UNIVERSITÉ DE MONTRÉAL

ESTIMATION DE TRAFIC ROUTIER PAR FILTRE DE KALMAN D'ENSEMBLE SOUS CONTRAINTE DE CONFIDENTIALITÉ DIFFÉRENTIELLE

HUBERT ANDRÉ
DÉPARTEMENT DE GÉNIE ÉLECTRIQUE
ÉCOLE POLYTECHNIQUE DE MONTRÉAL

MÉMOIRE PRÉSENTÉ EN VUE DE L'OBTENTION
DU DIPLÔME DE MAÎTRISE ÈS SCIENCES APPLIQUÉES
(GÉNIE ÉLECTRIQUE)
MAI 2017

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Ce mémoire intitulé :

ESTIMATION DE TRAFIC ROUTIER PAR FILTRE DE KALMAN D'ENSEMBLE SOUS CONTRAINTE DE CONFIDENTIALITÉ DIFFÉRENTIELLE

présenté par : ANDRÉ Hubert
en vue de l'obtention du diplôme de : Maîtrise ès sciences appliquées
a été dûment accepté par le jury d'examen constitué de :

M. GOURDEAU Richard, Ph. D., président
M. LE NY Jérôme, Ph. D., membre et directeur de recherche
M. MULLINS John, Ph. D., membre

DÉDICACE

À ceux qui se reconnaîtront...

REMERCIEMENTS

Je tiens à remercier Jérôme Le Ny, mon directeur de recherche, pour sa supervision ainsi que le centre de recherche GERAD. Je tiens également à remercier les membres de mon jury, Richard Gourdeau et John Mullins, pour avoir accepté de relire ce mémoire.

RÉSUMÉ

Le but de cette recherche est de proposer un algorithme numérique pour la publication en temps réel de certaines statistiques calculées à partir de données sensibles fournies par des utilisateurs. Afin de contrôler le gain possible d'information divulguée par ces statistiques sur les utilisateurs contributeurs, nous souhaitons que notre algorithme garantisse leur *confidentialité différentielle*, qui est une notion forte et formelle de la confidentialité. La confidentialité différentielle stipule que le gain d'information sur la participation d'un utilisateur individuel apporté par rapport à une connaissance préalable par la publication de données est borné. Cela signifie qu'il est impossible de savoir avec beaucoup de certitude si un utilisateur donné a participé à l'élaboration des données publiées, et cela pour tout utilisateur. L'anonymat des utilisateurs est donc garanti. Assurer un niveau de confidentialité différentielle se traduit généralement en dégradant les données, de manière maîtrisée. Le but de cette recherche est donc de proposer des mécanismes assurant la confidentialité différentielle et dégradant le moins possible les données. Nous supposons qu'un modèle dynamique stochastique expliquant la génération du jeu de données est disponible et à la disposition de tous, et nous cherchons à utiliser ses caractéristiques pour améliorer la qualité des données publiées.

Nous appliquons l'estimateur développé à la problématique de l'estimation de trafic routier. Les usagers de la route envoient leurs données en franchissant des capteurs statiques sur la route ou en fournissant leurs traces GPS. Un mécanisme différentiellement privé publie l'état du trafic sous la forme d'une carte spatio-temporelle de densité. Ici la confidentialité différentielle garantit que l'étude de la carte publiée au cours du temps ne peut pas révéler des informations sensibles trop précises sur les trajets des utilisateurs de la route.

Les mécanismes différentiellement confidentiels sont élaborés à base de filtres : filtre de Kalman étendu, filtre de Kalman d'ensemble, filtre à particules. Nous utilisons des jeux de données synthétiques et une certaine métrique pour comparer les différents filtres entre eux. Ensuite nous testons notre estimateur différentiellement confidentiel sur un jeu de données réelles, provenant du projet *Mobile Century*.

Cette recherche a permis de proposer un mécanisme différentiellement confidentiel à base de filtre de Kalman d'ensemble pour l'estimation de trafic routier. De plus, le fait que cet algorithme garantisse également la confidentialité différentielle avec des mesures provenant de traces GPS permet d'étendre le champs d'application à un réseau routier beaucoup plus large puisque nous ne nous restreignons plus aux routes équipées de capteurs fixes.

ABSTRACT

Road traffic estimation systems can rely nowadays on an increasing number and variety of sensors and data sources to provide better coverage and accuracy, from standard static detectors to, more recently, location traces obtained possibly from individual drivers' smartphones. Motivated by privacy concerns raised by such systems, this thesis discusses a methodology for estimating the macroscopic traffic state (density, velocity) along a road segment in real-time, while providing formal *differential privacy* guarantees to the individual drivers, a state-of-the-art notion of privacy that protects against adversaries with arbitrary side-information. This translates to the inability for an adversary to make a better guess for the participation of a specific individual, with the use of differentially private data. Differential privacy provides formal proof that the relative information gain for the adversary, with publication of differentially private data, is bounded.

Making data differentially private means randomizing in some way the data, thus making the published output less accurate. The goal of this research is to propose a numerical method to make private data differentially private for public release. Such methods are called differentially private mechanisms. The impact of the privacy constraint on estimation performance is mitigated by the use of a nonlinear model of the traffic dynamics, fused with the sensor measurements via an Ensemble Kalman Filter, a classical method for data assimilation.

The differentially private mechanism is applied to a road traffic estimation problem Road users send their data when they cross static sensors (position, occupancy), and through their smartphones (position, speed). The differentially private mechanism publishes the density map that is usable by any third party app, and the privacy guarantees will follow.

All the mechanisms are validated on synthetic data and tested on the Mobile Century dataset.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

DC	Différentiellement confidentiel.
FK	Filtre de Kalman.
FKE	Filtre de Kalman étendu.
FKEn	Filtre de Kalman d'ensemble.
LPV	Ligne de passage virtuelle.
LWR	Lighthill–Whitham–Richards
$o^l_{p,t}$	Mesures d'occupations pour le capteur à la position p, au temps t pour la voie l.
$o_{p,t}$	Occupation moyenne à la position p au temps t.
$c^l_{p,t}$	Mesures de comptes pour le capteur à la position p, au temps t pour la voie l.
$c_{p,t}$	Comptes moyens à la position p au temps t.
$v^l_{p,t}$	Mesures de vitesse pour le capteur à la position p, au temps t pour la voie l.
$V_{p,t}$	Moyenne géométrique des vitesses pour le capteur à la position p, au temps t.
v_0	Vitesse libre de trafic (vitesse d'une seule voiture sur une route vide).
ρ	Densité du trafic.
ρ_C	Densité critique à partir de laquelle les automobilistes adaptent leurs vitesses ($\rho > \rho_C \Rightarrow v < v_0$).
ρ_M	Densité maximale ou "congestionnée", "de bouchon", associée à une vitesse de trafic nulle.
w	Vitesse à laquelle la congestion se propage (dans le sens contraire du trafic, $w > 0$).
q	Flux du trafic.
q_M	Flux maximum du diagramme fondamental.
ΔT	Temps d'échantillonnage.
λ_p	Nombre de voies à la position p

LISTE DES ANNEXES

CHAPITRE 1 INTRODUCTION

L'estimation de la densité ainsi que l'estimation de la vitesse du trafic routier sont de grand intérêt pour des domaines variés. Une meilleure efficacité énergétique du transport. La diminution de la pollution. Des temps de trajets plus court. Ce dernier étant atteint en répartissant le trafic de manière à conserver un trafic fluide, grâce à des applications calculant les trajets en fonctions du réseau routier (distances, types de routes), mais aussi avec l'état en temps réel du trafic, c'est à dire en prenant en compte le niveau d'utilisation des routes sur le trajet. Cependant, pour avoir accès à cet estimé de l'état du trafic, les utilisateurs doivent généralement en contrepartie participer à l'élaboration de cet estimé en envoyant des données personnelles. Par ailleurs, les capteurs fixes à induction, ou des caméras de trafic, utilisent les données de tous les usagers de la route sans leurs consentements. Des méthodes doivent donc êtres développées pour garantir la protection de l'anonymat des usagers (l'association d'une trajectoire à un individu) ainsi que la qualité de l'estimé.

1.1 Éléments de la problématique

1.1.1 Contexte de l'étude

L'estimation en temps réel du trafic routier a bénéficié des importantes avancées de ces dernières années dans les technologies de détection. Les caméras de trafic combinées avec les algorithmes de vision par ordinateur peuvent jouer un rôle similaire aux traditionnels capteurs par boucle d'induction magnétique intégrés aux autoroutes, comptant les véhicules qui passent à un certain endroit et estimant leur vitesse. La révolution vient aussi de la facilité d'obtention de données mobiles envoyées par des appareils à l'intérieur des véhicules, mesurées par le système de localisation du réseau de téléphonie mobile, par les GPS à bord, ou par radio-identification. L'intégration de ces données mobiles permet de construire un estimé du trafic, non seulement sur certaines autoroutes équipées de capteurs fixes, mais sur l'ensemble du réseau, sans besoin de nouvelles infrastructures coûteuses.

1.1.2 Protection de l'anonymat

Cependant, un effet secondaire à la collecte d'informations personnelles de positions d'utilisateurs est la possibilité de suivi indésirable par des individus externes ayant accès aux données. C'est à dire la possibilité d'attribuer de manière correcte des trajectoires à des utilisateurs, même lorsque les moyens évidents d'identification comme les noms sont retirés des

bases de données. Même utiliser des capteurs statiques qui publient des grandeurs moyennées sur un intervalle de temps ne suffit pas pour garantir la confidentialité. En effet, imaginons un utilisateur démarrant sa voiture sur une route initialement vide. Nous pouvons suivre par exemple les comptes passant de zéro à un le long de la route et reconstruire la trajectoire. Le point de départ pouvant être le lieu d'habitation de l'utilisateur, et le point d'arrivée son lieu de travail. De manière plus sophistiquée, il est aussi possible de mener des attaques utilisant des méthodes statistiques pour des trafics à fortes densités de véhicules, et des informations complémentaires, comme d'autres bases de données publiques. Ainsi, avec seulement avec la publication de la densité agrégée du trafic (de Montjoye et al., 2013; Canepa and Claudel, 2013), il est possible de mener à bien des attaques sur les trajectoires des utilisateurs routiers. Des publications (Sweeney, 1997; Narayanan and Shmatikov, 2008) illustrent les risques d'exclure la possibilité de telles attaques en se basant seulement sur un raisonnement intuitif.

Le but de cette recherche est de développer une méthodologie pour la conception d'un estimateur en temps réel assurant des garanties de confidentialité formellement définies, la *confidentialité différentielle* (Dwork et al., 2006). Tandis que d'autres notions de confidentialité peuvent être considérées pour notre application, comme le k-anonymat (Sweeney, 2002; Hoh et al., 2012a) et ses nombreuses extensions (Li et al., 2007), la confidentialité différentielle est de plus en plus adoptée ces dernières années comme l'outil de pointe pour l'analyse confidentielle des données privées (Dwork and Roth, 2014). Les mécanismes différentiellement confidentiels publient des résultats (ici, les séries temporelles des états du trafic) qui sont randomisés d'une manière telle que leurs distributions ne dépendent que très peu des données de tout utilisateur particulier.

Nous pouvons faire une brève introduction à la définition mathématique de la confidentialité différentielle. La traduction mathématique de la définition est pour un mécanisme $\mathcal{M}$ (ϵ, δ)-différentiellement confidentiel : $\mathbb{P}(\mathcal{M}(d) \in s) \leq e^{\epsilon}\mathbb{P}(\mathcal{M}(d') \in s) + \delta$, pour toutes bases de données différant d'au plus un élément dites adjacentes (d, d'), pour tout ensemble d'arrivée s. Supposons deux événements dits adjacents E_1 et E_2, et un adversaire ayant une idée préalable du ratio $\frac{\mathbb{P}(E_1)}{\mathbb{P}(E_2)}$. Si cet adversaire reçoit une nouvelle information $\mathcal{A}$ venant de la publication d'un mécanisme ϵ- différentiellement confidentiel, alors nous pouvons montrer facilement que $e^{-\epsilon} \leq \frac{\mathbb{P}(E_1|\mathcal{A})}{\mathbb{P}(E_1|\mathcal{A})} / \frac{\mathbb{P}(E_1)}{\mathbb{P}(E_2)} \leq e^{\epsilon}$. C'est à dire que, pour un epsilon petit, la publication de données différentiellement confidentielles ne permet pas de discriminer entre les bases de données adjacentes. La conséquence au niveau individuel est que le fait qu'un individu choisisse ou non de fournir ses données ne change que très peu le risque qu'un adversaire (un individu malveillant voulant récupérer des données confidentielles), quelles que soient ses connaissances, puisse faire de meilleures déductions sur cet individu.

1.1.3 Considérations dynamiques

La difficulté cependant est de produire un estimé différentiellement confidentiel peu perturbé, c'est à dire un estimé du trafic suffisamment précis pour être utilisable. Il y a eu des travaux portant sur la protection de données de localisation à travers des mécanismes différentiellement confidentiels, mais aucun se concentrant sur un estimé basé sur un modèle dynamique, à l'exception du travail précédent celui-ci (Le Ny et al., 2014), que nous améliorons avec une méthode alternative d'assainissement (assainissement synonyme de rendre différentiellement confidentiel) des données de capteurs statiques, ainsi qu'avec la considérations des données mobiles (GPS). Pour étendre l'estimé du trafic à l'ensemble du réseau routier et ne pas se limiter aux points de mesures, ainsi que pour prendre en compte les erreurs de mesures et avoir un meilleur estimé, les mesures de capteurs de trafic sont intégrées avec un modèle hydrodynamique de trafic (Treiber and Kesting, 2013) à travers un filtre de Kalman d'ensemble (EnKF) produisant une sortie différentiellement confidentielle. La présence de capteurs mobiles permet de faire un échantillonnage spatial adaptatif qui permet d'améliorer encore l'estimé en utilisant la caractéristique mobile des capteurs GPS, ce qui constitue aussi une nouveauté par rapport aux travaux précédents. Le filtre de Kalman d'ensemble (Evensen, 2003), basé sur l'idée du filtre de Kalman (KF) et des méthodes de Monte-Carlo, est populaire pour l'estimation de trafic routier (Work et al., 2008), et mène à une implémentation plus simple et des résultats plus précis que les solutions (Le Ny et al., 2014) basées sur un filtre de Kalman étendu (EKF).

1.1.4 Base de données Mobile Century

Nous introduisons dans ce paragraphe un jeu de données publique de mesures de trafic sur lequel nous appliquerons notre estimateur. Le jeu de donné est dénommé *Mobile Century* et est disponible en accès libre à
`http://traffic.berkeley.edu/project/downloads/mobilecenturydata`.
Le jeu de données Mobile Century a été recueilli le 8 Février 2008, entre 10h et 18h (PST) sur l'Interstate 880,CA. Ce jeu de données est le résultat d'un projet de l'Université Berkeley et Nokia pour étudier l'estimation de trafic utilisant des données de GPS de téléphones mobiles. En plus des GPS, deux autres sources de données sont recueillies, des données de capteurs fixes par induction, et des données issues de traitement d'images de caméras hautes résolutions placées sur la route.

1.2 Objectifs de recherche

L'objectif de la recherche est la conception d'un algorithme différentiellement confidentiel, avec pour entrées les sources variées d'informations de capteurs, et en sortie l'état estimé en temps réel du trafic routier. Cet algorithme devra utiliser un modèle dynamique pour augmenter l'utilité (la précision) de l'estimateur. L'algorithme devra aussi traiter les données de capteurs mobiles. Cet algorithme sera évalué sur la simulation d'un trafic synthétique, et sera aussi comparé à des méthodes existantes d'estimation différentiellement confidentielles sur un jeu de données réel.

1.3 Plan du mémoire

Tout d'abord, le chapitre 1 a été consacré à l'introduction générale du problème. La revue de littérature est effectuée dans le chapire 2. L'énoncé du problème est présenté dans le chapitre 3. Ce chapitre comporte la présentation de la modélisation de la dynamique du système routier, ainsi que de la confidentialité différentielle. Ensuite, la modélisation des capteurs ainsi que les techniques d'assainissement des données de trafic sont étudiées dans le chapitre 4. Les différents filtres sont présentés au chapitre 5. Finalement, les filtres différentiellement confidentiels sont testés et comparés entre eux avec des données synthétiques et le jeu de données Mobile Century dans le chapitre 6. Finalement le chapitre 7 présente nos conclusions et ses limites, et les pistes restantes à étudier.

CHAPITRE 2 REVUE DE LITTÉRATURE

2.1 Introduction

Le sujet d'étude couvre principalement trois domaines de recherche. La dynamique du trafic routier, les techniques de filtrage, et la confidentialité différentielle.
L'estimation de l'état du trafic est abondamment traité dans la littérature. Cependant, l'utilisation de filtres utilisant les méthodes de Monte-Carlo comme le filtre à particules ou le filtre de Kalman d'ensemble est relativement récente, mais est très prometteuse avec l'accroissement de la puissance de calcul. L'utilisation et l'adaptation de ces filtres pour le trafic routier n'est pas encore un sujet clos. La confidentialité différentielle est aussi une notion assez nouvelle, et reste encore marginalement utilisée pour le problème du trafic, mais sa définition formelle et ses fortes garanties de sécurité personnelle des usagers en font aujourd'hui une définition puissante de confidentialité. Plusieurs mécanismes peuvent être développés pour atteindre la confidentialité différentielle. Cette revue de littérature a pour but d'introduire les avancées récentes et les mécanismes prometteurs qui ont été ou non adaptés au problème d'estimation de trafic.
Dans cette revue de littérature, nous aborderons l'état de l'art des techniques pour la modélisation de la dynamique du trafic (Section 2.2), l'état de l'art des techniques pour les filtres de Kalman d'ensemble et les filtres à particules (Section 2.3), et l'état de l'art des mécanismes pour atteindre la confidentialité différentielle (Section 2.4). La dernière section (Section 2.5) regroupe certaines des techniques des sections précédentes qui portent en particulier sur l'estimation du trafic routier.

2.2 Trafic

Le livre Treiber and Kesting (2013) est un ouvrage très complet qui traite de toutes les bases de la dynamique du trafic routier, ainsi que de sujets plus avancés. Ce manuel sera utilisé comme référence pour le modèle dynamique du premier ordre Lighthill–Whitham–Richards. Plusieurs autres travaux (Alecsandru et al., 2011; Sumalee et al., 2011; Staňková and De Schutter, 2010; Burger et al., 2013) ont étudié le modèle Lighthill–Whitham–Richards (LWR), sa pertinence et son application. L'article Fan and Seibold (2012) étudie la différence entre des modèles du premier et du second ordre.

Les travaux de Boel and Mihaylova (2004, 2006); Jabari and Liu (2012) introduisent de l'aléatoire dans la transmission de véhicules entre cellules dans le modèle. L'article Quek

et al. (2009) s'attaque aux limitations d'un modèle prédéfinit, avec un système de règles hybride neuronal flou. Nous pouvons aussi citer Celikoglu (2013) pour l'incorporation de la théorie des réseaux de neurones. Le problème des informations multisources pour le trafic est analysé dans Kong et al. (2009). Pour l'étude du cas particulier des téléphones mobiles, nous avons les travaux Hardjono et al. (2012); Handel et al. (2014); Hardjono et al. (2014). Notons que si les notions de confidentialité sont parfois abordées, les définitions de sécurité de l'information dans ces travaux sont très faibles voire inexistantes. Avec le développement massif des téléphones mobiles avec GPS intégrés, les chercheurs se sont mis à s'intéresser aux GPS personnels, ainsi qu'à l'influence de l'échantillonnage sur l'estimation (Zhang et al., 2013). La mise à l'échelle des estimateurs pour les grands réseaux est étudiée dans Chen et al. (2012a); Wang et al. (2011).

2.3 Filtre de Kalman d'ensemble et filtre à particules

Le filtre de Kalman (Harvey, 1990) et le filtre de Kalman étendu (Julier and Uhlmann, 1997), une de ses variantes pour les systèmes non linéaires, sont des filtres itératifs pour fusionner des données bruitées de capteurs avec des modèles stochastiques dynamiques de l'évolution d'un système. Le filtre de Kalman d'ensemble reprend des éléments du filtre de Kalman et des techniques de Monte-Carlo. Certaines grandeurs sont calculées à partir d'un ensemble de réalisations (Evensen, 2003; Burger et al., 2013). Le filtre à particules est un filtre qui demande généralement plus de ressources de calculs que d'autres filtres, mais il permet de modéliser des distributions complexes de paramètres. Ce filtre a gagné en popularité avec l'augmentation massive de la puissance de calcul des ordinateurs. De nombreux papiers de qualité sur ce filtre ont été publiés (Gordon et al., 1993; Cappé et al., 2007; Särkkä, 2013). Cependant, ce filtre s'adapte généralement mal aux problèmes de grandes dimensions. Ce soucis particulier est analysé dans (Snyder et al., 2008). Différents filtres à particules ont été développés (Wu et al., 2008; Okuma et al., 2004; Van Der Merwe et al., 2000; Nummiaro et al., 2003; Carpenter et al., 1999), certains pour des problèmes spécifiques de suivis de cibles multiples (Yi et al., 2013; Khan et al., 2004).

2.4 Confidentialité différentielle

La confidentialité différentielle est une notion très forte de confidentialité qui a été introduite dans (Dwork et al., 2006), puis développée avec (Dwork, 2006; Dwork and Roth, 2013). Des papiers étudient certains mécanismes particuliers comme le mécanisme Laplacien (Koufogiannis et al., 2015), ou certaines applications comme le comptage (Liebig, 2015), les réseaux

sociaux (Koufogiannis and Pappas, 2015), les réseaux électriques intelligents (Ratliff et al., 2014), ou encore des minimisations de risques (Bassily et al., 2014). Certaines études (Xiong, 2013) exploitent les caractéristiques des données pour atteindre une meilleure utilité. D'autres essais de définitions ont été entrepris basés sur la confidentialité différentielle (généralement pour l'assouplir) ou l'élargir (Kifer and Machanavajjhala, 2012), ou la lier avec l'entropie (Wang et al., 2014). La confidentialité différentielle pour les systèmes dynamiques est étudiée dans (Le Ny and Pappas, 2014; Le Ny, 2015). Le compromis entre l'utilité et la sécurité est analysé dans (Zhang et al.; Huang et al., 2014; Dong et al., 2014). La confidentialité différentielle a été incorporée à l'estimation bayesienne avec le travail de (Wang et al., 2015), utilisant la distribution non singulière a posteriori du filtre à particule comme moyen d'introduire la part d'aléatoire nécessaire aux mécanismes délivrant la confidentialité différentielle.

2.5 Études comportant au moins deux des sujets précédents

Quelques articles ont déjà été publiés à propos de l'estimation du trafic routier sous conditions de confidentialité différentielle, de l'estimation du trafic routier avec un FKEn ou un filtre à particules. Quelques articles ont été publiés sur la sécurité des informations privées pour le trafic routier avec des définitions différentes de confidentialité (Hoh et al., 2012b). Cependant l'état de l'art actuel pour la confidentialité individuelle est la confidentialité différentielle. Quelques papiers (Mihaylova et al., 2007; Chen et al., 2011; Chen and Rakha, 2014; Billot et al., 2010; Bi et al., 2013; Wu et al., 2015; Jun et al., 2012; Chen et al., 2012b) ont étudié l'estimation du trafic routier avec un filtre à particules, et particulièrement l'aspect haute dimension du problème (Mihaylova et al., 2012; Hegyi et al., 2007; Mihaylova et al., 2014; Pascale et al., 2014; Chen et al., 2014; Kong et al., 2013; Wang et al., 2009; Feng et al., 2015). Certaines recherches (Gisdakis et al., 2015; Cheng et al., 2006a,b) ont été faites pour les données mobiles associées avec filtre à particules. Le papier (Fan et al., 2013) propose un algorithme pour l'estimation confidentielle différentielle de systèmes dynamiques avec échantillonnage adaptatif.

2.6 Conclusion

La référence de base sur la théorie de la dynamique routière est le manuel (Treiber and Kesting, 2013). L'article (Burger et al., 2013) est utilisé ici comme référence pour la base théorique du filtre de Kalman d'ensemble. Le livre de Dwork and Roth (2013) est utilisé pour certaines techniques utiles pour la confidentialité différentielle. Le travail (Fan et al., 2013) a servi d'inspiration pour l'échantillonnage adaptatif. Cette brève revue de littérature

montre les différentes relations et liens entre les divers sujets liés au trafic, à l'estimation et à la confidentialité. Un estimateur différentiellement confidentiel pour le trafic routier, avec un mécanisme utilisant les propriétés de la distribution a posteriori d'un filtre à particules ou d'un filtre de Kalman d'ensemble n'a pas encore été conçu, ce que nous tentons donc de faire dans ce mémoire.

CHAPITRE 3 ÉNONCÉ DU PROBLÈME

Dans ce chapitre nous formulons l'énoncé du problème. Nous présentons le modèle dynamique utilisé pour le filtrage, et une méthode empirique pour déterminer certains de ses paramètres. Nous proposons également un modèle des capteurs. Nous formulons aussi la définition de la confidentialité différentielle.

3.1 La dynamique de trafic

Nous énonçons dans cette section le modèle dynamique et les hypothèses faites.

3.1.1 Introduction du modèle

Par soucis de simplicité, nous considérons l'estimation de l'état du trafic sur une route sans intersections. L'état du trafic peut être représenté de différentes façons : de manière microscopique, mésoscopique, et macroscopique.

Modèles microscopiques. Ces modèles prennent en considération les voitures individuelles et décrivent leurs comportements (accélération, freinage, changement de voies). Ces modèles de bas niveau décrivent au plus proche la réalité puisqu'ils peuvent être raffinés pour décrire précisément les comportements humains, temps de réaction, hésitations, inattentions, etc. Cependant, la grande dimension de ces modèles les rend peu maniables.

Modèles mésoscopiques. Ces modèles sont des modèles hybrides comportant des approches microscopiques et macroscopiques.

Modèles macroscopiques. Ces modèles font l'analogie entre le mouvement du trafic et le mouvement d'un fluide, d'où leurs noms de *modèles hydrodynamiques*. Les grandeurs sont localement agrégées. De manière générale, l'état du trafic au temps t et position x est caractérisé par une densité $\rho(x,t)$ (en véhicules par mètres) et une vitesse de trafic $v(x,t)$ ou un flux de trafic $q(x,t) := \rho(x,t)v(x,t)$. Ce sont à ces modèles macroscopiques que nous nous intéressons pour construire nos estimateurs.

