dot{f}_{k-1}(r+t/n))+u(Y_i-\theta_1^0'\dot{f}_{k-1}(r))] .[\dot{f}_{k}(r+t/n)-\dot{f}_{k-1}(r+t/n)](\theta_1-\theta_1^0) \tag{77}$$

$$U_{4,i,k}(\theta_1,t)=-u(Y_i-\theta_1^0'\dot{f}_{k-1}(r))[\dot{f}_{k}(r+t/n)-\dot{f}_{k-1}(r+t/n)](\theta_1-\theta_1^0) \tag{78}$$

Pour $U_{1,i,k}$, du fait de (63) et de $u\in Lip(1)$, on a

$$\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\left|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{1,i,k}(\theta_1,t)\right|]\le$$
$$\le C n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\left\|f_{k}(\theta_{1,\gamma},r+t/n)-f_{k-1}(\theta_{1,\gamma},r+t/n)\right\|d\gamma]$$
$$= C n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\int_0^1\left|h_{\alpha_k(\gamma)}\right|1_{r_k<X_i<r_k+t_k/n}]\le C n^{-1/2}\sum_{k=1}^{K}G_k(r_k+t_k/n)=O_{\mathsf{P}}(n^{-1/2})$$

Puis pour $U_{2,i,k}$, du fait de $G_{k,n}(r_k+t_k/n)=O_{\mathsf{P}}(n^{-1})$, de (63) et de $u\in Lip(1)$, on obtient :

$$\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\left|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{2,i,k}(\theta_1,t)\right|]\le C\left\|\theta_{1,\gamma}-\theta_1^0\right\|.\left\|\theta_1-\theta_1^0\right\|G_k(r_k+t_k/n)\le C n^{-1/2}n^{-1/2}=O_{\mathsf{P}}(n^{-1})$$

Pour $U_{3,i,k}$ et $U_{4,i,k}$ on obtient exactement de la même façon :

$$\mathsf{E}_{(\epsilon,X)}[\sup_{\|t\|\le B}\left|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{3,i,k}(\theta_1,t)\right|]=O_{\mathsf{P}}(n^{-1})$$

et

$$\mathsf{E}_{(\epsilon,X)}\left[\sup_{\|t\|\le B}\left|\sum_{i=1}^{n}\sum_{k=1}^{K}U_{4,i,k}(\theta_1,t)\right|\right]\le$$

$$C n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathsf{E}_{(\epsilon,X)}\left[u_{\theta^0}(\epsilon,X)\sum_{i=1}^{n}\sum_{k=1}^{K}\sup_{\|t\|\le B}\int_0^1\left\|f_{\_k}(\theta_{1,\gamma},r+t/n)-f_{\_k\text{-}1}(\theta_{1,\gamma},r+t/n)\right\|d\gamma\right]$$

$$=C n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\mathsf{E}_{(\epsilon,X)}[G_{k,n}(r_k+t_k/n)]=O_{\mathsf{P}}(n^{-1/2})$$

Ces quatre relations montrent (73).

Pour (74), la démonstration est similaire au cas à une seule rupture [Ciuperca, 2004].

■

d Distribution asymptotique de la séquence $\tilde{L}_n(t)$

Par rapport au cas linéaire, les processus de Poisson composé ont les même paramètres :

Pour k=1...K $(\{\tilde{L}_k^{(1)}(-t),t\ge 0\},\{\tilde{L}_k^{(2)}(t),t\ge 0\})$, sont toujours deux processus de Poisson indépendants avec le même taux $g(r_k)$ et tel que $l_k^{(1)}(0)=l_k^{(2)}(0)=0$.

Seule la distribution des sauts change en :

$$\log\frac{\varphi(\epsilon+(h_{\alpha_{k-1}^0}(X)-h_{\alpha_k^0}(X)))}{\varphi(\epsilon)}|X=r_j^+$$

$$\text{et } \log\frac{\varphi(\epsilon+(h_{\alpha_{k-1}^0}(X)-h_{\alpha_k^0}(X)))}{\varphi(\epsilon)}|X=r_j^- .$$

Les démonstrations portant sur le processus limite de la vraisemblance restent les même que dans le cas à une seule rupture [Ciuperca, 2004].

Théorème 12 : *Si on suppose que (C1) est vérifié et que* $g(r_k)>0$, $\forall k\ \ 1\le k\le K$, *alors* $\tilde{L}_n(t)$ *converge faiblement dans* $D[0,\infty)^K$ *vers la somme de K processus de Poisson composés indépendants :*

$$\{\tilde{L}(v)=\sum_{k=1}^{K}\tilde{l}_k(v_k);v\in\mathbb{R}_+^K\}\ \textit{avec}\ \tilde{L}_k(v_k)=\begin{cases}\tilde{L}_k^{(1)}(v_k) & si\ -t_k\ge 0\\ \tilde{L}_k^{(2)}(v_k) & si\ \ t_k\ge 0\end{cases}.$$

Théorème 13 : *Si on suppose que les conditions (C1),(C2),(C3) et (C4bis) sont vérifiées et que les sauts de ruptures sont non nulles, alors*

$$n(\hat{r}_{(n)}-r)\underset{n\to\infty}{\overset{\mathsf{P}}{\to}}M_-$$

L'ensemble des résultats développés ici montrent que pour un modèle de rupture, les propriétés des estimateurs par maximum de vraisemblance obtenus dans le cas d'une rupture unique par Koul et Qian et par Ciuperca sont généralisables au cas multi-phasiques.

Nous avons notamment montré que les estimateurs étaient n-consistants et que le processus de vraisemblance convergeaient vers un processus de Poisson composé.

Ces propriétés nous seront utiles pour notre analyse de l'activité de conduite (chapitre 3.5).

3 Analyse de l'activité de conduite par le modèle de Markov caché et les modèles de ruptures multi-phasiques: méthodologie, expérimentation et résultats.

La troisième partie de ce mémoire sera consacrée à la mise en relation des deux parties précédentes afin de construire une méthodologie d'analyse de l'activité de conduite.

Dans un premier temps, nous décrirons comment les résultats des études menées sur le conducteur et présentées lors de la première partie, nous permettent de structurer l'analyse de l'activité. En effet, ces recherches nous amènent à faire des choix importants. Ces choix seront alors déterminants tant pour segmenter l'activité de conduite, que pour étudier l'influence des caractéristiques des situations routières sur l'activité. Ceci nous permettra de déterminer *l'architecture de la base des données interprétées* que nous devrons recueillir et analyser (chapitre 3.1.1).

Puis, nous étudierons les moyens mis à notre disposition pour enregistrer et analyser le comportement. Nous examinerons alors les données disponibles et leurs degrés de fiabilité. Cela sera indispensable pour définir la *base des données objectives* (chapitre 3.1.2) pouvant être constituée.

Ces deux sections nous amèneront à exposer l'expérimentation effectuée afin d'enregistrer l'activité au niveau des données objectives et interprétées (chapitre 3.2).

Enfin, en s'appuyant sur les méthodes présentées dans la partie 2, nous décrirons les algorithmes et la méthodologie développés pour modéliser l'activité de conduite. Ceci nous permettra d'établir le *catalogue de modèles*, basé sur les modèles semi-Markovien Cachés, associées aux différentes *Situations de Conduites Vécues* (chapitre 3.3).

Par ailleurs, comme nous l'avons expliqué dans la première partie, les capteurs sont limités et bruités. Dès lors, il est impossible de distinguer, avec le panel de capteurs dont nous disposons, l'ensemble des Situations de Conduite Vécues.

Aussi, nous utiliserons les techniques de classification pour regrouper les situations proches. Ceci nous permettra alors de distinguer différentes catégories de Situation de Conduite Vécue. Ces catégories seront les Situations de Conduites Mesurées présentées dans la partie 1.

C'est sur ces catégories que nous testerons alors la capacité prédictive de la modélisation. (chapitre 3.4)

Enfin, nous verrons que l'utilisation hybride des méthodes de ruptures et des modèles construits permet d'envisager la segmentation automatique des flux de données sur la conduite (chapitre 3.5).

3.1 Méthodologie.

Etablir un système de catégorisation du comportement du conducteur, c'est chercher à établir des relations entre des situations vécues par le conducteur et des situations mesurées via les capteurs numériques (voir paragraphe 1.3.1). Dans notre cas, ces relations se concrétisent alors par un modèle mathématique de l'évolution des valeurs des capteurs dans chaque situation.

Pour réaliser cet objectif, nous avons choisi de nous appuyer sur deux types de données caractérisant les situations de conduite :

- **données objectives (caractéristiques des séquences de conduite mesurées)**
  Ces données sont des caractéristiques objectives mesurées sur l'activité de conduite, c'est-à-dire l'ensemble des valeurs numériques issues des capteurs portant sur l'activité de conduite,

- **données sur le comportement du conducteur interprétées d'un point de vue cognitif : (caractéristiques des séquences de conduite vécues)**
  Ces données sont formées par l'interprétation de la situation vécue en terme d'activité cognitive. Elles sont donc basées sur l'interprétation de l'objectif du conducteur et de la façon dont il perçoit son environnent à partir des enregistrements vidéo et de l'interprétation des valeurs de capteurs (voir chapitre 1.2.1.d et 1.3.1.a).

L'objectif de notre travail est alors de comprendre la relation entre ces deux visions de l'activité. Un **catalogue de modèles** est donc nécessaire pour faire le lien entre les enregistrements de l'activité et la situation vécue par le conducteur (illustration 3.1).

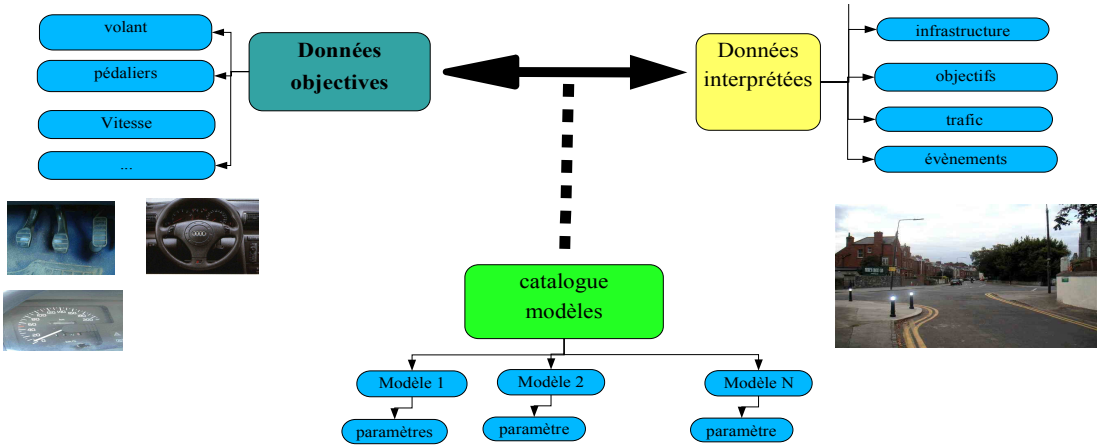


Illustration 3.1 Notre modélisation du conducteur se base sur l'existence de deux types de données : celles objectives et celles interprétées. Elle est définie comme un catalogue de modèles mettant en relation ces données.

3.1.1. Structurer l'analyse de l'activité de conduite

a Séquences de conduite : caractéristiques étudiées

Dans notre thèse, nous avons choisi de nous baser sur la théorie des frames de Minsky [Minsky, 1975]et nous nous sommes fortement inspirés de l'une de ses applications dans le domaine automobile via le modèle COSMODRIVE de Bellet [Bellet, 1998].

Nous avons vu dans la première partie que, selon cet auteur, lorsque le conducteur identifie un environnement routier, il sélectionne un *frame tactique* (actions à effectuer, zones de déplacements prévues et zones de perceptions) dans sa base de connaissances routières. La sélection de ce frame dépend de **l'objectif tactique** du conducteur, et du type **d'environnement routier** identifié par le conducteur.

Ainsi, pour cet auteur, l'activité de conduite est divisée en **séquences de conduite**, au cours desquelles le conducteur a un **objectif tactique homogène** (tourner, doubler...).

Lors de ces séquences, le frame générique est alors « instantié » (i.e : adapté) à la scène routière présente (telle qu'elle est perçue par le conducteur). Cette instanciation est fonction des particularités de la situation routière (virage plus ou moins sec, couleur du feu...).

Un frame peut donc regrouper des comportements différents. Par ce fait, il est difficile d'associer un modèle numérique des données à des situations aussi hétérogènes. Aussi, il nous est apparu important d'introduire 2 critères supplémentaires : la **situation initiale** et les **évènements** survenus dans la situation routière. S'ils nous sont utiles pour homogénéiser les comportements, ces critères peuvent aussi être vus comme des éléments importants d'instanciation du frame.

D'un autre coté, Tango et Montanari [Tango & Montanari, 2005] rappellent que les caractéristiques d'une situation de conduite peuvent être organisées en 5 grandes catégories.

1. l'environnement (météorologie, visibilité, « urbanité »...)
2. le conducteur (expérience, âge, « style » de conduite)
3. le véhicule (caractéristiques intrinsèques : vitesse max., poids...)
4. le trafic (vitesse et position des autres véhicules)
5. l'infrastructure (typologie et paramètres)

L'activité de conduite lors d'une situation de conduite sera donc fonction de l'objectif du conducteur, de la situation initiale, de l'environnement perçu, et de l'ensemble de ces caractéristiques.

Cependant, étudier l'influence de l'ensemble de ces facteurs nécessiterait un volume de données qu'il n'est pas possible, pour l'heure, de collecter.

C'est pourquoi, en nous basant sur de nombreux travaux antérieurs (voir partie 1.3.2: Pentland, Kumagai...), qui ont fait le lien entre des comportements, des objectifs et certaines infrastructures, nous avons décidé de nous focaliser sur l'influence du facteur « infrastructure » sur la variation de l'activité de conduite.

La connaissance qui découlera de cette recherche pourra alors être utilisée dans des études

ultérieures pour analyser l'influence des autres facteurs.

Dans cette optique, notre choix expérimental est de fixer les autres facteurs à une valeur constante :

1. En étudiant uniquement le comportement des conducteurs en ville, par temps clair et lorsque la visibilité est bonne,
2. En étudiant uniquement le comportement des conducteurs expérimentés et âgés d'environ 40 ans. Ceci nous permet dès lors de diminuer les effets d'expérience de conduite hétérogène,
3. En n'étudiant pas les différences de comportements suivant les véhicules. Le véhicule expérimental que nous utiliserons sera unique et nos données ne seront pas bruitées par les différences de conduite dûes à une disparité des véhicules,
4. En étudiant le comportement des conducteurs qu'en présence d'un trafic faible.

Aussi, en nous basant sur ces principes et sur la philosophie de COSMODRIVE, nous avons choisi, dans le cadre de cette thèse, d'étudier l'activité de conduite dans des **séquences** définies par **l'infrastructure (perçue), l'objectif,** la **situation initiale** et de prendre en compte les évènements potentiels (illustration 3.2).

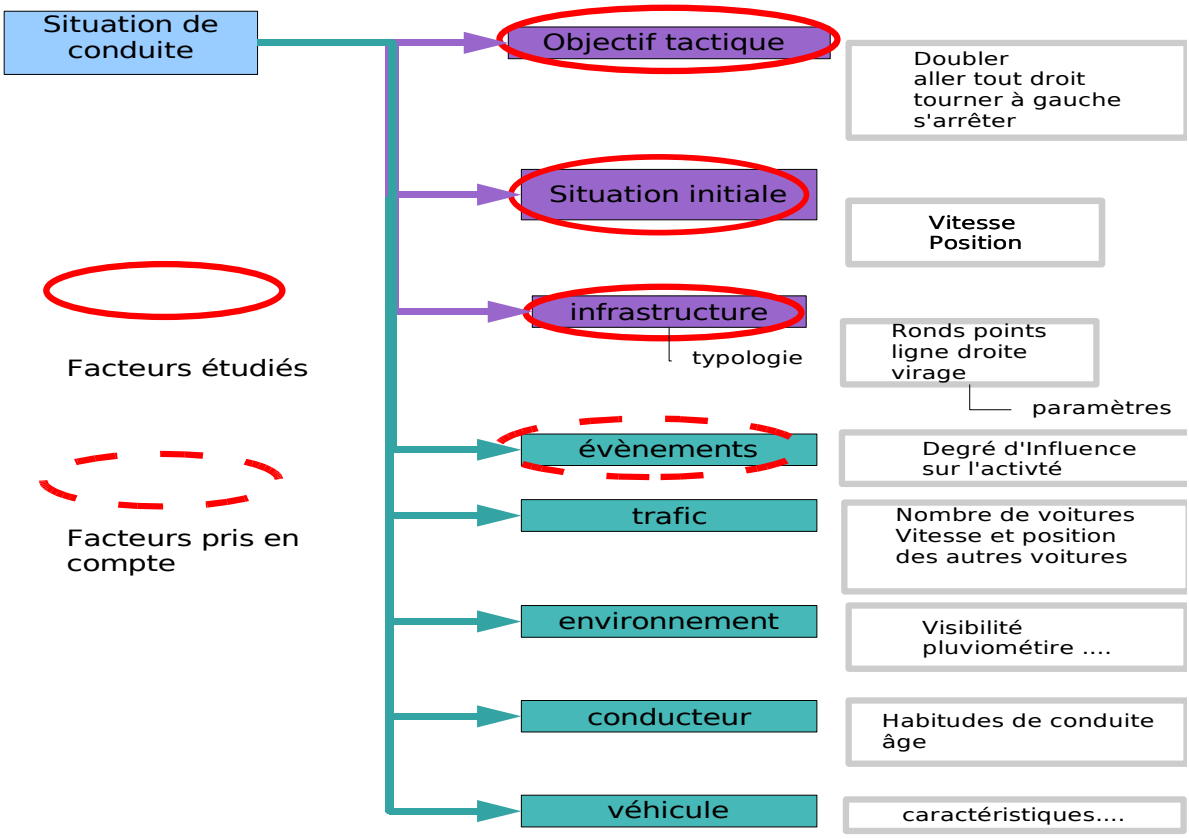


Illustration 3.2. : Ensemble des caractéristiques d'une situation de conduite. Dans cette thèse, nous n'étudierons que l'effet des 3 premiers facteurs sur le comportement.

Remarque : L'absence de dispositif à notre disposition pour mesurer la position du véhicule sur la voie fera que nous ne tiendrons pas compte de la position latérale du véhicule (écart à la ligne blanche).

b Sélection des séquences.

Par ailleurs, l'étude du modèle COSMODRIVE nous permet de supposer qu'en l'absence d'évènement extérieur important (i.e. : d'influence décisive sur l'activité), il existe, pour chaque situation de conduite, une *série de zones de déplacements* que le véhicule doit traverser pour arriver à l'état final souhaité et un *ensemble d'actions par défaut,* que le conducteur doit effectuer, associé à chacune de ces zones.

Les actions et les zones de déplacements-ci dépendent d'une part de son/ses objectifs (atteindre la zone voulue à la vitesse voulue...), de l'infrastructure traversée et de la situation initiale.

Cette série d'actions par défaut correspond, dans COSMODRIVE, à l'instanciation par défaut du frame de la situation, c'est-à-dire l'ensemble des actions que le conducteur pense réaliser avant même de s'engager dans la situation catégorisée.

- ***séquence prototypique :***

On nommera alors prototypique cette série d'actions/états correspondant à une situation sans évènement extérieur perturbant.

Salvucci et al [Salvucci et al, 2002] décrivent ainsi les différentes étapes typiques opérées lors d'un dépassement.. Ainsi, les conducteurs décélèrent un peu avant de doubler, puis accélèrent peu après avoir changé de voie, puis maintiennent leur vitesse élevée tant qu'ils ne sont pas revenus sur la voie.

- ***séquence atypique :***

Pourtant, des évènements, tant extérieurs comme un obstacle non prévu, qu' « internes » comme une configuration de l'infrastructure mal évaluée, peuvent perturber la structure de cette situation prototypique. Lorsque cette perturbation ne modifie pas l'enchaînement global des actions, mais change leur temporalité (état du conducteur et du véhicule se maintenant plus longtemps qu'habituellement), ou altère partiellement leur intensité (virage plus sec, accélération plus rapide...), nous parlerons de situations atypiques.

- ***séquence singulière :***

Lorsque la perturbation est si importante que l'enchaînement des actions est totalement bouleversé, la modélisation des actions lors d'une situation prototypique ne peut être appliquée. Nous parlerons alors de situation singulière.

« évitement d'un obstacle au dernier moment, gestion d'une situation accidentogène... » sont des exemples de ce genre de séquence.

Pour comprendre l'activité du conducteur, il est donc nécessaire de comprendre tout d'abord

comment est organisé l'enchaînement des actions dans une situation prototypique, puis de comprendre comment les évènements extérieurs agissent sur cette dernière, dans un premier temps de façon partielle (situation atypique) puis de façon totale (situation singulière) (illustration 3.3).

Dans le cadre de notre thèse, nous nous sommes focalisés sur la modélisation des situations prototypiques. Nous verrons alors que cette modélisation est utile pour comprendre l'activité dans les autres types de situations.

Cependant, afin de modéliser efficacement les situations non prototypiques, les modèles devront être complexifié en incluant des données provenant de capteurs spécifiques. Par ailleurs l'appui de travaux, comme ceux de Georgeon et al [Georgeon et al, 2006], spécifiquement dédiés à la modélisation cognitive à partir de l'analyse de l'activité, permettront de mieux comprendre l'enchaînement des phases d'actions et des zones de déplacements dans les situations singulières.

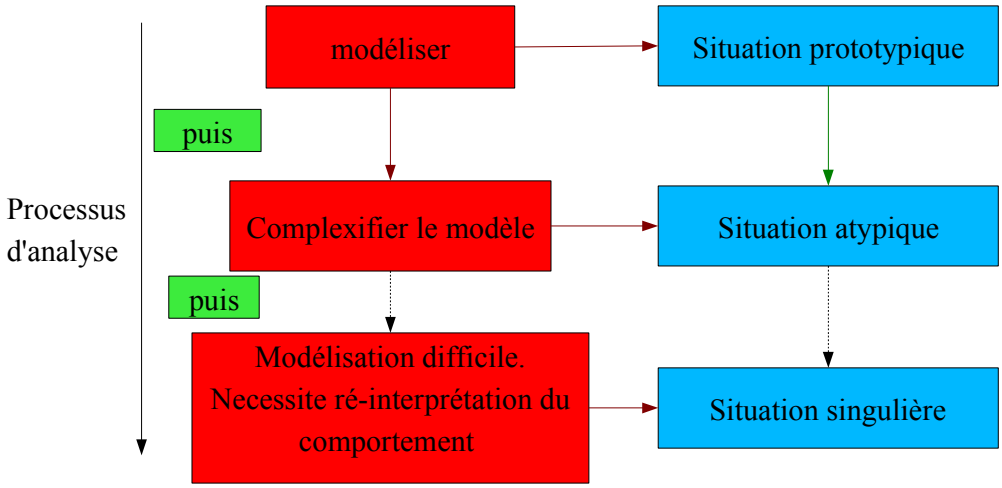


Illustration 3.3 : Processus général d'analyse de l'activité

Pour analyser et modéliser le comportement du conducteur, il faut avant tout pouvoir enregistrer son activité. La section suivante est donc centrée sur l'ensemble des moyens expérimentaux mis à notre disposition afin d'établir notre analyse.

3.1.2. Enregistrer l'activité de conduite : moyens expérimentaux

Le LESCOT dispose d'un véhicule instrumenté, MARGO (plateforme de MesuRe et d'Analyse pour l'erGonomie de la cOnduite). Ce véhicule a été développé au sein du laboratoire en 2004 afin d'obtenir des informations comportementales sur le conducteur (illustration 3.4).

Illustration 3.4 : MARGO : plateforme de MesuRe et d'Analyse pour l'erGonomie de la cOnduite

Il s'agit d'une Renault Scénic. Ce véhicule expérimental a été choisi pour sa représentativité dans le parc routier français : c'est un véhicule de type « intermédiaire » (plus imposant qu'une « compact » mais plus petit qu'une routière).

Les objectifs lors de l'élaboration du véhicule furent d'aménager une voiture où le recueil d'informations se devait à la fois d'être précis, complet et telle que l'instrumentation soit la moins intrusive possible [Bonnard et al, 2006] (illustration 3.5).

Illustration 3.5 : La micro-caméra embarquée n'interfère pas sur la tâche de conduite.

Les éléments numériques pouvant être enregistrés sur l'activité du conducteur sont notamment l'angle du volant, l'enfoncement des pédales, les clignotants et la dynamique du véhicule (vitesse et accélération).

L'ensemble de ces éléments résulte de choix techniques importants. Pour comprendre la nature des éléments enregistrés et la sensibilité des mesures effectuées, il importe de détailler le fonctionnement des principaux instruments de mesure.

a Le volant

Les mesures sur le volant sont réalisées par un capteur optique sur une roue dentée liée au volant par une courroie (illustrations 3.6 et 3.7). Dû au crénelage et à l'élasticité de cette courroie, ce dispositif a une sensibilité de +/- 0.1 degré.

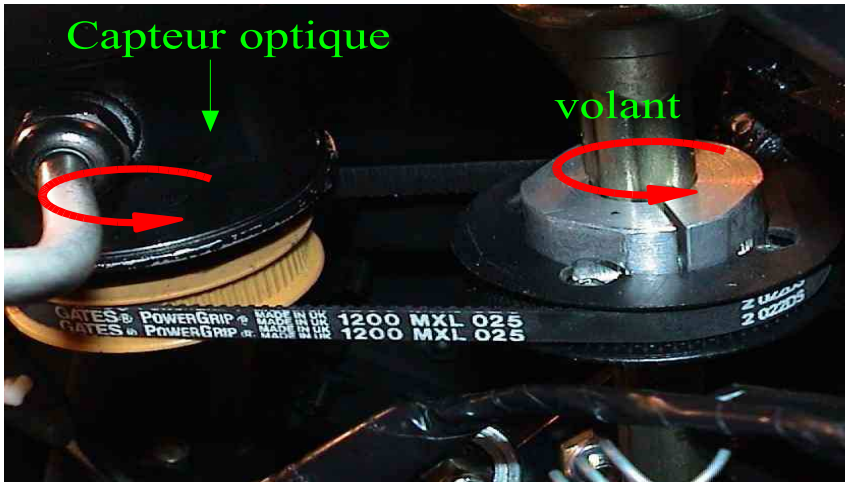


Illustration 3.6 : Prise de mesure de l'angle du volant par capteurs optiques

Illustration 3.7 : Prise de mesure de l'angle du volant par capteurs optiques (vue large)

b Les pédaliers.

Déterminer l'information à recueillir sur les pédaliers n'est pas un choix aussi anodin qu'il puisse y paraître. En effet, du fait de la course ellipsoïdale des pieds et du faible volume disponible pour installer des équipements, la mesure de l'enfoncement véritable de la pédale n'est pas aisé.

Le choix effectué au LESCOT a été d'assimiler l'angle d'enfoncement à la course d'un potentiomètre placé en arrière de la pédale (illustration 3.8). Cette méthode a l'avantage d'être plus simple à mettre en oeuvre qu'une mesure avec des roues dentées mais la mesure de l'angle en est très légèrement faussée.

Bien que cette inexactitude soit difficile à mesurer, étant donné la faiblesse de la course du potentiomètre, on peut la considérer comme négligeable par rapport à l'erreur de mesure des

capteurs.

Celle-ci est de 0.2% pour le seul potentiomètre (données constructeurs). On peut supposer que le reste du circuit d'acquisition entraîne une erreur de mesure (bruit sur les câbles, acquisition numérique) de plus faible importance.

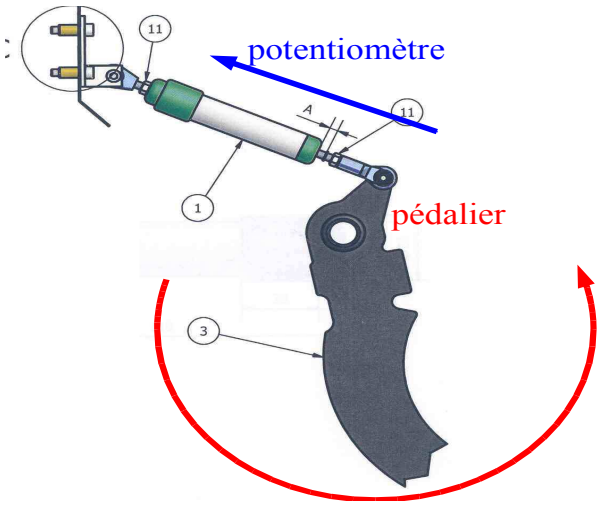


Illustration 3.8 : Mesure de l'angle du pédalier par potentiomètre

Ainsi, des tests ont montré la quasi-linéarité de la réponse du potentiomètre en fonction de l'angle du pédalier (Figure3.9).

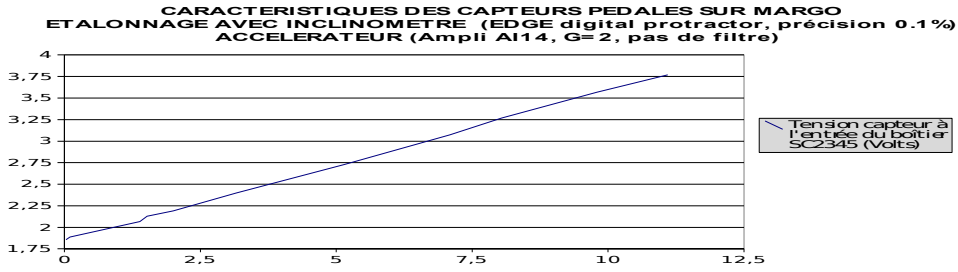


Figure 3.9 : Tension délivrée en fonction de l'angle du pédalier

c Dynamique du véhicule

La vitesse du véhicule est mesuré en estimant le nombre de tour de roue par unité de temps par l'utilisation d'un capteur optique sur une roue denté située sur l'axe moteur. La mesure de l'accélération dans les 3 dimensions se fait par un capteur dédié (illustration 3.10), de précision +/- 3%.

Illustration 3.10 : Mesure de l'accélération du véhicule

d Autres mesures

De plus, un laser à balayage, un télémètre, des caméras en stéréovision nous donnent des informations sur la présence d'obstacle via l'utilisation des algorithmes développées au LIVIC [Bellet et al, 2004]. Par ailleurs un GPS couplé à une base cartographique nous renseigne sur l'environnement du conducteur (contexte urbain).

Si les informations que donnent ces instruments sont importantes, leur fiabilité dans certaines situations est peu sûre. De plus, notre désir de fonder notre analyse sur des capteurs « basiques » , pour pouvoir être utilisé aisément, fera que nous n'utiliserons pas ce type de données.

L'ensemble de ces informations est enregistré en temps réel grâce au logiciel Labview © exécuté sur un ordinateur embarqué (illustration 3.11).

Illustration 3.11 : Ordinateurs embarqués permettant d'effectuer l'enregistrement des données.

Ce logiciel nous permet alors de récupérer l'ensemble de ces informations sous forme de fichiers textes.

En outre, quatre caméras filment la scène avant, le conducteur, la scène arrière et le processus

d'acquisition (illustration 3.12). Ces images sont alors recueillies sur un magnétoscope numérique.

Illustration 3.12 : 4 caméras filment en continu différentes vues de la scène de conduite

Enfin, une procédure de synchronisation permet de coupler, via l'envoi des time-codes par RS232, l'enregistrement vidéo avec l'enregistrement des valeurs numériques.

Ainsi, l'ensemble des données pouvant être recueilli sur le véhicule expérimental nous informe sur les actions du conducteur et sur la dynamique du véhicule avec une précision convenable pour notre étude. De plus, le couplage des données numériques avec la vidéo nous permet de comprendre dans quel environnement objectif évolue le conducteur et d'interpréter, notamment par l'étude des regards, comment le conducteur perçoit cet environnement.

Ces données nous seront utiles lors de l'expérimentation que nous présentons dans la prochaine section.

3.2 Expérimentation

3.2.1. Objet

L'objectif de notre expérimentation était de recueillir des données en nombre important sur le comportement du conducteur dans des situations de conduite diversifiées et de manière le plus écologique possible. L'expérimentation devaient se dérouler uniquement dans un contexte urbain (ville et autoroute) et rendre compte de la diversité des infrastructures (rond-point, virage, carrefour à feux, carrefour en T...), et des objectifs tactiques réalisables (changer de voies, tourner...).

Compte tenu des contraintes discutées dans la partie 3.1, les conducteurs étudiés devaient être expérimentés (au moins 10 ans de conduite régulière) et les conditions expérimentales devaient être une grande visibilité et un trafic peu important. Aussi, les expérimentations ont eu lieu en milieu de matinée (10h) ou en début d'après-midi (14h), de façon à évoluer en trafic peu important, et uniquement les jours présentant une pluviosité faible ou nulle et avec une visibilité importante.

Par ailleurs, les conducteurs devaient conduire le plus naturellement possible. Aussi l'expérimentation ne devait pas être trop longue (moins d'une heure). En effet, la conduite en milieu urbain requiert une attention importante, et il importe que la fatigue du conducteur n'altère pas sa conduite.

De même, au début de l'expérimentation, la consigne a été donnée de conduire « naturellement » (« doubler s'il en ressent le besoin », aller à « son » allure ...). De plus, afin d'interférer le moins possible avec la tâche de conduite, les directions de conduite ont été données par l'expérimentateur au plus tôt.

Enfin, le parcours ne devait pas être familier aux sujets afin d'éviter tout comportement singulier (voir remarque partie 1.1.4.b sur les lieux familiers).

Le nombre de sujets était fixé à 5.

Nous détaillons par la suite le parcours que nous avons défini comme lieu de l'expérimentation.

3.2.2. Le parcours

L'itinéraire est constitué d'une boucle, se déroulant sur les villes de Lyon et de Bron (illustration 3.12).

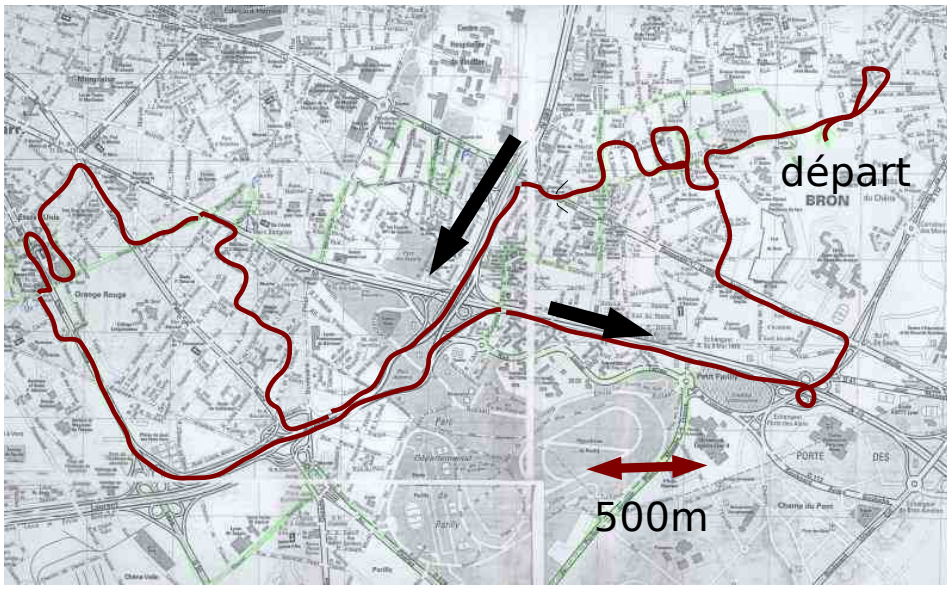


Illustration 3.13 : Parcours défini pour l'expérimentation sur les villes de Lyon Bron

Le parcours est caractérisé par la présence d'infrastructures variées (rond-point, intersection, carrefour en T, autoroute...). Néanmoins, cet itinéraire ne comporte pas toutes les situations possibles, en raison de deux contraintes que nous devions prendre en compte. D'une part, il nous fallait réduire au maximum le temps de passation, pour que la fatigue n'altère pas le comportement du conducteur. D'autre part, il nous fallait recueillir l'activité de conduite lors d'une même situation (couple infrastructure / objectif) au moins deux fois par conducteur, afin de constituer des échantillons assez importants pour pouvoir former des ensembles de test et d'apprentissage.

Le tableau 1 résume les infrastructures traversées couplées aux objectifs donnés aux conducteurs dans ces dernières.

Infrastructure / objectif donné	Fréquence
Lignes droites / aller tout droit	44
Intersection / tourner à gauche	8
Intersection / tourner à droite	8
Intersection / aller tout droit	16
ronds points / tourner à droite	2
ronds points / aller tout droit	3
carrefour en T voie principale / tourner à gauche	3
carrefour en T voie principale / tourner à droite	5
carrefour en T voie principale / aller tout droit	8
carrefour en T voie secondaire / tourner à gauche	3
carrefour en T voie secondaire / tourner à droite	3
virage léger gauche	2
virage gauche	2
virage droite	2

Tableau 1 : Fréquence moyenne des infrastructures/objectifs sur le parcours

La durée moyenne du parcours est de 45 minutes.

Les données de conduite sont recueillies à l'aide du véhicule expérimental du LESCOT et sont détaillées dans le paragraphe suivant.

3.2.3. Données recueillies.

Plus de quatre heures d'enregistrement vidéo et numérique ont été recueillies :

- Du coté de la vidéo, la scène avant, la scène arrière, et le conducteur sont filmés et enregistrés. L'enregistrement est ensuite numérisé au format Xvid©.
- Du coté des capteurs numériques, l'angle du volant, l'enfoncement des pédales, la vitesse, l'accélération longitudinale et latérale, les valeurs des comodos et l'état des phares sont enregistrés à une fréquence 50hz.

Le logiciel AMMAC (basé sur Matlab ©) a été conçu pour analyser ces données (illustration 3.14).

Sa conception, son élaboration et sa mise en oeuvre ont été effectuées pendant la thèse. Il nous a permis :

- de visualiser l'activité de conduite en permettant l'analyse simultanée de l'évolution des données numériques et vidéo,
- d'annoter des enregistrements afin de définir le début et la fin des séquences de conduite,
- de caractériser les séquences de conduite par une procédure itérative.
  - En effet, pour certaines caractéristiques, comme l'infrastructure, les différentes valeurs possibles peuvent être dérivées des travaux antérieurs (voir paragraphe 1.1.4.b).

  - Pour d'autres, comme l'objectif du conducteur, il est difficile de savoir a priori quelles sont les différentes possibilités. Dans ce dernier cas, seule l'analyse itérative des données peut fournir une liste exhaustive.

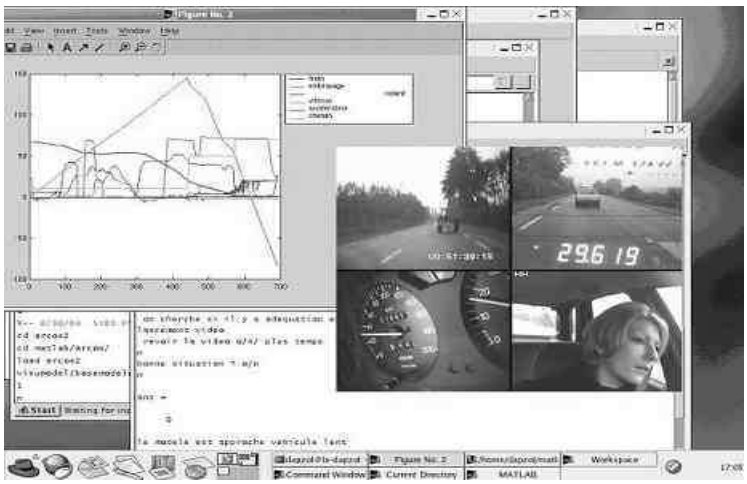


Illustration 3.14 : Logiciel AMMAC développé pour exploiter les données de conduite

Après analyse, chaque séquence de conduite est alors définie par 6 caractéristiques :

- Le contexte urbain (ville, autoroute, campagne),
- L'infrastructure traversée,
- L'objectif du conducteur,
- La vitesse initiale (nulle, faible, moyenne, forte),
- Le trafic (nul, moyen, élevé),
- Les particularités de la situation (événements survenus, inattention...)

Remarque : 1) Si les 5 premières variables sont à modalités, la 6ème nous renseigne sur l'ensemble des observations non catégorisables. Cette variable nous permet dès lors de classifier les séquences de conduite suivant leur typicité (prototypique, atypique, singulière),

2) Les différentes valeurs pour l'infrastructure et le contexte urbain ont été développées à partir de la classification des environnements routiers de Bellet (voir paragraphe 1.1.4.b) et en se situant respectivement au niveau 1 et 4 de la hiérarchie,

3) L'interprétation de l'objectif du conducteur se fait en fonction des principes énoncés dans le paragraphe 1.2.1.

Ainsi, 1209 séquences de conduite ont pu être extraites, interprétées (en terme de situation vécue par le conducteur) et catégorisées.

Le tableau en annexe 3 regroupe ces séquences en fonction du contexte urbain, de l'infrastructure vécue par le conducteur, de son objectif et de sa vitesse initiale.

3.3 Modélisation de l'évolution des signaux lors des situations de conduite par le Modèle Semi-Markovien Pondéré.

Afin de modéliser la relation entre des situations de conduite et l'enregistrement des capteurs, nous avons choisi de nous appuyer sur les chaînes de Markov cachées. Ce choix, comme nous l'avons expliqué dans la partie 1, s'est fait suivant les résultats des précédentes études sur la conduite (tant basées sur des analyses psychologiques que numériques) et sur les propriétés intrinsèques de ce type de modèle.

Les limites rencontrées dans les précédentes études utilisant les chaînes de Markov cachées pour l'analyse de la conduite étaient:

- le nombre restreint de situations étudiées
- la complexité des capteurs utilisés

Notre volonté d'appréhender un maximum de situations tout en utilisant des données issues de capteurs « simples », nous a amené à concevoir et développer:

- Le modèle Semi-Markovien caché pondéré (MSMCP), (chapitre 2.1.4.b),
- Des algorithmes d'apprentissage efficients dans le cadre de l'analyse de l'activité de conduite (chapitre 3.3.1.c),
- Un processus de modélisation spécifique. (chapitre 3.3.4).

Ces trois aspects seront décrits dans la partie suivante qui sera ainsi consacrée à la description du développement du catalogue de modèles associés aux différentes situations de conduite.

3.3.1. Modélisation par les chaînes de Markov cachées: principe, modèle, et algorithmes utilisés

Après un rappel sur des principes sous-tendant les chaînes de Markov cachées, nous présentons ici les caractéristiques générales de notre modélisation.

a Rappels sur les modèles de Markov cachées

Les modèles de Markov cachées (MMC), présentés dans la partie 2.1, sont fondés sur l'hypothèse que le signal étudié est divisé en plusieurs phases. Communément, on dit qu'il existe deux processus, l'un « Y », associé au signal étudié est dit le processus visible, et l'autre « S », passant par différents états et divisant ainsi le signal en différentes phases, est dit le processus invisible.

Nous avons vu qu'associés à ce genre de modèle, différents algorithmes existaient pour répondre à trois problèmes classiques (chapitre 2.1.2) visant

a) le calcul de la probabilité d'une observation,
b) le calcul de la séquence d'états cachés par laquelle est passé le processus S,

c) l'estimation des paramètres d'un modèle en fonction de données observées.

Par ailleurs, pour connaître la topologie d'un tel modèle (le nombre d'états et lien entre eux), il n'existe pas de solution simple. Seule l'utilisation d'un critère particulier (chapitre 2.1.3.a) et d'une approche itérative (chapitre 2.1.3.b) peut permettre d'accéder au modèle le plus adéquat pour représenter les données.

Dans cette optique, de nombreux auteurs ont élargi le concept des MMCs en développant des topologies ou des contraintes spécifiques (chapitre 2.1.4.a). Ainsi, les modèles semi-Markovien Cachés modélisent explicitement le temps passé par le processus S dans chaque état. C'est sur ce dernier modèle que nous nous sommes appuyés pour développer le Modèle Semi-Markovien Caché Pondéré qui intègre le concept de pondération et de modélisation explicite du temps passé dans chaque état. Nous avons alors établi pour ce modèle les solutions algorithmiques aux trois problèmes précédents (chapitre 2.1.4.b).

b Principe de la modélisation de l'évolution des signaux issus des capteurs par les modèles semi-Markovien cachés pondérés.

Capteurs utilisés.

Les données que nous utilisons pour analyser l'activité de conduite, seront le clignotant, l'angle du volant, l'accélération tangentielle, la vitesse, et les dérivées de l'angle du volant et de la vitesse. Ce choix se fonde sur la fiabilité de ces variables (au contraire d'autres mesures tel que la distance au véhicule précédent à la valeur plus incertaine) et sur l'interprétation aisée de leur valeur (voir partie 1.2).

Modèle utilisé

Dans le cadre de cette thèse, nous avons choisi de modéliser l'évolution des données issues des capteurs par les chaînes de Markov cachées.

En outre, des travaux préliminaires ont montré que l'efficacité de ces modèles pouvait être améliorée en intégrant deux contraintes:

1. Le temps passé dans chaque état est toujours supérieur à une seconde :

   En effet, comme nous l'avons vu dans la première partie, la conduite est une activité ayant une forte temporalité et en nous inspirant des travaux sur les schémas de conduite, on peut supposer que le couple véhicule-conducteur passe par des séries d'états stables. Aussi il nous est apparu important de modéliser cette contrainte en supposant que le temps passé dans chaque état est toujours supérieur à une seconde.

2. En fonction des situations, certaines variables sont plus porteuses d'information que d'autres.

   Par exemple, pour modéliser un « Tourner à Gauche », le volant semble a priori plus important que la vitesse. Aussi, pour palier à la complexité des situations de conduite, nous sommes amenés à utiliser le concept de pondération sur les variables. Ceci nous permettra de définir pour chaque chaque situation et pour chaque état quelles sont les variables les plus

pertinentes.

L'intégration de ces deux contraintes nous a semblé importante pour modéliser l'activité de conduite. Aussi, nous avons conçu et développé le Modèle Semi-Markovien Caché Pondéré (MSMCP) dont la partie théorique a été exposée dans la partie 2.1.4.b.

C'est ce modèle qui sera à la base de notre modélisation.

Relation entre les données capteurs et le modèle

Dans le cadre de l'analyse des données de conduite, les réalisations « y » du processus visible Y seront les données numériques issues des capteurs. Ces réalisations seront fonction du processus « invisible » S. Ce processus pourra s'apparenter à l'état dans lequel se trouve le couple « conducteur-véhicule » (illustration 3.15).

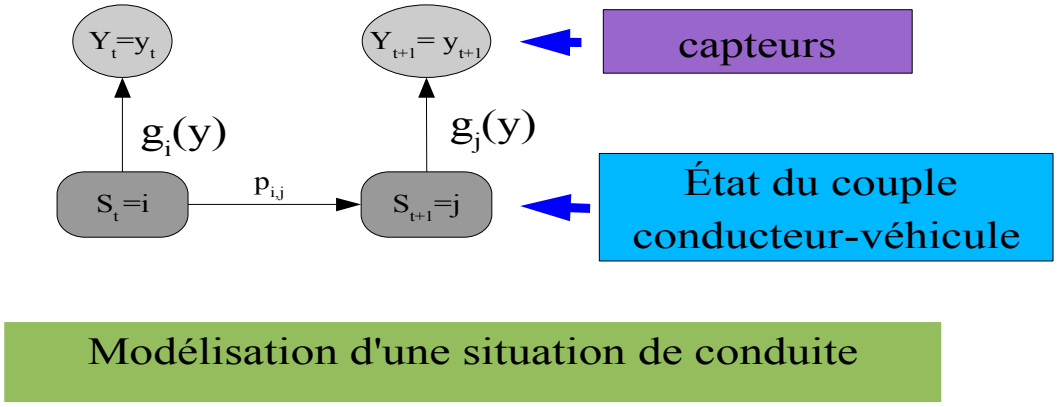


Illustration 3.15 : Modélisation d'une situation de conduite par chaînes de Markov cachées

Quelques définitions :

On appelle SC_d la d^{ième} situation de conduite modélisée et N le nombre total de situations de conduites modélisées.

De plus, on nomme une séquence de conduite $SC_{d,q}$ une réalisation d'une situation de conduite « SC_d » $d \in [1:N]$. Pour chaque situation de conduite, on dispose ainsi de N_d séquences de conduites $SC_{d,q}$ $q \in [1:N_d]$. Enfin, chacune de ces séquences de conduite $SC_{d,q}$ pourra être de durée variable $T_{d,q}$ $avec$ $d \in [1:N]$, $q \in [1:N_d]$.

Par ailleurs, à chaque situation « d » est associé un ensemble de séquences d'apprentissage SA_d et un ensemble de séquences de test ST_d . Le choix de ces ensembles sera discuté dans les paragraphes suivants.

Un modèle MSMCP de paramètre θ_d est associé à chaque situation de conduite SC_d .

avec $\theta_d=[\pi^d, p^d, m^d, v^d, \varphi^d, c^d]$ et

- π^d le vecteur de probabilité initial,
- p^i la matrice de transition,
- m^d et v^d la matrice des moyennes et des variances des lois normales associées à chaque état,
- ς^d le vecteur de paramètres caractérisant la loi de probabilité de durée des états,
- c^d la matrice des poids.

On nomme K^d le nombre d'états du modèle d.

De plus, pour faire face à la diversité des situations de conduite, nous avons défini deux types de topologies :

1. Linéaire : les états sont reliés à eux-mêmes et à l'état suivant (sauf le dernier état qui n'est relié qu'à lui-même.)
   La situation de conduite est alors supposée comme ayant une temporalité forte. C'est à dire que le conducteur a un comportement dans cette situation composé de plusieurs actions successives.
2. Totale : tous les états sont reliés entre eux.

   Dans ce cas, la situation de conduite est supposée « stable ». Il s'agit moins d'une succession déterminée d'actions que d'un ensemble d'actions possibles pour maintenir la stabilité de la situation. Par exemple lorsque le conducteur est dans la situation « rouler en vitesse stable en ligne droite » il opère des micro ajustements (accélérer, freiner...) qui n'ont pas d'ordre précis.

Par ailleurs, pour modéliser au mieux l'activité de conduite, nous avons développé des algorithmes d'apprentissage spécifiques. Ceux-ci seront discutés dans la prochaine partie.

c Algorithmes d'apprentissage

Afin d'estimer les paramètres des différents modèles, nous nous sommes appuyés sur deux algorithmes. Le premier, l'algorithme E-M, permet d'estimer les paramètres θ_d pour une topologie fixée. Le deuxième, conçu et développé pendant la thèse permet d'estimer la topologie adéquate en se basant sur un critère d'évaluation de modèle et des techniques de générations de modèles.

c.1 Apprendre les paramètres θ_d

Pour chacune des situations « d », les paramètres $\theta_d=[\pi^d, p^d, m^d, v^d, \varphi^d, c^d]$ recherchés sont ceux maximisant $p(SA_d|\theta_d)$. Pour trouver ces paramètres, nous avons adapté l'algorithme E-M pour les MSMCP (2.1.4.b.2).

Les observations « y » sont alors les enregistrements des capteurs lors des séquences de conduite choisies pour l'apprentissage et associées au modèle d.

Cependant, chaque séquence de conduite a une durée différente $T_{d,q}$ *avec* $d \in [1:N]$, $q \in [1:N_q]$. Aussi, l'apprentissage est équilibré de façon à ce que chacune ait la même importance dans l'apprentissage. Pour cela, pour le modèle « d », à l'étape (m) et pour chaque séquence $SC_{d,q}$ on défini $\theta_{d,q}^{(m+1)}$ maximisant $Q(\theta_d / \theta_d^{(m)}, SC_{d,q})$ avec Q tel que défini par la relation (8).

Puis on définit $\theta_d^{(m+1)} = \frac{1}{N_d} \sum_{q=1}^{N_q} \theta_{d,q}^{(m+1)}$, l'estimateur de θ_d à l'étape (m+1) de l'algorithme.

On itère l'algorithme (voir chapitre 2.1.1.c) jusqu'à ce que la vraisemblance n'augmente sensiblement plus $\frac{L(\theta_d^{(m+1)})}{L(\theta_d^{(m)})} < 1{,}0001$.

Cet algorithme a été implémenté en Matlab ©.

A titre d'exemple, pour un modèle à 6 états et pour 5 séquences d'apprentissage, 6 variables et $T_{q,d} \approx 300$, $\forall d = 1...5$, une itération requiert 1,5seconde[1].

c.2 Apprendre la topologie

Nous avons vu dans la partie 2 que pour connaître la topologie la plus efficiente pour un modèle, il fallait d'une part se baser sur des techniques de génération de modèles et d'autre part utiliser la vraisemblance pénalisée des données afin de choisir le modèle adéquat (paragraphe 2.1.2).

Dans notre cas, l'activité de conduite est extrêmement complexe (du fait d'un grand nombre de facteurs impliqués, tant au niveau de l'environnement qu'au niveau du conducteur) et n'est visible que partiellement avec les capteurs numériques. Aussi, à moins de posséder une large base de données sur le comportement, il est difficile de savoir si les ensembles d'apprentissage de chaque modèle ne constituent pas une vue biaisée d'une situation (présence d'événements inhabituels, action singulière du conducteur...) et si le modèle ne reflète pas uniquement ces singularités.