3.1.2 Diagramme fondamental

Pour les modèles hydrodynamiques du premier ordre (Treiber and Kesting, 2013), nous postulons une relation statique entre la densité et le flux du trafic, appelée le *diagramme fondamental*, qui est souvent supposée être triangulaire

$$q(\rho) = \begin{cases} v_0\rho, & \text{pour } \rho \leq \rho_C := \frac{w}{v_0+w}\rho_M \\ -w(\rho - \rho_M), & \text{for } \rho_C \leq \rho \leq \rho_M \end{cases} \tag{3.1}$$

dont les paramètres sont

- v_0 : la vitesse libre de trafic, c'est à dire la vitesse maximale si une voiture n'a pas à réduire sa vitesse dû à la présence d'autre véhicules. La vitesse libre v_0 devrait être proche de la vitesse limite autorisée.
- ρ_C : la densité critique entre un trafic libre (vitesse $v = v_0$) et un trafic congestionné (vitesse $v < v_0$).
- ρ_M : la densité maximale ou "congestionnée", "de bouchon", associée à une vitesse de trafic nulle.
- w : la vitesse à laquelle la congestion se propage, dans la direction inverse du sens du trafic. C'est aussi la vitesse de déplacement du front d'onde entre un trafic libre de densité critique qui entre en "collision" avec un trafic complètement congestionné, voir figure 3.2.
- q_M : le flux maximal pour le sommet du diagramme

La vitesse du trafic pour une densité ρ_B est notée v_B. La relation $q(x,t) := \rho(x,t)v(x,t)$ indique que la vitesse du trafic au point B est la pente de la droite passant par l'origine du repère (ρ, q) et par le point (ρ_B, q_B).

Une onde de choc est la propagation d'une discontinuité de la densité. Avec un diagramme fondamental triangulaire, il est possible d'observer la propagation d'ondes de choc en simulation de trafic. La vitesse de propagation d'une onde de choc entre un trafic de densité ρ_A et un trafic de densité ρ_B est notée v_{AB}. Un bilan sur le nombre de voiture entre les instant t et $t + \mathrm{d}t$ aux positions x et $x + \mathrm{d}x$ entourant l'onde de choc (voir figure 3.2) permet d'écrire :

$$\begin{aligned} (\rho_B - \rho_A)v_{AB}\,\mathrm{d}t &= (q_B - q_A)\,\mathrm{d}t \\ v_{AB} &= \frac{q_B - q_A}{\rho_B - \rho_A}. \end{aligned} \tag{3.2}$$

Cette vitesse s'interprète comme la pente de la droite allant du point (ρ_A, q_A) au point (ρ_B, q_B) sur le diagramme fondamental. Décrivons la figure 3.2. La courbe en trait plein représente

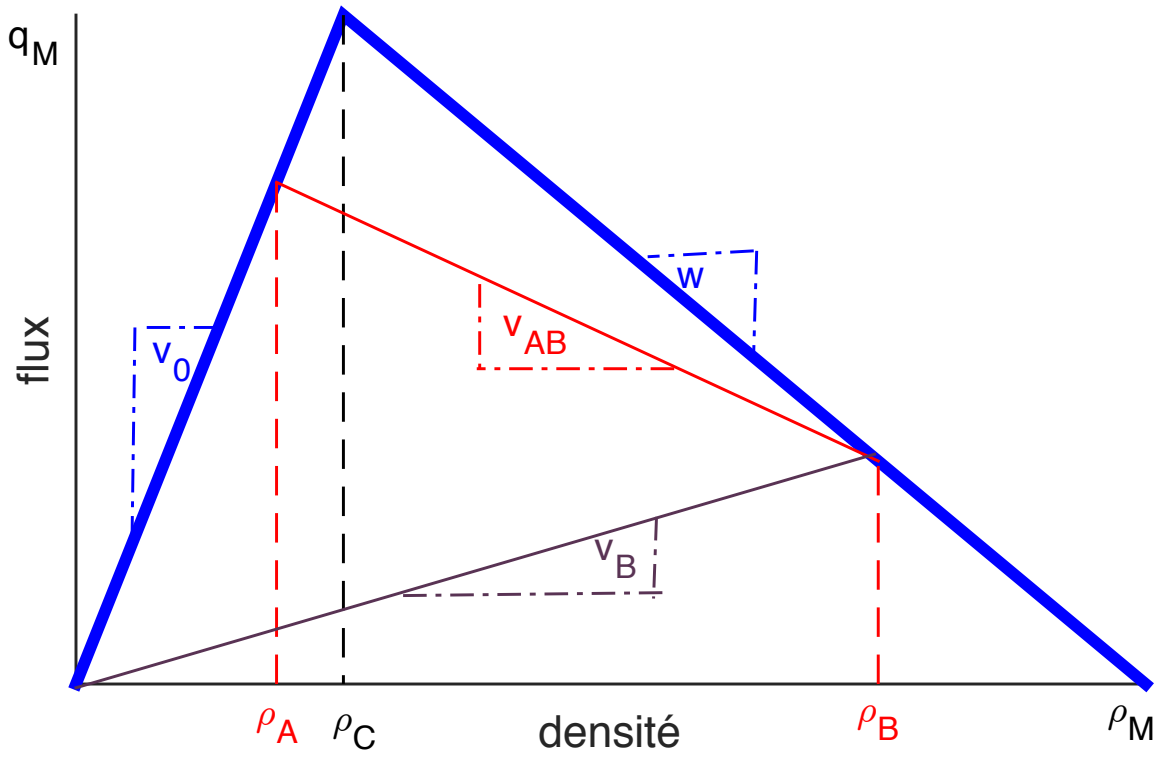

Figure 3.1 Diagramme fondamental

une onde de choc en densité au temps t. La courbe en pointillés représente cette même onde de choc après propagation de dx durant un temps dt. La vitesse V_{AB} est la vitesse du front de l'onde. Notons que nous pouvons avoir pour $\rho_B > \rho_A$, les flux $q_B > q_A$, mais aussi $q_A > q_B$. L'arrangement des densités de donne a priori pas le sens de la vitesse de propagation de l'onde.

Diagramme fondamental empirique

Nous pouvons déduire empiriquement les paramètres du diagramme fondamental pour une certaine route à l'aide de données mesurées. Cependant, nous observons que ces paramètres varient dans le temps, en fonction de l'heure de la journée ou de la portion observée. En effet, les conditions propres à la route, virages, visibilité, intersections, mais aussi l'ensoleillement ont des impacts non négligeable sur la façon générale de conduire, donc ont des effets sur le diagramme fondamental. Notons également que certaines journées et certaines heures connaissent un trafic plus dense en poids lourds, ce qui amène également à une modification

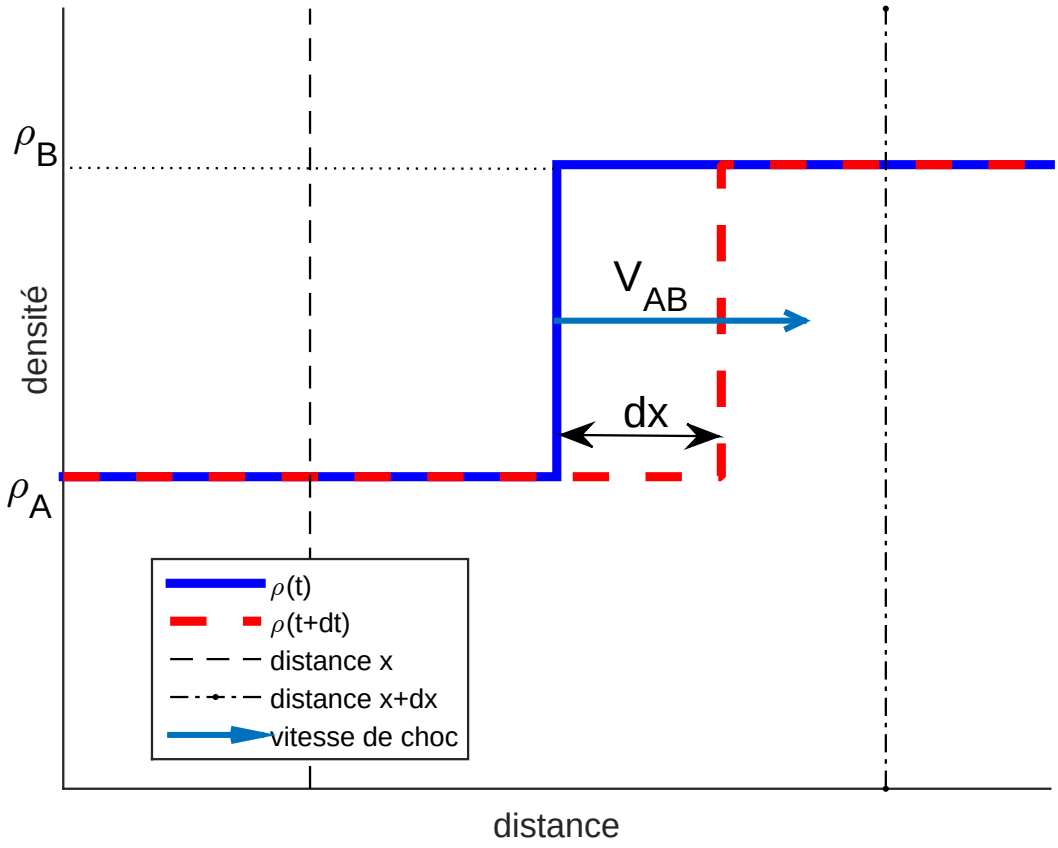

Figure 3.2 Propagation d'une onde de choc.

du diagramme. Les conditions météorologiques ont également un impact qui peut être mesuré puis déduit si nous disposons des données.

Le paragraphe ci-dessous est une citation traduite de (Dervisoglu et al., 2009) sur l'obtention empirique du diagramme fondamental :

Après que la densité critique (la densité au sommet du diagramme) eu été déterminée, les points flux-densité avec une densité supérieure à la densité critique (les points à droite du sommet) sont partionnés le long de l'axe horizontal (l'axe des densités) en intervalles qui ne s'intersectent pas contenant 10 points chacun. Horizontalement, chaque intervalle est caractérisé par sa "BinDensity", la moyenne des 10 valeurs de densités dans l'intervalle. Verticallement, chaque intervalle est caractérisé par "BinFlow", le plus grand flux non aberrant parmi les 10 valeurs de flux dans l'intervalle. Formellement, ce plus grand flux non aberrant

est déterminé comme suit :

$$Bin = \{f_1, f_2, \cdots, f_{10}\}$$
$$BinFlow = \max_{f_i} (f_i | f_i \in Bin, f_i < Q3 + 1.5IQR)$$

où $f_1, \cdots f_{10}$ *et* f_i *sont les valeurs de flux à l'intérieur de tels intervalles,* $Q3$ *est le* 75^e *percentile des données dans l'intervalle et* IQR *est définit comme la différence entre le* 25^e *percentile et le* 75^e *percentile des données.*

Pour trouver les deux segments de droites du diagramme, nous faisons ensuite deux régression par moindres carrés sur les paires BinDensity-BinFlow en contraignant de passer à travers le point de capacité maximal du diagramme. Il est utile de noter que les taux d'occupations (équivalent à une mesure de densité) ne sont pas uniques, et nous devons sélectionner le BinFlow maximum pour un Bin donné. Nous prenons aussi garde qu'au moins quatre valeurs soient disponibles pour la régression.

Résultats Les résultats numériques pour le jeu de données Mobile Century peuvent être trouvé dans le tableau 3.1. Certains paramètres n'ont pas pu être calculés à cause d'un manque de données dans certaines régions espace-temps. Nous pouvons cependant remplacer ces valeurs avec l'interpolation spatio-temporelle des autres valeurs trouvées. Un exemple de diagramme fondamental est montré sur la Figure 3.3.

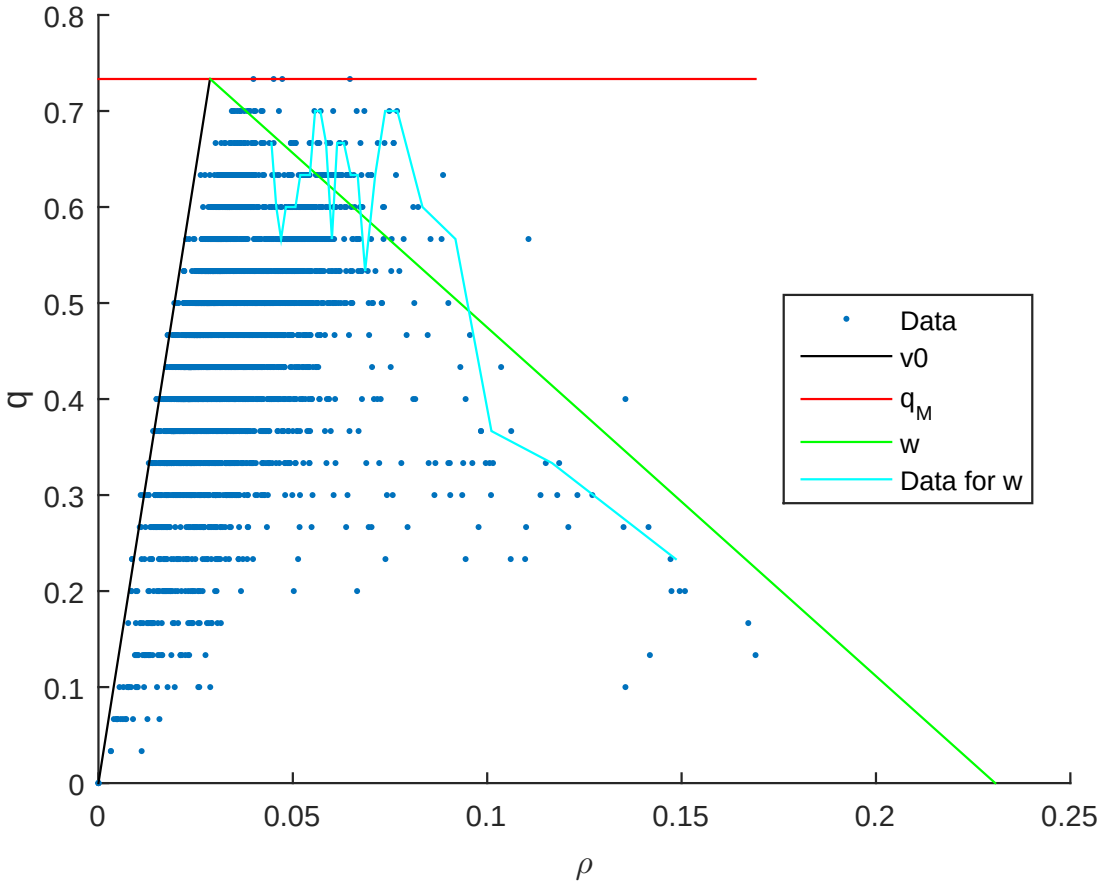

Figure 3.3 Diagramme fondamental du 27e capteur du jeu de données Mobile Century. La courbe *Data for w* représente les maximums des flux non aberrants.

Capteur	Vitesse libre $v0$ [miles/hr]	Flux max q_M [veh/min]	libre-congestionné Vitesse d'one de choc $-w$ [m/s]
1	57.56	50	NaN
2	59.46	50	NaN
3	65.60	54	-15.07
4	57.60	60	-16.64
5	60.60	46	-17.51
6	56.75	48	-16.20
7	56.96	48	-16.45
8	56.74	52	-18.97
9	57.92	50	-13.75
10	59.19	62	-29.68
11	56.85	50	-17.16
12	56.52	54	-23.15
13	56.70	50	-18.01
14	60.59	50	-22.91
15	61.18	48	-15.01
16	66.30	44	-12.28
17	NaN	46	NaN
18	56.79	46	-13.24
19	61.27	40	-8.78
20	56.20	42	-9.95
21	56.87	62	-11.05
22	57.18	44	-7.98
23	56.97	50	-29.75
24	58.21	58	NaN
25	56.74	50	-6.32
26	56.58	46	-11.56

Tableau 3.1 Paramètres du diagramme fondamental.

Diagramme fondamental inversible

Pour un diagramme fondamental triangulaire, nous avons une vitesse constante v_0 pour les densités $0 \leq \rho \leq \rho_C$. Nous ne pouvons pas alors inverser la relation densité-vitesse dans cette région. Or nous avons besoin, pour l'utilisation des données GPS, de pouvoir revenir

à la densité à partir des vitesses. Un autre diagramme utilisé dans la littérature (Work et al., 2008) est le diagramme parabolique. La concavité stricte du diagramme parabolique rend possible l'inversion de la relation densité-vitesse qui est une droite affine (le diagramme fondamental triangulaire étant non concave non stricte et non inversible pour les vitesses). L'inconvénient de ce diagramme fondamental est qu'il est symétrique et modélise mal la réalité. Nous proposons un nouveau diagramme fondamental hybride, voir les figures 3.4 et 3.5, reprenant le diagramme fondamental triangulaire pour la partie congestionnée du diagramme, et le diagramme parabolique pour la partie trafic libre.

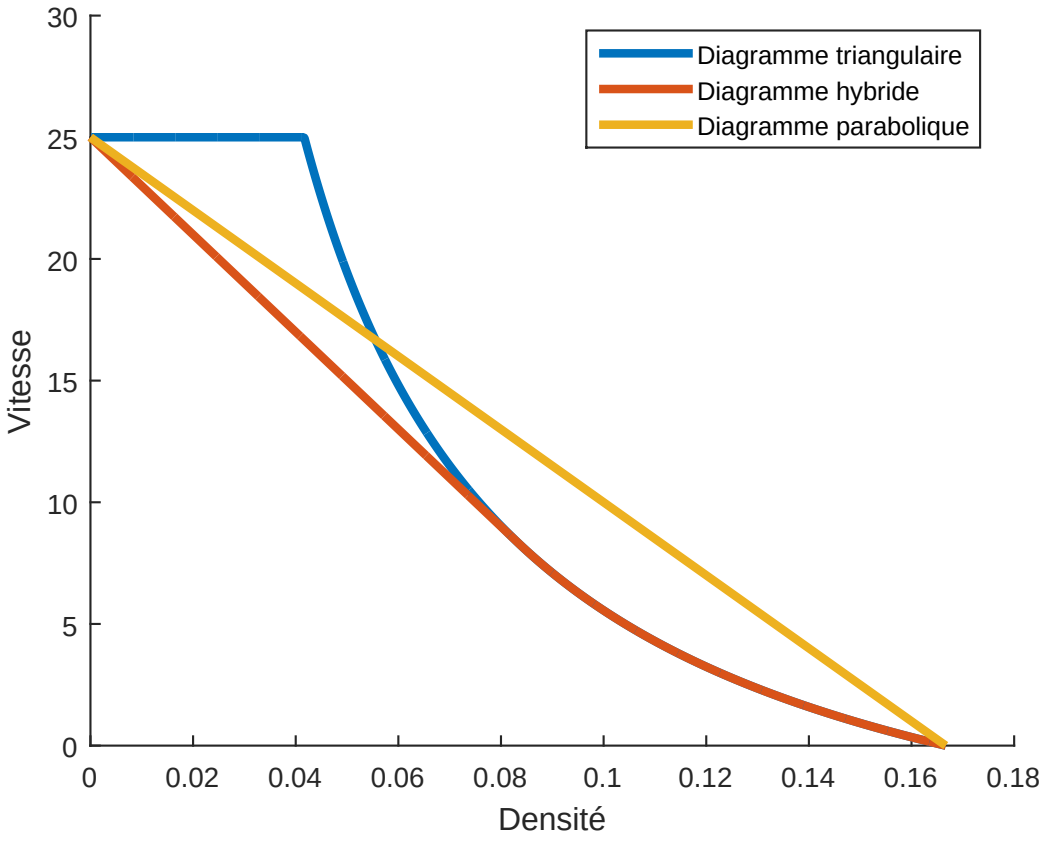

Figure 3.4 Diagrammes densité-vitesse.

Nous observons (Figure 3.5) que nous perdons la caractéristique de flux maximum pour le diagramme hybride. Nous pouvons faire le choix de conserver le flux maximum et de modifier la vitesse w par exemple.

Nous créons le diagramme fondamental hybride à partir du diagramme densité-vitesse issu du diagramme fondamental triangulaire. Nous gardons la partie de droite (de densité ρ supérieure

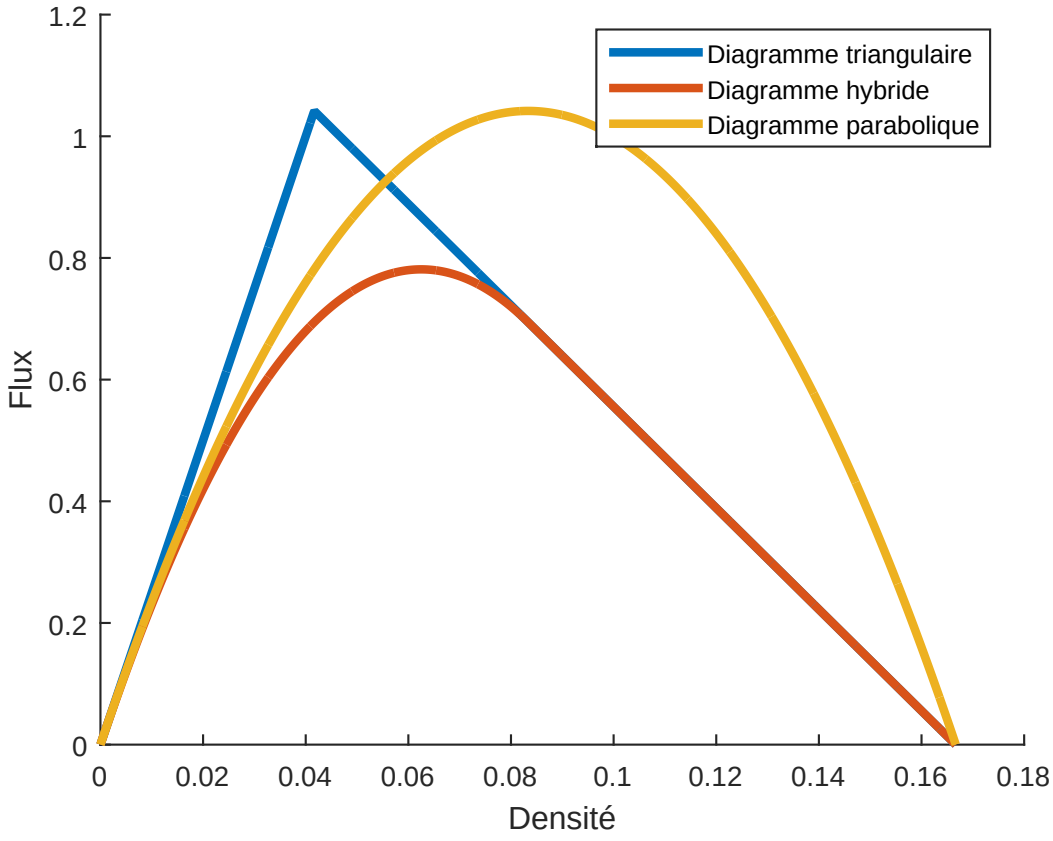

Figure 3.5 Diagrammes fondamentaux densité-flux.

à la densité critique ρ_C) du diagramme et prolongeons à gauche par une droite tangente passant par le point $(0, v_0)$. L'équation de la vitesse $v(\rho)$ pour le diagramme fondamental triangulaire est, pour $\rho \geq \rho_C$

$$\begin{aligned} v(\rho) &= \frac{q(\rho)}{\rho} \\ &= \frac{-w(\rho - \rho_M)}{\rho}. \end{aligned}$$

La droite $y(\rho)$ tangente au point $\rho = r$ à pour équation

$$\begin{aligned} y(\rho) &= \left.\frac{\mathrm{d}v}{\mathrm{d}\rho}\right|_{\rho=r} (\rho - r) + v(r) \\ &= -\frac{w\left(r^2 - 2\rho_M r + \rho_M \rho\right)}{r^2} \end{aligned}$$

Nous pouvons résoudre la condition à la limite $v(\rho = 0) = v_0$ pour obtenir le point tangent (ρ_h, v_h) entre la droite affine et le modèle initial

$$\begin{aligned} \rho_h &= \frac{2w\rho_M}{v_0 + w} \\ v_h &= \frac{1}{2}(v_0 - w). \end{aligned}$$

Nous utiliserons donc le diagramme hybride pour déduire les densités à partir des vitesses, mais nous utiliserons le diagramme fondamental triangulaire pour la dynamique du modèle. En effet, les diagrammes fondamentaux avec courbure non nulle produisent des ondes de choc de densité (voir figure 3.6 et ses explications). La vitesse de groupe v_g (la vitesse de propagation de l'enveloppe d'une onde) est, en se servant de la vitesse de propagation de deux fronts (voir équation (3.2)) infinitésimalement espacés

$$\begin{aligned} v_g &= \lim_{A \to B} v_{AB} \\ &= \lim_{A \to B} \frac{q(B) - q(A)}{\rho_B - \rho_A} \\ &= \frac{\mathrm{d}q}{\mathrm{d}\rho}. \end{aligned}$$

Lorsque le diagramme fondamental n'est pas une fonction affine par morceau ($\frac{\mathrm{d}q}{\mathrm{d}\rho}$=constante), la vitesse de propagation de l'onde dépend de la densité et il y a déformation de l'onde, alors que pour un diagramme fondamental triangulaire, l'onde se déplace en bloc. Le modèle parabolique est donc dispersif, et nous pouvons voir figure 3.6 le déplacement d'une onde pour un diagramme fondamental parabolique. Le profil des densités (l'onde) à l'instant t_1 est inférieur à la densité critique de flux maximal du diagramme fondamental parabolique, avec la densité maximal du profil égale à cette densité critique. La pente du flux au point critique est nulle, donc cette partie du profil ne se déplace pas ou peu, tandis que les densités plus faibles se déplacent d'autant plus rapidement (la dérivée du diagramme fondamental augmente en parcourant la courbe du point critique vers zéro). Plus la concentration est faible, plus la vitesse est rapide. Il advient que le haut du profil se déplace plus lentement que le bas ce qui a tendance à rendre le profil de plus en plus abrupte, il y a finalement création de l'onde de choc, qui n'est pas réaliste physiquement, d'où l'intérêt de garder le diagramme triangulaire pour la propagation de la densité, qui ne crée pas d'ondes de chocs.

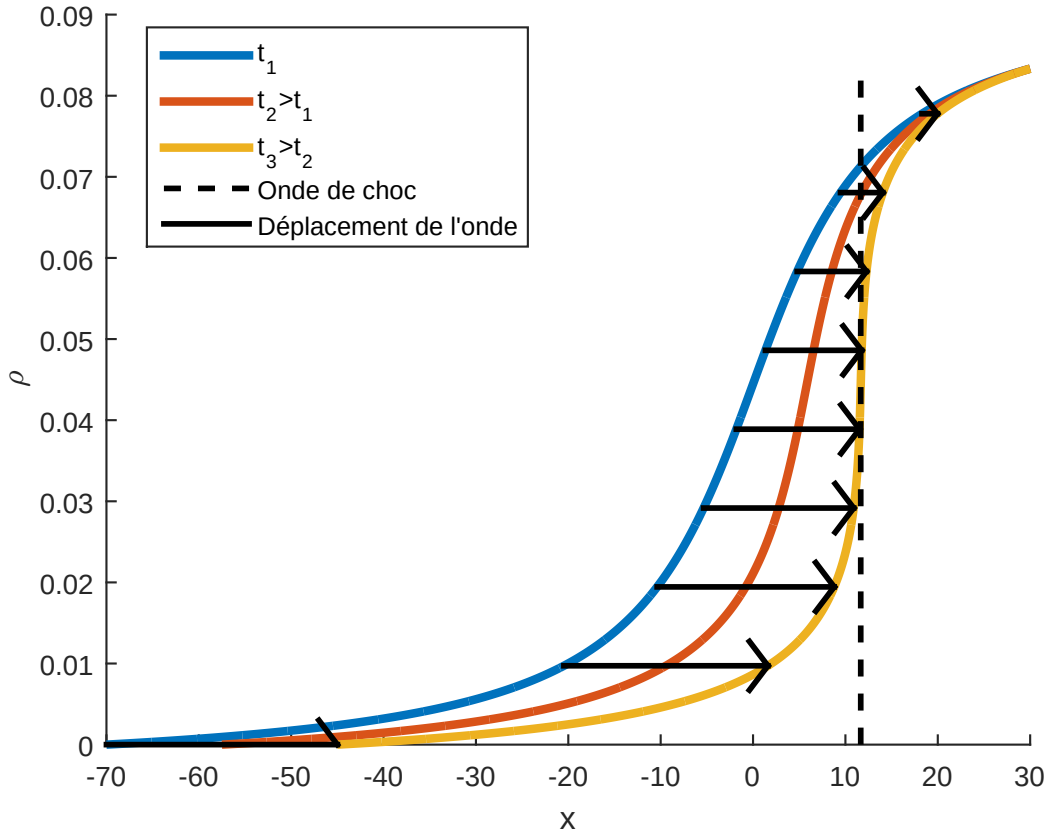

Figure 3.6 Processus de création d'une onde de choc pour diagramme fondamental parabolique.