Aussi, il ne nous semble pas opportun d'opter pour une validation automatique des modèles. Une expertise humaine, de par sa connaissance du comportement semble importante pour valider la cohérence et les propriétés généralisantes d'un modèle. Par ailleurs, l'un des principaux avantages du modèle de Markov caché est qu'il permet de diviser l'activité en différents états. Ces états peuvent alors être labellisés et leurs liens interprétés.

Nous choisissons de mettre en avant cette particularité des MMCs et cette capacité de l'expert à utiliser ses propres connaissances sur la conduite pour valider/infirmer un modèle.

Ces choix se retrouveront dans la méthodologie choisie telle que nous le verrons dans la partie , mais aussi dans l'apprentissage de la topologie.

Ainsi, afin d'augmenter la lisibilité des modèles, nous choisissons de nous inspirer des méthodes d'agglomération (en utilisant le critère de *l'état le moins instructif* [2.3.2.1 (a)]) et de division d'état (en utilisant le critère de *l'état le moins pertinent* [2.3.2.2 (b)]) afin de développer un algorithme spécifique d'apprentisage de la topologie.

En effet, ces méthodes s'appuyent sur des critères permettant d'interpréter avec plus de facilité

1 Pour le test, l'ordinateur utilisé est un Pentium 2800 Mhz

les différents états du modèle :

1. La première méthode (de *l'état le moins instructif*) permet de sélectionner un modèle où chaque état est « utile », c'est-à-dire tel que $\forall i \in [1:K]$,

$$card\{S_t = i\} > \frac{T}{2K} \text{ (Critère 1),} \tag{79}$$

avec K le nombre d'état du modèle et $card\{S_t = i\}$ approché avec l'algorithme de Viterbi (voir chapitre 2.1.2.b).

2. la deuxième méthode (de *l'état le moins pertinent)* permet de sélectionner un modèle où pour chaque état les observations ne sont pas dispersées autour des états parents, c'est-à-dire tel que $\forall i \in [1:K]$,

$$[p(A_i) > \frac{p(y_{1...T})}{2K}] \text{ (Critère 2)} \tag{80}$$

avec $A_i = \{y_t / S_t = i\}$ approché par l'algorithme de Viterbi.

Par ailleurs, nous choisissons d'utiliser le critère BIC pour valider un modèle. En effet, ce critère a l'avantage de nécessiter peu de calcul et de sélectionner un modèle prédictif (chapitre 2.1.2.a).

L'algorithme, conçu et mis en place lors de la thèse, permet de mixer ces trois types d'approche.

Ainsi, partant d'un modèle M_0, nous augmentons le nombre d'états si le premier critère n'est pas vérifié (i.e. si il existe i 0<i<K tel que $card\{S_t = i\} > \frac{T}{2K}$) .

Nous diminuons le nombre d'états si le premier critère est vérifié mais que le deuxième critère ne l'est pas (i.e. si il existe i 0<i<K tel que $\underset{i \in [1:K]}{argmin} [p(A_i) > p(y_{1...T})/K * 0.5]$).

Si les 2 critères sont vérifiés, si le critère BIC du modèle n'augmente plus, ou si le nombre d'itérations maximum est atteint, l'algorithme s'arrête (illustration 3.16).

L'ensemble de la procédure a été implémenté en Matlab©. Si l'algorithme donne des résultats corrects, comme nous le verrons par la suite, le fait de calculer $card\{S_t = i\} \quad \forall i \in [1:K]$, ralentit grandement son exécution. Pour un modèle complexe, les tests effectués montrent qu'une itération peut durer jusqu'à 10 minutes.

Cependant son utilisation est ponctuelle et sa relative lenteur n'est donc pas génante pour l'analyse.

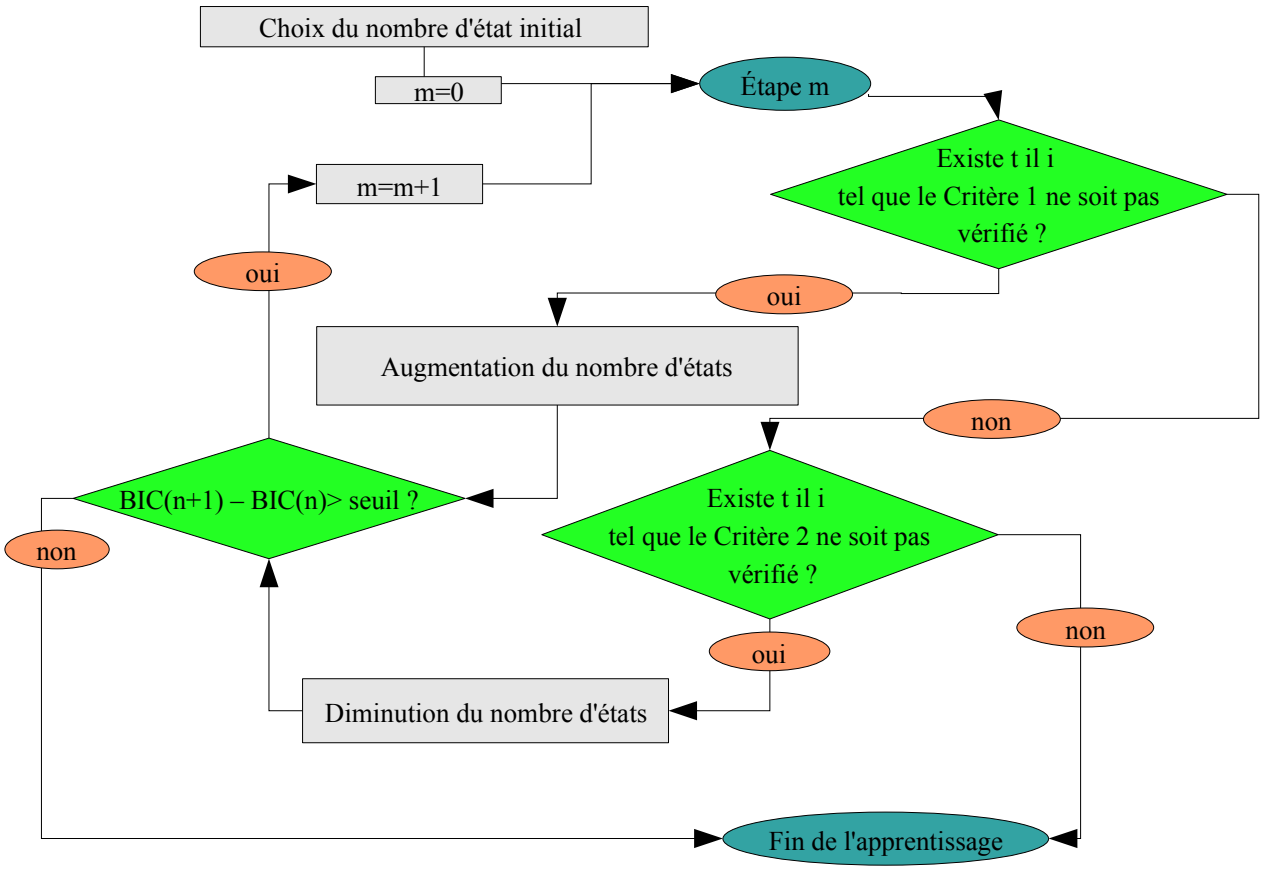


Illustration 3.16 : Apprentissage de la topologie : le nombre état initial est déterminé en fonction de la connaissance préalable que nous pouvons avoir de la situation.

Mise à jour de la topologie :

- Lorsque le nombre d'états est augmenté, l'état père est choisi comme l'état respectant le moins le critère 1. Puis un nouvel état est créé avec comme paramètre une variation aléatoire des paramètres de l'état père.
- la diminution d'états se fait en choisissant les états parents comme ceux respectant le moins le critère 2. Ces états sont éliminés et un nouvel état est créé avec comme paramètre, la moyenne de ceux des états parents.

3.3.2. Processus de modélisation : objectif, contraintes, méthode

Notre objectif est d'établir un ensemble de MSMCP modélisant l'évolution des capteurs lors des situations de conduite vécues catégorisées, suivant trois critères : *l'objectif du conducteur, l'infrastructure traversée* (telle qu'elle est *vécue* par le conducteur) et les *conditions initiales.*

Or, un des problèmes majeurs auxquels nous sommes confrontés est que la catégorisation première des séquences est incertaine. En effet, elle est effectuée sur des données partielles (voir chapitre 1) et sur une analyse sommaire des enregistrements vidéo. Des séquences peuvent ainsi être mal catégorisées et fausser le processus d'apprentissage.

Aussi, il est nécessaire que chaque séquence puisse être réaffectée à une catégorie diagnostiquée plus adéquate, au fur et à mesure de la modélisation.

Notre méthodologie souhaite tenir compte de cet état de fait en optant pour une double mise à jour d'une part sur les paramètres des modèles, d'autre part sur les séquences d'apprentissage des modèles.

En outre, pour pouvoir valider un modèle, il nous semble important d'aller au delà de la simple vraisemblance des données, et de ne retenir un modèle que s'il est interprétable en terme d'activité de conduite.

En conséquence, nous avons organisé notre apprentissage en trois étapes :

1. Sélection des séquences d'apprentissage.
2. Modélisation des Situations Vécues :

   A chaque situation de conduite est associée un MSMCP dont on optimise les paramètres. Cette étape permet de plus de remettre en question la catégorisation initiale des séquences.
3. Regroupement des Situations Vécues en Situations Mesurées

   On regroupe différentes Situations Vécues, proches du point de vue des capteurs, en Situation Mesurée.

Chaque Situation Mesurée comprendra alors un ensemble de Situations Vécues qui ne peuvent être pas distinguées avec les capteurs actuels.

La phase de *Validation* est constituée de deux étapes :

1. Test de la reconnaissance sur les séquences prototypiques
2. Test sur les séquences atypiques

3.3.3. Sélection des séquences.

Dans un premier temps, la liste des séquences extraites est triée en différentes situations homogènes suivant 3 critères :

- la vitesse initiale
- l'infrastructure vécue par le conducteur.
- l'objectif du conducteur

Puis nous retenons les situations les plus fréquentes, c'est-à-dire possédant plus de cinq séquences prototypiques (i.e. sans évènement externe venant perturber l'organisation initiale des mouvements du conducteur).

Les situations répondant aux critères précédants sont alors aux nombre de 36. Elles regroupent alors 718 séquences de conduite (tableau 2).

Puis, pour chacune de ces situations, on définit des séquences d'apprentissage. Celles-ci sont choisies de façon aléatoire parmi les N_d séquences de chaque situation. Pour limiter les temps de calcul, le nombre de séquence d'apprentissage est limité, soit à 5 si $N_d \geq 10$ soit à N_d /2 si $N_d < 10$.

Pour chaque situation, ces séquences seront donc les séquences SA_d , servant d'apprentissage aux modèle associée.

Les ensembles de tests ST_d seront l'ensemble des autres séquences.

Ainsi 126 séquences servent à l'apprentissage des modèles et 592 à leur validation.

La durée moyenne de chaque séquence est de 15s. Aussi pour chacun des 6 paramètres étudiés 15*50 Hz =750 données sont associées à chaque séquence.

Nombre de séquence: N_d	Nom de la situation : contexte / objectif / vitesse initiale
113	ville / ligne droite / suivre route / vitesse stable / vitesse moyenne
43	ville / ligne droite / suivre route / vitesse stable / vitesse forte
14	ville / ligne droite / accélérer / vitesse nulle
7	ville / ligne droite / accélérer / vitesse faible
94	ville / ligne droite / accélérer / vitesse moyenne
18	ville / ligne droite / freiner / vitesse moyenne
14	ville / ligne droite /s'arrêter en prévision de tourner droite / vitesse moyenne
7	ville / ligne droite /s'arrêter / vitesse faible
25	ville / ligne droite /s'arrêter / vitesse moyenne
9	ville / ligne droite / changer de voie : gauche vers droite / vitesse moyenne
6	ville / ligne droite / changer de voie : gauche vers droite / vitesse forte
16	ville / ligne droite /s'arrêter en prévision de tourner gauche / vitesse moyenne
14	ville / ligne droite / changer de voie droite vers gauche / vitesse moyenne
10	ville / intersection / tourner gauche / vitesse nulle
17	ville / intersection / tourner gauche / vitesse moyenne
32	ville / intersection / tourner droite / vitesse nulle
25	ville / intersection / tourner droite / vitesse moyenne
5	ville / intersection / aller tout droit / vitesse nulle
13	ville / intersection / aller tout droit / vitesse moyenne
22	ville / intersection / être arrêter en préparation de tourner droite / vitesse nulle
12	ville / intersection / être arrêter en prévision de tourner gauche / vitesse nulle
44	ville / intersection / être arrêter / vitesse nulle
8	ville / carrefour en T / voie secondaire / tourner gauche / vitesse nulle
9	ville / carrefour en T / voie secondaire / tourner droite / vitesse moyenne
5	ville / ronds points / tourner droite / vitesse moyenne
15	ville / ronds points / aller tout droit / vitesse moyenne
7	ville / virage léger gauche / suivre route / vitesse stable / vitesse moyenne
10	ville / virage gauche / suivre route / vitesse stable / vitesse moyenne
14	ville / carrefour en T / voie principale / tourner gauche / vitesse moyenne
22	ville / carrefour en T / voie principale / tourner droite / vitesse moyenne
8	ville / carrefour en T / voie principale / être arrêter / vitesse nulle
26	autoroute / ligne droite / suivre route / vitesse stable / vitesse moyenne
5	autoroute / ligne droite / accélérer / vitesse moyenne
6	autoroute / ligne droite / changer de voie : gauche vers droite / vitesse moyenne
10	autoroute / ligne droite / changer de voie droite vers gauche / vitesse moyenne
13	autoroute / virage léger droite / suivre route / vitesse stable / vitesse moyenne
718	Total

Tableau 2 : Séquence prototypique regroupée suivant : le contexte routier / l'infrastructure / l'objectif / la vitesse initiale

3.3.4. Modélisation des Situations de Conduites Vécues: une procédure itérative

De part les contraintes énoncées précédemment (catégorisation incertaine et validation par expertises), nous avons choisi de procéder par itérations successives pour modéliser les situations de conduite.

Aussi, pour modéliser l'évolution des capteurs dans *chaque situation* $d \in [1:N]$ nous procédons par étapes :

1. Dans une première phase nous initialisons la topologie du modèle en fonction de notre connaissance de la situation de conduite. Puis nous effectuons un apprentissage du paramètre θ_d via l'algorithme E-M (3.3.1.c.1)

2. Lors d'une deuxième phase, nous analysons la vraisemblance des séquences d'apprentissage. Si une séquence a une vraisemblance plus faible que les autres, cela peut être le fait, soit d'un mauvais apprentissage, soit d'un comportement mal interprété.
   - Dans le premier cas, un apprentissage de la topologie et un réapprentissage des paramètres sont nécessaires.
   - Dans le deuxième cas, la séquence doit être réanalysée en détail en la comparant notamment avec les autres séquences de la même situation. Si nécessaire, la séquence est alors recatégorisée comme appartenant à une autre situation.

3. Enfin, dans une dernière étape de validation, le modèle est soumis à une interprétation des états appris. Si ces états peuvent être expliqués, alors le modèle pourra être généralisable et est donc validé. Sinon, le modèle est construit « manuellement » (illustration 3.17).

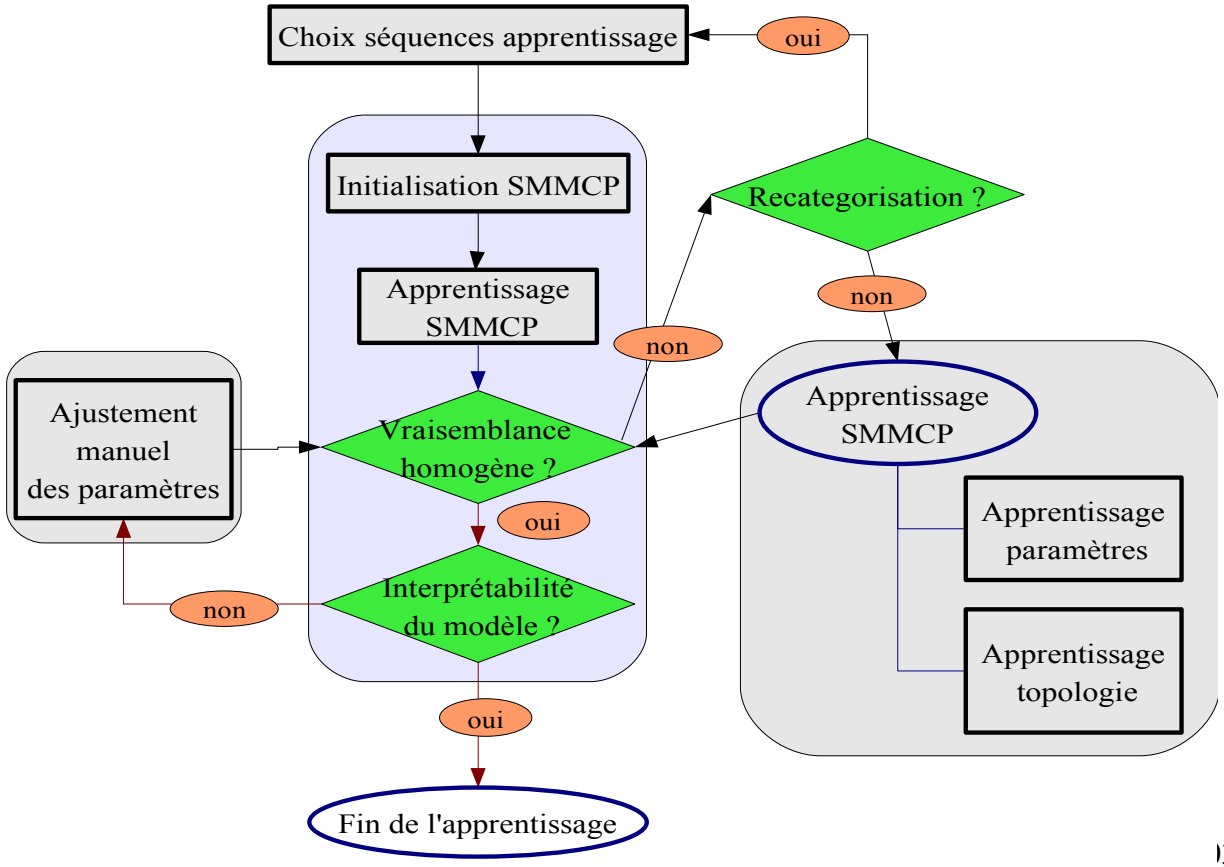


Illustration 3.17 : Processus d'apprentissage des modèles pour chaque situations.

Nous présentons l'application de ce processus en détaillant les différentes étapes (« trame principale », phases de «Recatégorisation», de « réapprentissage », et « d'ajustement manuel des paramètres ») et ce, à travers quatre exemples.

Nous illustrerons ces étapes à travers la modélisation de quatre situations particulières :

1. Ville / Rond-point / tourner à droite / vitesse moyenne.
2. Ville / Ligne droite /conduire à vitesse stable /vitesse moyenne.
3. Ville / Intersection / Tourner à gauche / vitesse moyenne.
4. Ville / Ligne droite /changer de voie gauche vers droite /vitesse moyenne.

a Trame principale : illustrée par la situation « *Ville / Rond-point / Tourner à droite / Vitesse moyenne* ».

Sur les 15 séquences de cette situation qui furent enregistrées, 5 ont été choisies aléatoirement pour l'apprentissage. Du fait de la forte contrainte dûe à l'infrastructure (rond-point), on suppose la topologie linéaire. En outre, on suppose que le modèle associé à cette situation est composé de 6 états.

Le processus d'apprentissage converge après 6 itérations. L'analyse qualitative de la vraisemblance des séquences par rapport au modèle ($\log(P(SC_{q,d}/\theta_d))$ montre une relative homogénéité dans leur valeur (tableau 3)

séquence	***1***	***2***	***3***	***4***	***5***
Log-vraisemblance	-9,94	-9,96	-12,57	-13,24	-7,51

Tableau 3 : log-vraisemblance des séquences d'apprentissage pour la situation Ville ligne droite / conduire à vitesse stable / vitesse moyenne

On utilise alors l'algorithme de Viterbi (2.1.1.b) pour visualiser le découpage le plus probable pour chaque séquence (Figure 3.18 et 3.19).

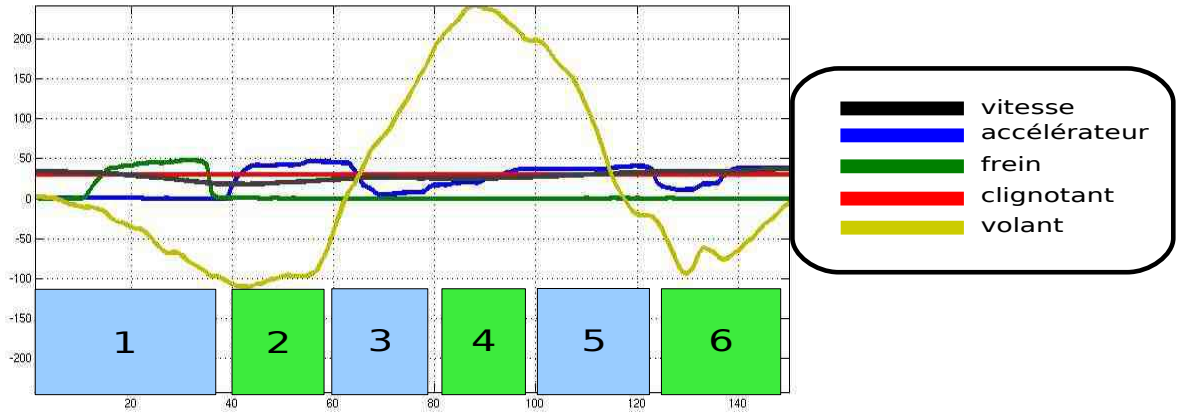


Illustration 3.18. : Utilisation de l'algorithme de Viterbi pour trouver le chemin le plus probable : séquence 1 de la situation « ville/ronds-points/aller tout droit /vitesse moyenne »

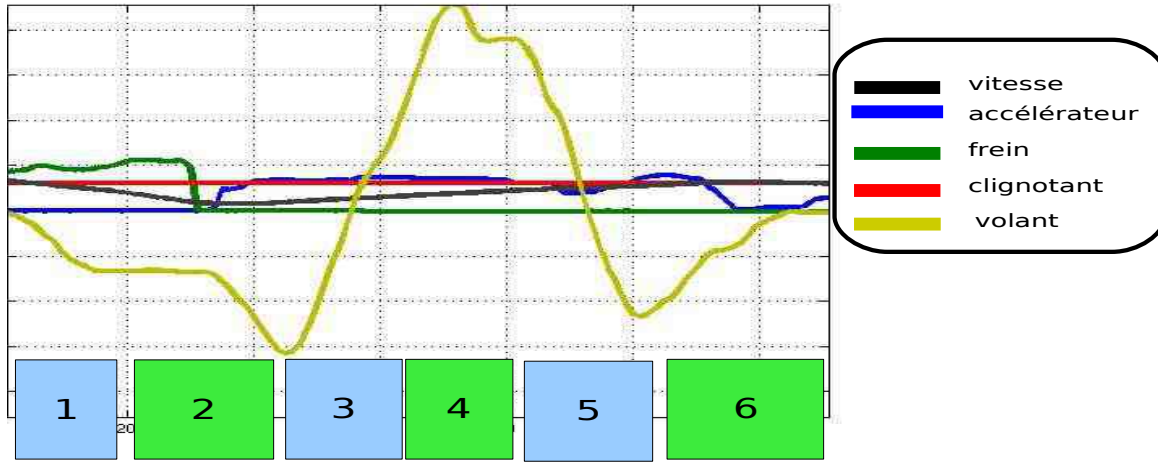


Illustration 3.19 : Utilisation de l'algorithme de Viterbi pour trouver le chemin le plus probable : séquence 4 de la situation « ville/ronds-points/aller tout droit /vitesse moyenne ». Ici une légère chicane fait durer un peu plus longtemps l'état 2.

Pour toutes les séquences, les différents états peuvent être labellisés de la même façon :

Première partie du rond point:

1. Engagement dans le rond-point.
2. Suivi de la première courbure du rond point.
3. Suivi dû à la deuxième courbure du rond point et augmentation de la vitesse.

Début de sortie :

4. Préparation de la sortie (l'angle au volant diminue).
5. Sortie du rond-point (le volant décrit une légère courbe vers la droite et la dérivée de la vitesse diminue)
6. Stabilisation : l'angle au volant revient à zéro et la vitesse se stabilise.

Pour l'ensemble des séquences, les différents états sont interprétables. Le modèle est donc validé.

La localisation de ces états dans l'infrastructure est illustrée par la Figure 3.20. Ils peuvent ainsi être interprétées, comme *des zones de déplacements* tels que les définisse le modèle COSMODRIVE (paragraphe 1.1.3.d).

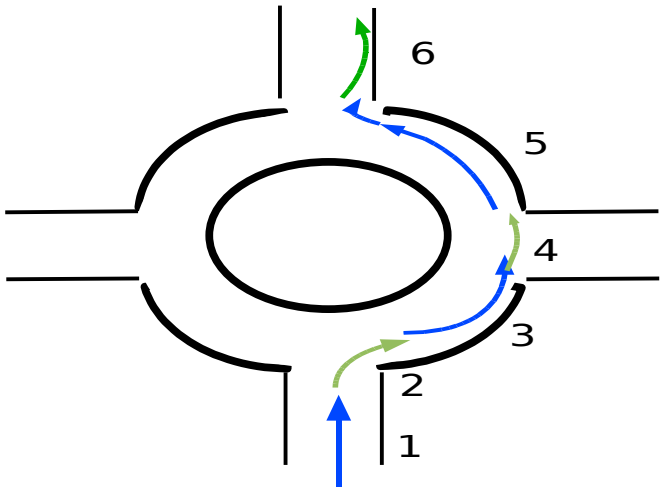


Figure 3.20 : Localisation des états du modèle associé à la situation : ville/ronds-points/aller tout droit /vitesse moyenne.

b Phase de recatégorisation *illustrée par la situation « Ville / Ligne droite / Conduire à vitesse stable / Vitesse moyenne »*

Sur les 132 séquences de cette situation enregistrées, cinq ont été choisies pour l'apprentissage. Du fait de la stabilité de cette situation, nous optons pour un MSMCP à un seul état.

L'apprentissage des paramètres se fait alors par l'algorithme présenté en 3.3.1.b.1.

La deuxième et la cinquième séquence ont des log-vraisemblances plus faibles que les autres (tableau 4).

	séquence				
	1	***2***	***3***	***4***	***5***
Log-vraisemblance	-2.3956	-2.7361	-2.3362	-2.2913	-4,3212

Tableau 4 : Log-vraisemblance des séquences d'apprentissage pour la situation Ville ligne droite / conduire à vitesse stable / vitesse moyenne

L'analyse conjointe des données numériques de ces séquences et les vidéos associées fait

ressortir que lors de la cinquième séquence, le conducteur effectue un écart rapide avec son volant sur la gauche (Figure 3.21).

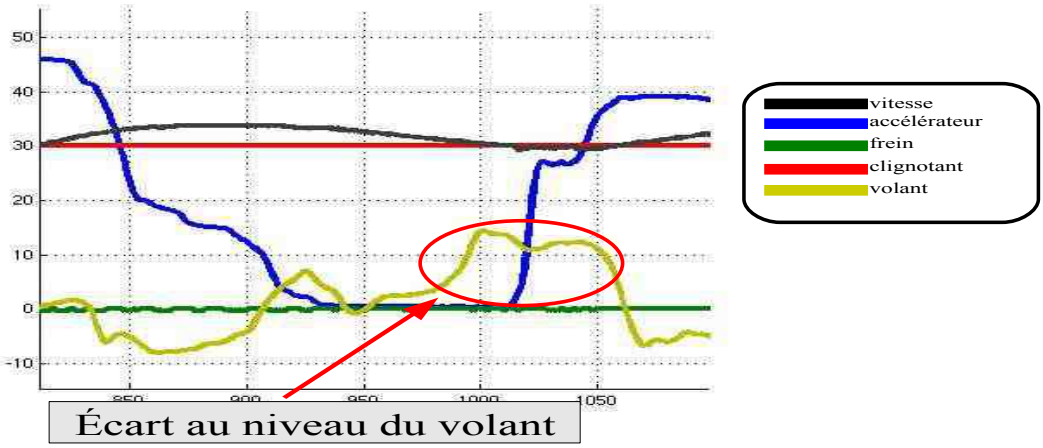


Figure 3.21 : Lors de la Séquence 5, l'infrastructure réelle est un carrefour en T. Cependant, du fait d'un comportement peu visible, l'infrastructure vécue par le conducteur avait tout d'abord été définie comme une ligne droite mais différents indices témoignent qu'il a bien pris en compte le carrefour.

Cet écart ayant une très faible influence sur la dynamique de la voiture, il n'a pu être noté via la vidéo. Il peut s'expliquer par la présence d'une intersection en T.

Deux autres indices témoignent que le conducteur a bien pris en compte l'intersection. D'une part, le conducteur arrête d'accélérer pendant 1s avant de passer l'intersection, d'autre part il a un très bref mouvement de tête au moment de son franchissement (Illustration 3.22).

Illustration 3.22 : Lors de la séquence 5, par une analyse détaillée de la vidéo, on peut voir que le conducteur regarde très brièvement si la voie est dégagée.

Ces éléments n'avaient pas pu être pris compte par une analyse rapide de la vidéo. Aussi la dernière séquence est recatégorisée. Par ailleurs, l'analyse approfondie de la séquence 2 ne donne pas lieu à un changement dans sa classification. Le modèle «Ligne droite / conduire à vitesse stable / vitesse moyenne » est donc réappris avec l'ensemble des 4 premières séquences.

		séquence				
		1	***2***	***3***	***4***	***5***
Log-vraisemblance	Avant la réaffectation	-2.3956	-2.7361	-2.3362	-2.2913	-4,3212
	Après	-2.4844	-2.5866	-2.3327	-2.2904	*[-5,4502]*

Tableau 5 : Evolution de la log-vraisemblance avant et après la recatégorisation de la séquence 5.

La vraisemblance des séquences est plus homogène. On peut donc en conclure que l'évolution des capteurs lors de ces séquences est plus homogène. Le modèle passe la première phase de validation.

Remarque : Comme la modélisation de cette situation n'est composée que d'un seul état, la question de l'interprétabilité du modèle ne se pose pas.

c Phase de réapprentissage illustrée par la situation « Ville / I*ntersection / Tourner à Gauche /vitesse moyenne »*

Dans un premier temps, en nous basant sur la modélisation de COSMODRIVE, nous avons supposé que ce modèle avait une topologie linéaire et comportait 5 états : «approche initiale » - « engagement dans l'intersection » - « dans l'intersection : tourner » - «sortie de l'intersection » - « stabilisation ».

Après l'apprentissage des paramètres du modèle, on remarque que la vraisemblance des séquences d'apprentissage n'est pas homogène. Pourtant, après revérification, toutes sont catégorisées correctement.

Aussi la topologie du modèle est mise à jour pour pouvoir être la plus adaptée à la complexité du modèle. Pour cela nous utilisons l'algorithme décrit en 3.3.1.b. c.2.

Après 5 itérations, l'algorithme converge en un modèle composé de 8 états (Figure 3.18).

A l'issue de cet algorithme, le critère BIC (voir chapitre 2.1.3.a) du modèle a diminué (tableau6).

	Modèle 5 états	**Modèle 8 états**
critère BIC	1503	1146

Tableau 6: évolution du critère BIC avant et après l'apprentissage de la topologie.

Le découpage en 8 états du nouveau modèle peut alors être labellisé de manière plus détaillée ainsi :

1. *« Intersection lointaine »* : seul le clignotant permet d'inférer que le conducteur a déjà pris en compte l'intersection.

2. *« Intersection proche »* le conducteur commence à freiner et tourne légèrement ses roues vers la droite.

Dans l'intersection:

3. *« Vérification »* le conducteur tourne ses roues vers la gauche, puis vérifie si la voie est libre.

4. *« Engagement »* Dans l'intersection le conducteur accélère et tourne.

5. *« Engagement fort »* le conducteur tourne fortement.

6. *« Contre braquage »* le conducteur remet son volant droit et accélère pour sortir de l'intersection.

7. *« Stabilisation »* le conducteur stabilise le véhicule par un léger accoup sur le volant.

8. *« Fin de l'intersection »* : le conducteur accélère en ligne droite.

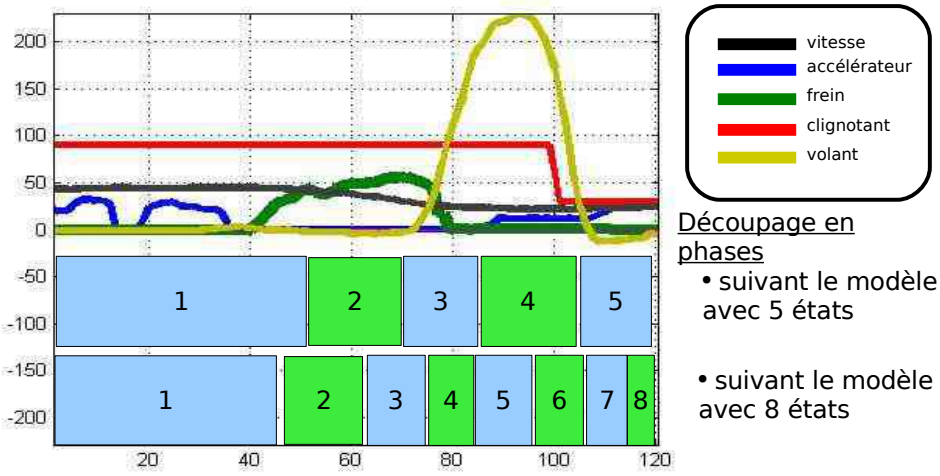


Illustration 3.23 : Tourner à gauche : différence de la modélisation entre les 2 topologies

Après apprentissage de la topologie, la deuxième séquence se distingue toujours par une vraisemblance plus faible (tableau 7).

Modèle à 8 état	Séquence				
	1	2	3	4	5
Log-vraisemblance	-13,32	-20,24	-9,72	-9,31	-10,65

Tableau 7 : log-vraisemblance des séquences pour le modèle à 8 états

Par ailleurs, sur cette séquence et au contraire des autres séquences, la labellisation proposée ci-dessus n'est pas applicable. Ceci s'explique par le fait que lors de cette séquence de conduite le conducteur n'a pas préparé sa manoeuvre. Aussi, la première phase dite « d'intersection lointaine » n'est pas présente (illustration 3.24).

Nous changeons donc manuellement les paramètres pour que la modélisation puisse intégrer le fait que la situation puisse commencer directement avec l'état « engagement » (illustration 3.24) : le vecteur de probabilité initiale π est alors défini ainsi : [0.5 ; 0.5 ; 0 ; 0 ; 0 ; 0 ; 0 ; 0 ; 0].

La vraisemblance de la séquence 2 diminue à-12,31.

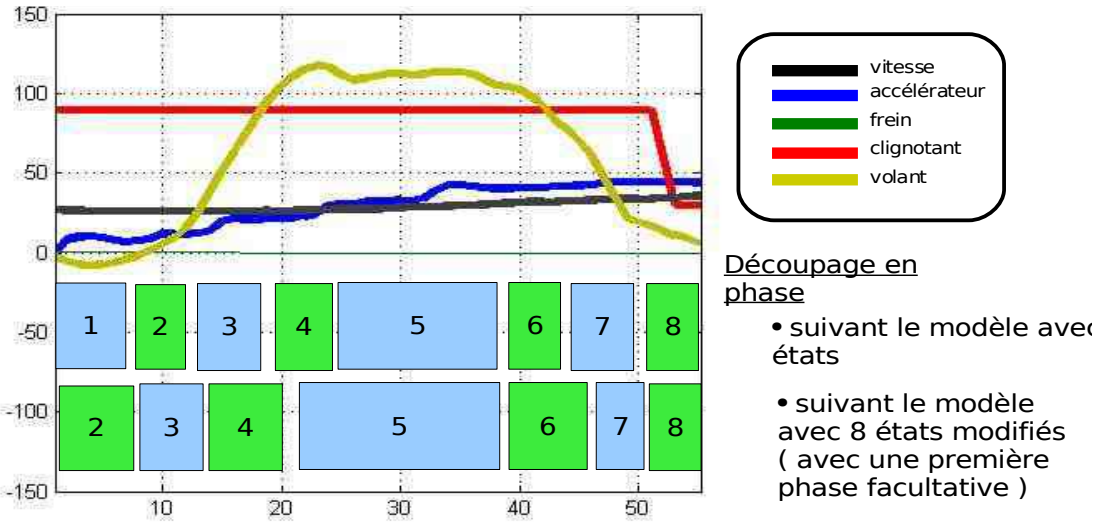


Illustration 3.24 : Séquence 2 : Tourner à gauche d'une intersection sans phase de préparation

d Phase d'ajustement manuelle des paramètres illustrée par la situation « *Ville / Ligne droite / Changer de voie de gauche vers droite / Vitesse moyenne* »

Pour cette situation, après la phase d'apprentissage, la vraisemblance des séquences semble homogène (tableau 5).

	séquence				
	1	***2***	***3***	***4***	***5***
Log-vraisemblance	-2.307	-3,2682	-3,0927	-2,6607	-2,9865

Tableau 8 : log-vraisemblance des séquences pour la situation « ville / changer de voie gauche vers droite /vitesse moyenne «

Pourtant, on constate que le modèle généré n'est pas interprétable : la labellisation des états est impossible. En effet, suivant les séquences, chaque état regroupe un ensemble non homogène de comportement. Ce fait s'explique par la grande hétérogénéité de cette situation.

En effet, pour chaque séquence, le conducteur adopte une stratégie différente : le conducteur rétablit ou non ses roues après le changement de voie (Figure 3.28), utilise ou non ses clignotants (Figure 3.26, 3.27), accélère ou freine en changeant de voie,etc...

L'apprentissage automatisé pour une situation si hétérogène n'est donc pas efficient.

Aussi nous construisons donc « manuellement » le modèle, en incluant les connaissances que nous pouvons avoir sur cette situation.

- En début de situation, utilisation possible des clignotants. *Il importe de faire commencer le modèle par 2 états possibles (avec, sans clignotant : phase 1 et 6) et d'inclure la possibilité que le conducteur allume son clignotant alors qu'il est déjà en cours de réalisation (phase 2)*
- Accroissement de la vitesse homogène. *Sur chaque état, la moyenne de la dérivée de la vitesse est définie égale à zéro.*
- Le volant passe par 3 états de moyenne respective : 0,-20,0 puis s'il y a rétablissement des roues : 2 phases suivent de moyenne 15 et 0.

La Figure 3.25 illustre le modèle ainsi construit

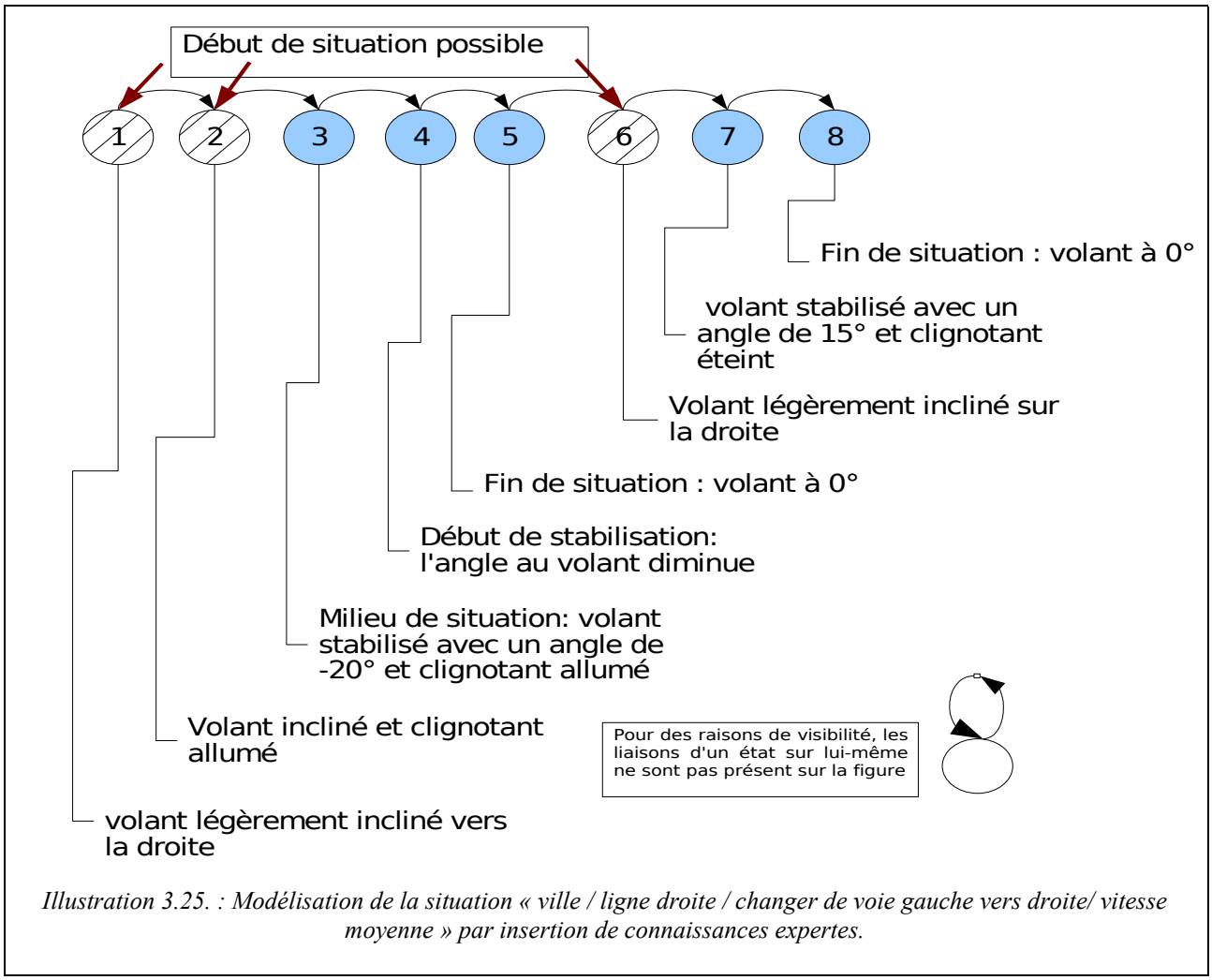


Illustration 3.25. : Modélisation de la situation « ville / ligne droite / changer de voie gauche vers droite/ vitesse moyenne » par insertion de connaissances expertes.

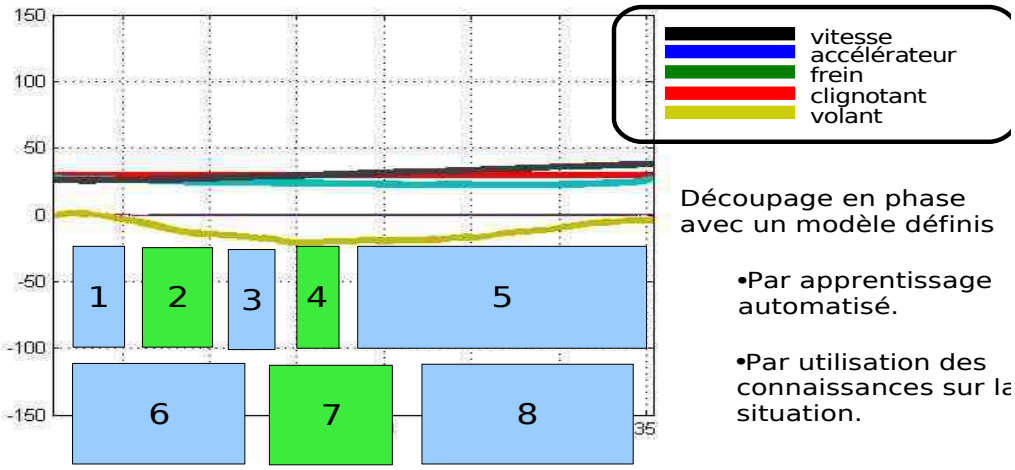


Figure 3.26 : Changer de voie sans utilisation du clignotant.

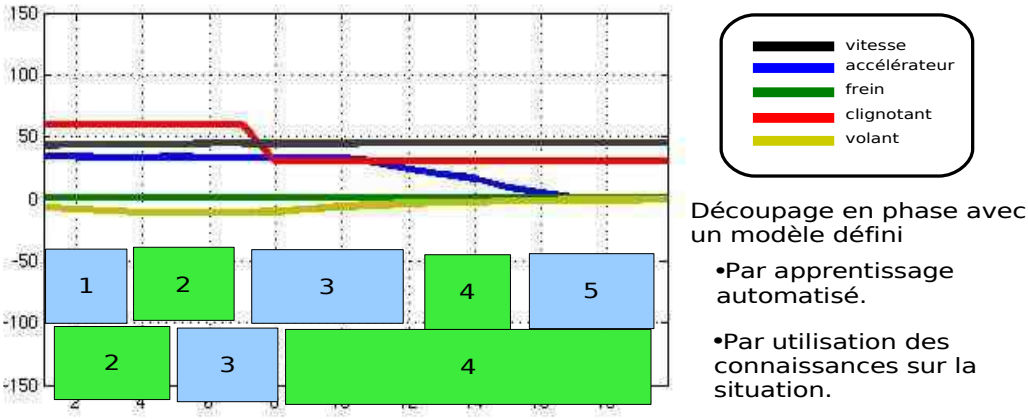


Figure 3.27 : Utilisation du clignotant pour changer de voie

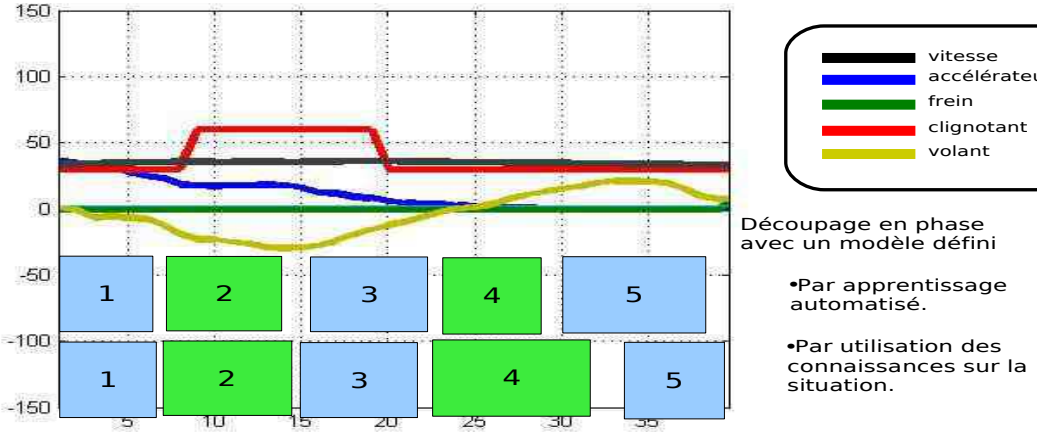


Figure 3.28 : Rétablissement des roues après le changement de voie : le conducteur tourne le volant légèrement sur la gauche après avoir changé de voie.

e Conclusion sur le processus d'apprentissage

Dans cette section, nous avons vu que le processus d'apprentissage se décomposait en en différentes phases d'apprentissage, de vérification et de labellisation.

Cette méthode a été programmée en Matlab. Du fait des différentes phases de validation et d'apprentissage, le processus de modélisation pour une situation peut être relativement long (à titre d'exemple, la modélisation d'une situation complexe type « changer de voie » peut prendre une heure et demi).

Ce processus d'apprentissage a été effectué pour l'ensemble des situations définies précédemment. Par l'utilisation conjointe des méthodes à apprentissage automatisé et d'une validation humaine des modèles générés, nous avons pu pour chacune des situations vécues modéliser l'évolution des paramètres associés.

Cependant, avec les technologies aujourd'hui disponibles, des situations se ressemblent. Aussi, il va nous falloir opérer un regroupement des situations trop proches.

3.3.5. Regroupement des Situations Vécues en Situations Mesurées

Avec les capteurs actuels, il est évident que toutes les situations vécues par le conducteur ne peuvent être distinguées. Le champ de perception technologique est en effet limité.

Ainsi, avec les capteurs que nous utilisons, les situations « Tourner à Gauche / Intersection / Vitesse nulle » et « Tourner à Gauche / Carrefour en T / Vitesse nulle » sont souvent confondues.

Il importe alors de regrouper certaines situations caractérisées suivant le point de vue du conducteur, en groupes de situations proches au niveau des capteurs. Cette proximité est donc ici fonction uniquement de l'évolution des capteurs. En employant la terminologie définie dans la première partie, nous choisissons de regrouper les *Situations de Conduite Vécues* en *Situations de Conduite Mesurées.*

Ce regroupement nous permettra alors de savoir :

- quelles sont les situations proches;

Ceci permet alors de s'interroger d'une part sur les instruments de mesure additionnels permettant de discriminer ces situations et d'autre part sur la proximité éventuelle de comportement dans les différentes situations.

- en fonction du nombre de regroupements effectués, (c'est à dire en fonction de la finesse de l'analyse souhaitée), quel sera le taux de reconnaissance du système.
- ou inversement, à quel niveau de détail il faut se placer pour avoir un taux de reconnaissance donné.

Pour opérer ce regroupement, nous choisissons d'utiliser une méthode conçue et développée lors de la thèse, s'inspirant des méthodes de Classification Ascendante Hiérarchique[1] sur les différents modèles et d'un indice de proximité spécifique.

Ainsi, chaque situation est tout d'abord considérée isolément. Puis, nous procédons itérativement. A chaque étape, les situations les plus proches sont regroupées, jusqu'à ce qu'il ne reste plus qu'un seul groupe de situation.

Nous prenons alors comme indice de proximité δ entre 2 situations, la pseudo-distance définie par Nechyba & Xu [Nechyba & Xu 1998] pour les modèles de Markov cachés,

$$\delta(M_1, M_2) = 1 - \sqrt{\frac{p(SA_1 \mid M_2)\, p(SA_2 \mid M_1)}{p(SA_1 \mid M_1)\, p(SA_2 \mid M_2)}} \tag{81}$$

avec SA_i les séquences d'apprentissage associées à M_i et si M représente plusieurs situations : $p(SA \mid M) = \min_{k \in [1:m]} p(SA \mid M_k)$.

Cette distance a ainsi l'avantage d'évaluer la proximité de modèles aux dimensions différentes en évaluant la vraisemblance des modèles vis à vis des séquences d'apprentissage associées à chacun.

Le processus consiste donc à considérer dans un premier temps l'ensemble des modèles et à calculer $\delta(M_i, M_j) \quad \forall i, j = [1:36]$. Puis les deux situations les plus proches sont regroupées. δ est recalculé sur le nouvel ensemble formé et les deux situations les plus proches sont une nouvelle fois regroupées. Ce processus est itéré jusqu'à ce qu'il ne reste qu'un seul groupe (Figure 3.29).

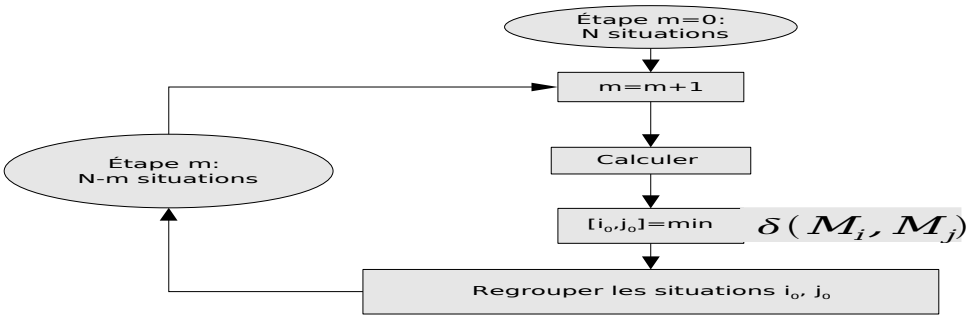


Figure 3.29 : Processus de regroupement des situations vécues en situations mesurées.

Le processus est effectué avec l'ensemble des 36 modèles construis précédemment.

Programmé en Matlab, le temps d'exécution de l'algorithme est de 2 minutes.

1 Technique statistique permettant d'opérer des hiérarchie de partition en classe des points d'un espace à N dimension. Usuellement toute hiérarchie minimise l'inertie intra-classe. Mais d'autres critères peuvent être aussi utilisés. Voir notamment à ce sujet Lebart et al [Lebart et al, 2000].

La Figure 3.44 donnent les regroupements ayant eu lieu.