Nous avons alors la fonction pour déduire la densité ρ en fonction de la vitesse V

$$\rho(V) = \begin{cases} \rho_M \frac{w}{V+w} & V \leq \frac{1}{2}(v_0 - w) \\ -4\rho_M \frac{w(V-V_0)}{(V_0+w)^2} & V > \frac{1}{2}(v_0 - w). \end{cases} \tag{3.3}$$

Nous voyons donc que la propagation d'une onde pour un diagramme fondamental non triangulaire provoque de manière autonome une onde de choc. Tandis que pour un diagramme fondamental triangulaire, l'onde peut se déplacer en bloc par morceaux (avec deux vitesses $v0$ ou w) et donc ne peut pas créer spontanément sans interventions extérieures une onde de choc. Cependant, une onde de choc déjà présente se propagera en bloc.

3.1.3 Modèle hydrodynamique

Cette section décrit l'obtention de l'équation hydrodynamique de conservation de "la masse" (ici le nombre de voitures).

Équation continue

Faisons un bilan de quantité de voitures sur une portion $\mathrm{d}x$ de route entre les instant t et $t + \mathrm{d}t$ à l'emplacement x, en notant $\mathrm{d}n_0$ la quantité initiale de voitures.

$$\begin{aligned}
\texttt{à l'instant } t\texttt{:} \quad & \rho(x,t)\,\mathrm{d}x && = \mathrm{d}n_0 \\
\texttt{à l'instant } t+\mathrm{d}t\texttt{:} \quad & \rho(x,t+\mathrm{d}t)dx && = \mathrm{d}n_0 + q(x,t)\,\mathrm{d}t - q(x+\mathrm{d}x,t)\,\mathrm{d}t \\
\texttt{différence:} \quad & \tfrac{\partial \rho}{\partial t}\,\mathrm{d}t\,\mathrm{d}x && = -\tfrac{\partial q}{\partial x}\,\mathrm{d}x\,\mathrm{d}t \\
\texttt{équation de conservation:} \quad & \tfrac{\partial \rho}{\partial t} + \tfrac{\partial q}{\partial x} = 0
\end{aligned}$$

Équation discrète

Nous décrivons l'évolution du trafic en temps discret, avec des périodes de temps de longueur τ. La route est discrétisée en N cellules de longueurs Δx_j, $1 \leq j \leq N$, avec la densité $\rho_{j,t}$ pour la cellule j à la période $t \geq 0$ assumée approximativement constante dans une cellule. Soit λ_j le nombre de voies et $f_{j,t}$ le *flux numérique* durant la période t (voir définition ci-dessous) à l'interface entre les cellules $j-1$ et j. La discrétisation perturbée de l'équation continue 3.1.3 de la densité suit une équation de continuité (Treiber and Kesting, 2013)

$$\rho_{j,t+1} = \rho_{j,t} + \left(\frac{\lambda_j}{\lambda_{j+1}} f_{j,t} - f_{j+1,t} \right) \frac{\tau}{\Delta x_j} + \nu_{j,t}, \tag{3.4}$$

Où ν_j est un bruit blanc gaussien dont la variance capture par exemple les erreurs dues aux hypothèses du diagramme fondamental. Dans le Modèle de Transmission par Cellule (MTC) (Daganzo, 1994), le flux numérique compatible avec le diagramme fondamental (3.1) est

$$f_{j,t} := \min\{v_0 \rho_{j-1,t}, v_0 \rho_C, -w(\rho_M - \rho_{j,t})\},$$

tel que (3.4) est alors une description stochastique, linéaire par morceaux de la représentation d'état de la dynamique du trafic. En effet, le flux à l'interface $[j, j+1]$ est limité soit par la capacité de flux de la cellule j (par exemple si la cellule $(j+1)$ était vide), soit par la capacité de réception de flux de la cellule $(j+1)$, voir (Treiber and Kesting, 2013).

Présentation d'une estimation espace-temps de densité

Le résultat d'une estimation de densité pour une route sans embranchement peut se représenter dans un diagramme temps-distance, voir Fig 3.7. Nous choisissons de représenter le temps en ordonnée et la distance en abscisse. L'état de la route à un instant donné est donc représenté par une ligne verticale, qui se déplace avec le temps. Nous pouvons y voir des fronts entre zones de faibles (couleur bleue) et fortes congestions (couleur rouge) se déplacer en amont du trafic (en remontant le trafic) à mesure que les voitures s'amassent sur la congestion et que les voitures en aval de la congestion se libèrent. Nous pouvons aussi distinguer la propagation des densités vers l'aval du trafic dans les zones de faibles densités, c'est à dire lorsque les voitures ne se font pas freiner entre elles à cause d'une forte densité, l'état de la route (la densité) se propage avec le flot des voitures. Certaines trajectoires de véhicules ont été tracées par des lignes noires (ces trajectoires seront utilisé pour la validation du mécanisme différentiellement confidentiel utilisant les données GPS). Dans les zones de faibles densités, la forte pente de ces trajectoires indique une vitesse élevée. Lorsque les véhicules entrent dans une zone de forte densité (en rouge), la pente presque nulle de la trajectoire indique des véhicules presque à l'arrêt.

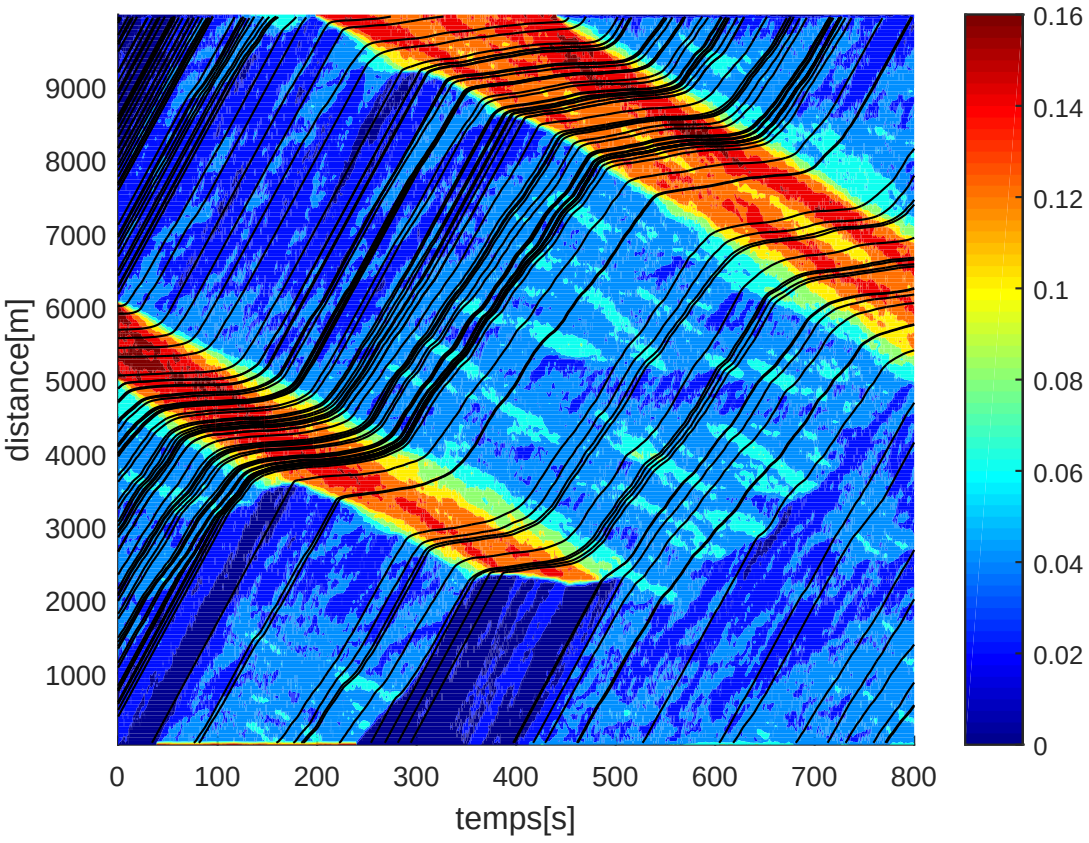

Figure 3.7 Exemple de représentation d'estimé de densité plus trajectoires de véhicules.

3.2 Les mesures de capteurs

Les capteurs statiques standards sont des boucles d'induction magnétique (Treiber and Kesting, 2013) placées à certaines positions fixes le long de la route, assumées être à la frontière entre deux cellules adjacentes (voir figure 3.8).

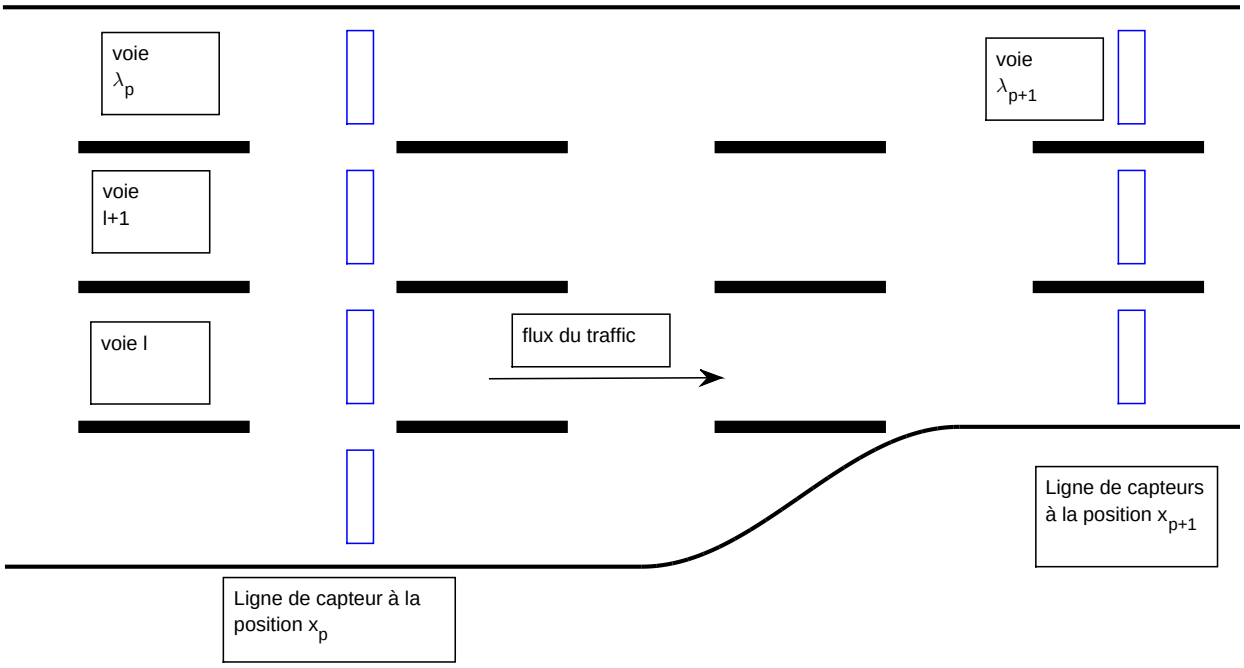

Figure 3.8 Schéma de route et positionnement des lignes de capteurs.

Un capteur à la frontière entre les cellules $(p-1)$ et (p) envoie avec une certaine fréquence les comptes $c^l_{p,t}$ des véhicules sur la voie l qui franchissent le capteur durant la dernière période d'échantillonnage. Nous notons $c_{p,t}$ les comptes moyens sur les voies. Les capteurs par boucle d'induction magnétique envoient également les taux d'occupation $o^l_{p,t} \in [0,1]$, qui est le pourcentage de temps durant lequel un véhicule était au dessus du capteur, sur la voie l durant la dernière période d'échantillonnage. De plus, certains véhicules sont équipés de systèmes capable de transmettre leurs vitesses et positions quand sollicités. Pour des raisons de confidentialité, nous utilisons une stratégie d'échantillonnage basée sur la position, introduite dans (Hoh et al., 2008) et appelée Lignes de Passage Virtuelles (LPV), où les véhicules envoient leurs vitesses seulement quand ils traversent des points de passage spécifiques le long de la route, encore une fois assumés être à l'interface entre deux cellules adjacentes. Dans (Le Ny et al., 2014), une stratégie d'estimation a été introduite qui reposait majoritairement sur la mesure des comptes pour l'estimation de la densité, et ne considérait pas les mesures de vitesses. À l'inverse, dans cette étude nous considérerons majoritairement les mesures d'occupation et de vitesses.

Nous considérons les modèles de mesure suivants à l'interface $(p-1) \to p$ avec un capteur statique et/ou un LPV.

$$\frac{1}{g_p \lambda_p} \sum_{l=1}^{\lambda_p} o_{p,t}^l = \rho_{p,t} + \mu_{p,t}^o, \tag{3.5}$$

$$\frac{1}{\lambda_p} \sum_{l=1}^{\lambda_p} c_{p,t}^l = c_{p,t} \tag{3.6}$$

$$\ln V_{p,t} = \ln\left(\frac{\rho_M}{\rho} - 1\right) - \ln w + \mu_{p,t}^v \text{ si } V_{p,t} < v_0, \tag{3.7}$$

Les différentes grandeurs étant

- g_p : le dénommé g-facteur (Jia et al., 2001) (longueur moyenne effective d'un véhicule à l'emplacement du capteur)
- μ_p^o, μ_p^v : des bruits gaussiens blancs
- λ_p le nombre de voies à l'emplacement p
- $V_{p,t}$ la vitesse à l'emplacement p au temps t. Le modèle de mesure de vitesse de l'équation (3.7) a été obtenu en inversant la deuxième relation dans (3.1). Les mesures de vitesses du trafic $V_{p,t}$ sont définies lorsqu'un véhicule traverse une LPV au temps t, comme les moyennes géométriques sur les derniers n véhicules qui ont envoyé leurs vitesses en traversant la LPV, c'est à dire,

$$V_{p,t} = \left(\prod_{i=1}^{n} v_p^{(i)}\right)^{\frac{1}{n}} \tag{3.8}$$

  Avec $v_p^{(i)}$ la $i^{\text{ème}}$ vitesse de la séquence $v_p = \{v_p^{(1)}, \cdots, v_p^{(n)}\}$ regroupant les vitesses des n derniers véhicules ayant traversé le capteur p. La valeur de n est discutée dans la section 4.2.1.

La figure 3.9 montre en rouge la détection du capteur d'occupation, pour le passage de trois voitures. En noir sont tracés les limites avant et arrière des voitures. La sortie du capteur est le ratio de temps pendant lequel le capteur détecte une voiture sur le temps ΔT de la période d'échantillonnage.

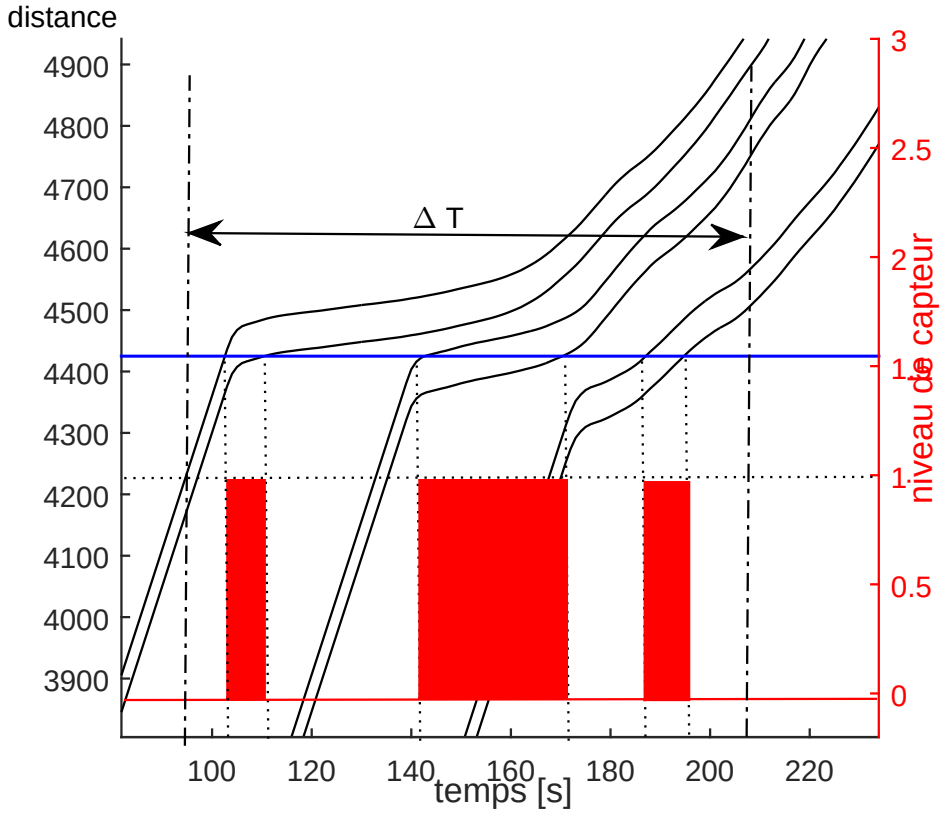

Figure 3.9 Sortie de capteur à induction pour des trajectoires de véhicules

L'estimation de la densité par l'équation (3.5) provient de l'approximation de vitesse quasi-constante durant l'intervalle de mesure et du modèle continu :

$$\rho_{p,t} + \mu^o_{p,t} = \frac{n_{p,t}}{\Delta X_{p,t}}$$

avec $n_{p,t}$ le nombre de voitures contenu dans l'espace $\Delta X_{p,t}$,

$$n_{p,t} = \frac{V_{p,t} \sum_{l=1}^{\lambda_p} o^l_{p,t}}{g}$$
$$\Delta X_{p,t} = V_{p,t} \Delta T$$

Les choix d'un modèle logarithmique dans (3.7) et d'une moyenne géométrique pour $V_{p,t}$ se

trouvent être pratiques pour notre mécanisme différentiellement confidentiel. Nous voulons maintenant produire un estimé de la densité ρ dans chaque cellule en temps réel, à partir du modèle dynamique (3.4) et des mesures (3.5), (3.7). Cependant, ces mesures saisissent des informations confidentielles sensibles sur la trajectoires des conducteurs, et donc un objectif supplémentaire ici est d'imposer pour l'estimé publié publiquement des garanties formelles de confidentialité pour le conducteur.

3.3 L'estimation différentiellement confidentielle

Nous avons pour but de produire un estimé différentiellement confidentiel (Dwork et al., 2006) du trafic en temps réel. Nous commençons avec une relation symétrique binaire sur notre espace de séquences de mesures $\{\mathbf{y}_t\}_{t\geq 0}$, avec $\mathbf{y}_t = \{o^l_{p,t}, v^{(i)}_{p,t}\}_{p,l,i}$, appelée la *relation d'adjacence* (voir définition 1), et qui intuitivement met en relation les séquences qui diffèrent par les données d'un individu unique. L'assainissement à travers un mécanisme différentiellement confidentiel devrait rendre difficile pour un adversaire de décider laquelle de deux séquences adjacentes $\{\mathbf{y}_t\}_{t\geq 0}$ ou $\{\tilde{\mathbf{y}}_t\}_{t\geq 0}$ a été utilisée pour produire un estimé donné.

Definition 1 (Adjacence pour le trafic) *Deux jeux de données sont adjacents s'il sont issus d'un même trafic, à l'exception de la trajectoire d'une voiture.*

Definition 2 ((ϵ, δ)-confidentialité différentielle) *Soit $\mathcal{Y}$ un espace équipé avec une relation symétrique binaire dénommée Adj, et soit $(\mathsf{R}, \mathcal{M})$ un espace mesurable. Soit $\epsilon, \delta > 0$. Une fonction randomisée M de $\mathcal{Y}$ dans R est (ϵ, δ)-différentiellement confidentielle pour Adj si pour tout $\mathbf{y}, \tilde{\mathbf{y}} \in \mathcal{Y}$ tel que Adj$(\mathbf{y}, \tilde{\mathbf{y}})$, nous avons*

$$P(M(\mathbf{y}) \in S) \leq e^{\epsilon} P(M(\tilde{\mathbf{y}}) \in S) + \delta, \forall S \in \mathcal{M}.$$

Pour mettre la Définition 2 dans le contexte de notre application, M est notre estimateur différentiellement confidentiel, qui doit nécessairement randomiser sa sortie (en pratique, en ajoutant du bruit) pour satisfaire la définition de la confidentialité différentielle, de telle manière que la distribution pour tous les signaux d'entrés est peu sensible aux différences entre signaux d'entrée adjacents.

3.3.1 Résistance au post-traitement

Une propriété cruciale de la confidentialité différentielle stipule que la confidentialité différentielle est *résistante au post-traitement*, c'est à dire manipuler un signal différentiellement

confidentiel ne peut pas affaiblir les garanties de confidentialité, tant que le signal d'entrée n'est pas ré-utilisé (voir Le Ny and Pappas, 2014, Theorem 1).

Theorem 1 *Soit $\mathcal{M}_1$ un mécanisme (ϵ, δ)-différentiellement confidentiel, et un autre mécanisme $\mathcal{M}_2$ ne dépendant pas du jeu de données d, alors $\mathcal{M}_2(\mathcal{M}_1(d))$ est (ϵ, δ)-différentiellement confidentiel.*

C'est à dire qu'un mécanisme (un filtre par exemple) accédant indirectement aux données rendues différentiellement confidentielles ne peut pas affaiblir les garanties de confidentialité.

3.3.2 Théorème de composition de mécanismes

Si nous utilisons les résultats de plusieurs mécanismes différentiellement confidentiels indépendant, chaque mécanismes introduisant du bruit de façon indépendante des autres, alors nous avons un théorème sur la garantie de confidentialité d'un algorithme utilisant toutes les sorties de ces mécanismes (Dwork and Roth, 2014).

Theorem 2 *Soient $\mathcal{M}_1, \cdots, \mathcal{M}_n$, n mécanismes indépendant, où $\mathcal{M}_i$ est (ϵ_i, δ_i)-différentiellement confidentiel. Alors le mécanisme $\mathcal{M} = (\mathcal{M}_1, \cdots, \mathcal{M}_n)$ est $(\sum_{i=1}^n \epsilon_i, \sum_{i=1}^n \delta_i)$-différentiellement confidentiel.*

3.3.3 Mécanisme gaussien

Nous présentons ici le mécanisme différentiellement confidentiel gaussien. Nous utilisons la norme-2 : soit $x \in \mathbb{R}^k$

$$\|x\|_2 = \sqrt{\left(\sum_{i=1}^{k} |x_i|^2\right)}$$

Soit deux jeux de données d et d', et un relation d'adjacence Adj. Nous définissons la sensibilité Δ_q d'une requête $q(d)$ comme

$$\Delta_q = \max_{d,d' : \mathtt{Adj}(d,d')} \|q(d) - q(d')\|_2$$

Définissons $K = \mathcal{Q}^{-1}(\delta)$ pour $\mathcal{Q}(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-\frac{u^2}{2}} du$ et $\kappa_{\delta,\epsilon} = \frac{1}{2\epsilon}\left(K + \sqrt{K^2 + 2\epsilon}\right)$.
La publication assainie est notée $\bar{q}(d)$. Nous avons le théorème suivant (Dwork, 2006)

Theorem 3 *Un mécanisme qui publie* $\bar{q}(d) = q(d) + w$*, où* w *sont des variables aléatoires gaussiennes iid avec pour variance* $\kappa^2_{\delta,\epsilon}\Delta^2_q$*, est* (ϵ, δ) *différentiellement confidentiel.*

CHAPITRE 4 ASSAINISSEMENT DES DONNÉES DE TRAFIC

Dans ce chapitre nous présentons les méthodes pour assainir les données. Assainir signifiant rendre les données différentiellement confidentielles au sens de la définition 2. Il n'y a pas de méthode unique, et celles-ci doivent être adaptées aux types de mesure. De plus nous verrons que nous pouvons améliorer l'utilité en diminuant la composante aléatoire des données assainies en ne protégeant pas des cas théoriquement possibles, mais qui ne surviennent pas en conditions réelles de trafic.

4.1 Mesures de taux d'occupation

Deux bases de données $\mathbf{y}$ et $\tilde{\mathbf{y}}$ sont dites adjacentes si elles sont générées par le même trafic à l'exception de la trajectoire d'une voiture unique. Notons qu'une voiture ne peut que se déplacer vers l'avant et se situe sur une voie spécifique à toute limite de cellule, donc une voiture ne peut influencer $o^l_{p,t}$ pour un p donné pour seulement une valeur unique de t et l, et elle ne peut traverser la ligne de passage virtuelle (LPV) seulement qu'une fois. Le taux d'occupation $o^l_{p,t}$ mesuré par un capteur statique est la somme des taux individuels d'occupation, pour les voitures passant au dessus du capteur durant la période t. A priori, il est possible qu'un seul véhicule ait une forte influence sur une mesure d'occupation, puisque pour une situation de trafic congestionné, le véhicule pourrait stationner sur le capteur pour la totalité de la période d'échantillonnage, tel que $o^l_{p,t} = 1$, tandis que sans ce véhicule $o^l_{p,t} = 0$ pour deux bases de données adjacentes.

Nous affirmons cependant que cette situation est peu probable et décidons de protéger seulement les voitures qui ont une influence bornée sur les mesures d'occupations, concept que nous intégrons dans notre relation d'adjacence. Entre bases de données adjacentes, toutes les mesures d'occupation doivent être identiques, à l'exception du fait que par le changement d'une trajectoire, la voiture correspondante peut traverser une ligne de capteurs à un temps et une voie différents. Quand il y a une différence entre des $o^l_{p,t}$ et $\tilde{o}^l_{p,t}$, due à l'influence d'une telle voiture, nous assumons cette différence bornée, c'est à dire,

$$|o^l_{p,t} - \tilde{o}^l_{p,t}| \leq \alpha$$

pour un α tel que discuté ci-dessous. Considérant maintenant la séquence avec les éléments

$O_{p,t} = \frac{1}{\lambda_p} \sum_{l=1}^{\lambda_p} o_{p,t}^l$ apparaissant dans (3.5), une analyse directe montre que

$$\begin{aligned} \|O - \tilde{O}\|_2^2 &= \sum_{p,t} \left(O_{p,t} - \tilde{O}_{p,t} \right)^2 \\ &= \sum_{p,t} \left(\frac{1}{\lambda_p} \sum_{l}^{\lambda_p} o_{p,t}^l - \tilde{o}_{p,t}^l \right)^2 \\ &\leq 2\alpha^2 \sum_{p=1}^{P_s} \frac{1}{\lambda_p^2} =: \Delta_o \end{aligned}$$

pour des séquences adjacentes, où P_s est le nombre d'interfaces pour lesquelles nous avons des détecteurs statiques. La présence du facteur 2 vient du fait que modifier une trajectoire de voiture équivaut à supprimer la trajectoire, puis d'en créer une nouvelle, ayant ainsi un double impact.

4.1.1 Choix de α

Dans le but de fixer α en particulier, nous raisonnons sur un modèle théorique de trafic microscopique constitué de voitures identiques et uniformément espacées, respectant le diagramme fondamental, et étudions l'impact d'une voiture sur le taux d'occupation. Un trafic idéal stationnaire respectant le diagramme fondamental est entièrement caractérisé par sa densité ρ. Pour un trafic idéal donné, le capteur d'occupation peut donner des mesures différentes en fonction du moment ou le capteur commence l'acquisition de la mesure. Ce phénomène est particulièrement important à faibles vitesses. En effet considérons un trafic à vitesse quasi-nulle, avec une forte densité ρ proche de ρ_M. Si la mesure du taux d'occupation démarre juste après le passage d'une voiture, aucune voiture ne retraversera le capteur et la mesure annoncée par le capteur sera 0. En revanche si la mesure démarre juste au moment où une voiture se présente au capteur, et que cette voiture reste au dessus du capteur pendant toute la durée d'acquisition, la mesure de taux d'occupation annoncée par le capteur sera de 1. Nous pouvons donc définir deux configurations. La première configuration est lorsque le début de la mesure commence juste avant le passage d'une voiture. Voir figure 4.1. Cette configuration donne un taux d'occupation mesuré maximum parmi toutes les configurations (configurations étant caractérisée par le début de la prise de mesure dans le trafic). La configuration 2 (voir figure 4.1) correspond à la configuration où la mesure débute juste après le passage d'une voiture. Cette configuration donne la mesure minimum du taux d'occupation pour un trafic donné.

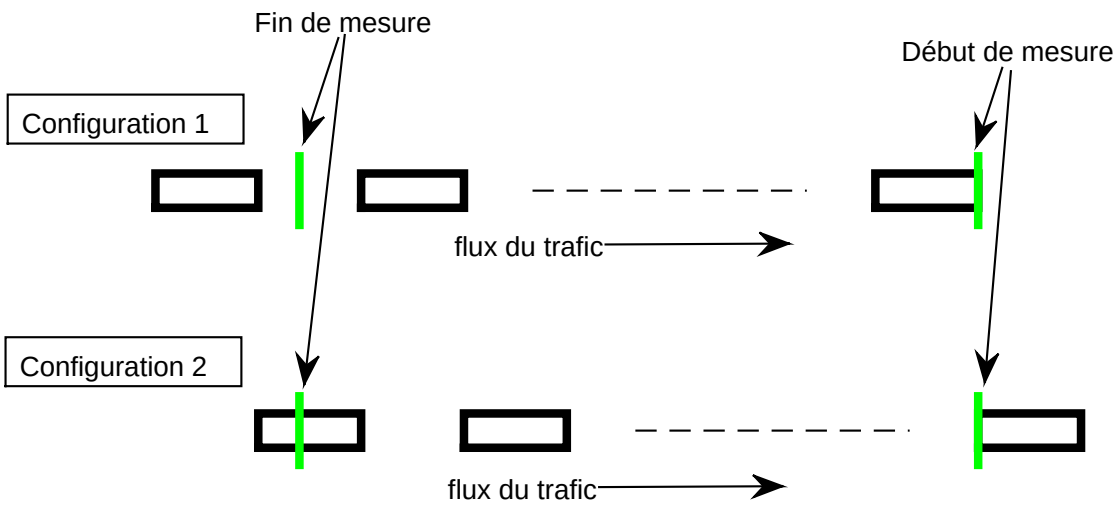

Figure 4.1 Configuration 1 et configuration 2 pour un même trafic

Les taux d'occupations mesurés dans la configuration 1, respectivement dans la configuration 2, sont nommés, $\mathcal{O}_1$, respectivement $\mathcal{O}_2$. Parce que la vitesse est constante durant les mesures, nous pouvons calculer le taux d'occupation comme le rapport des longueurs totales de voitures passant au dessus du capteur, sur la longueur total de trafic passant au dessus du capteur pendant ΔT. Nous appelons `floor(x)` la fonction partie entière de x. Nous définissons $d_\rho = \frac{1}{\rho}$ la distance pare choc avant à pare choc avant de deux voitures qui se suivent immédiatement. Le flux pour un trafic de densité ρ est donné par le diagramme fondamental

$$q_\rho = \min\left(v_0\rho, -w\left(\rho - \rho_M\right)\right)$$

Nous avons la vitesse

$$v_\rho = \min\left(v_0, \frac{q_\rho}{\rho}\right)$$

Le nombre entier de voitures passant au dessus du capteur est donc

$$\texttt{floor}\left(\frac{v_\rho \Delta T}{d_\rho}\right)$$

En appelant `mod(a,b)` la fonction `a modulo b`, la fraction de voiture restante est

$$\min\left(g, \texttt{mod}\left(v_\rho \Delta T, d_\rho\right)\right)$$

Nous avons alors le rapport des longueurs

$$\mathcal{O}_1 = \frac{\texttt{floor}\left(\frac{v_\rho \Delta T}{d_\rho}\right) g + \min\left(g, \texttt{mod}\left(v_\rho \Delta T, d_\rho\right)\right)}{v_\rho \Delta T}$$

Avec un raisonnement similaire, nous avons pour $\mathcal{O}_2$

$$\mathcal{O}_2 = \frac{\texttt{floor}\left(\frac{v_\rho \Delta T}{d_\rho}\right) g + \max\left(0, \texttt{mod}\left(v_\rho \Delta T, d_\rho\right) - (d_\rho - g)\right)}{v_\rho \Delta T}$$

Un trafic adjacent peut avoir une voiture de plus en insertion, si possible, ou une voiture de moins. Nous avons alors quatre cas de d'occupations pour les trafics adjacents $\mathcal{O}'_{10}$, $\mathcal{O}'_{11}$, $\mathcal{O}'_{20}$, $\mathcal{O}'_{21}$, :

$$\begin{aligned}
\mathcal{O}'_{10} &= \min\left(1, \mathcal{O}_1 + \frac{g/\lambda_p}{v_\rho \Delta T}\right) \\
\mathcal{O}'_{11} &= \max\left(0, \mathcal{O}_1 - \frac{g/\lambda_p}{v_\rho \Delta T}\right) \\
\mathcal{O}'_{20} &= \min\left(1, \mathcal{O}_2 + \frac{g/\lambda_p}{v_\rho \Delta T}\right) \\
\mathcal{O}'_{21} &= \max\left(0, \mathcal{O}_2 - \frac{g/\lambda_p}{v_\rho \Delta T}\right)
\end{aligned}$$

Nous avons donc quatre différences de taux d'occupations possibles $\Delta\mathcal{O}_{10}$, $\Delta\mathcal{O}_{11}$, $\Delta\mathcal{O}_{20}$, $\Delta\mathcal{O}_{21}$.

$$\begin{aligned}
\Delta\mathcal{O}_{10} &= |\mathcal{O}_1 - \mathcal{O}'_{10}| \\
\Delta\mathcal{O}_{11} &= |\mathcal{O}_1 - \mathcal{O}'_{11}| \\
\Delta\mathcal{O}_{20} &= |\mathcal{O}_2 - \mathcal{O}'_{20}| \\
\Delta\mathcal{O}_{21} &= |\mathcal{O}_2 - \mathcal{O}'_{21}|
\end{aligned}$$

L'écart maximum de taux d'occupation pour deux trafics idéaux adjacents de densité ρ est donc

$$\Delta\mathcal{O} = \max\left(\Delta\mathcal{O}_{10}, \Delta\mathcal{O}_{11}, \Delta\mathcal{O}_{20}, \Delta\mathcal{O}_{21}\right)$$

Nous pouvons voir à la figure 4.2 que $\Delta\mathcal{O}$ est une fonction croissante de la densité. Étant donné que le mécanisme différentiellement confidentiel ajoute un bruit qui augmente avec $\Delta\mathcal{O}$, nous voulons limiter la protection aux $\delta\mathcal{O}$ faibles, donc aux densités faibles, ce qui aurait pour conséquence d'introduire moins de bruit, donc d'augmenter l'utilité. Cependant, nous voulons que restreindre $\Delta\mathcal{O}$ n'impose pas une relation d'adjacence trop sévère pour les automobilistes. Nous utilisons donc le jeu de données Mobile Century avec l'estimation $\rho = \frac{o}{g}$, et nous souhaitons placer $\Delta\mathcal{O} = \alpha$, pour une densité théorique ρ_α tel que 98% des données estimées de densités soient inférieures à ρ_α. Nous choisissons alors $\alpha = 0.015$, pour $\rho_\alpha = 0.081$, voir figure 4.2. Le rare cas où nous recevons une donnée telle que $\frac{o}{g} \geq \rho_\alpha$, le trafic est en forte congestion et nous pouvons ignorer la donnée, laissant le modèle dynamique propager la congestion.

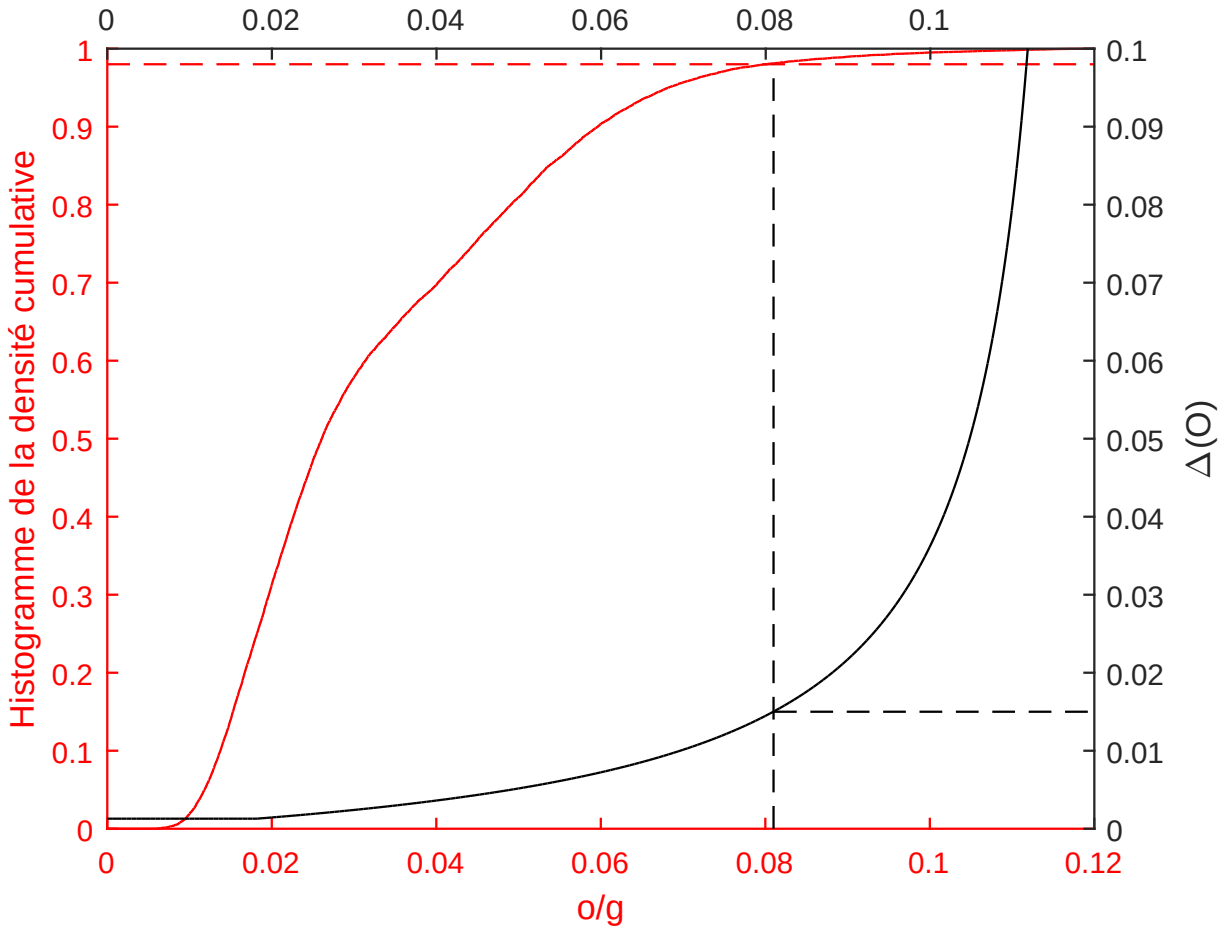

Figure 4.2 Détermination de la borne α.

Nous pouvons alors assainir les données avec le théorème 3 en utilisant la sensibilité $\Delta_o = 2\alpha^2 \sum_{p=1}^{P_s} \frac{1}{\lambda_p^2}$.

4.2 Mesures de vitesses

De manière similaire aux taux d'occupations, nous protégeons seulement contre des variations relatives bornées pour les voitures envoyant leurs données de vitesses quand elles traversent les LPV. Spécifiquement, quand une trajectoire de voiture change, elle pourrait rester dans le même groupe de voitures de taille n utilisé pour calculer $V_{p,t}$ dans (3.7) ou elle pourrait être échangée avec une autre voiture, mais dans tous les cas, nous faisons l'hypothèse que la vitesse mesurée $V_{p,t}$ satisfait

$$\frac{|V_{p,t} - \tilde{V}_{p,t}|}{\min\{V_{p,t}, \tilde{V}_{p,t}\}} \leq \gamma$$

Nous discutons de la valeur à attribuer à γ section 4.2.1. Nous avons alors pour le vecteur L avec pour composantes $\ln V_{p,t}$ « »

$$\begin{aligned}
\|L - \tilde{L}\|_2^2 &= \left\|\ln\left(V_{p,t}\right) - \ln\left(\tilde{V}_{p,t}\right)\right\| \\
&= \sum_{p=1}^{P_v}\sum_{t=1}^{\infty}\left|\ln\left(\frac{V_{p,t}}{\tilde{V}_{p,t}}\right)\right|^2 \\
&\leq \sum_{p=1}^{P_v}\sum_{t=1}^{\infty}\left|\frac{V_{p,t}}{\tilde{V}_{p,t}} - 1\right|^2 \\
&\leq \sum_{p=1}^{P_v}\sum_{t=1}^{\infty}\left|\frac{V_{p,t} - \tilde{V}_{p,t}}{\tilde{V}_{p,t}}\right|^2 \\
&\leq 2P_v\gamma^2 =: \Delta_v
\end{aligned} \tag{4.1}$$

où P_v est le nombre de LPV. Notons le facteur 2 pour les mêmes raisons que précédemment, la modification d'une trajectoire a un double impact sur les mesures.

Les valeurs pour les constantes α et γ doivent être choisies basées sur un compromis acceptable entre la confidentialité et la performance de l'estimation. Des valeurs plus élevées de α et γ donneront une meilleure confidentialité (rester dans l'adjacence en permettant des écarts plus importants) mais au prix d'un bruit plus important pour assainir les données. Cependant, des valeurs faibles de α et γ restreignent l'adjacence.

4.2.1 Choix de γ

Rappelons la formule de calcul : $V_{p,t} = \left(\prod_{i=1}^{n} v_{p,t}^{(i)}\right)^{\frac{1}{n}}$ (3.8). Pour la détermination de γ, nous faisons d'abord un calcul sur un trafic idéal, de manière similaire à la section 4.1.1. Supposons donc un trafic à vitesse constante v, et une vitesse $V_{p,t}$ calculée avec la contribution de n véhicules. Supposons qu'une voiture parmi les n ait une vitesse différente $v_\beta = \beta v$ du reste du lot, avec une vitesse définit par un rapport de proportionnalité β par rapport a la vitesse du trafic v. Nous avons alors

$$V_{p,t}(\beta, n) = \left(\left(\prod_{i=1}^{n-1} v\right)\beta v\right)^{\frac{1}{n}} = \beta^{\frac{1}{n}} v$$

Nous pouvons tracer la fonction

$$\begin{aligned}\Gamma &= \frac{|v - \beta^{\frac{1}{n}} v|}{\min\left\{v, \beta^{\frac{1}{n}} v\right\}} \\ &= \frac{|1 - \beta^{\frac{1}{n}}|}{\min\left\{1, \beta^{\frac{1}{n}}\right\}}\end{aligned}$$

voir figure 4.3.

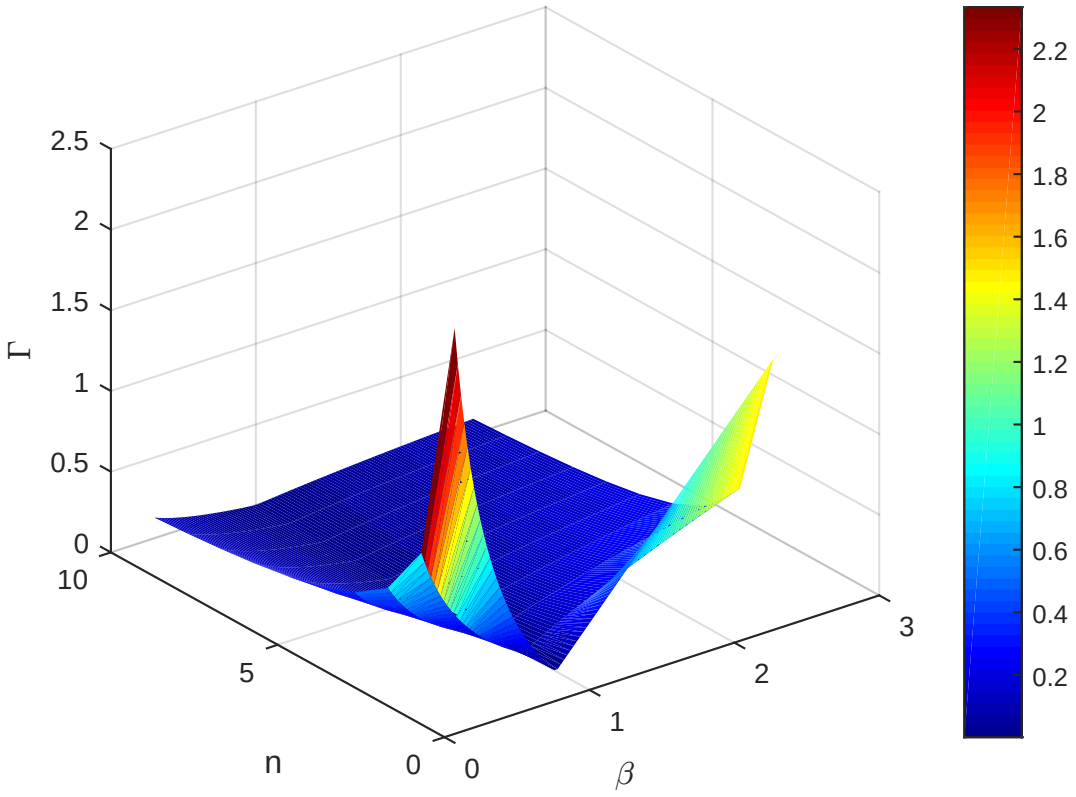

Figure 4.3 Variation relative de vitesse pour trafic idéal

Plus nous prenons un n élevé, plus l'impact d'une voiture est réduit, et plus nous pouvons offrir des garanties de confidentialité élevée, mais plus nous avons besoin d'attendre pour réunir ces n voitures, et plus le risque de calculer une moyenne avec des valeurs trop différentes de vitesses est grand. En choisissant $n = 5$, nous pouvons voir Figure 4.4 et Figure 4.5, en

utilisant le jeu de données Mobile Century que l'attente moyenne est de 1.7 minutes, et que la moyenne des déviations standards des vitesses pour des paquets de 5 voitures est de 3.8 m/s.

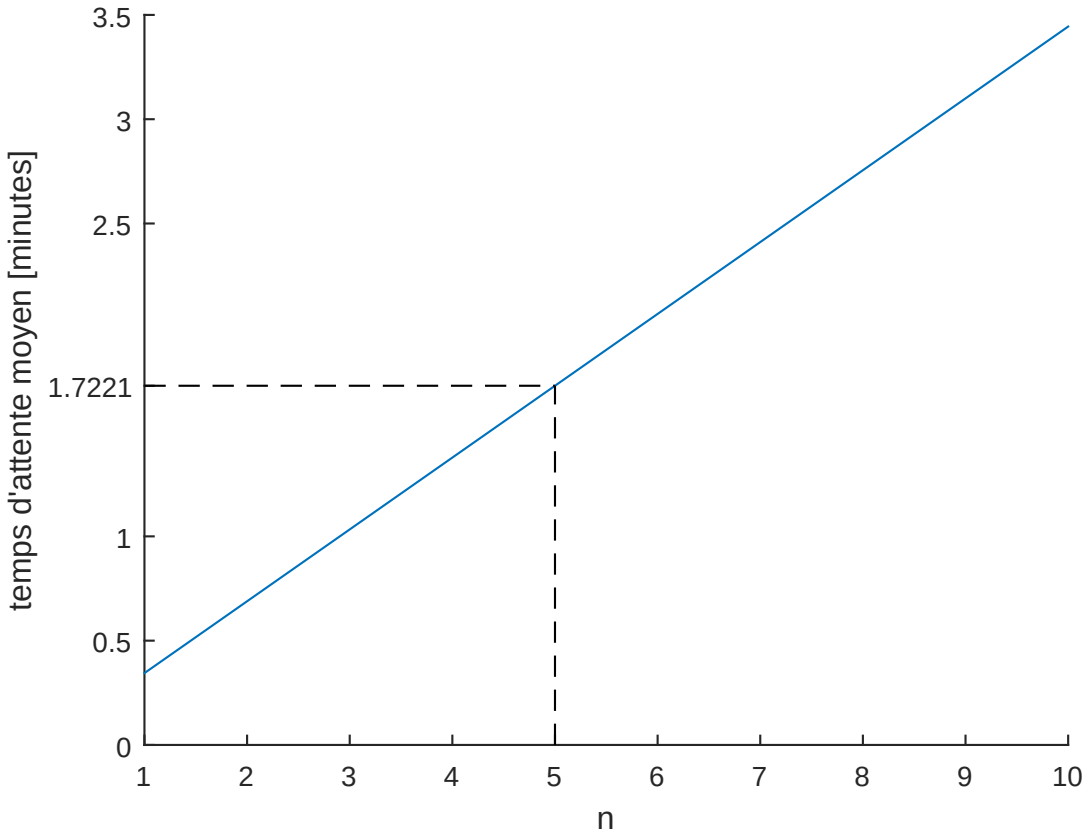

Figure 4.4 Temps d'attente moyen de n véhicule pour une LPV.

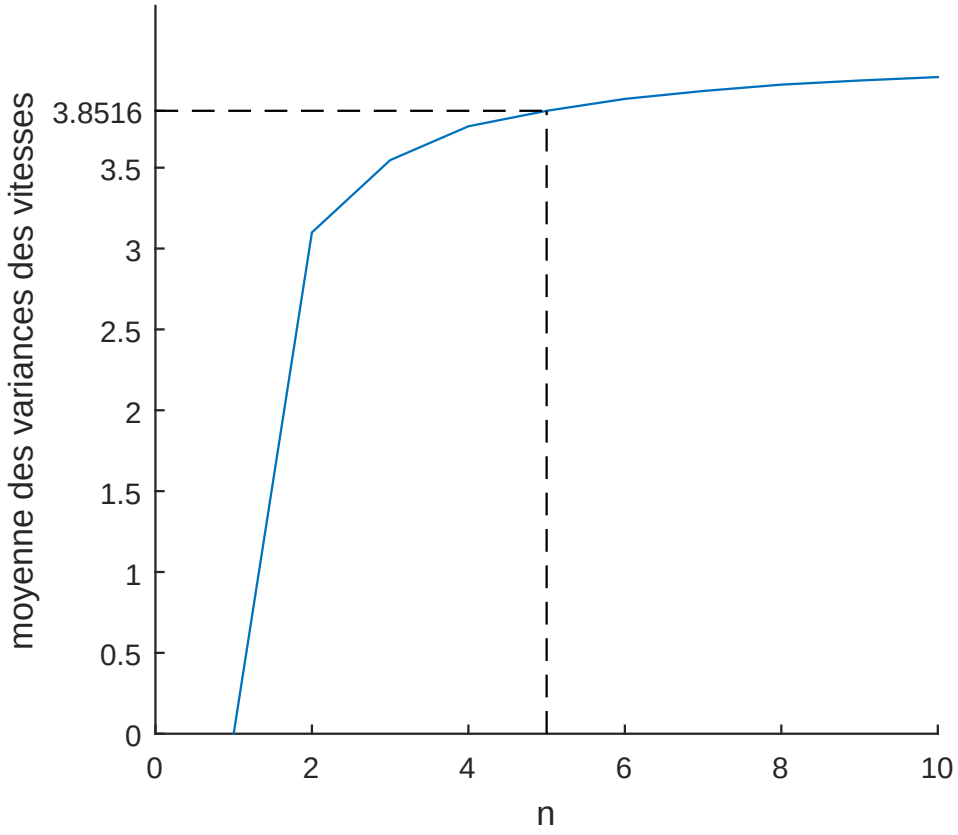

Figure 4.5 Variance des vitesse pour des paquets de n voitures pour le Mobile Century Dataset

Pour $n = 5$, en se référant à la Figure 4.6, nous pouvons choisir $\gamma = 0.1$. En conséquence un utilisateur est protégé s'il ne dévie pas de plus d'un facteur de 0.62 ou 1.61 de la moyenne géométrique des vitesses des 4 voitures le suivant et le précédant.