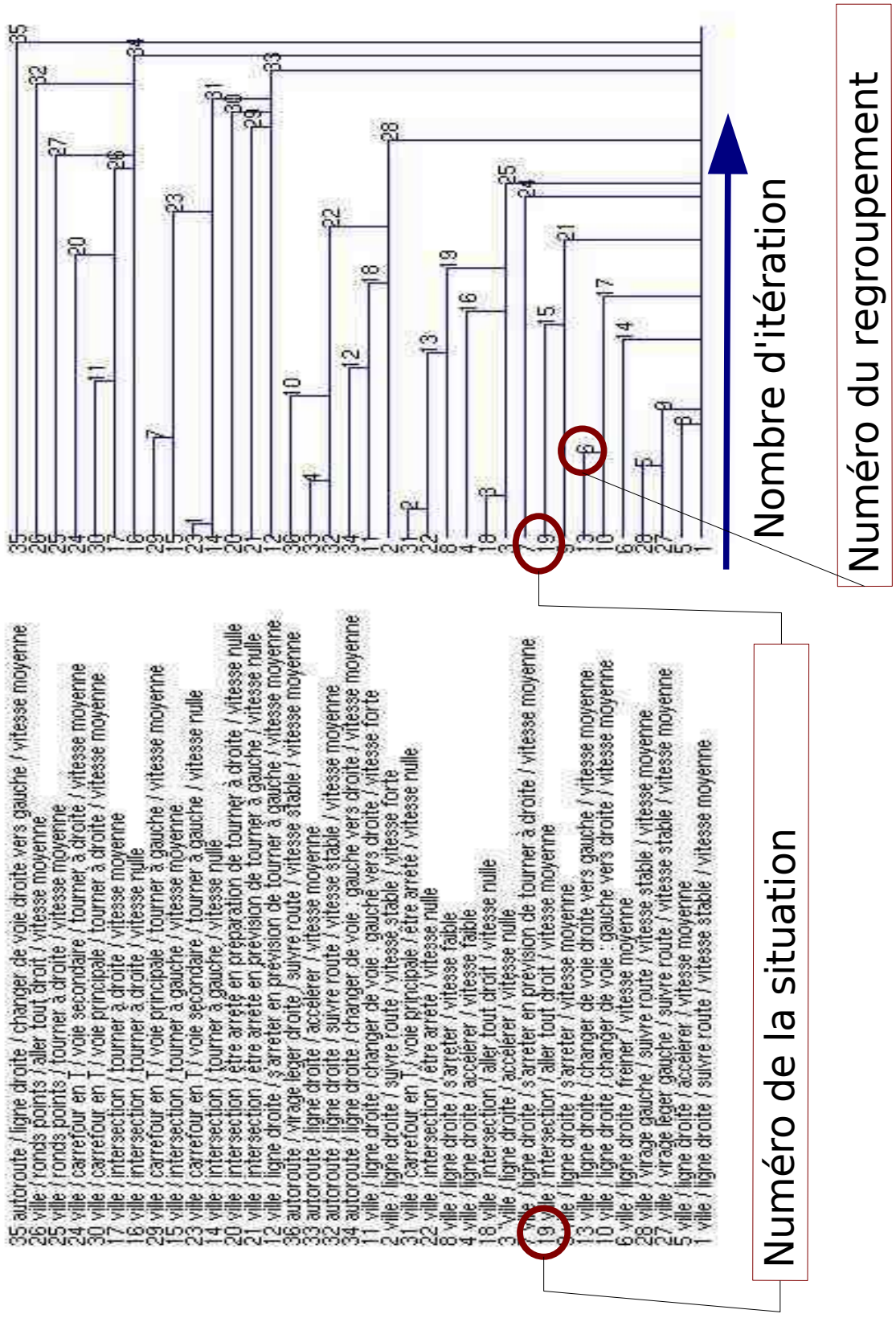


Figure 3.30 : Arbre de regroupement des différents situations modélisées. Au début du processus, chaque situation correspond à un groupe. On a donc 36 groupes. Puis à chaque étape les 2 groupes de modèles ayant la distance la plus faible sont agrégés.

On peut noter alors trois sortes de regroupements :

- Ceux associant des situations où les comportements sont similaires. Ainsi, plusieurs situations *homogènes* suivant *l'objectif,* mais *différentes* suivant *l'infrastructure* sont regroupées. Des exemples de ce type de regroupement sont :
  - « Carrefour en T » et «Carrefour à 4 voies » lors d'un « Tourner à gauche » ou lors d'un « Tourner à droite ».

  - « Accélérer en ligne droite » et «Aller tout droit à une intersection » avec une vitesse initiale nulle.

  En effet, le comportement dans ces situations est similaire du point de vue des actions de conduite. Il pourrait seulement se distinguer par une étude des regards des conducteurs mis en relation avec la durée des différentes états composant les situations. Par ailleurs, ces regroupements pourraient indiquer que la classification première des situations vécues n'était pas optimale. Des situations préalablement définies comme différentes seraient peut-être à regrouper. Faire verbaliser les conducteurs sur leurs comportements dans ces situations pourrait permettre de résoudre cette ambiguïté.

  Par ailleurs, si on s'intéresse uniquement au diagnostic de la situation et non au similiarité de comportement, l'utilisation de donnée cartographique permettrait de lever ces confusions.

- Ceux associant des situations où les stratégies menées par le conducteur peuvent se recouper.
  Ainsi, « Autoroute / Aller à vitesse stable » et « Autoroute / Accélérer » sont regroupées. Ceci peut s'expliquer par le fait que pour rester à vitesse stable, le conducteur peut passer par des phases d'accélération et de freinage.

- Ceux associant des situations où les technologies utilisées ne sont pas assez efficientes pour les discriminer complètement.

  Ainsi, les situations de changement de voies sont regroupées entre elles puis avec la ligne droite. En effet ces situations sont hétérogènes au niveau des stratégies pouvant être choisies par le conducteur. Parfois dans une ligne droite légèrement incurvée, le comportement du conducteur peut être le même que lors d'un changement de voie « lent ». Aussi, il est difficile de les différencier sans utiliser en plus un capteur de positionnement latéral.

Le cas des situations où l'infrastructure est le rond-point (lorsque l'objectif est d'aller soit tout droit ou soit à droite) est un cas particulier. Ces situations sont associées lors du 27ème et 32ème regroupement, aux groupes comprenant les situations où le conducteur tourne à droite (dans un carrefour en T ou à 4 voix).

Ceci peut s'expliquer par le fait que le comportement du conducteur est similaire lors de la première partie d'un « rond-point », à celui lors d'un « Tourner à droite ». Cependant, cette ressemblance n'est que partielle : l'infrastructure l'oblige à suivre une trajectoire différente de celle d'un « Tourner à droite lors d'une intersection ». Cette dissimilarité entre les comportements explique pourquoi ce regroupement ne se fait que tardivement.

Ce processus de regroupement nous permet dès lors de définir des groupes de situations proches. C'est sur ces groupes que la validité de la modélisation va être testée. Ceci se fera lors de la prochaine section.

3.4 Résultats

Pour tester la validité de notre base de modèles, nous déterminons pour chaque modèle la probabilité qu'il ait engendré chacune des séquences de conduite non utilisées dans l'apprentissage (soit 592 séquences). Ainsi, nous calculons, pour tous les modèles d, $d \in [1:N]$, et pour chaque séquence $SC_{d,q}$ $q \in [1:N_d]$, la probabilité $\mathrm{p}(SC_{d,q}|M_d)$.

Une séquence appartenant à d_0 est dite reconnue si la situation la plus probable est la situation d_0 : $\mathrm{p}(SC_{d_0,q}|M_d) = \max_{d \in [1:D]} \mathrm{p}(SC_{d,q}|M_d)$ .Dans le cas d'un regroupement de situations, une situation est dite reconnue si la situation la plus probable appartient au groupe contenant la situation d_0 .

Dans tous les cas, il est aisé de calculer, en utilisant l'algorithme Forward-Backward pour les MSMCP (2.1.3.b.2), la probabilité que la séquence soit reconnue avant t_0 $\forall t_0 < T_{d,q}$ (Figure 3.31), avec $T_{d,q}$ la durée de la séquence.

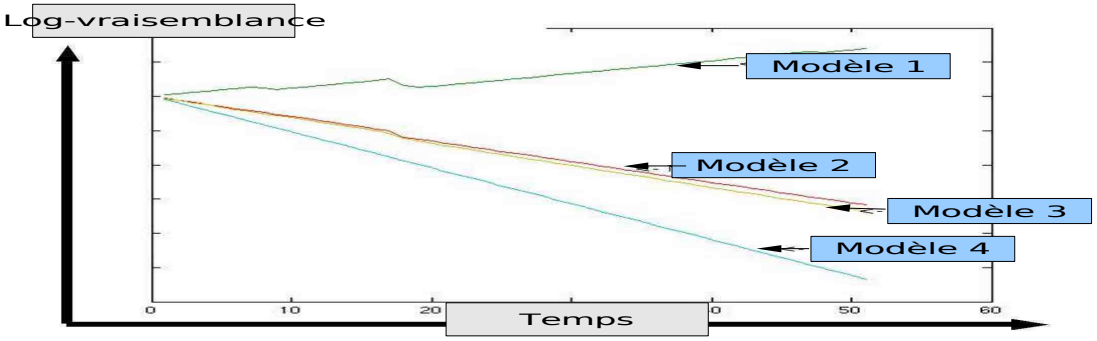


Figure 3.31 : Evolution des probabilités d'appartenance d'une séquence à différents modèles en fonction du temps. Le modèle 1 est ici celui qui a la plus forte probabilité d'avoir généré la séquence.

Nous utilisons cette possibilité pour tester notre modélisation suivant deux critères : la reconnaissance a posteriori et « en ligne ».

3.4.1. Reconnaissance a posteriori

Le premier critère, dit de « reconnaissance a posteriori » est calculé sur la séquence complète.

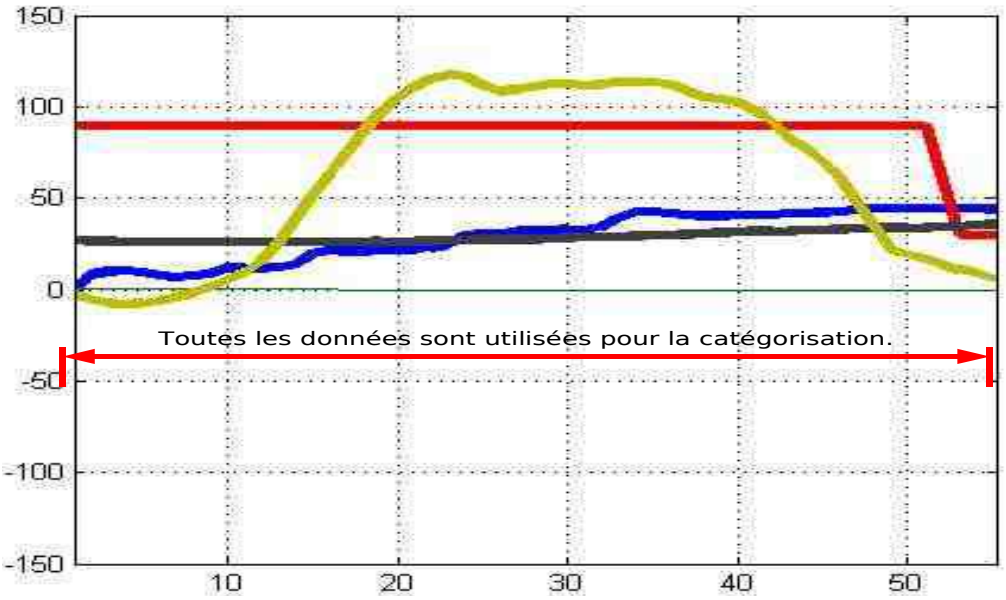


illustration3.32 : Pour le calcul de la reconnaissance a posteriori, toutes les données de chaque séquence sont utilisées pour les catégoriser.

Ainsi, pour chaque séquence, est calculée, à partir de l'ensemble des données la composant (illustration 3.32), la probabilité d'appartenance à chaque modèle. Puis nous comparons la situation la plus probable avec la vraie situation. Cela nous permet alors de mesurer la capacité de la modélisation à discerner les différences entre les situations. Cette comparaison est effectuée pour toutes les séquences n'ayant pas servi à l'apprentissage (soit 592 séquences).

Nous donnons ici le graphique représentant le taux de reconnaissance en fonction du nombre de regroupements effectués. Pour éviter qu'une situation ayant un grand nombre de séquences ne pondère trop les résultats, ce taux est calculé de 2 manières :

- la première est le taux effectif, c'est à dire simplement le nombre de séquences reconnues dans la bonne catégorie par rapport aux nombres de séquences totales
- la deuxième est le taux normé, c'est à dire la moyenne des taux de reconnaissance pour chaque catégorie.

Les résultats sont donnés dans le graphique 3.33.

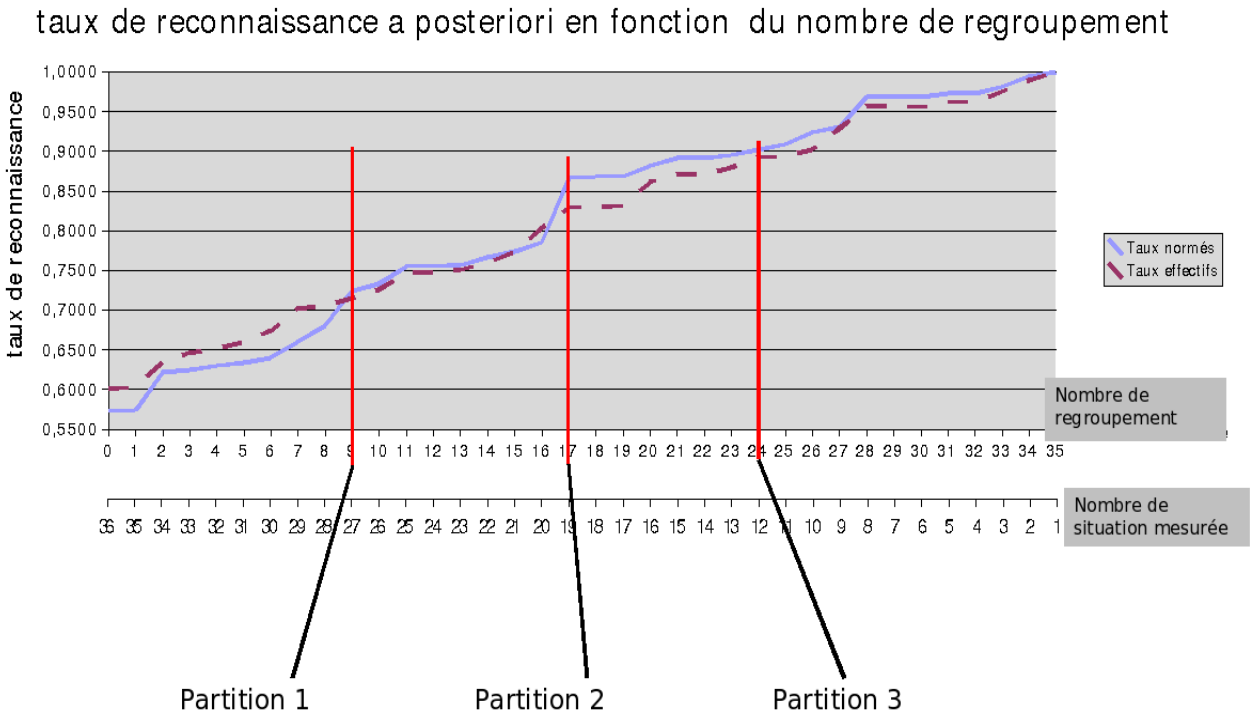


Figure 3.33 : Taux de reconnaissance a posteriori en fonction du nombre de regroupement.

Remarquons tout d'abord que les 2 courbes sont proches entre le 10ième et le 17ième regroupement. A ce moment là, les erreurs sont donc réparties uniformément entre chaque situation. Avant le 10ième regroupement, des situations peu reconnues (comme les situations de changement de voies) ont un faible taux de reconnaissance et peu de séquences. Le fait qu'elle soit composée de peu de séquences baisse le taux normé par rapport au taux effectif. Après le 17ième regroupement, ces situations sont regroupées et sont reconnues. Le taux normé passe alors au dessus du taux effectif.

Au niveau du 10ième et du 18ième regroupements, on assiste une rupture dans l'évolution du taux de reconnaissance. Aussi, on considère les partitions formées à ces deux niveaux :

1. La première que nous nommerons « **partition étendue** », est donc l'ensemble des catégories formées avec le dixième regroupement. Elle est formée de 26 « situations mesurées ». Le taux de reconnaissance de ces situations est de 75,35%.

2. La deuxième, que nous nommerons « **partition intermédiaire**», est l'ensemble des catégories formées avec le dix-huitième regroupement. Elle est formée de 18 « situations mesurées ». Le taux de reconnaissance de ces situations est alors de 87,88%.

L'étude d'une autre partition nous semble importante. Il s'agit de la première partition formée pour laquelle le taux de reconnaissance dépasse 90%. La Figure 3.33 montre que cette partition est créée après le 24ième regroupement.

3. Cette troisième partition, que nous nommerons « **partition réduite**», est donc l'ensemble des catégories formées avec le vingt-quatrième regroupement. Elle est formée de 12 « situations mesurées ». Le taux de reconnaissance de ces situations est alors de 90,53%.

Pour valider la méthodologie choisie, nous avons comparé les résultats de notre modélisation avec d'autres types de modélisation : sans les poids (Figure 3.34), sans les contraintes temporelles (Figure 3.35) et sans l'utilisation d'une méthodologie spécifique d'apprentissage (Figure 3.36).

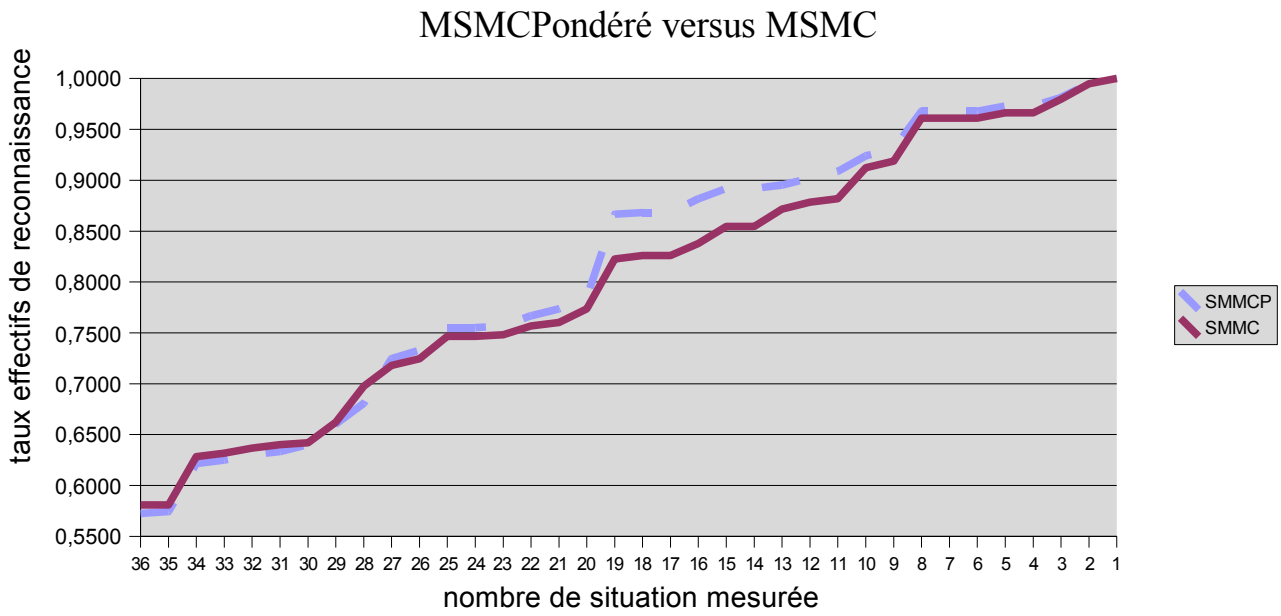
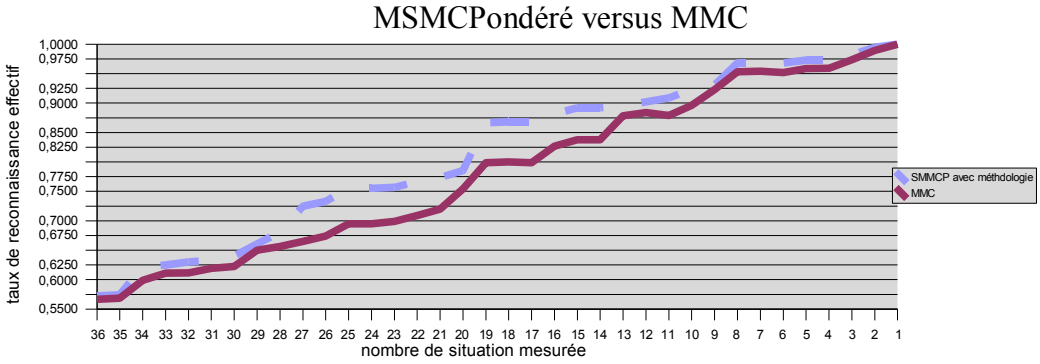


Figure 3.34 : Performance d'une modélisation avec / sans le concept de pondération

Le taux de reconnaissance basé sur notre modélisation est ici supérieur ou égal à celui d'une modélisation sans la notion de poids sur les variables (Figure 3.34).

Figure 3.35 : Performance d'une modélisation avec / sans le contraintes temporelles



Le taux de reconnaissance basé sur notre modélisation est ici supérieur en tout point à celui d'une modélisation sans la notion de contrainte sur la durée des états (Figure 3.35).

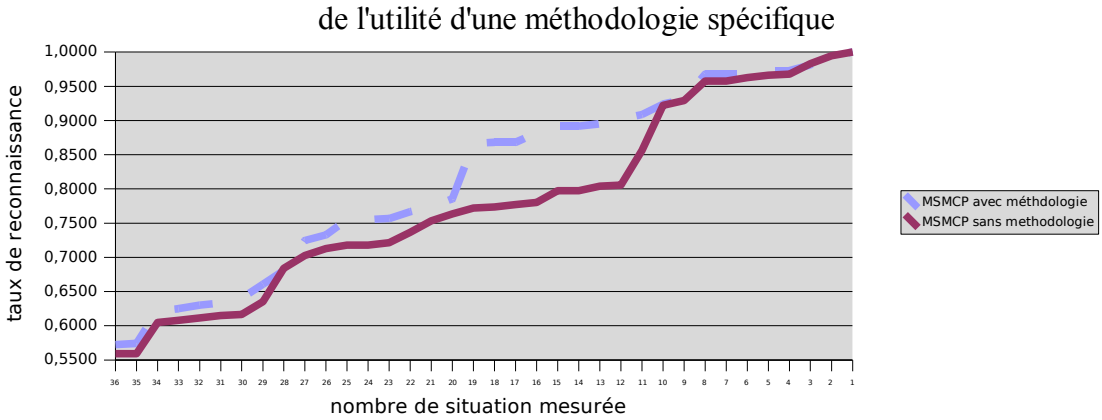


Figure 3.36 : Comparaison des taux de reconnaissance en utilisant et en n'utilisant pas une méthodologie spécifique

Le taux de reconnaissance en utilisant notre méthodologie est supérieur en tout point à celui utilisant une méthode d'apprentissage simple (sans validation par interprétabilité des modèles, sans ajustement manuel des paramètres, et sans apprentissage de la topologie).

Dans tous les cas, notre modélisation, basée à la fois sur des contraintes temporelles, des niveaux de pondération sur les variables, et sur une méthode spécifique d'apprentissage, produit des résultats supérieurs aux autres modélisations n'intégrant pas l'ensemble de ces caractéristiques (Figure 3.37).

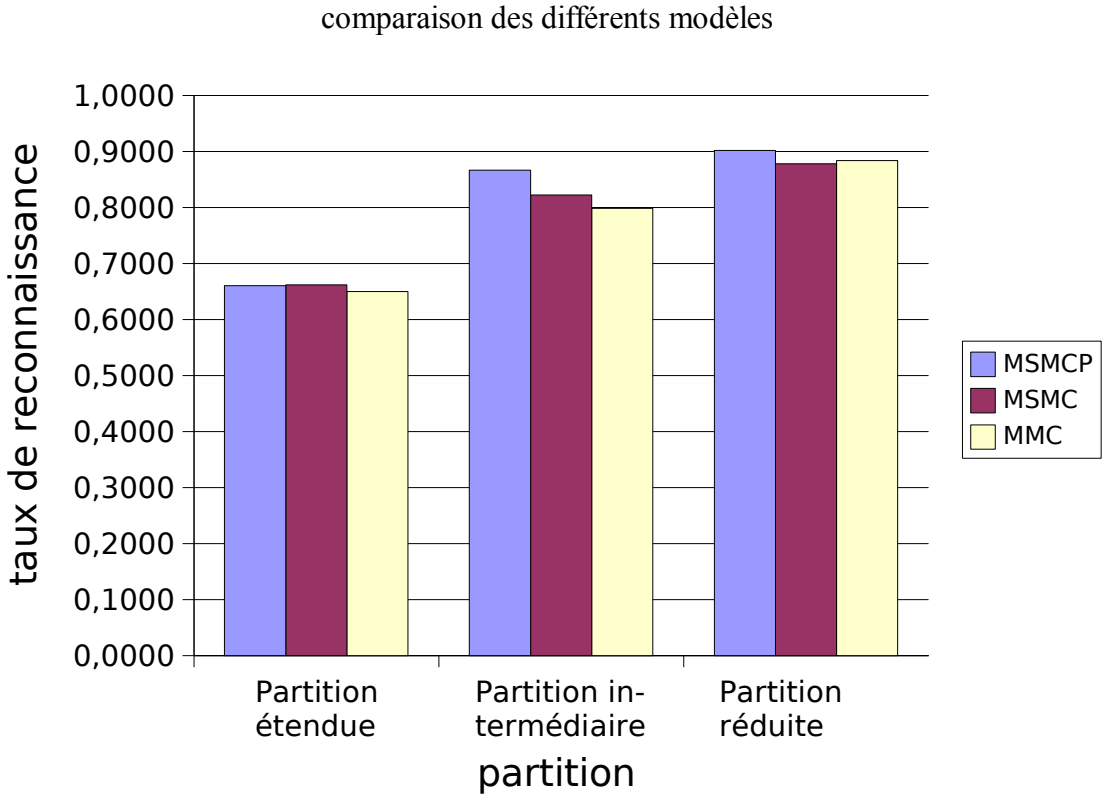


Figure 3.37 : Comparaison des taux de reconnaissance sur les 3 partitions (étendue, intermédiaire, et réduite) pour les différents modèles

La notion de pondération apparaît donc importante pour opérer un diagnostic sur les partitions intermédiaires et réduites. Pour la partition étendue, certaines situations sont très proches. Aussi, l'apprentissage a donné une matrice de poids quasiment identique pour ces situations et le concept de pondération ne peut permettre dans ce cas de les distinguer.

Par ailleurs, si le modèle MSMCP est plus performant pour 12 situations de conduite que les modèles de Markov caché « simple », c'est surtout pour 18 situations que la différence devient importante (9%).