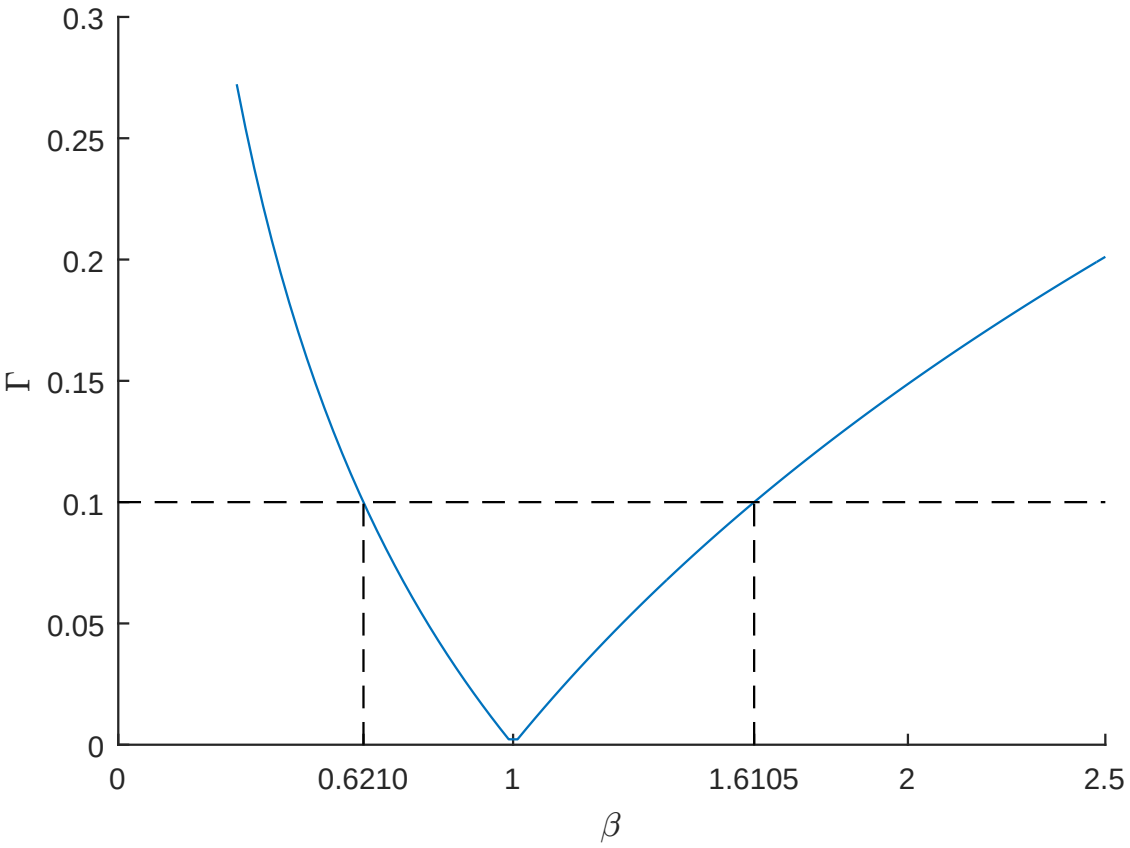

Figure 4.6 Variation relative de vitesse géométrique dûe à une voiture parmi cinq.

Nous pouvons alors appliquer le théorème 3 sur L les logarithmes des vitesses. Ce mécanisme donne le résultat différentiellement confidentiel $r_{p,t} = \log(V_{p,t}) + w_{p,t}$, $w_{p,t}$ étant une variable aléatoire gaussienne. En prenant l'exponentielle, il vient

$$\exp(r_{p,t}) = V_{p,t} \exp(w_{p,t})$$

Nous ne pouvons cependant pas simplement prendre l'exponentiel $\exp(r_{p,t})$ pour estimer la vitesse car l'exponentiel d'une loi normale de moyenne nulle ($\exp(w_{p,t})$), n'est pas à moyenne nulle.

$$\mathbb{E}[V_{p,t} \exp(w_{p,t})] = V_{p,t}\mathbb{E}[\exp(w_{p,t})]$$

Nous pouvons montrer que l'espérance m de l'exponentielle d'une loi gaussienne de moyenne

μ et de variance σ^2 est $m = \exp\left(\mu + \frac{\sigma^2}{2}\right)$. Nous devons donc diviser la sortie du mécanisme gaussien par m pour ne pas avoir de biais systématique sur la moyenne. Nous pouvons écrire le mécanisme final :

Un mécanisme publiant la séquence $\tilde{L}_{p,t} = \frac{\exp\left(L_{p,t}+w^v_{p,t}\right)}{\exp\left(\frac{\kappa^2_{\delta,\epsilon}\Delta^2_v}{2}\right)}$*, où* $w^v_{p,t}$ *sont des variables aléatoires gaussiennes iid avec variance* $\kappa^2_{\delta,\epsilon}\Delta^2_v$*, est* (ϵ, δ)*-différentiellement confidentiel.*

Se reporter au théorème 3 (dans la section 3.3.3) pour la signification des différents paramètres.

4.3 Mesures de comptes

Notons $c^l_{p,t}$ la mesure de compte pour la position p au temps t et la voie l, et notons $c_{p,t}$ la valeur moyenne des comptes par voie, avec λ_p le nombre de voies à la position p. Nous avons

$$c_{p,t} = \frac{1}{\lambda_p}\sum_{l=1}^{\lambda_p} c^l_{p,t}$$

Notons C^∞ et $\tilde{C}^\infty$ la collection de tous les comptes pour deux jeux de données adjacents. Nous avons, en notant P le nombre de capteurs, en prenant en compte le fait que modifier une trajectoire peut modifier au plus $2P$ valeurs de $c^l_{p,t}$

$$|c^l_{p,t} - c'^l_{p,t}| = 0 \text{ ou } 1$$

En ramenant à la moyenne des comptes pour une ligne de capteur constituée de λ_p voies nous avons

$$|c_{p,t} - \tilde{c}_{p,t}| = 0 \text{ ou } \frac{1}{\lambda_p}$$

Et en prenant en compte tout le jeu de données, une voiture peut modifier tous les capteurs successivement par son passage et son absence, nous avons (avec P le nombre de voies à la position P)

$$\sum_{t=0}^{\infty}\sum_{p=1}^{P}\left|c_{p,t} - c'_{p,t}\right| \leq \sum_{p=1}^{P}\frac{2}{\lambda_p}$$

Nous pouvons alors calculer la norme

$$\begin{aligned}\|C^{\infty} - \tilde{C}^{\infty}\|_2 &= \sqrt{\sum_{t=0}^{\infty}\sum_{p=1}^{P} |c_{p,t} - \tilde{c}_{p,t}|} \\ &\leq \sqrt{\sum_{p=1}^{P} \frac{2}{\lambda_p^2}}\end{aligned}$$

Nous pouvons alors utiliser le mécanisme gaussien, voir théorème 3 pour l'assainissement des comptes, avec la sensibilité $\Delta_c = \sqrt{\sum_{p=1}^{P} \frac{2}{\lambda_p^2}}$.

4.4 Utilisation de sources multiples

Il est possible d'utiliser plusieurs capteurs pour la même information. Autrement dit, dans notre cas nous pouvons utiliser dans notre algorithme de filtrage les données assainies de taux d'occupation, de comptes, de vitesses. Cependant, d'après le théorème 2, pour des paramètres de confidentialité fixés (ϵ, δ), nous devons dégrader d'avantage les sources de mesure (prendre des ϵ_i plus petit pour garder $\sum_i \epsilon_i = \epsilon$).

CHAPITRE 5 FILTRAGE DIFFÉRENTIELLEMENT CONFIDENTIEL

Dans ce chapitre nous présentons différents filtres différentiellement confidentiels. Les filtres utilisent un modèle de trafic pour propager l'état du trafic entre les mesures, et corrige l'état avec les mesures.

5.1 Filtre de Kalman étendu

Le filtre de Kalman est un filtre optimal pour les systèmes linéaires avec des bruits additifs indépendants gaussiens pour le modèle et les mesures. Cependant, les modèles de trafics sont non-linéaires, ou linéaires par morceaux. De même, le modèle de mesure de vitesses comme moyenne géométrique est un modèle de mesure non linéaire. Le filtre de Kalman étendu utilise une linéarisation de la dynamique autour de l'estimé.

5.1.1 Algorithme général

De manière générale, pour un état x_k à l'instant k tel que

$$\begin{aligned}\boldsymbol{x}_k &= f(\boldsymbol{x}_{k-1}, \boldsymbol{u}_{k-1}) + \boldsymbol{w}_{k-1}\\ \boldsymbol{z}_k &= h(\boldsymbol{x}_k) + \boldsymbol{v}_k\end{aligned}$$

avec w_k et v_k sont des bruits gaussiens de moyennes nulles et de covariances respectives Q_k et R_k, u_k un vecteur de contrôle. Les fonctions f et h sont utilisées respectivement pour calculer l'état prédit et la mesure prédite. Nous notons l'état prédit $\hat{x}_{k|k-1}$ pour l'instant k sachant l'état à l'instant $k-1$. Nous utilisons respectivement les matrices jacobiennes $\boldsymbol{F}_k$ et $\boldsymbol{H}_k$, c'est à dire

$$\boldsymbol{F}_{k-1} = \left.\frac{\partial f}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}_{k-1|k-1}, \boldsymbol{u}_k}, \boldsymbol{H}_k = \left.\frac{\partial h}{\partial \boldsymbol{x}}\right|_{\hat{\boldsymbol{x}}_{k|k-1}}$$

Nous avons l'algorithme général du filtre de Kalman étendu

$$
\begin{aligned}
\texttt{Prédiction de l'état:} \quad & \hat{\boldsymbol{x}}_{k|k-1} = f(\hat{\boldsymbol{x}}_{k-1|k-1}, \boldsymbol{u}_{k-1}) \\
\texttt{Prédiction de la covariance:} \quad & \boldsymbol{P}_{k|k-1} = \boldsymbol{F}_{k-1}\boldsymbol{P}_{k-1|k-1}\boldsymbol{F}_{k-1}^{\top} + \boldsymbol{L}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{L}_{k-1}^{T} \\
\texttt{Innovation:} \quad & \tilde{\boldsymbol{y}}_k = \boldsymbol{z}_k - h(\hat{\boldsymbol{x}}_{k|k-1}) \\
\texttt{Innovation de la covariance:} \quad & \boldsymbol{S}_k = \boldsymbol{H}_k\boldsymbol{P}_{k|k-1}\boldsymbol{H}_k^{\top} + \boldsymbol{M}_k\boldsymbol{R}_k\boldsymbol{M}_k^{T} \\
\texttt{Gain de Kalman:} \quad & \boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{H}_k^{\top}\boldsymbol{S}_k^{-1} \\
\texttt{Mis à jour de l'état:} \quad & \hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k\tilde{\boldsymbol{y}}_k \\
\texttt{Mis à jour de la covariance:} \quad & \boldsymbol{P}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{P}_{k|k-1}
\end{aligned}
$$

avec

$$
\boldsymbol{L}_{k-1} = \left.\frac{\partial f}{\partial \boldsymbol{w}}\right|_{\hat{\boldsymbol{x}}_{k-1|k-1}, \boldsymbol{u}_{k-1}}, \boldsymbol{M}_k = \left.\frac{\partial h}{\partial \boldsymbol{v}}\right|_{\hat{\boldsymbol{x}}_{k|k-1}}
$$

5.1.2 Filtre de Kalman étendu différentiellement confidentiel

En utilisant le théorème 1, si nous rendons différentiellememt confidentielles les mesures avant leur utilisation par le filtre, alors l'estimé du filtre sera aussi différentiellement confidentiel. Nous pouvons utiliser diverses sources pour une même mesure (par exemple taux d'occupation et vitesse) en utilisant le théorème 2.

Assainir les mesures comme décrit dans la section précédente mène à de nouvelles valeurs $\bar{O}_{p,t}$, $\bar{L}_{p,t}$, qui suivent le même modèle de mesure que dans (3.5), (3.7), mis à part les bruits additionnels gaussiens $w^o_{p,t}$, $w^v_{p,t}$ qui sont simplement ajoutés aux bruits gaussiens $\mu^o_{p,t}$ et $\mu^v_{p,t}$, qui en fait ne font simplement qu'augmenter leurs variances par les termes additifs du Théorème 3. Ces mesures perturbées sont ensuite combinées avec le modèle dynamique (3.4) via un filtre de Kalman étendu ou un filtre de Kalman d'ensemble (Evensen, 2003) (filtre traité dans la section 5.2) pour fournir l'estimé du trafic. Cette étape d'assainissement équivaut à une étape de post-traitement, donc les garanties de différentielle confidentialité obtenues après assainissement des mesures sont toujours valables.

5.1.3 Calcul de la matrice F

Dans l'algorithme du filtre de Kalman étendu, nous devons calculer le jacobien F. Nous donnons ici l'algorithme 1 pour former cette matrice.

Algorithm 1 Algorithme du calcul de la matrice F du filtre de Kalman étendu

for $j = 2 \cdots N$ **do**
  F=zeros(N,N)
  if $\min\{v_0\rho_{j-1,t}, v_0\rho_C, -w(\rho_M - \rho_{j,t})\} = v_0\rho_{j-1,t}$ **then**
    $F(j, j-1) = F(j, j-1) + \frac{\lambda_j}{\lambda_{j+1}\frac{v_0\tau}{\Delta x_j}}$
  else
    if $\min\{v_0\rho_{j-1,t}, v_0\rho_C, -w(\rho_M - \rho_{j,t})\} = -w(\rho_M - \rho_{j,t})$ **then**
      $F(j, j) = F(j, j) + \frac{\lambda_j}{\lambda_{j+1}\frac{v_0\tau}{\Delta x_j}}$
    end if
  end if
  if $\min\{v_0\rho_{j,t}, v_0\rho_C, -w(\rho_M - \rho_{j+1,t})\} = v_0\rho_{j,t}$ **then**
    $F(j, j) = F(j, j) + \frac{v_0\tau}{\Delta x_{j+1}}$
  else
    if $\min\{v_0\rho_{j,t}, v_0\rho_C, -w(\rho_M - \rho_{j+1,t})\} = -w(\rho_M - \rho_{j+1,t})$ **then**
      $F(j, j+1) = F(j, j+1) + \frac{v_0\tau}{\Delta x_{j+1}}$
    end if
  end if
end for

5.2 Filtre de Kalman d'ensemble

Le filtre de Kalman d'ensemble est similaire au filtre de Kalman, mais il utilise un ensemble de particules, c'est à dire un ensemble de vecteurs d'états représentant une densité de répartition de l'état du système, pour calculer la moyenne et la covariance de l'erreur utilisées pour calculer le gain de Kalman. Le filtre de Kalman étendu linéarise la dynamique autour de l'estimé, donc si l'état du trafic est proche de la densité critique ρ_C, une mauvaise estimation mène à une linéarisation complètement faussée.

5.2.1 Algorithme du filtre de Kalman d'ensemble

L'algorithme du filtre de Kalman d'ensemble est décrit dans l'algorithme 3 pour un système d'état générique de dynamique $x_{t+1} = f(x_t, \omega_t)$, un système de mesure linéaire $y_t = Hx_t + v_t$, où v_t a une covariance connue R_t. En appliquant le théorème 1, la sortie de l'algorithme 3 est confidentielle. Pour notre cas, le modèle de mesure de vitesse (3.7) est non linéaire, c'est à dire $y_t = h(x_t) + v_t$, d'où l'utilisation d'une extension du filtre de Kalman d'ensemble expliquée

plus en détails dans (Evensen, 2003) et dans la section 5.2.2. La sortie du filtre de Kalman d'ensemble est un ensemble de particules. La littérature suggère de construire l'estimé par la moyenne de l'ensemble. Cette méthode n'est pas adapté pour le trafic, car les valeurs des états étant bornés (les taux d'occupation entre 0 et une occupation maximale), et l'état réel pouvant facilement atteindre ces bornes, moyenner empêche l'estimé d'atteindre ces bornes. Pour résoudre ce problème, nous construisons l'estimé à partir du nuage de particules avec l'algorithme 2.

Algorithm 2 Éstimé à partir de l'ensemble du filtre de Kalman

for $k = 1, \cdots, \dim\left(x_t^k\right)$ **do**
    Séparer les particules en `ceil`$(\frac{np}{4})$ intervalles I_j de longueurs identiques, contenant n_j particules.
    Sélectionner l'intervalle I_M contenant le nombre maximale de particules n_M.
    La composante k de l'estimé est la moyenne des composantes k des n_M particules de l'intervalle I_M
end for

Algorithm 3 Filtre de Kalman d'ensemble

for $k = 1 \cdots n$ **do**
    $x_0^k \sim \pi_0$ ▷ Tirage de n échantillons de la distribution π_0
end for
for $t \geq 0$ **do**
    for $k = 1 \cdots n$ **do**
        $x_t^k \leftarrow f(x_{t-1}^k, \omega_{t-1}^k)$ ▷ Prédiction du modèle
    end for
    $\bar{x}_t \leftarrow \frac{1}{n}\sum_{k=1}^n x_t^k$ ▷ Moyenne de l'ensemble
    $E = [x_t^1 - \bar{x}_t, \cdots, x_t^n - \bar{x}_t]$ ▷ Deviation à la moyenne
    $P \leftarrow \frac{1}{K-1}E(E)^T$ ▷ Matrice de covariance
    $K_t \leftarrow P(H_t)^T\left[HP_t(H)^T + R_t\right]^{-1}$ ▷ gain de Kalman
    $\hat{y}_t \leftarrow M(y_t)$ ▷ Appliquer le mécanisme confidentiel
    for $k = 1 \cdots n$ **do**
        $\xi^k \sim N(0, R_t)$
        $x_t^k \leftarrow x_t^k + K_t\left[\hat{y}_t - Hx_t^k + \xi^k\right]$ ▷ Mise à jour
    end for
    Publier $\bar{x}_t$ en applicant l'algorithme 2 ▷ Estimé
end for

5.2.2 Opérateur de mesure non linéaire

Nous avons une mesure de vitesse non linéaire. Nous ne pouvons donc pas utiliser la formulation avec la matrice d'observation H. Supposons que notre modèle non linéaire de mesure puisse s'écrire, pour une mesure y et un état x

$$y = h(x) + \nu$$

avec ν un certain bruit de mesure. Supposons que nous avons n réalisations (ou particules) dans notre ensemble, en notant le vecteur d'état $x^{(k)}$ de la réalisation k, nous augmentons tous les vecteurs d'états tels que

$$\hat{x^{(k)}} = \begin{bmatrix} x^k \\ h\left(x^{(k)}\right) \end{bmatrix}$$

Nous pouvons alors utiliser un observateur linéaire directe qui extrait la valeur $h(x)$ de $\hat{x}$.

5.2.3 Filtre de Kalman d'ensemble utilisant les vitesses GPS

Nous avons déterminé section 4.2.1 que nous calculons la vitesse à envoyer au filtre de Kalman d'ensemble à partir des vitesses des 5 dernières voitures traversant la LPV. La fréquence des envois dépend donc du flux des voitures et ne se fait pas à intervalles régulier comme pour les capteurs fixes. Les envois de mesures de vitesses ne se font pas simultanément comme pour les capteurs fixes, mais individuellement pour chaque LPV. L'algorithme d'envoi des vitesses est reporté dans l'algorithme 4.

Les tailles des différentes matrices de Kalman doivent être redimensionnées à chaque nouvelles mesures suivant la taille du vecteur de mesure. La matrice de la covarariance des erreurs de mesures peut être calculée à partir d'un trafic synthétique type. Un exemple d'estimation est montré figure 5.1.

Algorithm 4 Algorithme d'envoi des vitesse

Initialiser la vitesse $v = \texttt{NaN}_{[n,m]}$ de taille $[n, m]$ avec n le nombre de LPV et m le nombre de voitures pour le calcul de la moyenne géométrique.
Initialiser la position $\texttt{pos} = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}_{[1,n]}$ avec $\texttt{pos}(k)$ la position de la première valeur NaN de la ligne k de v.
for $t = 0, \cdots, T$ **do**
    Densité de sortie $\rho = \begin{bmatrix} \texttt{NaN} & \cdots & \texttt{NaN} \end{bmatrix}$ de taille n le nombre de LPV.
    Calculer les distances $d_t^{j,k}$ de toutes les trajectoires de voitures j aux différentes LPV k au temps t : $d_t^{j,k} = x_t^j - p^k$, avec x_t^j la position de la voiture j au temps t et p^k la position de la LPV k.
    Calculer les distance $d_{t-1}^{j,k}$
    Sélectionner les ζ couples $(j_z, k_z), 1 \leq z \leq \zeta$ pour lesquels $d_t^{j,k}$ et $d_{t-1}^{j,k}$ changent de signe (traversée de la LPV k).
    for $z = 1, \cdots, \zeta$ **do**
        $v(k_z, \texttt{pos}(k_z)) = v_t^{j_z}$, avec $v_t^{j_z}$ la vitesse de la voiture j_z au temps t.
        if $\texttt{pos}(k_z) = m$ **then**
            $\texttt{pos}(k_z) = 1$
            Calculer moyenne géométrique $V = (\prod_{i=1}^{m} v(k_z, i))^{\frac{1}{m}}$
            Rendre la vitesse différentiellement confidentielle $V = \frac{Ve^w}{e^{\frac{\kappa_{\epsilon,\delta}^2 \Delta_v^2}{2}}}$ avec w suivant une loi gaussienne de moyenne nulle et de variance $\kappa_{\epsilon,\delta}^2 \Delta_v^2$. Voir équation (4.1) pour l'expression de Δ_v et voir 3 pour $\kappa_{\epsilon,\delta}$.
            $\rho(k_z) = f(V)$, avec la fonction 3.3 utilisant le modèle hybride.
        else
            $\texttt{pos}(k_z) = \texttt{pos}(k_z) + 1$
        end if
    end for
end for

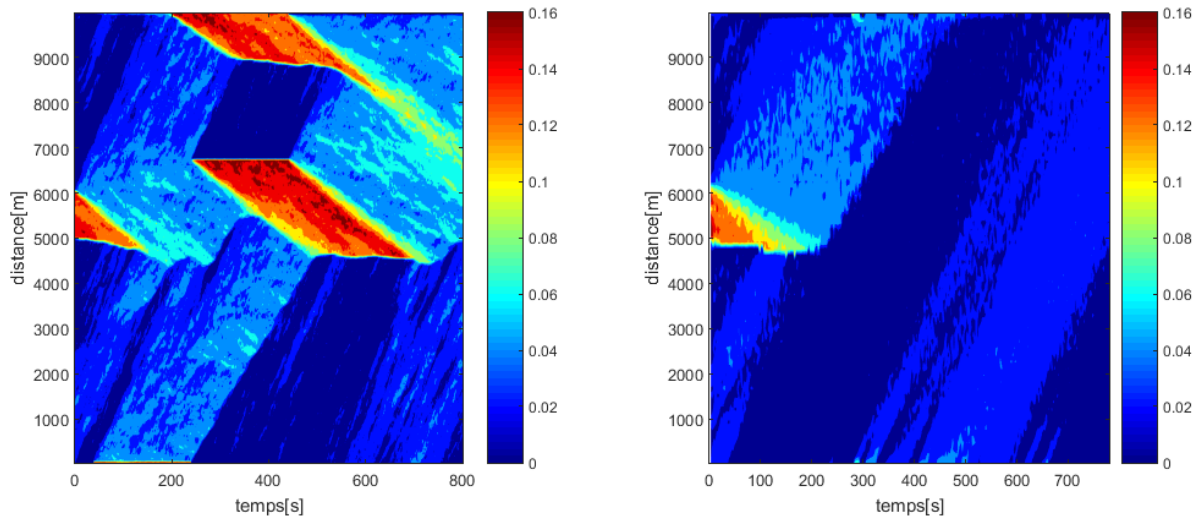

Figure 5.1 Trafic simulé (gauche) et trafic estimé avec filtre de Kalman d'ensemble et mesures GPS de vitesses.

Le fait que l'envoi des vitesses soit asynchrone ne compromet pas les garanties de confidentialité.

5.2.4 Ajustement des positions des LPV

Comme nous l'avons montré dans la Section 4.2, l'assainissement des données de vitesse requiert l'addition d'un bruit de variance proportionnel au nombre de LPV. Il est donc important de travailler avec un nombre limité de LPV et de les positionner de manière à optimiser l'utilité des mesures. Nous utilisons un algorithme où la position des LPV est changée en temps réel basée sur la taille des innovations $\|y_t - H\bar{x}_t\|$ dans l'étape de mise à jour des mesures du filtre de Kalman d'ensemble, tout en prenant garde à ce qu'une voiture ne puisse pas envoyer deux fois sa vitesse pour la même LPV. Nous ajustons $r_1, \cdots, r_{P_v}$, les positions des LPV, utilisant une approche basée sur la minimisation d'une fonction de potentielle à chaque étape. Étant donné les positions actuelles $\hat{r}_1, \cdots, \hat{r}_{P_v}$, nous obtenons les nouvelles positions en minimisant (localement) une fonction du type

$$\sum_{s=1}^{P_v} \left[\left(\sum_{q=1}^{P_v} f_A(\hat{r}_q, r_s; h_q) \right) + \left(\sum_{\sigma \neq s} f_R(r_s, r_\sigma) \right) + f_B(r_s) \right].$$

Ici $f_A(\hat{r}_q, r_s; h_q)$ est un champs attracteur de r_s vers $\hat{r}_q$ avec une force augmentant avec h_q, la norme de l'innovation pour la mesure à la position $\hat{r}_q$. Le second terme est un terme répulsif maintenant un espace suffisant entre les capteurs, et f_B est un terme qui empêche les positions des capteurs de dériver au delà des limites de la route. Nous pouvons prendre par exemple

$$\begin{aligned} f_A(\hat{r}_q, r_s; h_q) &= -h_q k_{A_1} \exp\left(-\frac{|r_s - \hat{r}_q|}{k_{A_2}}\right) \\ f_R(r_s, r_\sigma) &= k_{R_1} \frac{1}{|r_s - r_\sigma|^{k_{R_2}}} \\ f_B(r_s) &= \max(0, -k_B(r_s - b_1), k_B(r_s - b_2)) \end{aligned}$$

avec $k_{A_1}, k_{A_2}, k_{B_1}, k_{B_2}$ des constantes, et b_1, b_2 respectivement la limite amont et aval de la route.

Considération de confidentialité

Le calcul de l'innovation se fait avec les données assainies, différentiellement confidentielles, la confidentialité est donc préservée.

Test

Un résultat de test de l'algorithme de placement des LPV est visible Figure 5.2. Nous pouvons voir les cercles noirs le long de la route, correspondant aux anciennes positions des LPV. Après l'arrivée des nouvelles mesures, nous avons les innovations pour chaque positions, représentées par les lignes verticales bleue, dont la hauteur est proportionnelle à l'innovation. Les croix rouges montrent où nous devons placer les nouvelles LPV. Nous observons bien que les nouvelles positions se rapprochent des points de plus grandes innovations.

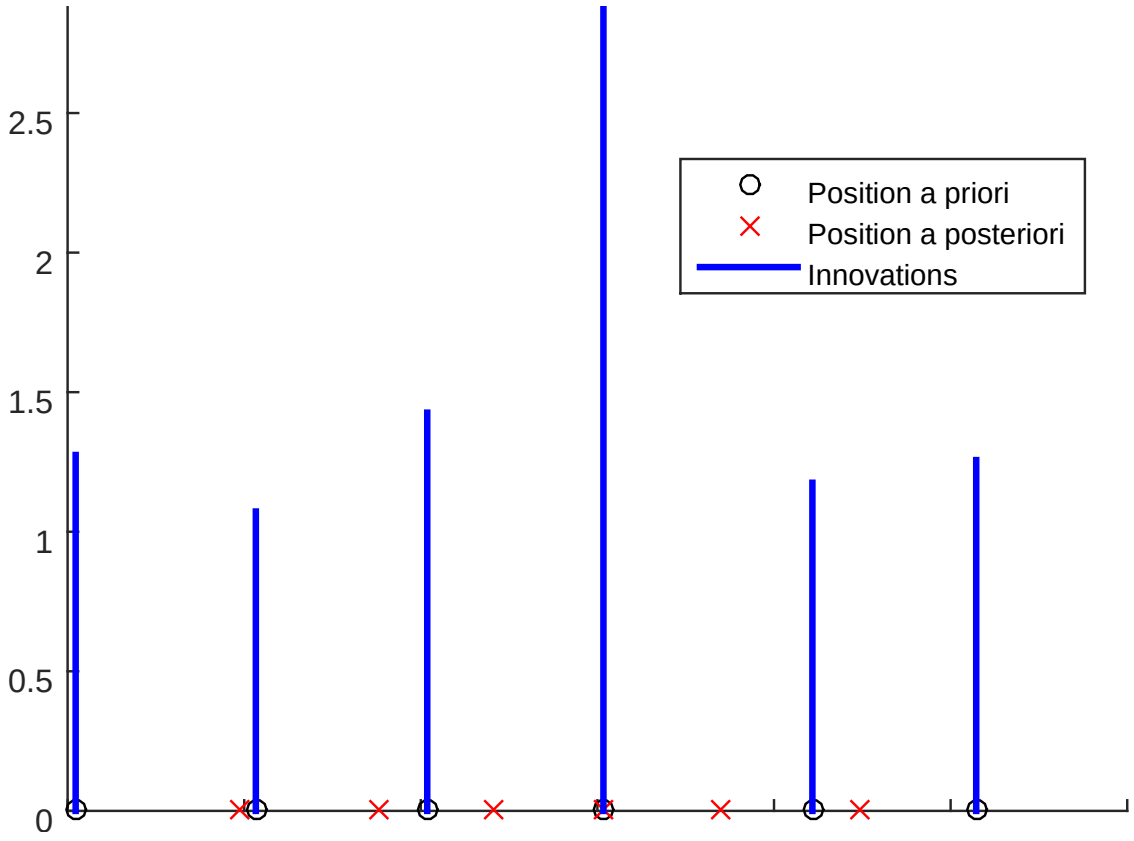

Figure 5.2 Nouvelles positions des LPV

5.3 Filtre bayésien, cas continu

Cette section et la section suivante présentent deux nouveaux mécanismes différentiellement confidentiels. Le mécanisme présenté dans cette section est basé sur du tirage dans une distribution continue (voir théorème 4). La version discrétisée du mécanisme précédant, sous la forme d'un filtre à particule bootstrap n'a pas encore d'équivalent différentiellement confidentiel (voir section 5.4.1). Cependant, un mécanisme reposant sur la modification des poids dans le filtre à particules est proposé (voir section 5.4.2). Il est à noter que le mécanisme développé dans cette section ne s'applique pas au problème de trafic dû a sa nature discrète, et que le mécanisme de perturbation des poids développé section 5.4.2 offre des performance inférieure au filtre de Kalman d'ensemble (comparaison faite section 6.4) et ne sera donc pas retenu comme choix final de filtre. Nous souhaitons tout de même présenter ces approches, mais le lecteur peut poursuivre au prochain chapitre s'il le souhaite.

5.3.1 Cadre Bayesien

Étant donné les variables aléatoires temporelles $X_k \in \mathbb{R}^p$ et $Y_k \in \mathbb{R}^q$, pour $k \geq 0$ considérons la réalisation temporelle du vecteur d'état $\{x_k\}_{k\geq 0}$ et considérons un signal mesuré temporel discret $\{y_k\}_{k\geq 0} = y^k$ une réalisation temporelle de Y^k, pour lesquelles nous avons un modèle d'état de la forme

$$\begin{aligned} x_{k+1} = a(x_k, \omega_k) \quad &\leftrightarrow \quad f_{X_k}(x|x_{k-1}) \\ y_k = b(x_k, \nu_k) \quad &\leftrightarrow \quad g(y|x_k), \end{aligned} \tag{5.1}$$

où a, b sont des fonctions connues potentiellement non linéaires du vecteur d'état $x_k \in \mathbb{R}^n$ et des vecteurs bruits ω_k, ν_k. Chaque processus de bruits ω^∞, ν^∞ est une séquence de variables aléatoires indépendantes et identiquement distribuées avec des distributions distinctes et généralement non gaussiennes. Nous notons $f_{X_f}(x|x_i)$ la fonction de densité de transition d'un état x_i vers un autre état X_f. L'état initial x_0 est aléatoirement distribué selon une distribution connue de densité π_0.

5.3.2 Bornes sur le ratio des densités d'observation

Le mécanisme présenté dans cette section requiert que la fonction $(y, y') \mapsto \frac{g(y|x)}{g(y'|x)}$, avec $\|y - y'\|_1 \leq \delta$ soit bornée par $[e^{-\gamma}; e^{\gamma}]$ pour tout x et j.

$$e^{-\gamma} \leq \frac{g(y|x)}{g(y'|x)} \leq e^{\gamma} \tag{5.2}$$

Dépendamment de la forme de la fonction $g(y|x)$ de notre modèle, ce n'est pas forcément le cas. Cependant, cette propriété peut être obtenue par une modification de $g(y|x)$. La fonction de densité $g(y|x)$ représente usuellement le processus d'acquisition des données. Nous pouvons artificiellement déformer cette fonction pour satisfaire nos besoins. Cette propriété sur le ratio de la densité d'observation sera utilisée ultérieurement pour le mécanisme différentiellement confidentiel. De manière à minimiser les modifications de la fonction de densité $g(y|x)$ nous faisons des modifications seulement lorsque nécessaire. Nous pourrions satisfaire la condition (5.2) en faisant un produit de convolution entre la densité de probabilité et une distribution de Laplace, mais notre méthode est supérieure car la densité n'est pas modifiée dans son ensemble. L'idée générale est décrite dans l'algorithme 5. La densité de probabilité est appelée p, et nous avons $|x - x'| \leq \delta$. Soit ϵ un petit nombre.

Algorithm 5 $e^{-\gamma} \leq \frac{p(x)}{p(x')} \leq e^{\gamma}$

1: Calculer $R(x, x') \mapsto \frac{p(x)}{p(x')}$ pour tous les couples (x, x') possibles tels que $|x - x'| \leq \delta$
2: **while** $e^{-\gamma} \not\leq R(x, x') \not\leq e^{\gamma}$ **do**
3: Trouver les intervalles I tels que $e^{-\gamma} \not\leq R(x, x') \not\leq e^{\gamma}$ pour $x' \in I$
4: $p(I) \leftarrow p(I)e^{\epsilon}$
5: $p \leftarrow \frac{p(u)}{\int p(u)\,\mathrm{d}u}$
6: $R(x, x') \mapsto \frac{p(x)}{p(x')}$
7: **end while**

La figure 5.3 illustre un exemple d'une fonction densité d'observation modifiée pour satisfaire la condition (5.2), pour un $\delta = 1$, et $\exp(\gamma) = 2$, $\exp(-\gamma) = 0.5$. Nous pouvons voir en bleue la distribution originale, et en rouge la distribution modifiée. Nous observons que la distribution modifiée garde globalement la même forme, et conserve les détails de la distribution originale, comme le pic inversé au sommet de la distribution. Si nous avions appliqué une convolution avec une distribution laplacienne, ce pic aurait disparu.

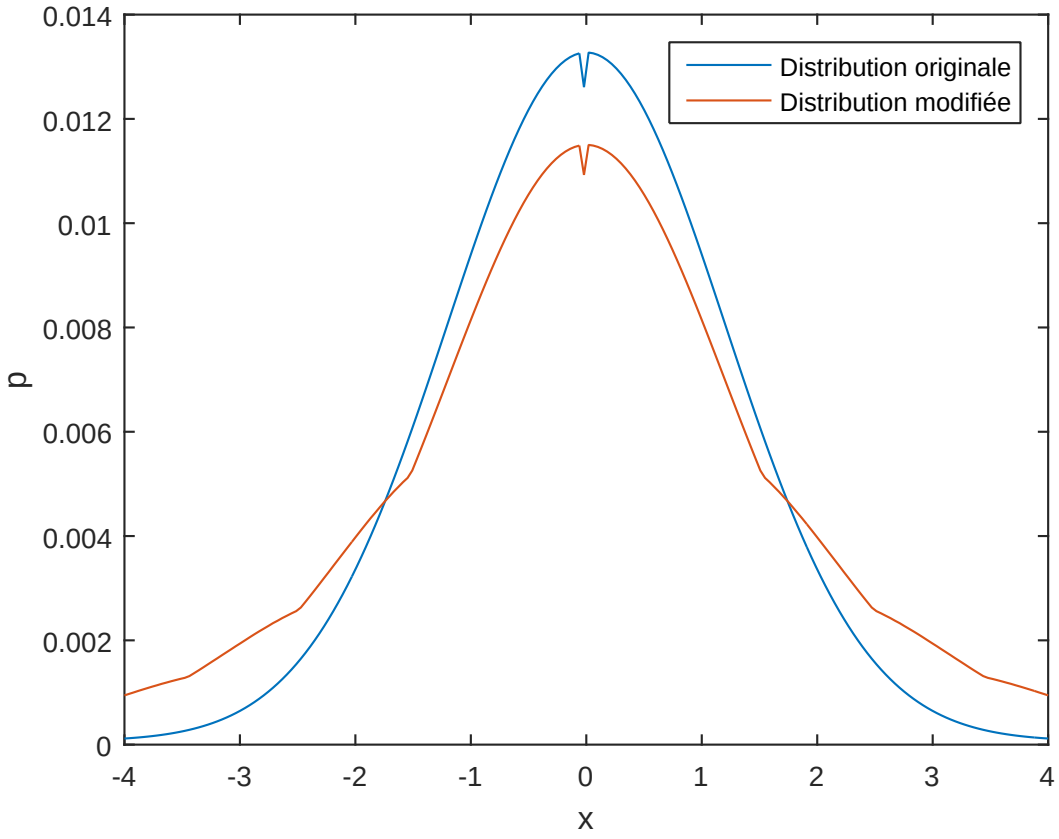

Figure 5.3 Nouvelle distribution après modification

Nous avons tracé figure 5.4 le ratio pour $|x - x'| = 1$. Nous voyons en rouge le ratio non borné de la distribution originale, et en bleu le ratio borné par $[e^{-\gamma}, e^{\gamma}]$.

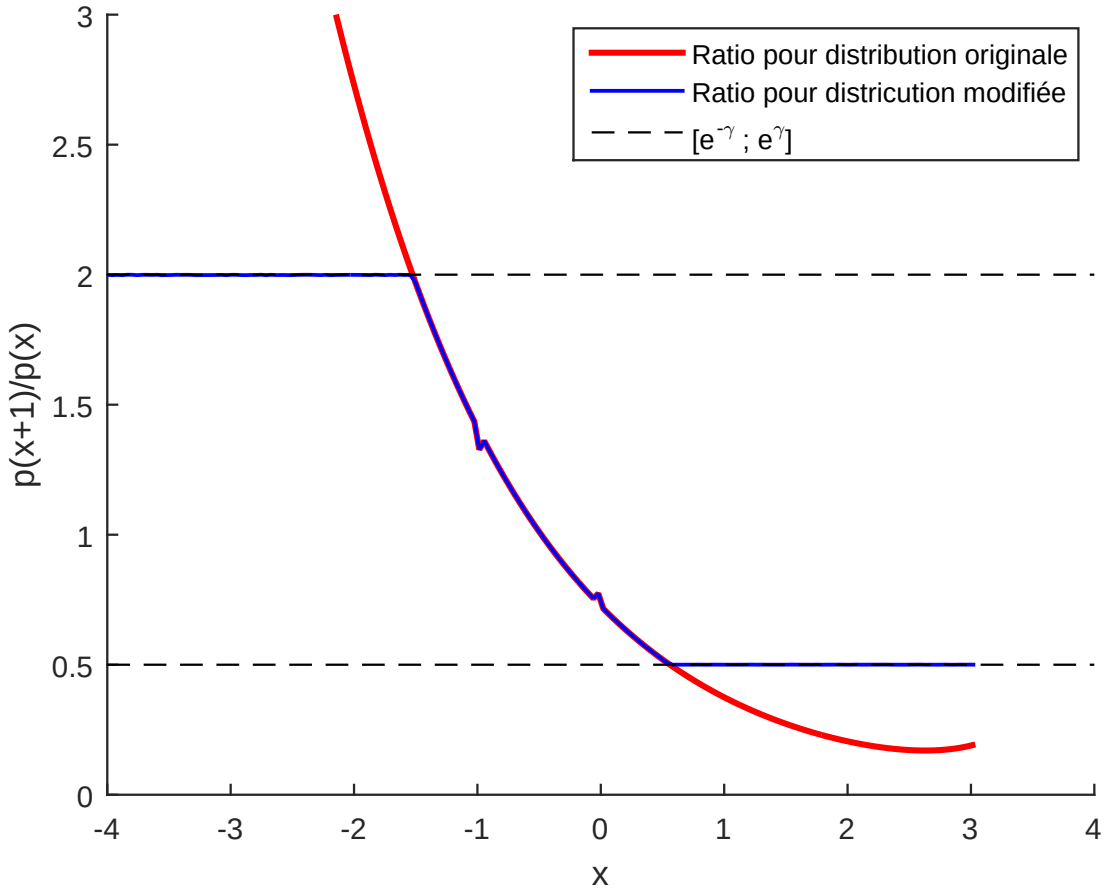

Figure 5.4 Ratio de la distribution pour $x' - x = 1$

Nous pouvons vérifier avec la figure 5.5 que le ratio est bien borné sur l'ensemble du domaine.

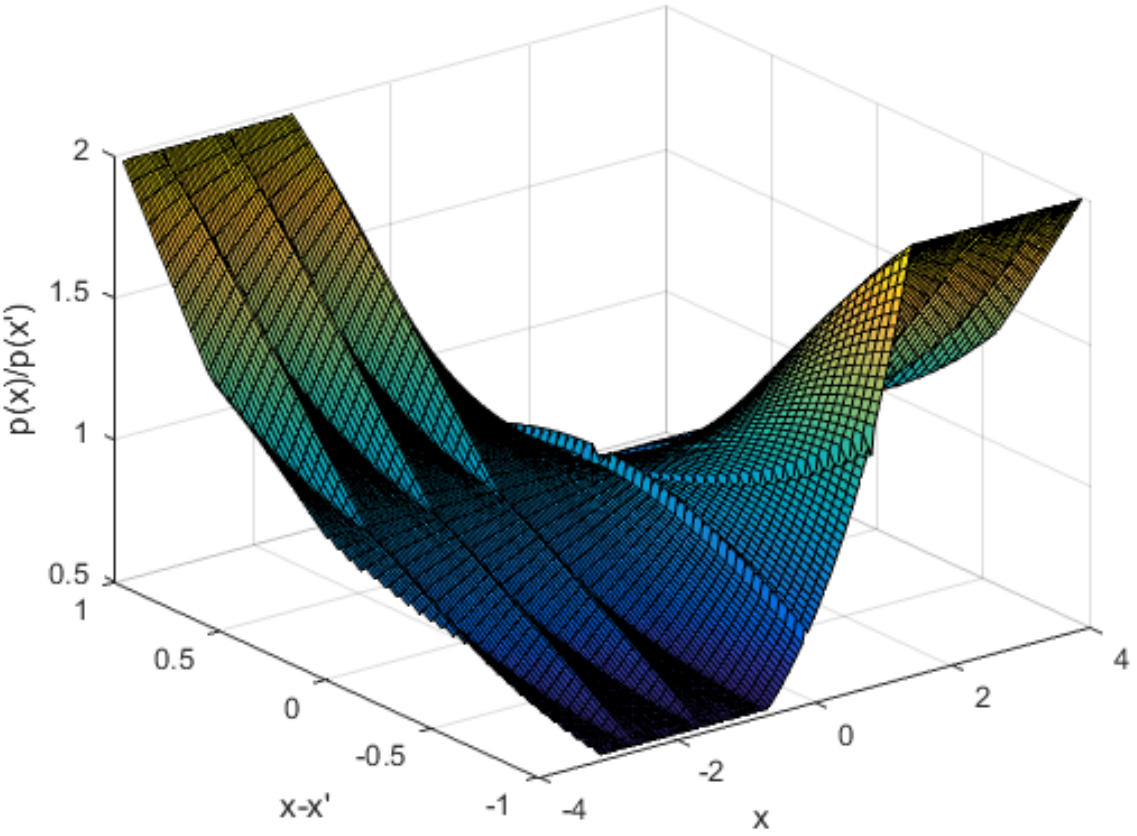

Figure 5.5 Ratio de la distribution modifié pour $|x - x'| \leq \delta$

5.3.3 Mécanisme différentiellement confidentiel

Nous avons la relation d'adjacence suivante. Deux jeux de données $(y, \tilde{y})$ sont adjacents si $\|y^\infty - \tilde{y}^\infty\|_1 \leq \Delta$, avec $\|\mathbf{y}^\infty\|_1 = \sum_{t=0}^{\infty} |y_t|$. De manière analogue à l'application au trafic, supposons que les jeux de données $(y^\infty, \tilde{y}^\infty)$ ne peuvent pas différer d'un nombre β maximal d'entrées, donc une relation valide d'adjacente peut être :

$$\begin{aligned} &\text{Adj}(\mathbf{y}^\infty, \tilde{\mathbf{y}}^\infty) \Leftrightarrow \\ &\exists\ l_1, \ldots, l_\beta \in \mathbb{N} \text{ s.t. } y_k = \tilde{y}_k, \forall k \notin \{l_i\}_{i=1}^{\beta}. \end{aligned} \tag{5.3}$$

Le mécanisme ci-dessous adapte une idée de (Wang et al., 2015).

Theorem 4 *Après application de l'algorithme 5, faire un tirage d'un échantillon de* $f_{X_k}(x|\mathbf{y}^k)$

est $2\beta\gamma$-différentiellement confidentiel. Alternativement, faire le tirage d'un échantillon dans la distribution $\alpha f_{X_k}(x|\mathbf{y}^k)^{\epsilon/(4\beta\gamma)}$ (avec α une constante normative) est ϵ-différentiellement confidentiel.

Proof: Le théorème de Bayes dit que

$$f_{X_k}(x|Y^k = \mathbf{y}^k) = \frac{g(\mathbf{y}^k|X_k = x)f_{X_k}(x)}{g(y^k)}$$
$$= \frac{g(y|x_k)\prod_{m=0}^{k-1}\int g(y_m|X = x_m)f_{X_m}(x_m|x_k)\mathrm{d}x_m}{\int g(y|x_k)\prod_{m=0}^{k-1}\int g(y_m|X = x_m)f_{X_m}(x_m|x_k)\mathrm{d}x_m\mathrm{d}x_k}$$

$$f_{X_k}(x|Y^k = \mathbf{y}^k) \leq e^{2\beta\gamma} f_{X_k}(x|Y^k = \tilde{\mathbf{y}}^k)$$

Échantillonner dans $f_{X_k}(x|\mathbf{y}^k)$ est $2\beta\gamma$-différentiellement confidentiel. Notons p_1, p_2 deux distributions telles que $p_1 \leq e^{\epsilon_1} p_2$.

$$p_1^\zeta \leq e^{\epsilon_1\zeta} p_2^\zeta$$
$$\alpha p_1^\zeta \leq \frac{\alpha}{\beta} e^{\epsilon_1\zeta}\beta p_2^\zeta$$

Avec α, β des constantes normatives telles que $\alpha p_1, \beta p_2$ soient des distributions. Nous avons

$$\frac{\alpha}{\beta} = \frac{\int p_1^\zeta}{\int p_2^\zeta} \leq e_1^{\epsilon\zeta}$$

donc nous avons

$$\alpha p_1^\zeta \leq e^{\epsilon_1 2\zeta}\beta p_2^\zeta$$

Si nous voulons ϵ_2-différentiellement confidentiel, nous devons choisir

$$\gamma = \frac{\epsilon_2}{2\epsilon_1} \tag{5.4}$$

Faire un tirage dans la distribution

$$\alpha f_{X_k}(x|\mathbf{y}^k)^{\frac{\epsilon}{4\beta\gamma}}$$

avec α une constante normative est ϵ-différentiellement confidentiel. $\square$

Nous avons deux moyens d'atteindre la ϵ-confidentialité différentielle. Nous pouvons utiliser seulement le paramètre γ, dans l'expression (5.2), que nous pouvons ajuster avec l'algorithme 5 pour avoir un degré de confidentialité $\epsilon = 2\beta\gamma$. Nous appelons cette méthode le mécanisme 1. Nous pouvons également utiliser γ puis en ajustant le niveau de confidentialité en élevant à la puissance $\epsilon/(4\beta\gamma)$ le postérieure. Nous appelons cette méthode le mécanisme 2. Notons que nous perdons un facteur 2 avec la deuxième approche, lorsque nous mettons à la puissance $\frac{\epsilon_2}{2\epsilon_1}$ dans (5.4), et que nous avons un degré de liberté pour obtenir la ϵ-confidentialité différentielle. La méthode à utiliser devrait être celle qui donne la distribution postérieure avec la plus faible variance dans le but de réduire le facteur aléatoire du tirage, et être également celle qui donne en moyenne l'estimé le plus proche de la réalité.

Nous faisons un test en prenant en exemple la distribution d'observation telle que montrée dans la figure 5.3. Nous définissons arbitrairement une distribution antérieure bi-modale, un état réel ainsi que sa mesure à travers la distribution d'observation modifiée. Nous pouvons voir à la figure 5.6 les différentes distributions. Le postérieur 1 correspondant au postérieur du mécanisme 1, et le postérieur 2 celui du mécanisme 2. Pour la valeur de γ choisie pour cette figure, la puissance à laquelle est élevée la distribution 1 est inférieure à 1. Après normalisation, la distribution 2 est donc plus aplatie que la distribution 1.

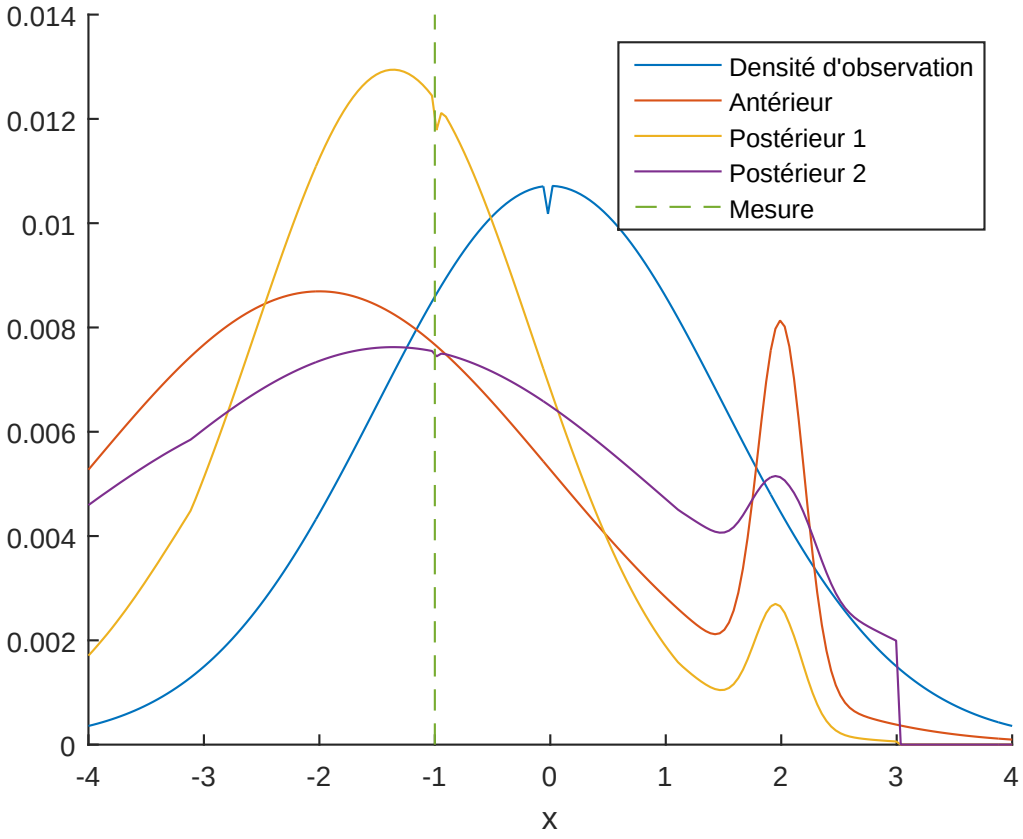

Figure 5.6 Exemple de distributions d'observation, d'antérieur et de postérieurs 1 et 2

Nous faisons varier γ, et calculons les distributions 2 successives pour obtenir un niveau fixé ϵ de confidentialité. Nous calculons aussi la distribution 1, en choisissant le γ pour obtenir le même niveau de confidentialité final ϵ. Nous pouvons ensuite analyser les postérieurs pour déterminer la meilleure stratégie quand au choix du mécanisme 1 ou du mécanisme 2, et du choix de γ pour le mécanisme 2. Les critères de choix sont

— une faible variance du postérieur, pour limiter le facteur aléatoire du tirage

— un faible écart du maximum de la distribution à la valeur réelle, pour une bonne précision.

Nous pouvons voir figure 5.7 la variation de ces critères en fonction de γ pour le mécanisme 2, ainsi que les valeurs pour le mécanisme 1.

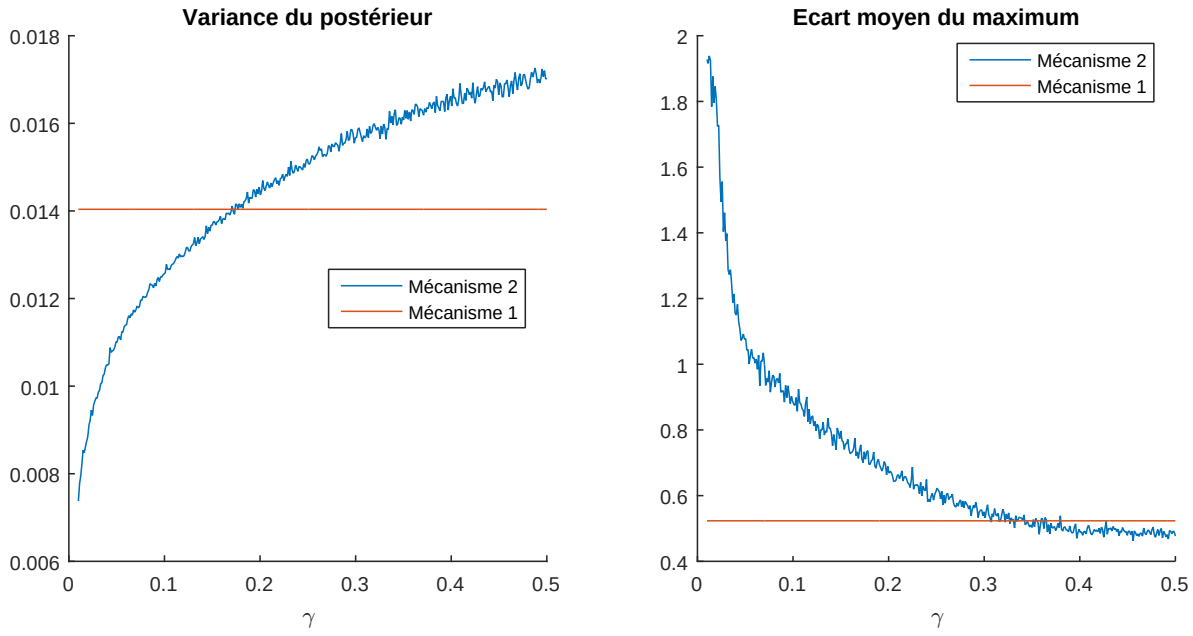

Figure 5.7 Performance des mécanismes pour un exemple de distributions

Lorsque γ est pris faible, la densité d'observation s'uniformise. La mesure a alors peu d'importance dans la mise à jour de l'antérieur vers le postérieur. Ensuite, pour le mécanisme 2, l'antérieur sera mis à une puissance élevée, ce qui fait baisser la variance après normalisation. Cependant l'estimé risque d'être imprécis.