Ce fait confirme notre choix en matière de modélisation. En effet, si les modèles de Markov simples peuvent permettre de classifier un nombre restreint de situations (comme dans les précédentes études), l'augmentation des situations étudiées engendre une complexité supplémentaire. L'intégration de caractéristiques spécifiques à la conduite (pondération, durée sur les états, processus de modélisation particulier) peut alors apporter une solution à ce problème.

Par la suite, nous testerons la capacité de la base de modèle à fournir un diagnostic précoce.

3.4.2. Reconnaissance en ligne

Le second critère dit de «reconnaissance en ligne » est calculé en utilisant seulement les t_0

premières secondes ($t_0 = 1,2$) de chaque séquence (illustration 3.38). Ce critère est utilisé pour tester la capacité du modèle à fournir un diagnostic précoce.

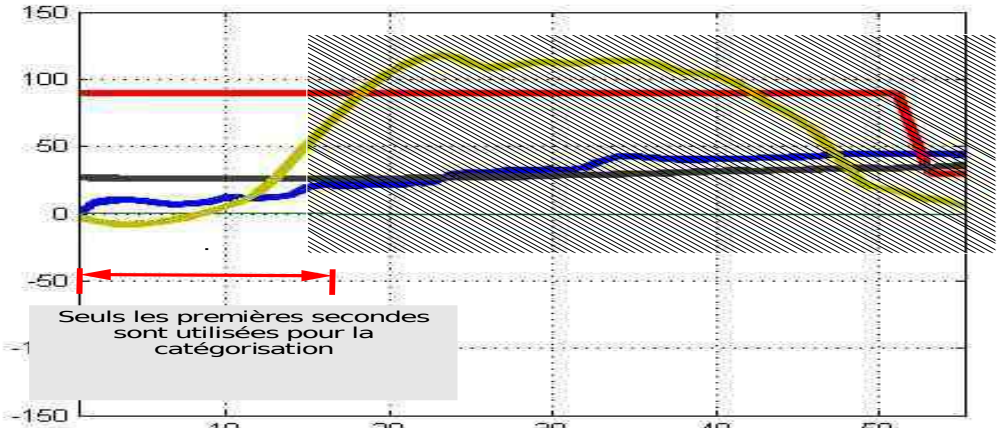


Illustration 3.38 : Reconnaissance en ligne : seules les t_0 premières secondes de chaque séquence sont utilisées pour le diagnostic.

Les taux de reconnaissance sont satisfaisants (75 % pour la partition intermédiaire, pour une reconnaissance en 1s). Le taux « en ligne », 2 secondes après le début de la situation, est proche des taux de reconnaissance a posteriori (Figures 3.39 et 3.40). Cela indique que la modélisation effectuée a pu, en partie, capter dès les premières secondes les différences entre situations.

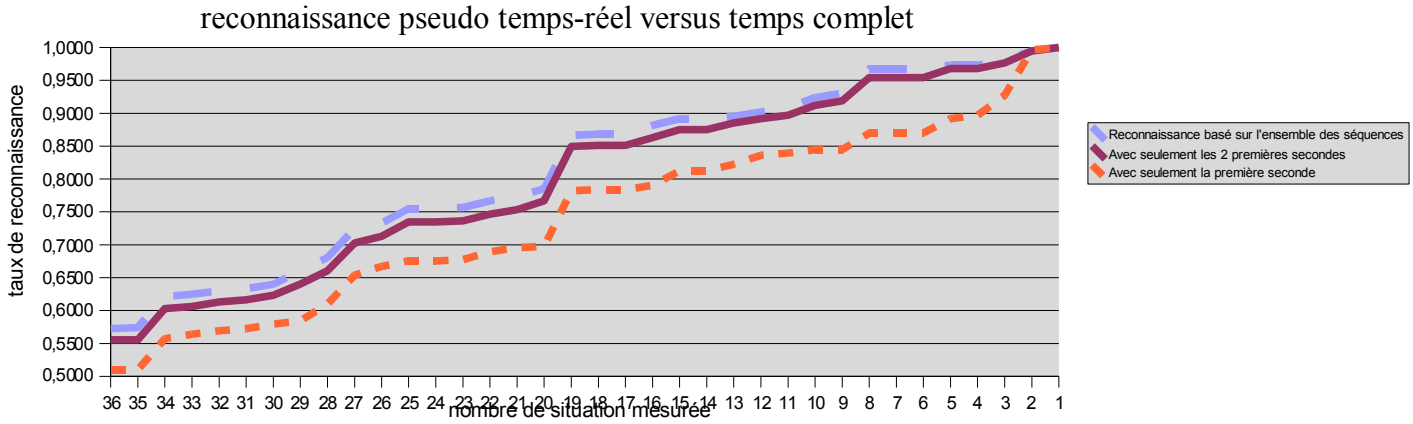


Figure 3.39 : Comparaison des taux de reconnaissance en fonction des regroupements pour une reconnaissance basée soit sur la première seconde, soit sur la deuxième seconde, soit sur la totalité de chaque séquence.

Pour une reconnaissance en une seconde, le taux est inférieur de 15% à celui a posteriori.

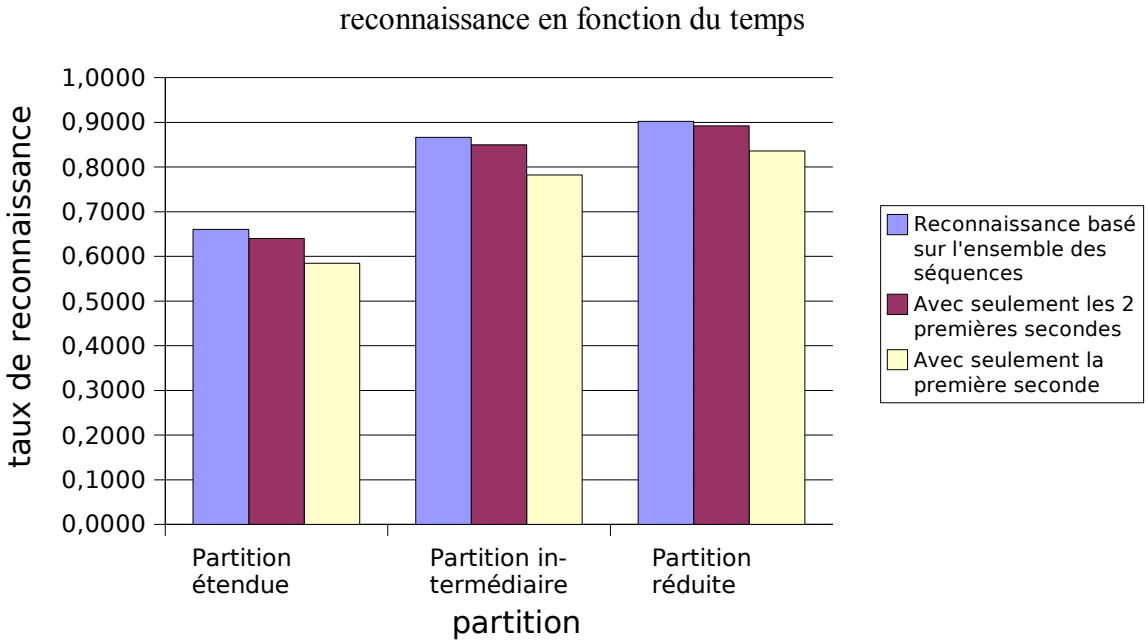


Figure 3.40 : Reconnaissance en ligne pour les 3 partitions

De plus, les confusions semblent moins dûes à un problème de modélisation qu'à une difficulté intrinsèque de discriminer certaines situations dès les premières secondes. En effet, pour certaines situations, les capteurs utilisés ne nous permettent pas de prédire au plus tôt. Pour d'autres situations, seule l'analyse de l'état final du véhicule permet de les classifier correctement.

Ainsi, les séquences où le conducteur tourne à une intersection sans mettre son clignotant, et celles où le conducteur va s'arrêter sont mal catégorisées dès la première seconde. Une étude oculométrique pourrait permettre d'augmenter les taux de reconnaissance.

L'une des voies possibles pour discriminer les comportements dans ces situations serait l'utilisation de données oculométriques.. L'amélioration des taux de reconnaissance pourraient aussi se faire en utilisant des données cartographiques.

Par ailleurs, avec des capteurs moins complexes, et un nombre de situations étudiées plus élevé, nous nous rapprochons des taux de reconnaissance des études précédentes [Pentland&Liu 1999] [Kumagai et al 2003].

3.4.3. Séquences non prototypiques

Nous avons testé la capacité prédictive de la base de modèle sur les séquences atypiques et singulières en fonction des différents regroupements. Les résultats sont donnés dans le graphique ci-dessous (Figure 3.41). On note que si le taux de reconnaissance est inférieur à celui des séquences prototypiques (49% de bonne reconnaissance pour une partition étendue), il reste correct, notamment pour la partition réduite avec près 80 % de bonne reconnaissance pour les situations non prototypiques en 2s.

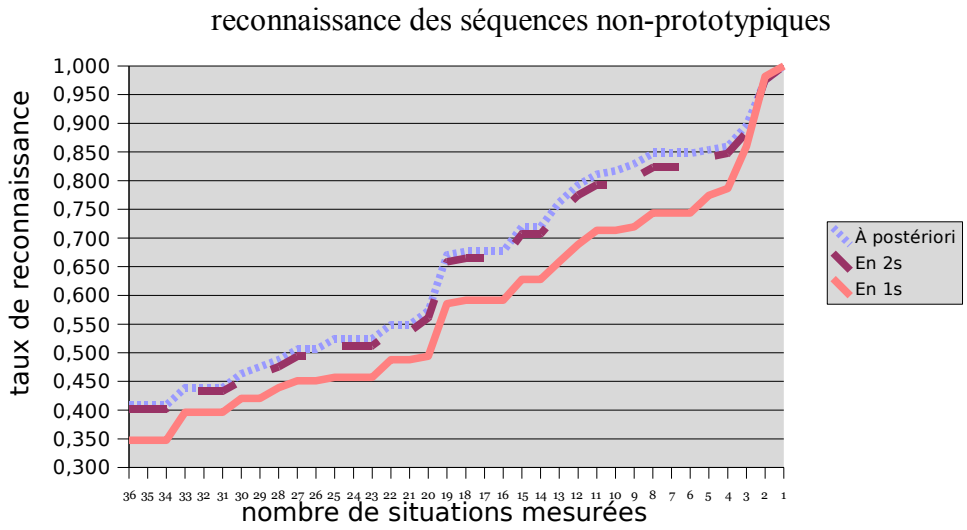


Figure 3.41 : Taux de reconnaissance des séquences non-prototypiques en fonction du nombre regroupement effectué.

On compare ensuite des taux de reconnaissance des situations atypiques et prototypiques a posteriori (Figure 3.42) et en ligne (Figure 3.43).

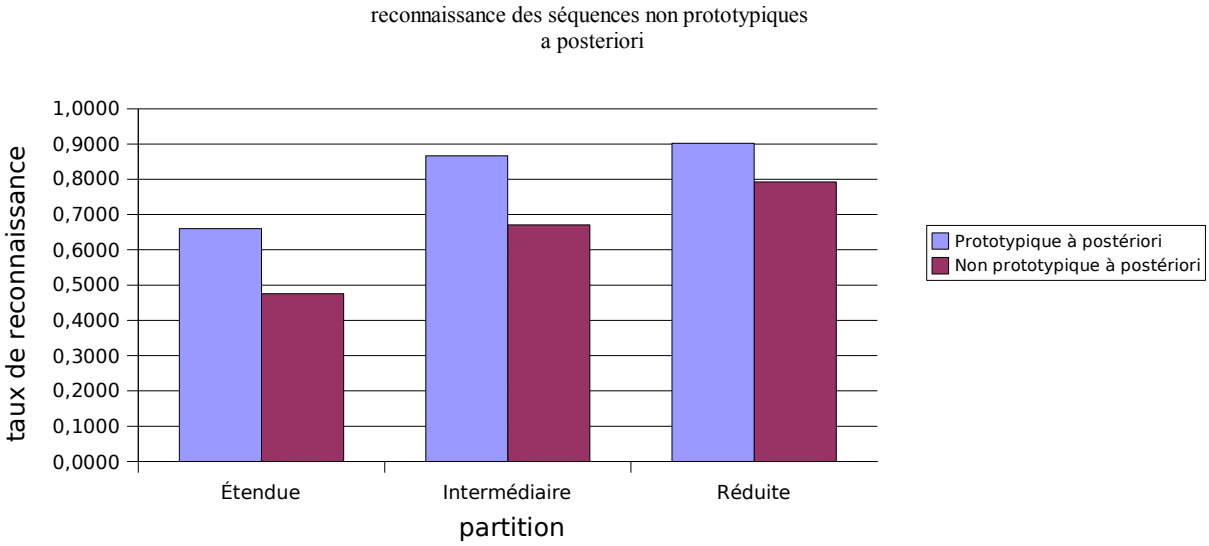


Figure 3.42 : Comparaison des taux de reconnaissance des séquences non-prototypiques et prototypiques a posteriori

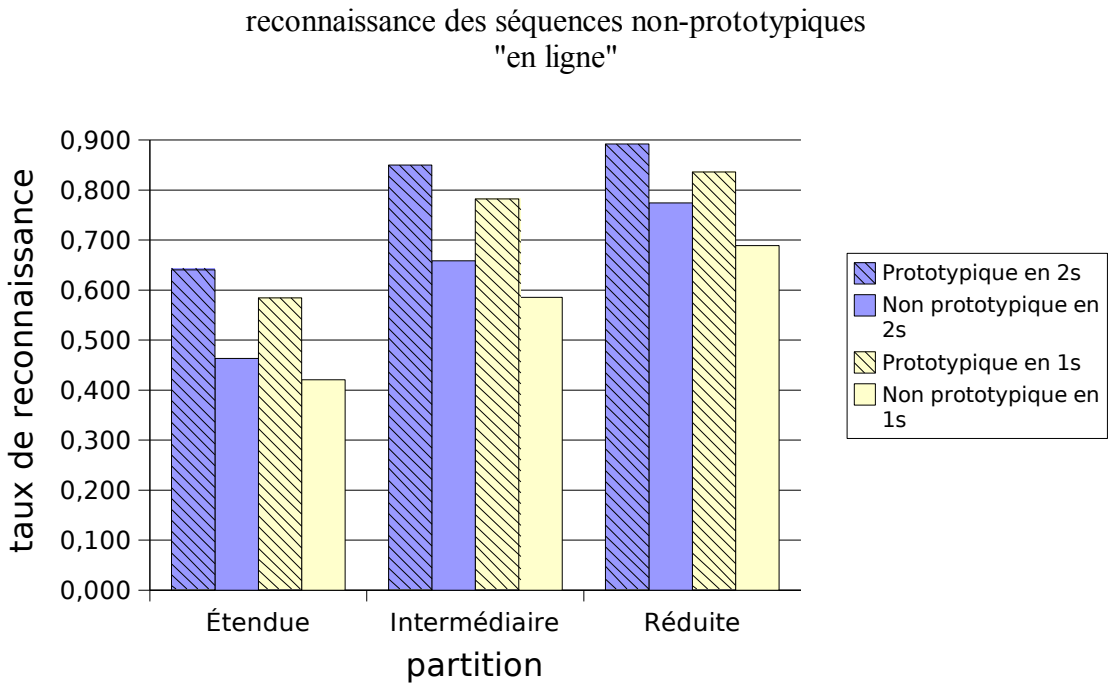


Figure 3.43 : Comparaison des taux de reconnaissance des séquences non-prototypiques et prototypiques "en ligne"

La différence de reconnaissance entre les situations prototypiques et les autres que ce soit en temps réel ou a posteriori est approximativement constante et de l'ordre de 15-20%. Ceci implique que, si pour une partie des séquences non protypiques l'activité de conduite reste la même que pour les séquences prototypiques, pour une autre partie le changement provoqué dans l'activité par les évènements extérieurs est trop important pour pouvoir être analysé avec le catalogue de modèles construits.

Pour affiner ces résultats, une étude supplémentaire devrait être menée pour comprendre si les confusions viennent :

- d'un déficit dans la modélisation (n'intégrant pas certaines stratégies),
- d'une ressemblance importante au niveau des capteurs entre différentes situations,
- de l'interprétation de l'activité du conducteur dans ces situations (l'objectif et l'environnement perçu, préalablement catégorisés sont ils justifiés ?).

Dans cette section, nous avons vu qu'au problème de déterminer à quelle situation appartient une séquence de données, les modèles de Markov cachés apporte une solution pertinente. Cependant, ce genre de modèle n'est pas adéquat pour prendre en compte les changements de situations.

Aussi, pour résoudre ce type de problème, nous verrons, dans la prochaine section, l'utilité des

modèles de ruptures multi-phasiques, modèles dont la partie théorique a été étudiée dans la partie 2.2

3.4.4. Retrouver une situation dans un flux continu de données par l'utilisation des modèles de ruptures multi-phasiques.

a Problématique

Généralement, lors d'expérimentations, des flux importants de données sont recueillis.

Puis, ces données sont segmentées et étiquetées par l'analyste, mais, en fonction de l'expérimentation, ce travail peut prendre un temps considérable. Aussi, le pré-traitement de ces données, en indiquant automatiquement *les moments* où le conducteur a effectué telle action ou rencontré telle situation particulière, serait un apport important pour toute étude sur l'activité de conduite.

Or, les modèles construits dans la partie précédente nous permettent de catégoriser des séquences *préalablement découpées dans un ensemble de données*.

Aussi, pour constituer un module de séquençage et d'étiquetage automatique, nous proposons de définir une fenêtre d'analyse du flux de données, de la faire progresser du début à la fin de l'expérimentation et d'étudier l'évolution de la vraisemblance des modèles sur les données délimitées par cette fenêtre.

On suppose alors que la vraisemblance des modèles augmentera lorsque la fenêtre se rapprochera d'une situation préalablement modélisée, sera stable lorsque la fenêtre sera proche de la situation, puis diminuera lorsqu'elle s'éloignera de la situation.

L'étude des ruptures dans l'évolution de la vraisemblance semble, dès lors, une approche intéressante.

Or, lors de la partie 2.2, nous avons montré que, pour les modèles linéaires et non linéaires de rupture, les estimateurs par maximum de vraisemblance étaient consistants.

Aussi, il nous semble possible d'appliquer ces résultats à l'étude de l'évolution de la vraisemblance d'un modèle associé à une situation sur les flux de données. Cela nous permettra de trouver des ruptures dans son évolution et les paramètres caractérisant les phases formées.

Les règles définies précédemment rendront possible l'évaluation des moments où les situations de conduite recherchées se sont réalisées (illustration 3.23).

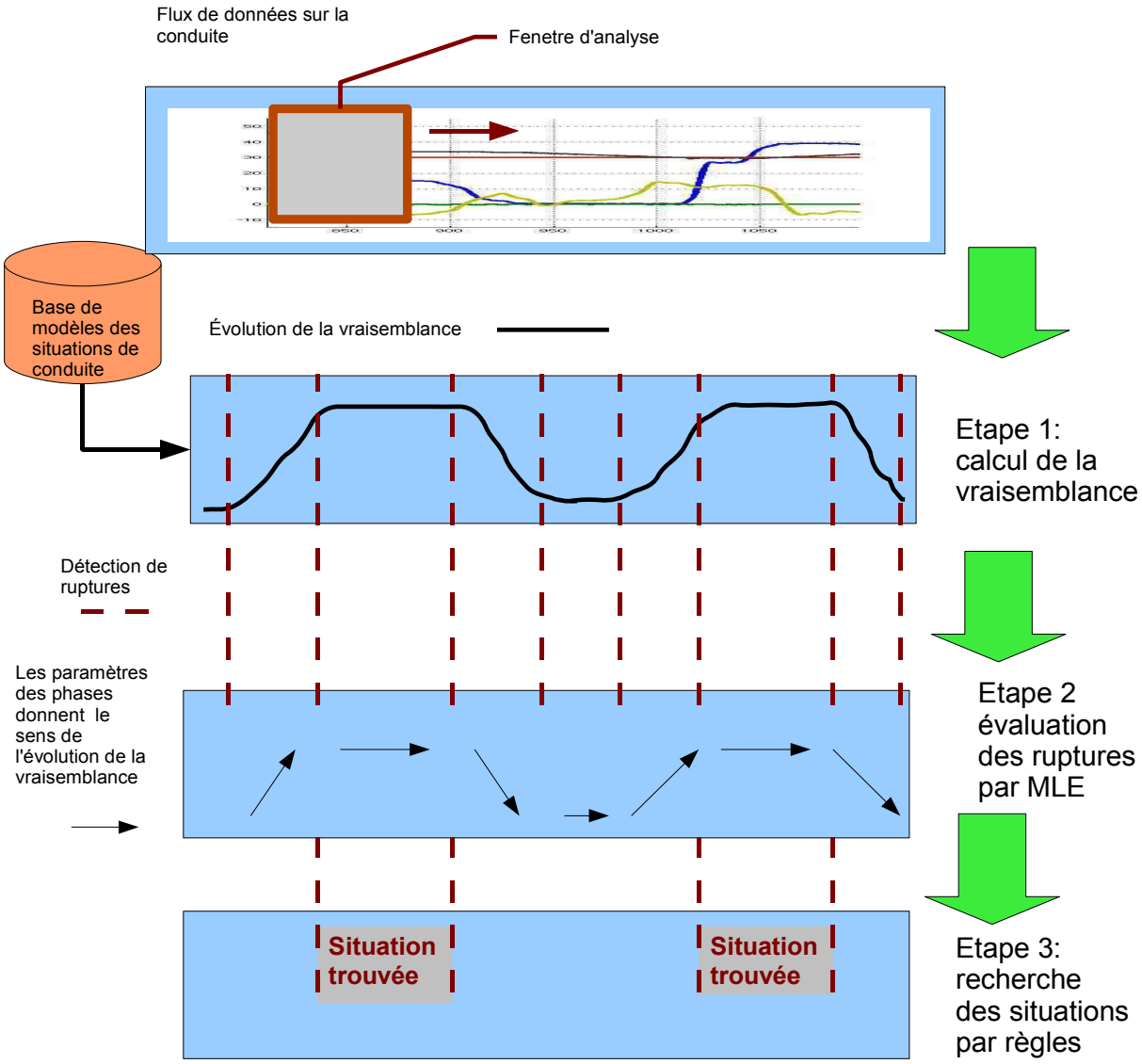


Illustration 3.44 : Processus d'analyse des flux de données par analyse des ruptures de la vraisemblance des modèles

b Exemple

Pendant un laps de temps particulier, on suppose qu'a eu lieu D fois une situation de conduite particulière (tourner à gauche, doubler, ou « conduire en ligne droite à vitesse stable »). On recherche où ont eu lieu les différentes réalisations de cette situation.

Pour cela, nous avons choisi d'étudier l'évolution de la vraisemblance des données par rapport au modèle M_0 représentant la situation recherchée, et d'analyser les ruptures dans son évolution.

Ainsi, on définit une fenêtre de longueur « 2 l » [t-l,...t+l] et on étudie la fonction $P(y_{t-l},...y_{t+l}|M_0)=f_\theta(t)$, avec M le modèle représentant la situation recherchée.

On suppose alors que $f_\theta(t)$ est une fonction multi-phasique linéaire c'est-à-dire $f_\theta(t)=\sum_{k=0}^{K} a_k+b_k*t\ 1_{(r_k<t\le r_{k+1})}$ avec $K \gg D$.

Grâce aux résultats obtenus dans la partie 2 sur les propriétés des estimateurs $\{\hat{a}_k,\hat{b}_k,\hat{r}_k,k\in[1:K]\}$, nous pouvons utiliser la méthode définie en 2.2.1.b pour estimer les paramètres du modèle.

Enfin, pour analyser ces paramètres, nous faisons les 3 hypothèses suivantes :

- si $b_k>0$, la probabilité $P(y_{t-l},...y_{t+l}|M_0)$ augmente donc la fenêtre se rapproche de la zone où la situation a été réalisée.
- si $b_k<0$, la probabilité $P(y_{t-l},...y_{t+l}|M_0)$ diminue, donc la fenêtre s'éloigne de la zone où la situation a été réalisée.
- Si $b_k\approx 0$, la probabilité $P(y_{t-l},...y_{t+l}|M_0)$ est constante la fenêtre doit être proche de la zone où la situation a été réalisée.

Aussi s'il existe k tel que $b_{k-1}>0$, $b_k\approx 0$, $b_{k+1}<0$, on peut affirmer qu'entre $r_k et r_{k+1}$ la situation s'est réalisée.

Les premières simulations montrent que cette méthode donne des résultats corrects.

La Figure 3.45 montre une séquence de conduite où le conducteur a évité un obstacle par la droite puis « a roulé à vitesse stable sur une ligne droite » puis a « tourné à gauche à une intersection ». La log-vraisemblance est calculée en fonction du modèle, M, associé à la situation *« ligne droite / aller tout droit / conduire à vitesse stable »,* et la taille de la fenêtre « 2.l » est pris égale à deux secondes.

La log-vraisemblance augmente bien quand la fenêtre s'approche de cette situation, puis est stable lors de cette dernière et enfin diminue.

Les ruptures trouvées dans l'évolution de la vraisemblance découpent bien la séquence de conduite en plusieurs parties. Les règles définies ci-dessus permettent de situer correctement la situation « rouler à vitesse stable sur une ligne droite » dans l'ensemble de données.