Lorsque γ est pris élevé, la densité d'observation est peu modifiée et peut demeurer précise. La mesure a alors de l'importance (autant d'importance que cette densité d'observation est "pointue") dans la mise à jour de l'antérieur vers le postérieur. Ensuite, pour le mécanisme 2, l'antérieur sera mis à une puissance faible, ce qui fait augmenter la variance après normalisation (aplatissement du postérieur).

C'est donc à l'utilisateur de choisir en fonction des spécifications qu'il recherche entre le mécanisme 1 et le mécanisme 2, et choisir un γ si il choisit le mécanisme 2.

5.4 Cas discret : Filtre à particules bootstrap

La discrétisation du postérieure bayesien mène au filtre à particules. L'algorithme général du filtre à particules est décrit dans l'Algorithme 6

Algorithm 6 Algorithme du filtre à particules

for $i = 1, \cdots, N$ **do**

  Tirage $\tilde{x}_0^{(i)} \sim q_0(x_0|y_0)$

  Assignation des poids d'importance

$$\tilde{\omega}_0^i = \frac{g(y_0|\tilde{x}_0^{(i)})}{q_0(\tilde{x}_0^{(i)}|y_0)}$$

end for

for do$t = 1, \cdots, T$

  if Re-tirage **then**

    Sélectionner N indices $j_i \in \{1, \cdots, M\}$ en fonction des poids

$$\{\omega_{t-1}^j\}_{1 \leq j \leq N}$$

    Mettre $x_{t-1}^{(i)} = \tilde{x}_{t-1}^{(j_i)}$ et $\omega_{t-1}^{(i)} = 1/N, i = 1, \cdots N$.

  else

    $x_{t-1}^{(i)} = \tilde{x}_{t-1}^{(i)}, i = 1, \cdots N$.

  end if

  for do$i = 1, \cdots, N$

    Propager

$$\tilde{x}_t^{(i)} \sim q_t\left(\tilde{x}_t^{(i)}|x_{t-1}^{(i)}, y_t\right)$$

    Calculer les poids

$$\tilde{\omega}_t^{(i)} = \tilde{\omega}_{t-1}^{(i)} \frac{g(y_t|\tilde{x}_t^{(i)}) f(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})}{q_t(\tilde{x}_t^{(i)}|x_{t-1}^i, y_t)}$$

  end for

  Normaliser les poids

$$\omega_t^{(i)} = \omega_t^{(i)} / \sum_{j=1}^{N} \tilde{\omega}_t^{(j)}$$

end for

5.4.1 Mécanisme d'échantillonnage du postérieur

De manière similaire au mécanisme bayesien continu, nous pouvons créer un mécanisme qui utilise l'incertitude des mesures pour le filtre à particules, en calculant et suivant ces incertitudes pour garantir la confidentialité différentielle. Les distributions pour le filtre à particules consistent en un nuage de points, des particules, pondérées, représentant les états. Les particules se propagent suivant une fonction d'importance, et les poids sont mis à jour. La fonction d'importance est une distribution sur l'espace d'états qui favorise les états supposés avoir une plus grande importance sur l'estimé. En effet, pour le cas continu, l'ensemble de l'espace d'état est pris en compte pour le calcul du postérieur. Cependant, pour le cas discret, nous devons discrétiser cet espace en un nombre fini de particules. Il est préférable d'avoir des particules qui décrivent la distribution, avec des poids non nuls, plutôt que d'avoir la majorité des particules sur des états très improbables de poids quasi-zéro, avec seulement quelques particules de poids élevés. Le biais introduit est corrigé en divisant le poids des particules par la densité de la fonction d'importance. Nous choisissons la fonction d'importance q_t comme étant la fonction densité de transition $f(\tilde{x}_t^{(i)}|\tilde{x}_{t-1}^{(i)})$, aussi appelé distribution antérieure. En effet, il est probable que les particules de poids non nuls soient autour de la distribution de l'état propagé de l'estimé. Nous devons aussi effectuer des re-tirages des particules. L'ensemble des particules est re-tiré parmi les particules existantes, en fonction de leurs anciens poids, et les nouveaux poids sont réinitialisé à $\frac{1}{n}$ si n est le nombre de particules. Cela garanti qu'un grand nombre de particules puissent se propager des états probables, et que les états improbables (de poids faibles) soient retirés de l'estimation.

Nous commençons par N tirages initiaux et l'attribution des poids, la distribution postérieure est

$$P(\cdot|\tilde{x}_0^{(1)}, \cdots, \tilde{x}_0^{(N)}, y_0) = \sum_{i=1}^{N} \delta(\cdot - \tilde{x}_0^{(i)}) \frac{g(y_0|\tilde{x}_0^i)}{\sum_{k=1}^{N} g(y_0|\tilde{x}_0^i)}$$

Il en vient que la distribution postérieure avant le tirage initial est

$$\begin{aligned} P(\cdot|y_0) &= \int_{x_0^1} \cdots \int_{x_0^N} \sum_{i=1}^{N} \delta(\cdot - \tilde{x}_0^{(i)}) \frac{g(y_0|\tilde{x}_0^i)}{\sum_{k=1}^{N} g(y_0|\tilde{x}_0^i)} \prod_{k=1}^{N} \pi_0\left(x_0^{(k)}\right) \mathtt{d}x_0^1 \cdots \mathtt{d}x_0^N \\ &\leq e^{2L\|y_0 - y_0'\|} P(\cdot|y_0') \end{aligned}$$

L'étape de re-tirage re-attribut à chaque particule l'état d'une autre particule. Nous avons

donc N^N combinaison possible

$$P_2(\cdot|y_0) = \int_{x_0^1} \cdots \int_{x_0^N} \sum_{i_1=1}^{N} \cdots \sum_{i_N=1}^{N} \frac{1}{N} \left[\delta(\cdot - x_0^{(i_1)}) + \cdots + \delta(\cdot - x_0^{(i_N)})\right] \times \frac{g(y_0|x_0^{i_1}) \cdots g(y_0|x_0^{i_N})}{\sum_{k=1}^{N} g(y_0|x_0^k)} \prod_{k=1}^{N} \pi_0\left(x_0^{(k)}\right) \mathtt{d}x_0^1 \cdots \mathtt{d}x_0^N$$

Nous pouvons ensuite propager les particules

$$P_3(\cdot|y_0) = \int_{x_1^1} \cdots \int_{x_1^N} \int_{x_0^1} \cdots \int_{x_0^N} \sum_{i_1=1}^{N} \cdots \sum_{i_N=1}^{N} \frac{1}{N} \left[\delta\left(\cdot - x_1^{(1)}\right) + \cdots + \delta\left(\cdot - x_1^{(N)}\right)\right] \times \prod_{k=1}^{N} f\left(x_1^{(k)}|x_0^{(i_k)}\right) \frac{g(y_0|x_0^{i_1}) \cdots g(y_0|x_0^{i_N})}{\sum_{k=1}^{N} g(y_0|x_0^k)} \prod_{k=1}^{N} \pi_0\left(x_0^{(k)}\right) \mathtt{d}x_0^1 \cdots \mathtt{d}x_0^N \mathtt{d}x_1^1 \cdots \mathtt{d}x_1^N$$

Nous pouvons finalement calculer les nouveaux poids après l'obtention de la mesure y_1

$$\begin{aligned} P(\cdot|y_0, y_1) = & \int_{x_1^1} \cdots \int_{x_1^N} \int_{x_0^1} \cdots \int_{x_0^N} \sum_{i_1=1}^{N} \cdots \sum_{i_N=1}^{N} \left[\frac{g(y_1|\tilde{x}_1^1)}{\sum_{k=1}^{N} g(y_1|\tilde{x}_0^{(k)})} \delta\left(\cdot - x_1^{(1)}\right) + \cdots + \right. \\ & \left. \frac{g(y_1|\tilde{x}_1^N)}{\sum_{k=1}^{N} g(y_1|\tilde{x}_0^i)} \delta\left(\cdot - x_1^{(N)}\right)\right] \times \prod_{k=1}^{N} f\left(x_1^{(k)}|x_0^{(i_k)}\right) \frac{g(y_0|x_0^{i_1}) \cdots g(y_0|x_0^{i_N})}{\sum_{k=1}^{N} g(y_0|x_0^k)} \times \\ & \prod_{k=1}^{N} \pi_0\left(x_0^{(k)}\right) \mathtt{d}x_0^1 \cdots \mathtt{d}x_0^N \mathtt{d}x_1^1 \cdots \mathtt{d}x_1^N \\ & \leq e^{(N+3)L\|y_{0,1} - y'_{0,1}\|} P(\cdot|y'_0) \end{aligned}$$

Nous remarquons que sans hypothèses supplémentaires, nous ne pouvons pas assurer un degré de confidentialité différentielle faible avec ce mécanisme, principalement dû à l'étape de re-tirage qui inclus une dépendance supplémentaire de l'estimé aux mesures.

5.4.2 Mécanisme de perturbation des poids

Nous pouvons atteindre la confidentialité différentielle en rendant les poids dans le filtre à particules différentiellement confidentiels. Nous utilisons un mécanisme multiplicatif pour conserver la positivité et laisser proche de zéro les poids très faibles. Sachant que pour le filtre à particules `bootstrap` la valeur des poids est $\omega_t^{(i)} = \frac{g(y_t|\tilde{x}_t^{(i)})}{\sum_{k=1}^{N} g(y_t|\tilde{x}_t^{(k)})}$. Pour le mécanisme

multiplicatif nous devons étudier la norme-2 des différences des logarithmes. Nous avons

$$\begin{aligned}\log\left(w_t^{(i)}\right) - \log\left(w_t'^{(i)}\right) &= \log\left(\frac{w_t^{(i)}}{w_t'^{(i)}}\right) \\ &= \log\left(\frac{g\left(y_t|\tilde{x}_t^{(i)}\right)}{g\left(y_t'|\tilde{x}_t^{(i)}\right)}\frac{\sum_{k=1}^{N} g\left(y_t'|\tilde{x}_t^{(k)}\right)}{\sum_{k=1}^{N} g\left(y_t|\tilde{x}_t^{(k)}\right)}\right) \\ &\leq 2\gamma\end{aligned}$$

En considérant l'ensemble des poids pour toutes les particules et tous les temps, avec un nombre N de particules, un nombre n_s de capteurs, en considérant qu'une voiture peut changer $2n_s$ valeurs entre deux jeux de données adjacents (en se retirant n_s fois du jeu de données initial et s'ajoutant n_s fois dans le jeu de données adjacent), nous déduisons la sensibilité logarithmique pour la norme-2, avec l'hypothèse $e^{-\gamma} \leq \frac{g(y_t|\tilde{x}_t^{(i)})}{g(y_t'|\tilde{x}_t^{(i)})} \leq e^{\gamma}$

$$\Delta_{\log\omega} = 2\sqrt{2n_sN}\gamma$$

Nous récapitulons ce mécanisme dans l'algorithme 7. Nous avons également pour ce mécanisme un paramètre d'ajustement, γ. Prendre un γ faible réduit le bruit à ajouter, mais dégrade la distribution d'observation.

Algorithm 7 Algorithme du mécanisme différentiellement confidentiel de modification de poids du filtre à particules

if $e^{-\gamma} \nleq \frac{g(y_t|\tilde{x}_t^{(i)})}{g(y_t'|\tilde{x}_t^{(i)})} \nleq e^{\gamma}$ **then**
    Appliquer algorithme 5
end if
for $i = 1, \cdots, N$ **do**
    Tirage $\tilde{x}_0^{(i)} \sim q_0(x_0|y_0)$
    Assignation des poids d'importance

$$\tilde{\omega}_0^i = \frac{g(y_0|\tilde{x}_0^{(i)})}{q_0(\tilde{x}_0^{(i)}|y_0)}$$

    Perturbation des poids

$$\begin{aligned} \nu^{(i)} &\sim \mathcal{N}\left(0, \kappa_{\delta,\epsilon}\sqrt{2n_s N}2\gamma\right) \text{ voir théorème 3} \\ \mu &= \exp\left(\kappa_{\delta,\epsilon}^2 4n_s N\gamma^2\right) \\ \tilde{\omega}_0^i &\leftarrow \frac{\exp\left(\log\left(\tilde{\omega}_0^i\right) + \nu^{(i)}\right)}{\mu} \end{aligned}$$

end for
Normaliser les poids

$$\omega_0^{(i)} = \omega_0^{(i)} / \sum_{j=1}^{N} \tilde{\omega}_0^{(j)}$$

for $t = 1, \cdots, T$ **do**
    if Re-tirage **then**
        Sélectionner N indices $j_i \in \{1, \cdots, M\}$ en fonction des poids

$$\{\omega_{t-1}^j\}_{1 \leq j \leq N}$$

        Mettre $x_{t-1}^{(i)} = \tilde{x}_{t-1}^{(j_i)}$ et $\omega_{t-1}^{(i)} = 1/N, i = 1, \cdots N$.
    else
        $x_{t-1}^{(i)} = \tilde{x}_{t-1}^{(i)}, i = 1, \cdots N$.
    end if

17: **for** $i = 1, \cdots, N$ **do**

18: Propager

$$\tilde{x}_t^{(i)} \sim q_t\left(\tilde{x}_t^{(i)} | x_{t-1}^{(i)}, y_t\right)$$

19: Calculer les poids

$$\tilde{\omega}_t^{(i)} = \tilde{\omega}_{t-1}^{(i)} \frac{g(y_t|\tilde{x}_t^{(i)}) f(\tilde{x}_t^{(i)}|x_{t-1}^{(i)})}{q_t(\tilde{x}_t^{(i)}|x_{t-1}^{i}, y_t)}$$

20: Perturbation des poids

$$\begin{aligned} \nu^{(i)} &\sim \mathcal{N}\left(0, \kappa_{\delta,\epsilon}\sqrt{2n_s N} 2\gamma\right) \text{ voir théorème 3} \\ \mu &= \exp\left(\kappa_{\delta,\epsilon}^2 4 n_s N \gamma^2\right) \\ \tilde{\omega}_t^i &\leftarrow \frac{\exp\left(\log\left(\tilde{\omega}_t^i\right) + \nu^{(i)}\right)}{\mu} \end{aligned}$$

21: **end for**

22: Normaliser les poids

$$\omega_t^{(i)} = \omega_t^{(i)} / \sum_{j=1}^{N} \tilde{\omega}_t^{(j)}$$

23: **end for**

CHAPITRE 6 RÉSULTATS DE SIMULATIONS

Dans ce chapitre nous allons évaluer l'utilité de l'estimateur énoncée (section 6.2) puis sélectionner le meilleur filtre (section 6.4), en estimant la densité d'un trafic généré par simulation (avec les valeurs numériques présentées section 6.1). Le nombre de particules nécessaire pour chaque filtre est préalablement déterminé dans la section 6.3. Ensuite nous nous pencherons sur le nombre optimal de capteurs pour l'estimation dans la section 6.5. Nous étudierons ensuite l'impact du choix des paramètres de confidentialité (section 6.6), puis l'échantillonnage spatial dynamique 6.7. La section suivante (section 6.8) traitera de l'estimation de la densité à partir des trajectoires. La section 6.9 traite de l'application de l'estimateur confidentiel au jeu de données Mobile Century. Enfin nous traiterons de l'estimation de trafic généré par des modèles d'ordre supérieurs dans la section 6.10.

6.1 Valeurs numériques

Notre filtre différentiellement confidentiel est premièrement validé sur des données synthétiques (simulées) pour une route à une voie, ce qui nous permet de comparer la performance de l'estimateur par rapport à la vérité simulée, voir la figure 6.1. Pour l'exemple de la figure 6.1 le trafic à l'état initial de la simulation est fluide, à l'exception d'une zone congestionnée entre 5000 et 6000 mètres. Nous voyons les voitures qui s'amassent sur le front amont de la congestion, qui remonte la route avec le temps, tandis que le front aval se résorbe dans le trafic fluide, et donc remonte aussi le trafic. Nous simulons une seconde congestion entre 220 et 450 secondes sur la sortie de la route.

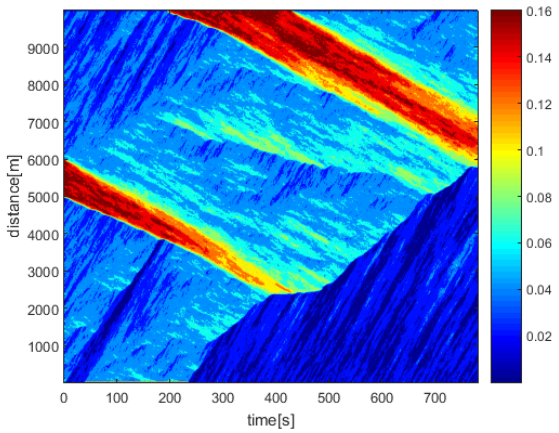

Figure 6.1 Exemple de densité simulée

Les paramètres de simulation sont :

— $P_s = 10$ capteurs statiques

— $\tau = 0.5$ s

— $\Delta x_p = 25$ m

Les paramètres du diagramme fondamental sont

— $v_0 = 90$ km/h

— $w = 30$km/h

— $g = 6$ m

— $\rho_M = 1/7$ vehicles/m

Pour le digramme fondamental. Les mesures de taux d'occupation sont obtenues périodiquement toutes les 30 s. Les paramètres de confidentialité différentiels sont :

— $\epsilon = \ln(2 + P_s)$

— $\delta = 0.05$

— $\alpha = 0.015$ (relation d'adjacence)

— $\gamma = 0.4$ (relation d'adjacence)

6.2 Utilité de l'estimateur

L'utilité de l'estimé $\tilde{\rho}$ peut être obtenu lorsque l'on connaît la vérité simulée ρ pour un jeu de donné synthétique. Ce critère u se calcule comme la moyenne empirique du carré des erreurs

$$u = \mathbb{E}\left(\|\tilde{\rho} - \rho\|^2\right)$$

6.3 Nombres de particules nécessaires

Le nombre de particules pour le filtre de Kalman d'ensemble ou le filtre à particules doit être déterminé. Un nombre élevé de particules augmente la puissance de calcul nécessaire pour rester un estimateur en temps réel. Mais un nombre élevé de particules signifie généralement un meilleur estimé. Cependant, l'utilité de l'estimateur converge après un certain nombre n_M de particules. Ajouter encore des particules n'améliore pas significativement l'estimé dont la partie aléatoire est alors majoritairement due au erreurs de mesures et aux approximations du modèle. L'exception est pour le mécanisme de perturbation des poids pour le filtre à particules, où augmenter le nombre de particules augmente le bruit introduit.

6.3.1 Filtre de Kalman d'ensemble

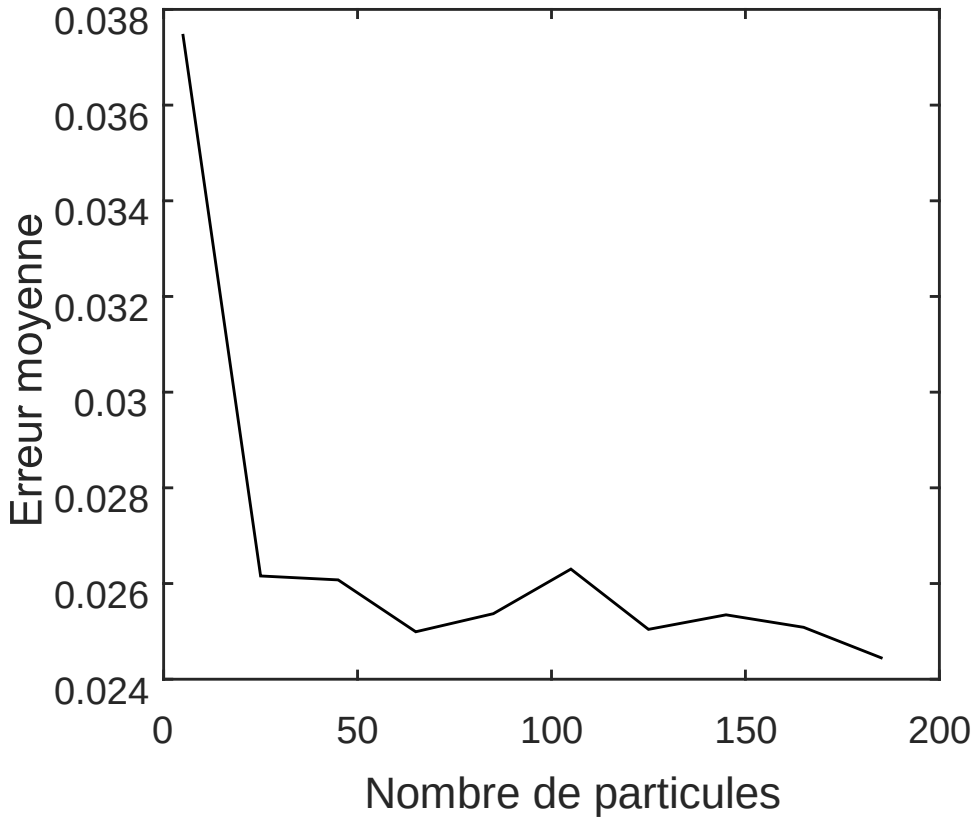

Figure 6.2 Erreur moyenne du filtre de Kalman d'ensemble privé pour un certain nombre de particules

Nous calculons l'utilité en moyenne sur 10 simulations en faisant varier le nombre de particules pour chaque itérations. Nous observons avec la Figure 6.2 que l'utilité converge pour environ 60 particules.

6.3.2 Filtre à particules

Nous faisons un travail identique pour le filtre à particules non-privé, voir Figure 6.4. Nous observons une convergence de l'utilité après 160 particules.

6.3.3 Filtre à particules avec perturbation des poids

Le mécanisme différentiellement confidentiel modifiant les poids des particules dans le filtre introduit un bruit multiplicatif augmentant avec le nombre de particules, or le nombre de particules est un facteur de qualité d'estimation dans le filtre à particules. Nous cherchons donc le nombre optimal de particules. Pour un $\gamma = \log(2)$ fixé, nous avons l'utilité reportée figure 6.3 pour un nombre variable de particules.

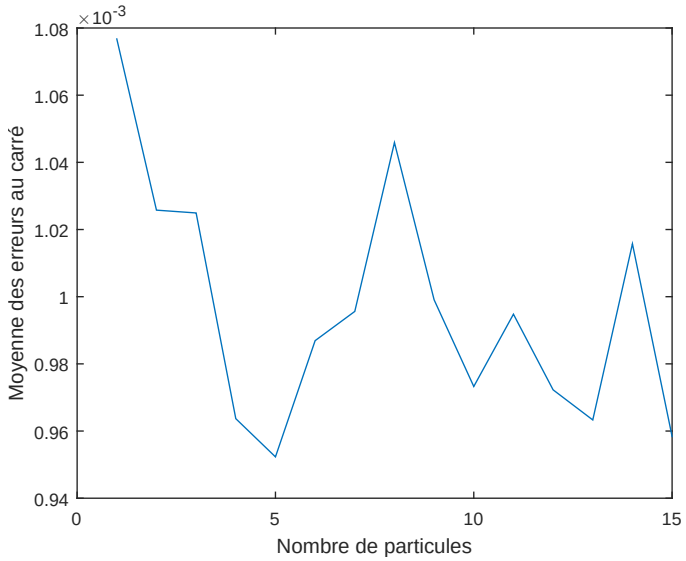

Figure 6.3 Erreur moyenne du filtre du filtre à particules Bootstrap avec mécanisme de modification des poids

La racine carrée de l'utilité est du même ordre de gradeur que la valeur de la densité critique. Nous concluons donc que ce mécanisme, différentiellement privé, n'apporte pas la précision nécessaire pour une utilisation pratique de la densité estimée. En effet le filtre à particules nécessite un nombre assez élevé de particules pour avoir une estimation correcte, et ce nombre élevé de particules n'est pas compatible avec une insertion raisonnable de bruit pour le mécanisme différentiellement confidentiel.

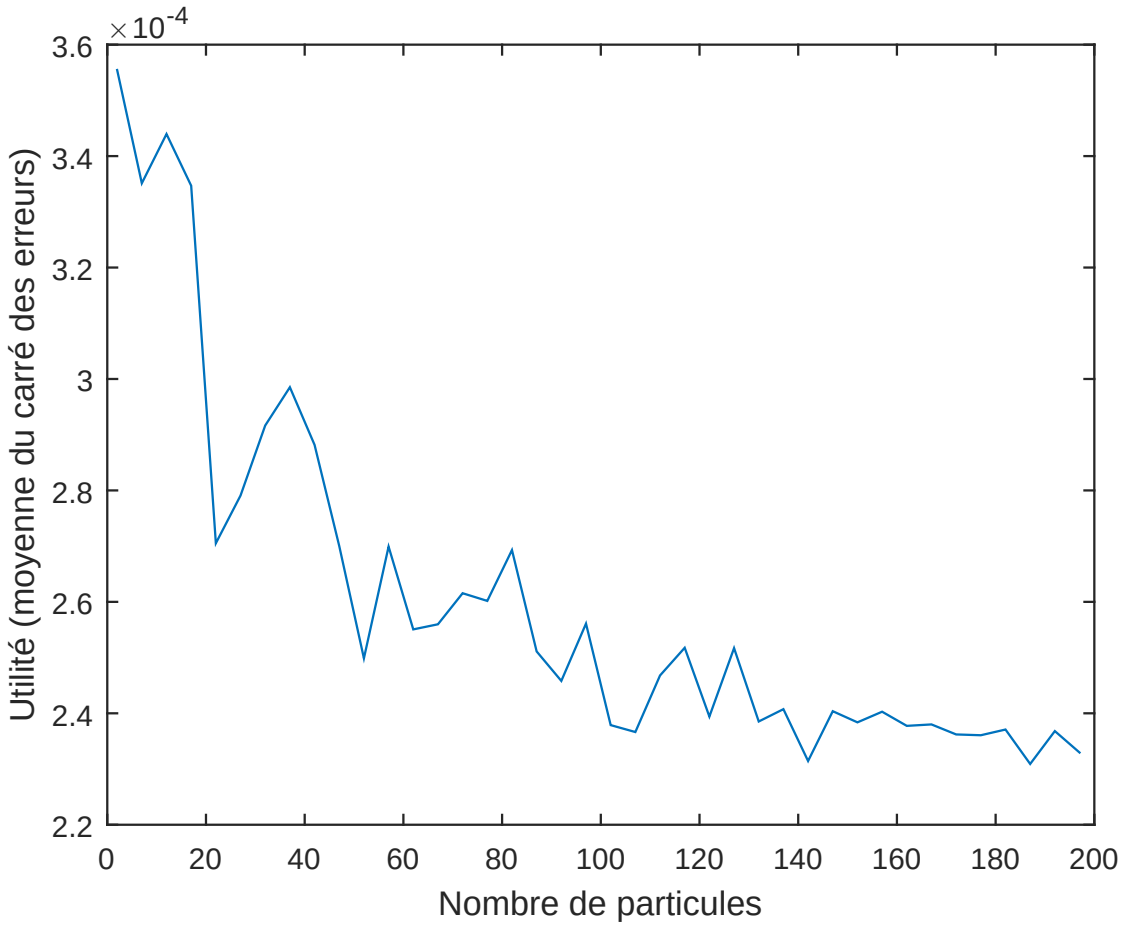

Figure 6.4 Erreur moyenne du filtre du filtre à particules Bootstrap

6.4 Comparaison des filtres différentiellement confidentiels avec les mesures d'occupations

Nous comparons l'utilité des différents filtres dans le tableau 6.1. Les valeurs sont moyennées sur 30 simulations.