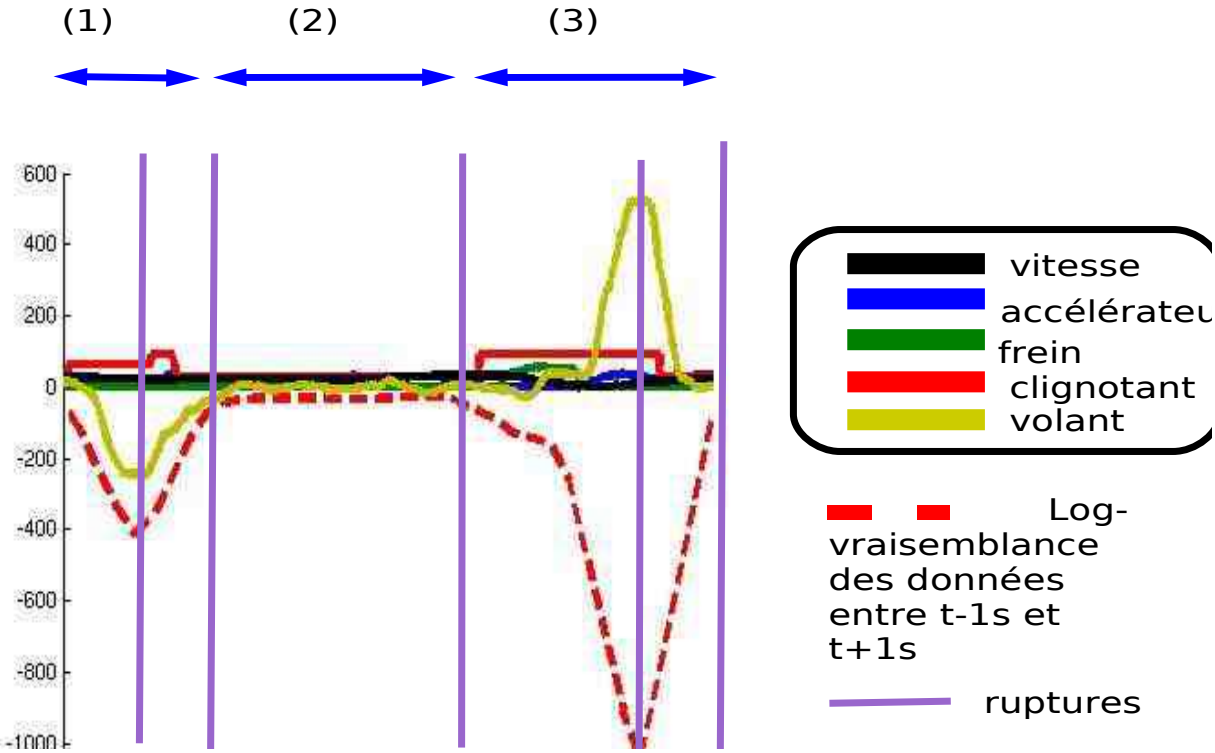


Illustration 3.45 : Evolution de la log vraisemblance du modèle « conduire en ligne droite à vitesse moyenne » lors d'une séquence de conduite comprenant 3 situations : (1) « éviter un obstacle », (2) « conduire en ligne droite à vitesse moyenne » et (3) « tourner à gauche à une intersection ».

Pour pouvoir être validé et pour pouvoir trouver la largeur de la fenêtre la plus adéquate, une étude de plus grande ampleur serait nécessaire. Cependant, elle implique un nombre de vérifications importantes et n'a donc pas pu être réalisée lors de cette thèse.

Deux problèmes seraient alors à résoudre.

D'une part, le temps de calcul de la log-vraisemblance à chaque pas de la fenêtre est important et rend difficile une utilisation simple de ce type de méthode. Cependant, l'évaluation de la vraisemblance sur des segments éloignés est indépendants. Aussi, la parrallélisation des calculs pourrait être une voie possible pour améliorer la rapidité de la méthode.

D'autre part, si on recherche plusieurs situations simultanément, des conflits peuvent apparaître. A un temps donné, deux situations peuvent être reconnues. Dans ce cas, la comparaison des vraisemblance et l'établissement de règles expertes (du type « si l'état 5 n'est pas atteint la situation M n'a pu avoir lieu ») pourraient permettre d'éliminer les confusions.

3.4.5. Synthèse des résultats

L'expérimentation effectuée a permis de recueillir 1209 séquences de conduites. Sur la totalité de ces séquences, 718 séquences prototypiques et 160 non-prototypiques correspondaient aux 36 situations étudiées (paragraphe 3.3.3). Les autres situations comportaient trop peu de séquences pour pouvoir être étudié.

A l'aide du processus de modélisation développé, les 36 situations ont été modélisées par des MSMCP en utilisant 126 des séquences prototypiques.

Nous avons alors vu que les apports de cette thèse se situent à différents niveaux:

1/ Interprétabilité possible des modèles (paragraphe 3.3.4) .

De part les contraintes fixées dans le processus de modélisation, les 36 modèles construits peuvent être interprétables. Par la suite, la mise en correspondance de l'interprétation des états et de connaissance sur l'activité de conduite pourra enrichir la connaissance sur le comportement du conducteur.

2/ Proximité entre situations (paragraphe 3.3.5)

Grâce à la méthode de classification développée, nous avons défini des proximités entre situations et entre groupes de situations. Nous avons alors déterminé une hiérarchie de partition de de situation de conduite.

Ces partitions font sens et les proximités peuvent être interprétés comme une proximité en terme comportementale ou comme un déficit technologique.

3/ Des taux de reconnaissance correctes (paragraphes 3.41 et 3.4.2)

Nous avons définis 3 partitions, étendue (comprenant 24 groupes de situation), intermédiaire (comprenant 18 groupes de situation) et réduite (comprenant 12 groupes de situation).

Pour ces 3 partions, les taux de reconnaissance a posteriori, c'est-à-dire avec l'ensemble des données de chaque séquence, sont correctes (respectivement 75%, 87% et 90%).

Les taux « en ligne », 2 secondes après le début de la séquence, sont quasiment équivalent au taux à posteriori. alors que les taux « en ligne », 1 seconde après le début de la séquence sont inférieur de 15% aux taux a posteriori.

Ces taux furent calculé sur les 592 séquences prototypiques n'ayant pas servi à l'apprentissage.

Avec un nombre de situations plus élevé et des capteurs plus basiques, les taux de reconnaissance se rapprochent des études précédentes. Par exemple, Pentland et al s'interresaient respectivement à 6 situations et obtenaient un taux de reconnaissance de 95% et Kumagai et al *** s'intéressaient à 3 situations et obtenaient un taux de 98,3 %.

Ces chiffres témoignent que la modélisation a pu capter, et ce dès les premiers instants, les caractéristiques propres de chaque situations.

4/ Validité du modèle et du processus de modélisation choisis(paragraphe 3.4.1) .

Pour analyser l'activité de conduite nous avions choisi de développer le modèle semi-markovien caché pondéré (MSMCP) tenant compte d'une part des contraintes temporelles inhérente à l'activité de conduite et intégrant la notion de pondération. De plus nous avions opté pour un processus de modélisation spécifique basé sur la validation des modèles par expertise et sur un

apprentissage spécifique de la topologie.

Les taux de reconnaissance en utilisant ce modèle et ce processus sont supérieur à ceux n'intégrant pas l'ensemble de ces contraintes.

Par exemple, en utilisant un modèle de Markov caché simple pour modéliser les situations de conduite, les taux de reconnaissance chutent en moyenne de 6%.

Dans tous les cas, le modèle de Markov caché semble une solution adéquate pour modéliser l'activité de conduite.

5/ Reconnaissance des séquences non-prototypiques (paragraphe 3.4.3)

Le taux de reconnaissance a posteriori ou « en ligne » sur les 180 séquences non-prototypiques est inférieur de 15% à celui des séquences prototypiques.

Si pour une partition étendue, le taux de reconnaissance de 49% ne permet pas d'envisager une utilisation efficiente. Pour une partition réduite, il atteint les 75% et rend possible le diagnostic des séquences même complexe.

6/ Possiblité de retrouver une situation dans un flux continu de données (paragraphe 3.4.4)

Lors de la partie 2.2, nous avons étudier les estimateurs des modèles de ruptures multi-phasiques. L'utilisation hybride de ces modèles avec les modèles de Markov cachés associés aux différentes situations de conduite nous permet dès lors d'envisager la segmentation et la labellisation automatique de données portant sur la conduite.

Les premiers résultats sont prometteurs. Pourtant, afin de valider la méthode proposée, une étude de plus grande ampleur devra être menée.

L'ensemble de ces résultats, basé sur une importante expérimentation, valide les choix effectués tant au niveau des modèles développés qu'au niveau de la méthodologie adoptée.

Conclusion et perspectives

L'objectif de cette thèse était d'établir un cadre d'analyse des données recueillies sur les véhicules et de modéliser leur évolution pour de les mettre en correspondance avec des comportements de conduite.

Cet objectif a été atteint en établissant une méthodologie spécifique d'analyse du comportement. Cette méthodologie :

1. a été structurée par les connaissances sur la cognition du conducteur issues des sciences humaines,
2. s'est basée sur des résultats théoriques en mathématique,
3. a été organisée autour d'un processus de modélisation hybride mêlant apprentissage automatisé et validation experte.
4. a été validée par une expérimentation importante.

En effet, les connaissances sur la cognition du conducteur issues des sciences humaines nous ont permis de caractériser l'activité de conduite comme une succession de situations de conduite définies principalement suivant l'homogénéité de l'objectif tactique du conducteur, de la situation initiale, de l'infrastructure telle qu'elle est vécue par le conducteur, et des événements pouvant y survenir.

Par ailleurs, la mise en parallèle des sciences de l'ingénieur et des sciences humaines nous a permis de faire la distinction entre des situations de conduite réelles, des situations vécues, et des situations mesurées et de comprendre l'articulation de ces trois visions de l'activité. De plus, le découpage par phase de l'activité de conduite, illustré notamment dans COSMODRIVE, et les résultats de différentes études sur l'analyse des données de conduite, nous ont permis de choisir un modèle probabiliste adapté à l'étude du comportement du conducteur : le modèle de Markov caché.

Le choix de ce modèle nous a amené à développer deux résultats théoriques. D'une part, nous avons étendu le modèle de Markov caché au « modèle Semi-Markovien caché pondéré ». Ceci nous a permis d'intégrer des contraintes temporelles liées à l'activité de conduite, et la notion de pondération associée aux différentes variables. Nous avons alors présenté les différents algorithmes permettant l'utilisation de ce genre de modèle. D'autre part, le besoin de segmenter les données de

conduite, nous a amené à établir des résultats théoriques sur les estimateurs maximum de vraisemblance des paramètres des modèles de régression multi-phasiques.

Enfin, nous avons élaboré une méthode spécifique d'apprentissage des modèles mêlant :

1. un apprentissage automatisé des paramètres et de la topologie,
2. une validation et spécification des modèles par expertise humaine basée sur la labellisation des différentes états des modèles de Markov.

Cette double approche, reposant sur l'apport réciproque du formalisme MMC et sur les résultats des expertises humaines a alors amélioré l'adéquation des modèles aux situations rencontrées. Cette analyse constitue un apprentissage supervisé de l'activité de conduite.

Pour valider les méthodes choisies, cette recherche a nécessité un travail expérimental important, tant dans le recueil de données (1200 séquences de conduite ont été enregistrées et catégorisées), que dans la conception et la réalisation des outils de dépouillement et d'analyse.

Ces données ont permis de construire un catalogue de modèles de l'évolution des capteurs pour 36 situations de conduite.

Nous avons regroupé ces modèles en situations de conduites mesurées grâce à des méthodes issues de la classification hiérarchique. Ceci nous a permis de déterminer dans quelle situation, ou dans quel groupe de situations, se trouve le conducteur à partir d'un enregistrement inconnu.

Le taux de reconnaissance, qu'il soit calculé a posteriori ou « en ligne » (1 ou 2s après le début de la séquence) est satisfaisant (75% pour une partition étendue, 90% pour une partition réduite) et supérieur à ceux basés sur un autre type de modélisation.

Ceci confirme la pertinence de nos choix en matière de pondération des variables, de contrainte temporelle sur les états, et de nécessité d'une méthodologie spécifique d'apprentissage des modèles.

Les apports de cette thèse sont donc multiples. Par rapport aux précédentes études, nous avons élargi le spectre des situations étudiées de façon importante en étudiant simultanément 36 situations.

De plus, nous avons montré que :

- les chaînes de Markov cachées sont adéquates pour modéliser un grand nombre de situations de conduite.
- l'utilisation de ce modèle pour l'analyse de l'activité de conduite nécessite l'intégration de contraintes spécifiques pour être efficace. Nous avons défini théoriquement ces contraintes et mis en place les algorithmes nécessaires à leur utilisation.
- l'utilisation conjointe des méthodes d'apprentissage automatisées et d'une expertise humaine dans le cadre de la modélisation de l'activité de conduite est efficiente.

De plus, nous avons montré, entre autre, que les estimateurs des modèles de ruptures multi-phasiques étaient consistants. Ceci nous permet d'envisager l'utilisation hybride des modèles de Markov généré et des modèles de ruptures pour rechercher automatiquement des situations dans un ensemble de données.

Notre approche en terme de modélisation de l'activité de conduite permet d'ouvrir de nouvelles perspectives en terme de recherches.

Premièrement, au niveau de la segmentation d'un parcours de conduite, la méthode mise en place doit être approfondie. La taille de la fenêtre d'analyse devra varier en fonction des situations recherchées. Cette variation pourra faire l'objet d'études supplémentaires.

Deuxièmement, le dispositif méthodologique mis en place permet de façon aisée l'acquisition et l'analyse des comportements de conduite. La classification des situations de conduite et la labellisation des états permettent de disposer de paramètres plus synthétiques sur le comportement. Elles pourront ainsi être utilisés comme facteurs descriptifs dans les recherches étudiant la différence de comportement entre populations et/ou étudiant le comportement dans des situations à risques (Tourner à Gauche, Doubler).

Pour affiner ces analyses, les caractéristiques des conducteurs pourront être intégrées dans la conception des modèles, et les modèles associées à chaque type de conducteur comparées entre eux à l'aide d'une métrique adéquate.

En outre, l'ensemble des analyses sur l'activité de conduite pourront être enrichies par l'intégration de nouvelles données (oculométre, télémètre, GPS). Ces données étant échantillonnées différemment, les modèles développés devront être complexifiés pour pouvoir les intégrer.

Troisièmement, la base de modèles développée ainsi que la méthodologie mise en place pourront s'intégrer dans la conception de système d'assistance adaptatif (pouvant s'adapter au comportement du conducteur) au niveau des modules de diagnostic du comportement. Ceci pourra alors se faire à deux niveaux, d'une part, en proposant un outil de diagnostic via la base de modèles.

Pour améliorer l'efficacité prédictive des modèles, un apprentissage des paramètres et de la topologie basé sur le concept de discrimination (MMI, BEC) devra être développé. De plus, avec l'accroissement du nombre de situations étudiées, le temps de calcul nécessaire pour le diagnostic augmentera. Pour résoudre ce problème, des solutions algorithmiques devront être apportées.

Par ailleurs, l'utilisation des connaissance en sciences humaines sur certaines situations de conduite, nous permettra de prendre en compte les différentes stratégies pouvant être adoptées par le conducteur, pour une même situation de conduite. Ceci permettra d'augmenter les taux de reconnaissance et de mener des analyses à grandes échelles sur le comportement.

Quatrièmement, le processus de classification des situations proches développées permettra de s'interroger sur la similarité en terme comportementale entre certaines situations de conduite et sur les apports technologiques possibles pour les distinguer.

Cinquièmement, pour les besoins de notre étude, nous avons considéré les séquences de conduite comme indépendantes. Or, en pratique, il existe une dépendance entre des séquences de conduite successives. L'intégration de cette dépendance constituera une étape importante pour la compréhension de l'activité de conduite. Elle pourra se faire notamment en s'inspirant des modèles développés pour la reconnaissance de la parole (comme, par exemple, le modèle de Markov caché hiérarchique).

Bibliographie

Akaike H (1973) : *Information theory as an extension of the maximum likelihood principle* In Second International Symposium On Information Theory Akademia kiado, Budapest p267-281 *B.N. petrov and F. Csaki*

Bailly B (2004) : *Conscience de la situation : aspects fondamentaux, méthodes et applications pour la formation des conducteurs* .Thèse de Doctorat. Université Claude Bernard Lyon 2

Baum L & Eagon J (1967) : *An inequality with applications to statistical estimation for probabilistic functions of markov process and to a model of ecology*. bulletin AMS 73 p360-363

Baum L & Petrie T.(1966) : *Statistical inference for probabilistic functions of finite state markov chains*. The Annals of Mathematical Statistics 37 p1554-1563

Baum L. Petrie T., Souled G., and Weiss N.(1970) *A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains* Ann. Math. Statist., 41 p164-171

Bellet T (1998) : *Modélisation et simulation cognitive de l'opérateur humain: une application à la conduite automobile* .Thèse de Doctorat. Université paris V

Bellet T, Tattegrain-Veste H & Bonnard A (2004) : *Conception d'un module de gestion adaptative de la chm pour l'anticollision, et spécification des besoins en matière de technologies de perception* Rapport ARCOS R040122-4/LESCOT

Bellet T & Tattegrain-Veste H (2004) : *Cognitive engineering for adaptive anti-collision systems.* ITS in Europe Congress

Bengio Y & Frasconi P (1995) : *An input output HMM architecture* In Advances In Neural Information Processing Systems *The MIT Press Tesauro and Touretzky and T. Leen ed* p427-434

Bengio Y & Frasconi P (1996) : Input-output HMM's for sequence processing. *IEEE Transactions on Neural Networks* 7 p1231-1249

Bhattachary P (1994) : Some aspect of change-point analysis. *IMS lecture notes monograph series* 23 p28-56

Bonnard A, Deleurence P, Piechnik B, Goupil C, Chanut O & Hélène Tattegrain-Veste. H (2006) : *Conception et réalisation d'une plateforme de Mesure et d'Analyse pour l'eRGonomie de la cOnduite :MARGO* Rapport INRETS/LESCOT (à paraître)

Bouchard G Celeux Gilles (2004) *Model sélection in supervised classification.* Rapport de Recherche n5391 INRIA

Brand M, Nuria N. &. Pentland A. (1996) : *Coupled hidden markov models for complex action recognition* In Proceedings Of IEEE Conference On Computer Vision And Pattern Recognition Puerto Rico

Bruyas, M-P, Chapon A, Lelekov-Boissard T, Letisserand, D ; Duraz M ; Aillerie I (2005): *Evaluation de l impact de communications vocales sur la conduite automobile* Recherche Transport et Securité 5

Cacciabue P & Hollnagel E (2005) : *Modelling driving performance: a review of criteria, variables and parameters* In Proceedings Of The International Workshop On Modelling Driver Behaviour In Automative Environments. p185-197

Carlstein E (1988) : *Non parametric estimation of a change-point.* ann. statist. 16 188-197

Carsten O (2005) : *From driver models to modelling the driver what do we keally need to know about the driver ?* In Proceedings Of The International Workshop On Modelling Driver Behaviour In Automative Environments.

Ciuperca G (2004) : Maximum likelihood estimator in a two-phase non-linear regression model. statistic and decision 22(4) p335-349

Csorgo M & Horvath L (1997) : *Limit Theorems In Change-Point Analysis*. wiley Ed.

Dapzol N (2003) : *Rapport de stage : méthodologie d'analyse des situations de conduites* INRETS

Dempster A, Laird N & RUbin D. (1977) : *Maximum likelihood from incomplete data using the em algorithm*. journal of the royal statistical society 39B

Durand J (2003) : *Modèles à structure cachée: inférence, sélection de modèles et applications*. Thèse de Doctorat. Université Joseph Fournier

Etienne V & Marin-Lamellet C (2005) : *Déficits attentionnels dans la maladie d'alzheimer et conduite automobile* In 8ième Réunion Francophone Sur La Maladie D'Alzheimer Et Les Syndromes Apparentés.

Eubank RL & Speckman PL (1994) : *Nonparametric Estimation Of Functions With Jump Discontinuities, Change-Point Problem*. Muller H.G., Carlstein E., Siegmun D. Eds. Hayward

Facer M & Muller H.G. (2003) : *Nonparametric estimation of the location of a maximum in a response surface*. Multivariate Anal 1 p191-217.

Feder P (1975) : *On asymptotic distribution theory in segmented regression problems-identified case.* Ann. Statist. 3 p49-83.

Feder P (1975) : *The log likelihood ratio in segmented regression*. Ann. Statist. 3 p84-97

Fine Y, Singer Y. &. Tishby. N. (1998) : *The hierarchical hidden markov model: analysis and applications.* Machine Learning 32 p41-62

Fleury D, Dubois D. & Morvant C. (1993) : *Représentations catégorielles; perception et/ou action ? Contribution à partir d'une analyse des situations routirères* In Représentations pour l'action, Toulouse, Octares Editions, p77-93

Forbes, Huang, Kanazawa & Russel (1995) : T*he BATMOBILE: towards a bayesian automated taxi. IJCAI* pp1878-1885

Ge X & Smyth P (2001) : *Segmental semi-markov models for endpoint detection in plasma etching*. IEEE Transactions on Semiconductor Engineering

Georgeon O, Mille A, Bellet, T.(2006) *Musette-Abstract: un outil et une méthodologie pour analyser une activité humaine médiée par un artefact technique complexe* Ingénierie des connaissances; (à paraître)

Gill R (2004) : *Maximum likelihood estimation in generalized broken-line regression*. In Canadian Journal of Statistics 3 p227-238

Gill R & Baron M (2004) : *Consistent estimation in generalized broken-line regression.* In Journal of Statitical Planning and Inference 126 p441-460

Heiga Z, Keiichi T, Takashi M, Takao K & Tadashi K (2004) : *Hidden semi-markov model based speech synthesis* In International Workshop On Spoken Language Translation

Jurgensohn T (2005) : *Control theory models of the driver* In International Workshop On Modelling Driver Behaviour In Automotive Environments

Kamal M & Hasegawa-Johnson M (2003) : *Maximum conditional mutual information projection for speech recognition* Department of Electrical And Computer Engineering, University of Illinois.

Kampfe B (2005) : *Impact of skill acquisition in interacting with hmi on driver's situation awareness* In International Workshop On Modelling Driver Behaviour In Automotive Environments p292-295

Koul H & Qian L (2002) : *Asymptotics of maximum likelihood estimator in a two-phase linear regression model*. Journal of Statitical Planning and Inference 108 p99-119

Kuge N, Yamamura T, Shimoyama O & Liu. A (2000) : *A driver behavior recognition method based on a driver model framework* In SAE 2000 World Congress, Session: Human Centered Driver Assistance Systems.

Kumagaï, Sakaguchi, Okuwa & Akamatsu (2003) : *Prediction of driving behavior through probabilistic inference* In Proceedings Of The Eight International Conference On Engineering

Application Of Neural Networks

Kwong S, Chau C, Man K & Tang K. (2001) : O*ptimisation of hmm topology and its model parameters by genetic algorithms*. pattern recognition 34 p509-522

La documentation Française (2004) : *La Sécurité Routière En France, Bilan De L'Année 2003*. ONISR ed.

Lebarbier E & Mary-Huard T (2004) : *Le critère bic: fondements théoriques et interprétations* rapport INRIA

Lebart L., Morineau A. & Piron M. (2000) *Statistique exploratoire multidimensionnelle*. Dunod Ed ; pp155-175.

Levinson (1985) : *Structural methods in automatic speech recognition. IEEE* 73 p1625-1650

Loader C (1996) : *Change-point estimation nonparametric regression*. Ann. Statist. 4 p1667-1678

Logan B & Moreno P (1998) : *Factorial hmms for acoustic modeling* In Proceedings ICASSP p813-816

Mazet (1991) : *Perception et action dans la catégorisation: le cas de l'environnement urbain et routier*. Paris, Thèse de doctorat, Université de Paris V & Ecole Pratique des Hautes Etudes.

McDonough & Waibel (2003) : *Maximum mutual information speaker adapted training with semi-tied covariance matrices In Proceedings I*CASSP, Beijing, China

McKnight & Adam (1970) : Driver education task analysis. *Human Resources Research Organisation HumPRO 1*

Michon J. A (1985) : *A critical review of driver behaviour models*. In *Human Behavior And Traffic Safety*. . Evans L and Schwing R.G. (Ed.). Plenum Press; p485-520

Minsky M. (1975) : A Framework for representing knowledge. In the psychology of Computer Vision. . In P.H. Winston (ed); p211-277

Morel M, Collet C, Petit C, Holler S, Bruyas M, Boy & G (2005) : *Attention sharing between driving and added tasks: a physiological and behavioural study.* In proceedings of the international workshop on modelling driver behaviour in automotive environments

Murphy K (2002) : *Dynamic bayesian networks: representation, inference and learning*. Thèse de Doctorat. Computer Science Division Berkeley

Murphy K (2002) : *Hidden semi-markov models* Departments of computer science and statistics; Tech Report University of British Columbia

Murphy K & Paskin M (2001) : *Linear time inference in hierarchical hmms* In Proceedings Of Neural Information Processing Systems

Naatanen R & Summala H (1974) : *A model for the role of motivational factors in driver's decision-making* Accident Analysis and Prevention 6 p243-261

Nechyba, M, Xu, Y, (1998) : *Stochastics similarity for validating human control strategy models.* In proceedings IEEE Trans. on Robotics and Automation 14 (4) p437-451

Oliver N & Pentland A (2000) : *Graphical modeles for driver behavior recognition in a smart car* In IEEE Conference On Intelligent Vehicles Detroit.

Pal C & Hu M (2001) : *Methodologies for constructing and training large hiearchical hidden markov models for sequence analysis.* Computational Molecular Biology

Peltier M (1993) : *Un système adaptatif de diagnostic d'évolution basé sur la reconnaissance de formes floues, application au diagnostic du comportement d'un automobiliste* .Thèse de Doctorat. Université de technologie de Compiègne

Pentland A & Liu A (1999) : *Modeling and prediction of human behavior*. Neural Computation 11 p229-242

Pribe C & Rogers S (1999) : *Learning to associate observed driver behavior with traffic controls.* Transport Research Record 1679 95-100

Rabinner (1989) : *A tutorial on hidden markov models and selected applications in speech*

recognition. In Proceedings of the IEEE 77 (2), p257-286 .

Reschenhoffer E (1996) : *Prediction with vague prior knowledge*. communications in statistics theory and methods 25 p601-608

Reyes-Gomez, Raj & Ellis (2003) : *Multi-channel source separation by beamforming trained with factorial hmms* In IEEE Workshop on Applications of Signal Processing to Audio and Acoustic

Rezek I, Sykacek P. & Roberts. S. (2000) : *Learning interaction dynamics with coupled hidden markov models* In IEEE Proceedings - Science, Measurement And Technology

Rukhin A & Vajda I (1997) : *Change-point estimation as a nonlinear regression problem.* Statistics 3 p181-200

Salvucci D., Liu,A. (2002) : *The time course of a lane change: driver control and eye-movement behavior*. Transportation Research Part F 5 p123-132

Schulze U (1987) : *Multiphase regression: stability testing, estimation, hypothesis testing* Thèse de Doctorat; Akademie Berlin

Shelton C, Kim A. &. Poggio T. (2002): *Input output hidden markov models for modeling stock order flows* Tech report Institute of Technology of Massachusetts

Stenger B, Ramesh V, Paragios N, Coetzee F & Buhmann J (2001) : *Topology free hidden markov models: application to background modeling* In Proceedings International Conference On Computer Vision.

Stolcke A & Omohundro S (1994) : *Inducing probabilistic grammars by bayesian model merging* In International Conference On Grammatical Inference 862 p106-118

Tango F & Montanari R (2005) : *Modeling traffic and real situations* In Proceedings Of The International Workshop On Modelling Driver Behaviour In Automative Environments. p133-149

Tattegrain-Veste H, Bellet T, Chanut O & Legorrec G (2001) : *UML et génération de code: application à la modélisation cognitive du conducteur automobile.* In Génie logiciel

Tattegrain-Veste H, Bruyas, MP. , Bellet T, Forzy J, Simoes A, Carvalhais J, Lockwood P, Boudy J, Baligand B, Damiani S & Opitz M (2004) : *Vers une gestion centralisée des informations vocales en fonction du contexte de conduite : le projet cemvocas.* Recherche Transports Sécurité N82

Thomsen R (2002) : *Evolving the topology of hidden markov models using evolutionary algorithms* In Proceedings Of The 7Th International Conference On Parallel Problem Solving From Nature Springer-Verlag London

Van der Geer S (1988) : *Regression analysis and empirical processes. CWI Tract.* 45

Van Der Horst R (2005) : *Time-related measures for modelling risk in driver behaviour* In International Workshop On Modelling Driver Behaviour In Automotive Environments p113-123

Van Der Molen H & Botticher A.M.T. (1988) : *A hierarchical risk model for traffic participants*. Ergonomics 31 p537-555.

Viterbi AJ (1967) : Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory* 13(2) p260–267

Weir D. & Chao K.(2005) Review of control theory models for directional and speed control. In Proceedings Of The International Workshop On Modelling Driver Behaviour In Automative Environments. p25-37

Wiell Janssen (2004) :*Inventory of available tools and methodologies for driver behaviour analysis humanist.* In Proceedings Of The International Workshop On Modelling Driver Behaviour In Automative Environments.

Wu J & Chu C (1993) : *Nonparametric function estimation and bandwidth selection for discontinuous regression functions*. Statistica Sinica p557-576

Zhong S & Ghosh J (2001) : *A new formulation of coupled hidden markov models Department of electrical and computer engineering* Tech Report University of Texas Austin

Zhong S & Ghosh J (2002) : *HMMs and coupled HMMs for multi-channel eeg classification*

In IEEE International Joint Conference On Neural Networks 2 p1154-1159

Index des figures et des illustrations

4 Annexe

4.1 Détails de l'algorithme de Baum-Welch

L'algorithme consiste, pour l'étape m, à choisir $\theta^{(m+1)}$ maximisant $Q(\theta|\theta^{(m)})$ avec Q définit comme en (8). Baum et al [Baum et al, 1967] donnent alors les formules de réestimations :

- *pour les probabilités de transition :* $p=\mathrm{P}(S_t=i, S_{t+1}=j)$

La réestimation de p à l'étape m+1, $p^{(m+1)}$ est égale au quotient du nombre attendu de transition de l'état i à l'état j par le nombre attendu de transition partant de l'état i, soit

$$p_{i,j}^{(m+1)}=\sum_{t=1}^{T}\mathrm{P}(S_t=i, S_{t+1}=j/y)=\frac{\sum_{t=1}^{T}\alpha(t,i)\,p_{i,j}^{(m)}\,g_j(y_{(t+1)})\,\beta(t+1,j)}{\sum_{t=1}^{T}\alpha(t,i)\,\beta(t,i)} \qquad (82)$$

$$\forall\, i,j\in[1:K]$$

- *pour les probabilités initiales* $\pi_i=P(S_1=i)$

$\pi_i^{(m+1)}$ est la réestimation de π_i et est égale au nombre attendu de fois où S_1=i

$$\pi_i^{(m+1)}=\mathrm{P}(S_1=i/y)=\frac{\pi_i\,\partial P(y)/\partial\pi_i}{\sum_k \pi_k\,\partial P(y)/\partial\pi_k}=\frac{\alpha(1,j)\,\beta(1,j)}{\sum_j \alpha(1,j)\,\beta(1,j)} \qquad (83)$$

$$\forall\, i\in[1:K]$$

- *Calcul des paramètres de* g_i

Deux cas peuvent se présenter :

➔ *la densité (Yn/Sn) est une loi multinomiale de paramètres* $\{r_{1,i}, r_{2,i}, \ldots r_{R,i}\}$

Selon Rabinner, on réestime alors g_k (j) par le quotient entre le nombre attendu de fois où $(S_t=k)\cap(y_t=j)$ sur le nombre de fois où la suite S a pour valeur j.

$$r_{a,i}=\frac{\sum_{t,y_t=k} \mathrm{P}(S_t=i/y_t)}{\sum_{t} \mathrm{P}(S_t=i/y_t)}=\frac{r_{a,i}\partial \mathrm{P}(y)/\partial r_{a,i}}{\sum_{b=1}^{R} p_{i,k}\partial \mathrm{P}(y)/\partial r_{b,i}} \\ =\frac{\sum_{t,y_t=k} \alpha(t,i)\beta(t,i)}{\sum_{t} \alpha(t,i)\beta(t,i)} \forall i\in[1:K], a,b\in[1:R] \quad (84)$$

➔ *La densité est un mélange de M lois normales* $g_j(y)=\sum_{k=1}^{M} c_{jk}N(y,\mu_{jk}U_{jk}) \forall j\in[1:s]$

avec y l'observation c_{jk} le coefficient de la mixture, μ_{jk} la moyenne pour la k ième mixture et U_{jk} la matrice de covariance pour l'état j.

on a alors l'égalité suivante : $\sum_{k=1}^{M} c_{jk}=1 ;\ c_{jk}\geq 0\ ,\ 1\leq j\leq N,\ 1\leq k\leq M$.

De même, la convergence est assurée par les formules de ré estimation ci-dessous [Rabinner, 1989]. En notant

$$\gamma_t(j,k)=\frac{\alpha(t,j)\beta(t,j,).c_{jk}N(y_t,\mu_{jk},U_{jk})}{\sum_{i}\alpha(t,i)\beta(t,i)\sum_{m=1}^{M} c_{\Im}N(y_t,\mu_{im},U_{im})} \quad (85)$$

Les coefficients de mixture c_{ik} sont réestimés par

$$c_{jk}^{(m+1)}=\frac{\sum_{t=1}^{T}\gamma_t(j,k)}{\sum_{m=1}^{M}\sum_{t=1}^{T}\gamma_t(j,m)} \quad (86)$$

puis les moyennes μ_{jk} par

$$\mu_{jk}^{(m+1)}=\frac{\sum_{t=1}^{T}\gamma_t(j,k).y_t}{\sum_{t=1}^{T}\gamma_t(j,k)} \quad (87)$$

et enfin la matrice de covariance U_{jk} par

$$U_{jk}^{(m+1)}=\frac{\sum_{t=1}^{T}\gamma_t(j,k).(y_t-\mu_{jk}).(y_t-\mu_{jk})^T}{\sum_{t=1}^{T}\gamma_t(j,k)} \quad (88)$$

, $\forall j \in [1:K], k \in [1:M]$.

4.2 Modèle semi-Markovien caché (MSMC)

4.2.1. Calcul de la vraisemblance

Selon Rabinner [Rabinner, 1989,] pour calculer la vraisemblance dans le cas d'un MSMC, on pose

- $\alpha(t,i) = \mathbf{P}(y_1, y_2, \ldots y_t, S_i \text{ finis à } t \mid \theta)$.
- D la durée maximum pour tous les états : $D \in \mathbb{N}$ tel que $D = min(A)$ $A = \{ d \ / \ \forall t > d, i \in [1:K], h_i(t) = 0 \}$.

On a alors :

$$\alpha(t,j) = \sum_{i=1}^{K} \sum_{d=1}^{D} \alpha(t-d,i) p_{i,j} h_j(d) \prod_{s=t-d+1}^{t} g_j(y_s) \quad \forall t > D$$

avec

$$\alpha(1,i) = \pi(i) h_i(1) b_i(y_1)$$

$$\alpha(2,i) = \pi(i) h_i(2) \prod_{t=1}^{2} g_i(y_t) + \sum_{j=1 \ j \neq i}^{n} \alpha(1,j) h_i(1) g_i(y_t)$$

$$\alpha(3,i) = \pi(i) h_i(3) \prod_{t=1}^{3} g_i(O_t) + \sum_{d=1}^{2} \sum_{j=1 \ j \neq i}^{K} \alpha(3-d,j) p_{j,i} h_i(1) \prod_{t=4-d}^{3} g_i(y_t)$$

...

$$\alpha(t,i) = \pi(i) h_i(t) \prod_{s=1}^{t} g_i(y_s) + \sum_{d=1}^{t-1} \sum_{j=1 \ j \neq i}^{K} \alpha(t-d,j) p_{j,i} h_i(d) \prod_{s=t-d+1}^{t} g_i(y_s)$$

Comme dans le cas général, l'égalité suivante est vérifié $\mathbf{P}(y \mid \theta) = \sum_{i=1}^{K} \alpha_T(i)$.

4.2.2. Algorithme de Viterbi pour le modèle Semi-Markovien caché

Ge & Smyth [Ge & Smyth, 2001] ont adapté pour les MSMC l'algorithme de Viterbi. Ainsi, pour connaître la suite $s_1^* \ldots s_T^*$ maximisant $\mathrm{p}(S_{1\ldots T}=s_{1\ldots T}/y_{1\ldots T})$, on définit $r_t:[1:K]\rightarrow[1:K]$ par

$$r_t(i)=max_s\{\mathrm{P}(s/y_{1\ldots t}) \ /s=s_{1\ldots t-1}, s_t=i\} \tag{89}$$

La fonction $r_t(i)$ satisfait alors l'égalité suivante :

$$r_t(i)=\max_{d_i\in 1:D}(\max_{j\in 1:K} a(t,d,j))h_i(d)\mathrm{P}(y_{t-d_i+1\ldots y_t}/s_{t-d_i+1\ldots t}=i)$$

$\forall i\in[1:K], t\in[1:T]$ avec $a(t,d,j,i)=r_{t-d_i}(j)p_{ji}$ *si* $t-d>1$; $=\pi_i$ *sinon*.

Comme dans le cas des MMCs simples, ceci permet de trouver s_T^*. En effet, l'égalité suivante est vérifiée, $\max_{i\in[1:K]}(r_T(i))=s_T^*$

Puis on définit, pour $\forall i\in[1:K], t\in[1:T]$ la fonction $prec_t(i)$ par :

$$prec_d(i):[1:K]\rightarrow[1:D]*[1:K]$$

$$prec_d(i)=\underset{d\in 1:D}{argmax}\,\underset{i\in 1:K}{argmax}[p_{j,i}.r_{t-d}(j)h_i(d)]\ .$$

Le reste de la suite se trouve alors aisément par l'égalité suivante :

$\forall t\in[1:T-1]$ si $[d_0, i_0]=prec_t(s_t^*)$ alors $s_{t-d0}^*\ldots s_{t-1}^*=i_0$.

4.2.3. Formule de réestimations pour les MSMC

Pour les MSMC, pour calculer les formules de réestimations, Rabinner [Rabinner, 1989] rappelle qu'il est nécessaire de définir :

$$\alpha^*(t,i)=\mathrm{P}(y_1, y_2, \ldots y_t, S_i \text{ commence à } t+1 \mid \theta)$$

$$\beta(t,i)=\mathrm{P}(y_{t+1}, y_{t+2}, \ldots y_T, / S_i \text{ finis à } t \mid \theta)$$

$$\beta^*(t,i)=\mathrm{P}(y_{t+1}, y_{t+2}, \ldots y_T \mid S_i \text{ commence à } t+1 \ /\theta)$$

les relations entre $\alpha, \alpha^*, \beta, \beta^*$ sont alors les suivantes.

$$\alpha^*(t,j)=\sum_{i=1}^{K}\alpha(t,i)p_{ij}$$

$$\alpha(t,i)=\sum_{d=1}^{D}\alpha^*(t-d,i)h_i(d)\prod_{s=t-d+1}^{t}g_i(y_s)$$

$$\beta^*(t,i)=\sum_{i=1}^{K}\beta(t,i)p_{ij}$$

$$\beta(t,i)=\sum_{d=1}^{D}\beta^*(t+d,i)h_i(d)\prod_{s=t+1}^{t+d}g_j(O_s)$$

Les formules de réestimations sont alors égales à :

$$\pi_i^{(m+1)}=\frac{\pi_i^{(m)}\beta^*(0,i)}{\mathrm{P}(y/\theta)}=\frac{\pi_i^{(m)}\beta^*(0,i)}{\sum_{i=1}^{K}\beta^*(0,i)}$$

$$p_{ij}^{(m+1)}=\frac{\sum_{t=1}^{T}\alpha(t,i)p_{ij}^{(m)}\beta^*(t,j)}{\sum_{j=1}^{K}\sum_{t=1}^{T}\alpha(t,i)p_{ij}^{(m)}\beta*(t,j)}$$

$$h_i^{(m+1)}(d)=\frac{\sum_{t=1}^{T}\alpha^*(t,i)h_i^{(m)}(d)\beta(t+d,i)\prod_{s=t+1}^{t+d}g_j(y_s)}{\sum_{d=1}^{D}\sum_{t=1}^{T}\alpha^*(t,i)h_i^{(m)}(d)\beta(t+d,i)\prod_{s=t+1}^{t+d}g_j(y_s)}$$

- Selon Rabinner, si $g_i(y)$ est une loi multinomiale de paramètre $\{b_i(q),q\in[1:Q]\}\ Q\in\mathbb{N}$,la formule de réestimation ci dessous est vérifiée ([Rabinner, 1989]) :

$$g_i^{(m+1)}(q)=\frac{\sum_{t=1}^{T}\sum_{T,O_t=k}[\sum_{\tau<t}\alpha^*(\tau,i)h_i(d)\beta(\tau,i)-\sum_{\tau<t}\alpha(\tau,i)h_i(d)\beta(\tau,i)]}{\sum_{k=1}^{M}\sum_{t=1}^{T}\sum_{T,O_t=k}[\sum_{\tau<t}\alpha^*(\tau,i)h_i(d)\beta(\tau,i)-\sum_{\tau<t}\alpha(\tau,i)h_i(d)\beta(\tau,i)]}$$

- Selon Heiga et al, si $g_i(y)$ est une loi normale de paramètre (m_i,σ_i) , les formules de réestimations ci dessous sont vérifiées ([Heiga *et al.*, 2004])

$$m_j = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \eta_t^d(j)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j)} \tag{90}$$

$$\sigma_j = \frac{\sum_{t=1}^{T} \sum_{d=1}^{t} \xi_t^d(j)}{\sum_{t=1}^{T} \sum_{d=1}^{t} \gamma_t^d(j)} \tag{91}$$

$$avec \quad \gamma_t^d(j) = \sum_{\{i=1,\ i\neq j\}}^{K} \alpha(t-d,i)\, p_{i,j}\, h_j(d)\, \beta(t,j) . \prod_{r=t-d+1}^{t} g_j(y_r)$$

$$\eta_t^d(j) = \sum_{\{i=1,\ i\neq j\}}^{K} \alpha(t-d,i)\, p_{i,j}\, h_j(d)\, \beta(t,j) \prod_{r=t-d+1}^{t} g_j(y_t) . y_r$$

$$\xi_t^d(j) = \sum_{\{i=1,\ i\neq j\}}^{K} \alpha_{t-d}(i)\, p_{i,j}\, h_j(d)\, \beta_t(j) \prod_{r=t-d+1,\ r\neq s}^{t} g_j(y_r) . (y_r - m_j)^2$$

4.3 Séquences de conduite enregistrées

Numéro de la situation	Nombre de Séquences	Nom de la situation: contexte /infrastructure / objectif / situation initiale
1	4	ville / ligne droite / suivre route / vitesse stable / vitesse nulle
2	4	ville / ligne droite / suivre route / vitesse stable / vitesse faible
3	135	ville / ligne droite / suivre route / vitesse stable / vitesse moyenne
4	47	ville / ligne droite / suivre route / vitesse stable / vitesse forte
5	16	ville / ligne droite / accélérer / vitesse nulle
6	9	ville / ligne droite / accélérer / vitesse faible
7	102	ville / ligne droite / accélérer / vitesse moyenne
8	4	ville / ligne droite / accélérer / vitesse forte
9	1	ville / ligne droite / freiner / vitesse faible
10	26	ville / ligne droite / freiner / vitesse moyenne
11	4	ville / ligne droite / freiner / vitesse forte
12	1	ville / ligne droite / doubler / vitesse nulle
13	2	ville / ligne droite / doubler / vitesse moyenne
14	2	ville / ligne droite / s'arrêter en prévision de tourner droite / vitesse faible
15	20	ville / ligne droite / s'arrêter en prévision de tourner droite / vitesse moyenne
16	4	ville / ligne droite / s'arrêter en prévision de tourner droite / vitesse forte
17	8	ville / ligne droite / s'arrêter / vitesse faible
18	29	ville / ligne droite / s'arrêter / vitesse moyenne
19	5	ville / ligne droite / s'arrêter / vitesse forte
20	1	ville / ligne droite / changer de voie : gauche vers droite / vitesse nulle
21	1	ville / ligne droite / changer de voie : gauche vers droite / vitesse faible
22	22	ville / ligne droite / changer de voie : gauche vers droite / vitesse moyenne
23	9	ville / ligne droite / changer de voie : gauche vers droite / vitesse forte
24	2	ville / ligne droite / s'arrêter en prévision de tourner gauche / vitesse faible
25	24	ville / ligne droite / s'arrêter en prévision de tourner gauche / vitesse moyenne
26	5	ville / ligne droite / s'arrêter en prévision de tourner gauche / vitesse forte
27	1	ville / ligne droite / changer de voie droite vers gauche / vitesse nulle
28	2	ville / ligne droite / changer de voie droite vers gauche / vitesse faible
29	22	ville / ligne droite / changer de voie droite vers gauche / vitesse moyenne
30	5	ville / ligne droite / changer de voie droite vers gauche / vitesse forte
31	1	ville / ligne droite / être arrêté / vitesse nulle
32	20	ville / carrefour en croix / tourner gauche / vitesse nulle
33	2	ville / carrefour en croix / tourner gauche / vitesse faible
34	20	ville / carrefour en croix / tourner gauche / vitesse moyenne
35	37	ville / carrefour en croix / tourner droite / vitesse nulle
36	2	ville / carrefour en croix / tourner droite / vitesse faible
37	32	ville / carrefour en croix / tourner droite / vitesse moyenne
38	13	ville / carrefour en croix / aller tout droit / vitesse nulle
39	1	ville / carrefour en croix / aller tout droit / vitesse faible
40	14	ville / carrefour en croix / aller tout droit / vitesse moyenne
41	3	ville / carrefour en croix / aller tout droit / vitesse forte
42	24	ville / carrefour en croix / être arrêté en préparation de tourner droite / vitesse nulle
43	12	ville / carrefour en croix / être arrêté en prévision de tourner gauche / vitesse nulle
44	44	ville / carrefour en croix / être arrêté / vitesse nulle
45	2	ville / carrefour en croix / être arrêté / vitesse faible
46	1	ville / carrefour en croix / être arrêté / vitesse moyenne
47	11	ville / carrefour en T / voie secondaire / tourner gauche / vitesse nulle
48	1	ville / carrefour en T / voie secondaire / tourner gauche / vitesse faible
49	11	ville / carrefour en T / voie secondaire / tourner gauche / vitesse moyenne
50	1	ville / carrefour en T / voie secondaire / tourner gauche / vitesse forte
51	1	ville / carrefour en T / voie secondaire / tourner droite / vitesse nulle
52	1	ville / carrefour en T / voie secondaire / tourner droite / vitesse faible
53	9	ville / carrefour en T / voie secondaire / tourner droite / vitesse moyenne
54	1	ville / carrefour en T / voie secondaire / s'arrêter en prévision de tourner gauche / vitesse mo

Numéro de la situation	Nombre de Séquences	Nom de la situation: contexte /infrastructure / objectif / situation initiale
54	1	ville / carrefour en T / voie secondaire / s'arrêter en prévision de tourner gauche / vitesse moyenne
55	1	ville / carrefour en T / voie secondaire / être arrêté en prévision de tourner gauche / vitesse nulle
56	2	ville / ronds points / tourner gauche / vitesse moyenne
57	1	ville / ronds points / tourner droite / vitesse nulle
58	1	ville / ronds points / tourner droite / vitesse faible
59	11	ville / ronds points / tourner droite / vitesse moyenne
60	2	ville / ronds points / tourner droite / vitesse forte
61	2	ville / ronds points / aller tout droit / vitesse nulle
62	17	ville / ronds points / aller tout droit / vitesse moyenne
63	1	ville / ronds points / aller tout droit / vitesse forte
64	8	ville / virage léger gauche / suivre route / vitesse stable / vitesse moyenne
65	4	ville / virage léger gauche / suivre route / vitesse stable / vitesse forte
66	2	ville / virage léger gauche / accélérer / vitesse moyenne
67	1	ville / virage léger gauche / accélérer / vitesse forte
68	2	ville / virage léger gauche / s inserer dans un flux plus rapide / vitesse forte
69	1	ville / virage léger gauche / changer de voie : gauche vers droite / vitesse forte
70	1	ville / virage léger gauche / changer de voie droite vers gauche / vitesse forte
71	1	ville / virage léger gauche / s'arrêter / vitesse moyenne
72	11	ville / virage gauche / suivre route / vitesse stable / vitesse moyenne
73	3	ville / virage gauche / suivre route / vitesse stable / vitesse forte
74	3	ville / virage gauche / accélérer / vitesse nulle
75	3	ville / virage gauche / accélérer / vitesse moyenne
76	1	ville / virage gauche / changer de voie : gauche vers droite / vitesse forte
77	4	ville / virage léger droite / suivre route / vitesse stable / vitesse moyenne
78	2	ville / virage léger droite / suivre route / vitesse stable / vitesse forte
79	1	ville / virage léger droite / accélérer / vitesse forte
80	1	ville / virage léger droite / freiner / vitesse forte
81	1	ville / virage léger droite / changer de voie : gauche vers droite / vitesse forte
82	1	ville / virage droite / suivre route / vitesse stable / vitesse faible
83	5	ville / virage droite / suivre route / vitesse stable / vitesse moyenne
84	5	ville / virage droite / suivre route / vitesse stable / vitesse forte
85	1	ville / virage droite / accélérer / vitesse moyenne
86	1	ville / virage droite / accélérer / vitesse forte
87	1	ville / virage droite / freiner / vitesse forte
88	7	ville / carrefour en T / voie principale / tourner gauche / vitesse nulle
89	17	ville / carrefour en T / voie principale / tourner gauche / vitesse moyenne
90	4	ville / carrefour en T / voie principale / tourner droite / vitesse nulle
91	2	ville / carrefour en T / voie principale / tourner droite / vitesse faible
92	24	ville / carrefour en T / voie principale / tourner droite / vitesse moyenne
93	1	ville / carrefour en T / voie principale / aller tout droit / vitesse nulle
94	4	ville / carrefour en T / voie principale / aller tout droit / vitesse moyenne
95	8	ville / carrefour en T / voie principale / être arrêté / vitesse nulle
96	2	ville / chicane / suivre route / vitesse stable / vitesse forte
97	3	ville / chicane / freiner / vitesse moyenne
100	27	autoroute / ligne droite / suivre route / vitesse stable / vitesse moyenne
101	1	autoroute / ligne droite / suivre route / vitesse stable / vitesse forte
102	3	autoroute / ligne droite / accélérer / vitesse faible
103	5	autoroute / ligne droite / accélérer / vitesse moyenne
104	3	autoroute / ligne droite / freiner / vitesse moyenne
105	1	autoroute / ligne droite / changer de voie : gauche vers droite / vitesse faible
106	15	autoroute / ligne droite / changer de voie : gauche vers droite / vitesse moyenne
107	3	autoroute / ligne droite / changer de voie : gauche vers droite / vitesse forte

Numéro de la situation	Nombre de Séquences	Nom de la situation: contexte /infrastructure / objectif / situation initiale
108	4	autoroute / ligne droite / changer de voie droite vers gauche / vitesse faible
109	17	autoroute / ligne droite / changer de voie droite vers gauche / vitesse moyenne
110	2	autoroute / virage léger gauche / suivre route / vitesse stable / vitesse moyenne
111	2	autoroute / virage léger gauche / changer de voie droite vers gauche / vitesse moyenne
112	1	autoroute / virage gauche / suivre route / vitesse stable / vitesse faible
113	4	autoroute / virage gauche / suivre route / vitesse stable / vitesse moyenne
114	1	autoroute / virage gauche / changer de voie : gauche vers droite / vitesse moyenne
115	13	autoroute / virage léger droite / suivre route / vitesse stable / vitesse moyenne
116	4	autoroute / virage droite / suivre route / vitesse stable / vitesse faible
117	4	autoroute / virage droite / suivre route / vitesse stable / vitesse moyenne
118	1	autoroute / virage droite / accélérer / vitesse moyenne
119	1	autoroute / insertion / suivre route / vitesse stable / vitesse faible
120	2	autoroute / insertion / suivre route / vitesse stable / vitesse moyenne

Résumé

L'objet de cette thèse est d'établir un cadre d'analyse des données recueillies sur les véhicules en vue de les mettre en correspondance avec des comportements. Pour cela, nous avons défini le modèle Semi-Markovien caché pondéré pour modéliser les signaux issus des capteurs et établi des résultats théoriques sur les modèles de régressions multi-phasiques dans le cas linéaire et non-linéaire. Puis, nous avons établi une méthodologie d'analyse de l'activité basée sur un apprentissage semi-automatique, et structurée par les résultats des modèles cognitifs du conducteur. Pour valider cette méthodologie, nous avons effectué une expérimentation où furent enregistrées 1209 séquences de conduite. Ces données nous ont permis d'implémenter des modèles de Markov cachées décrivant l'évolution des capteurs associés à des situations de conduite caractérisées par l'objectif, l'infrastructure perçue par le conducteur et la vitesse initiale.

Les modèles générés nous permettent dès lors de catégoriser, avec un taux satisfaisant, à quelle situation ou à quel groupe de situation appartient une séquence inconnue. Par ailleurs, nous illustrons l'utilité des modèles markoviens conjugués aux modèles multi-phasiques pour la recherche automatique de situation dans un ensemble de données.

Mot clés:

Modèle semi-Markovien cachés pondéré, modèles multi-phasiques, apprentissage semi-automatique, catégorisation du comportement du conducteur, analyse de l'activité de conduite analyse hiérarchique.

The objective of this study is to establish an analysis framework in order to link the data recorded on vehicles with driver's behavior. To model sensors data evolution, we defined the weight hidden semi Markov model, and we presented theoretical results in a multi-phase random regression model. Then we developed a methodology to analyze driving activity. This was based on a semi-automatic learning process and structured by cognitive research results. To assess this methodology an experiment was conducted where 1209 driving sequences were recorded. This data was used to build Markov models linked with driving situations defined by the objective, the infrastructure seen by the driver and the initial speed.

These models allow us to categorize, with a satisfactory rate, which situation or which group of situations the driver is in. Moreover, we showed the utility of the Markov models mixed with multi-phase models in order to automatically research a situation in a data set.

Key-Word:

Weight hidden semi Markov model, multi-phase random regressions models, semi-automatic learning process, driver behavior categorization, driving activity analysis, hierarchic analysis.