Nous déduisons que le filtre qui donne le meilleur estimé pour un niveau de confidentialité fixé est le filtre de Kalman d'ensemble.

FKEn	FP-mesures	FP-poids
6.0390e-04	9.2980e-04	9.5540e-04

Tableau 6.1 Utilité des filtres différentiellement confidentiels

6.5 Nombre de capteurs pour le filtre de Kalman d'ensemble

Pour des paramètres (ϵ, δ) fixés de confidentialité, nous avons le choix sur le nombre de capteurs à incorporer dans le filtre. En effet, si nous utilisons tous les capteurs disponibles, nous avons beaucoup d'informations le long de la route, mais le mécanisme ajoute un bruit aux mesures d'autant plus grand que le nombre de capteurs est élevé. Dans le cas idéal, il faudrait disposer pour l'étalonnage du dispositif d'une densité obtenue tout au long d'un trafic type, et faire une analyse des placements des capteurs. De manière intuitive, les capteurs seront les plus significatifs aux endroits de la route où le trafic est susceptible de dévier le plus par rapport aux prédictions du modèle. Pour notre trafic synthétique (voir figure 6.1) nous reportons l'utilité de l'estimateur pour un nombre variable de capteurs équidistants. Nous reportons les résultat figure 6.5. Pour ce trafic particulier, le trafic est seulement contraint

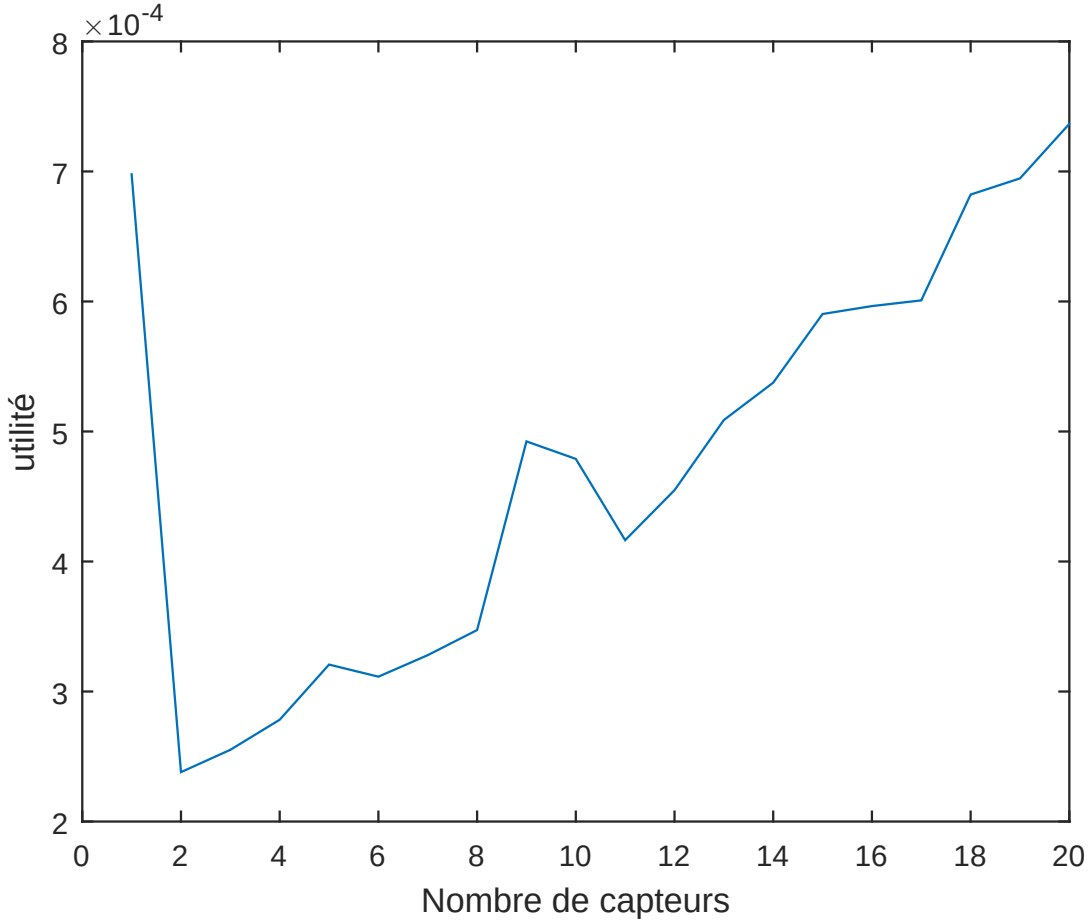

Figure 6.5 Erreur moyenne du filtre du filtre pour trafic sans incidents

au début et à la fin de la route, et la dynamique propage ces conditions aux limites. Pour ce

genre de trafic simple, la meilleure estimation est pour deux capteurs, positionnés au début et à la fin de la route. Les données de ces capteurs sont alors peu perturbés, et la dynamique du modèle reflète assez fidèlement le trafic entre ces deux capteurs. Nous synthétisons un trafic plus imprévisible avec un blocage temporaire soudain au milieu de la route, voir figure 6.6.

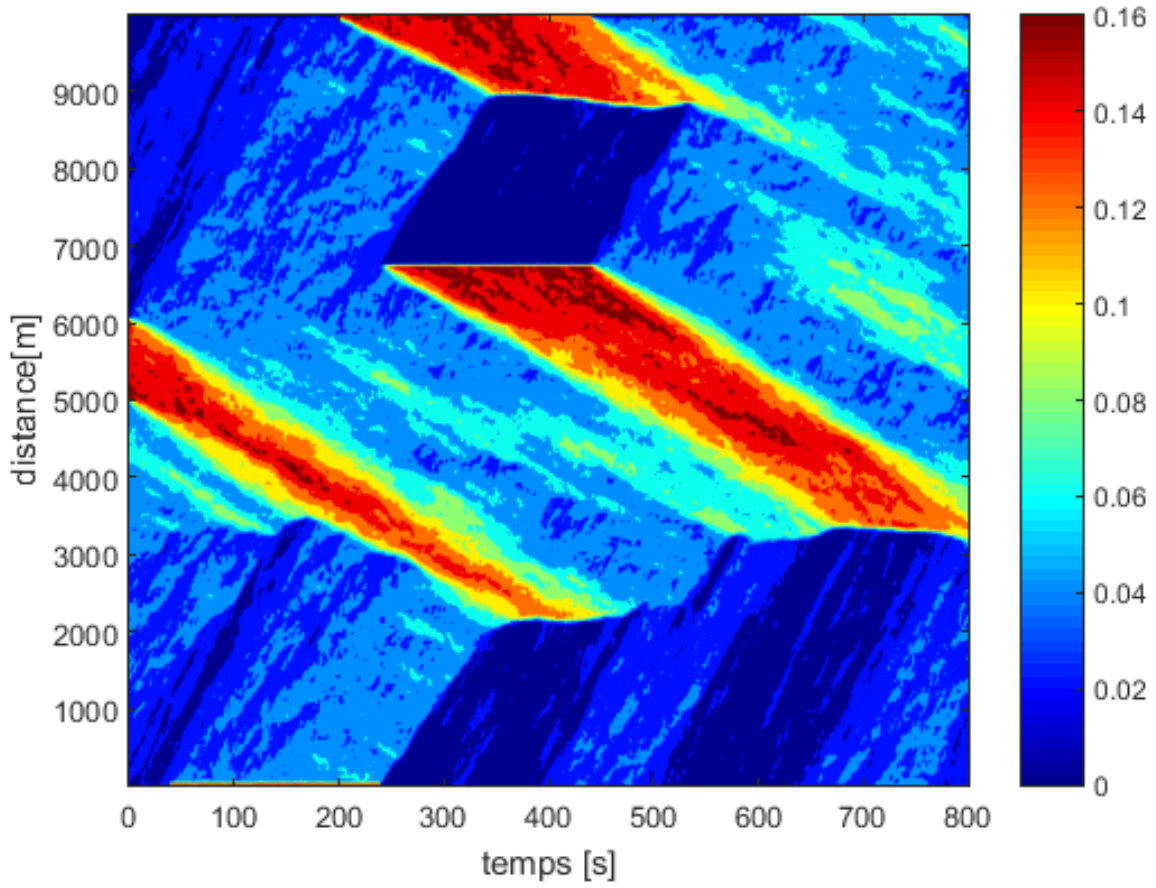

Figure 6.6 Trafic avec blocage du trafic temporaire.

Nous calculons de nouveau l'utilité de l'estimateur pour un nombre variable de capteurs équidistants. L'utilité de l'estimateur est reportée dans la figure 6.7.

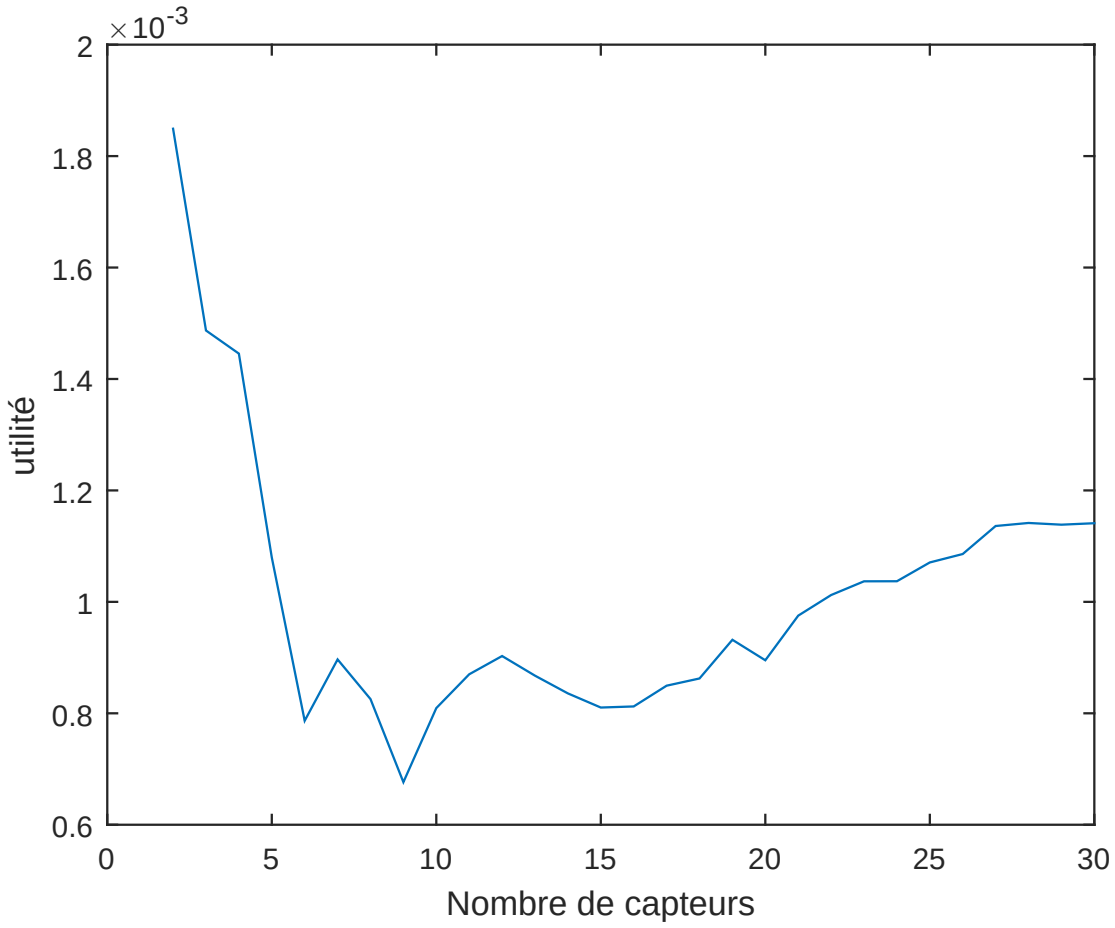

Figure 6.7 Erreur moyenne du filtre du filtre pour trafic avec incident

Nous avons la meilleure utilité pour 9 capteurs.

6.6 Choix des paramètres de confidentialité

Le choix des paramètres de confidentialité dépend principalement de la politique de protection voulue. Nous proposons cependant dans le tableau 6.2 les valeurs des écarts types du bruit gaussien à appliquer aux mesures de taux d'occupation pour avoir la (ϵ, δ)-confidentialité différentielle pour l'application au trafic.

δ \ ϵ	0.0100	1.1378	1.6544	1.9933	2.2460	2.4476	2.6152	2.7587
0	261.9698	1.9838	1.4008	1.1784	1.0547	0.9737	0.9154	0.8709
0.0062	15.9874	0.1325	0.0966	0.0828	0.0752	0.0701	0.0665	0.0637
0.0125	14.3496	0.1212	0.0889	0.0764	0.0695	0.0649	0.0616	0.0591
0.0187	13.3202	0.1142	0.0840	0.0724	0.0659	0.0617	0.0586	0.0562
0.0250	12.5516	0.1090	0.0805	0.0695	0.0634	0.0593	0.0564	0.0541
0.0313	11.9306	0.1049	0.0777	0.0672	0.0613	0.0574	0.0546	0.0525
0.0375	11.4052	0.1014	0.0753	0.0652	0.0596	0.0558	0.0531	0.0511
0.0438	10.9472	0.0984	0.0733	0.0635	0.0581	0.0545	0.0519	0.0499
0.0500	10.5393	0.0957	0.0714	0.0620	0.0568	0.0533	0.0507	0.0488
0.0562	10.1704	0.0933	0.0698	0.0607	0.0556	0.0522	0.0497	0.0478
0.0625	9.8325	0.0912	0.0683	0.0595	0.0545	0.0512	0.0488	0.0470
0.0687	9.5201	0.0892	0.0670	0.0584	0.0535	0.0503	0.0480	0.0462
0.0750	9.2289	0.0873	0.0657	0.0573	0.0526	0.0495	0.0472	0.0455
0.0813	8.9557	0.0856	0.0646	0.0564	0.0518	0.0487	0.0465	0.0448
0.0875	8.6980	0.0840	0.0635	0.0555	0.0510	0.0480	0.0458	0.0441
0.0938	8.4537	0.0825	0.0625	0.0546	0.0502	0.0473	0.0452	0.0435
0.1000	8.2212	0.0811	0.0615	0.0538	0.0495	0.0466	0.0446	0.0430
0.1063	7.9991	0.0797	0.0606	0.0531	0.0488	0.0460	0.0440	0.0424
0.1125	7.7863	0.0784	0.0597	0.0524	0.0482	0.0454	0.0434	0.0419
0.1188	7.5819	0.0772	0.0589	0.0517	0.0476	0.0449	0.0429	0.0414
0.1250	7.3849	0.0760	0.0581	0.0510	0.0470	0.0444	0.0424	0.0409
0.1313	7.1947	0.0748	0.0573	0.0504	0.0465	0.0439	0.0420	0.0405
0.1375	7.0107	0.0737	0.0565	0.0498	0.0459	0.0434	0.0415	0.0401
0.1437	6.8324	0.0727	0.0558	0.0492	0.0454	0.0429	0.0411	0.0396
0.1500	6.6593	0.0717	0.0551	0.0486	0.0449	0.0424	0.0406	0.0392

Tableau 6.2 Écarts types du bruit gaussien à appliquer aux mesures de taux d'occupation pour avoir la (ϵ, δ)-confidentialité différentielle

6.7 Échantillonnage spatial dynamique

Nous appliquons notre algorithme d'échantillonnage spatial pour proposer un meilleur positionnement des LPV. Un résultats de simulation est montré figure 6.8.

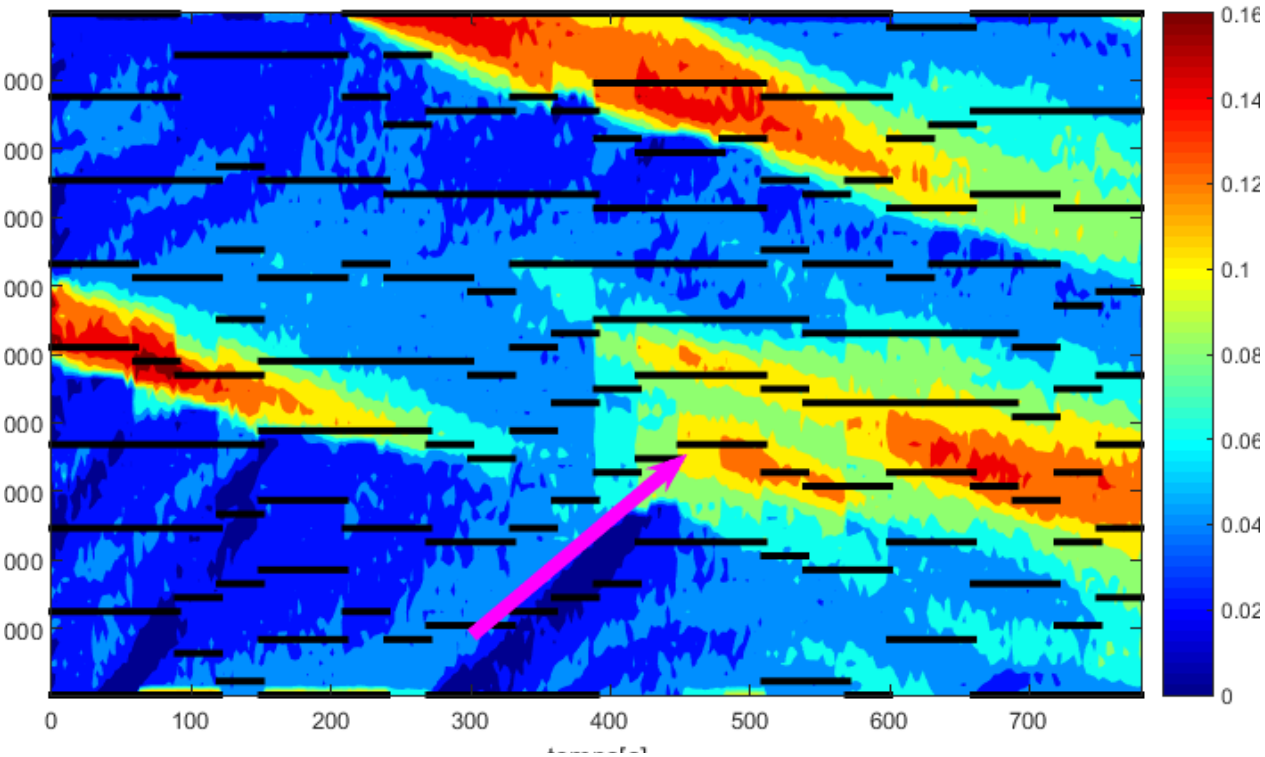

Figure 6.8 Positionnement des capteurs dynamiques

Les positions des capteurs sont représentés par les lignes horizontales noires. Nous observons par exemple la migration de certains capteurs vers une zone de plus grande incertitude (représentée par la flèche magenta).

6.8 Estimation du trafic à partir de trajectoires

À partir des densités synthétiques, nous pouvons avoir le champs des vitesses avec la relation densité-vitesse inversible décrite section 3.1.2. Nous pouvons alors générer des trajectoires en intégrant ces vitesses. Nous appliquons ensuite le mécanisme différentiellement confidentiel décrit section 5.2.3. Un résultat de simulation est représenté figure 6.9. La figure de gauche représente les données synthétiques simulées sous la forme de lignes noires représentant les trajectoires. La densité du trafic est représentée par la carte de couleur, mais ces données ne sont pas accessibles pour l'estimateur. La figure de droite représente l'estimation. Les mesures sont représentées par les points rouges. Nous voyons que l'estimé est proche de la réalité simulée lorsqu'il y a des mesures. Nous voyons que la congestion brusque empêche la prise de mesure et que cette congestion brutale et courte est non détectée. Cet exemple

montre l'importance d'avoir la possibilité de LPV mobiles qu'offre les données GPS

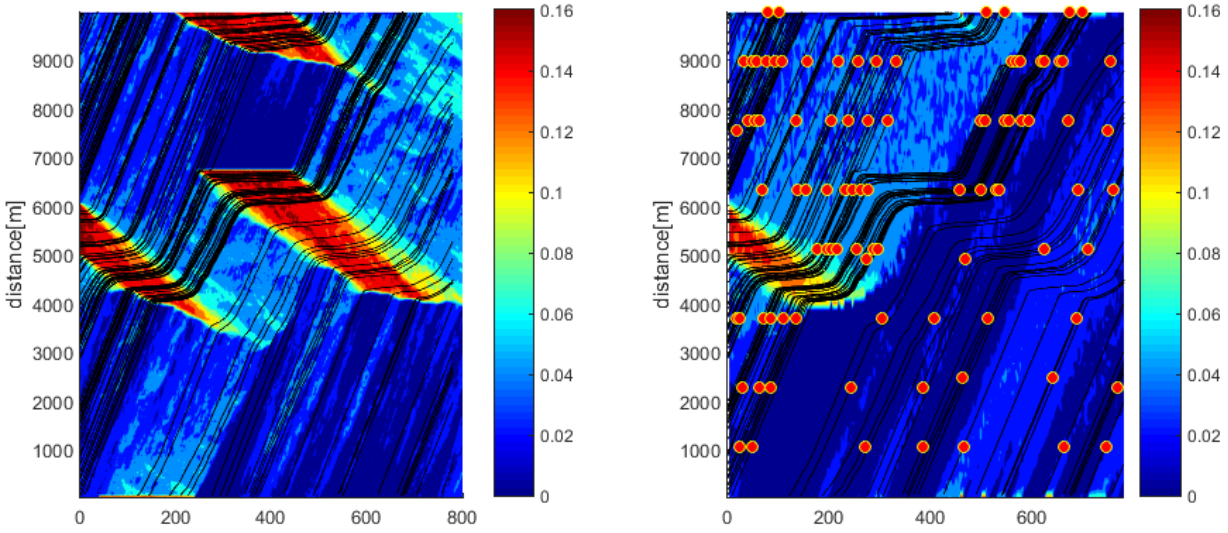

Figure 6.9 Estimation par mesures GPS, avec aperçu des trajectoires, des LPV et des mesures

6.9 Application au jeu de données Mobile Century

Nous pouvons appliquer notre estimateur différentiellement confidentiel au jeu de données Mobile Century. Nous présentons dans la section suivante les données disponibles avec le jeu de données Mobile Centuy.

Description des données

Le détail du jeu de donné est disponible en Annexe.

Données des détecteurs à boucle d'induction magnétique Le jeu de données contient les mesures de boucles d'induction magnétiques placées le long de la route. Les entrées de la base de données sont associées aux envois des capteurs. Les propriétés de ces entrées sont

- Identifiant de la ligne de capteurs.
- Temps unix (secondes).
- Comptes des véhicules pour chaque voies.
- Taux d'occupation des capteurs pour chaque voies.

La position en longitude et latitude des capteurs est aussi fournie, ainsi que la position kilométrique de la ligne de capteurs sur la route.

Lignes de Passage Virtuelles (LPV) Le jeu de données Mobile Century comprends aussi des entrées de vitesses que les GPS des téléphones mobiles envoient lorsque ceux-ci traversent des lignes des passages virtuelles (LPV). Les propriétés de ces entrées sont

— Identifiant du véhicule
— Temps unix (secondes)
— Coordonnées de la LPV
— Vitesse du véhicule

Temps de trajets Le jeu de données comprend des temps de trajets entre deux intersections le long de la route. Ces données ont été obtenues apès traitement d'images de caméras surplombant la route. Les propriétés de ces entrées sont

— Temps unix du départ (boulevard Stevenson)
— Temps unix d'arrivée (route Decoto)

Trajectoires individuelles Le jeu de données contient aussi des entrées de vitesses individuelles. Les propriétés sont

— Identifiant du véhicule
— Temps unix
— Latitude et longitude de l'envoi
— Repère kilométrique le long de la route
— Vitesse du véhicule

Le jeu de données contient aussi des trajectoires brutes non traitées issues des téléphones Nokia N95.

6.9.1 Mesures d'occupation

Nous utilisons les données de taux d'occupation fournies par les capteurs statiques pour estimer la densité du trafic. Nous appliquons donc le mécanisme décrit par l'algorithme 3. Le résultat graphique est produit figure 6.10.

6.9.2 Mesures GPS

En plus des mesures de capteurs statiques, nous disposons également de 1388 trajectoires de voitures individuelles envoyant leurs positions et vitesses toutes les 4 ou 5 secondes. Nous pouvons apercevoir ces trajectoire figure 6.11, avec le niveau de couleur représentant une densité calculée avec la formule (3.3).

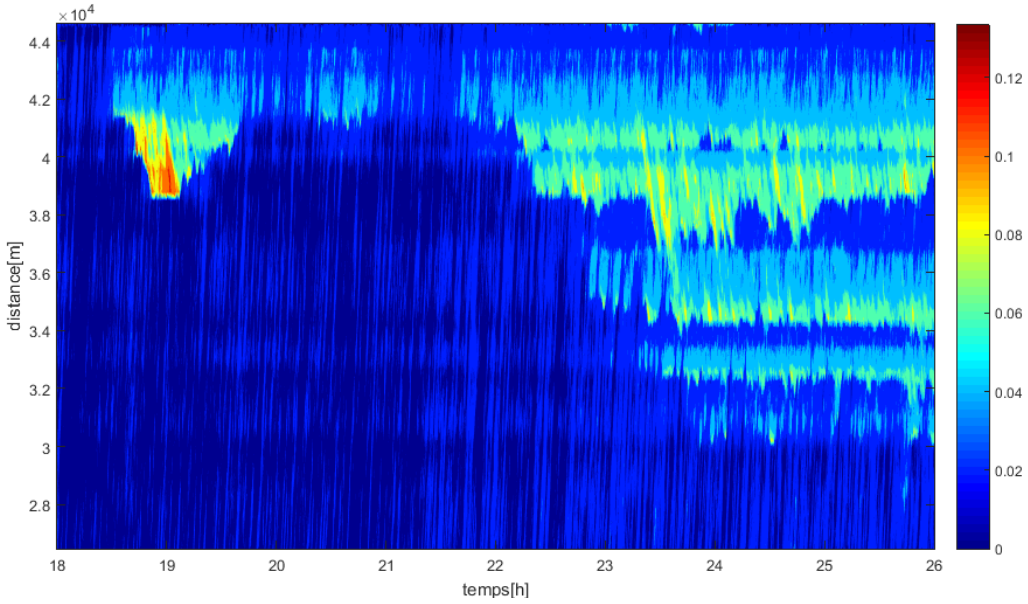

Figure 6.10 Estimation différentiellement confidentielle de densité du jeu de données Mobile Century à partir des mesures de taux d'occupation

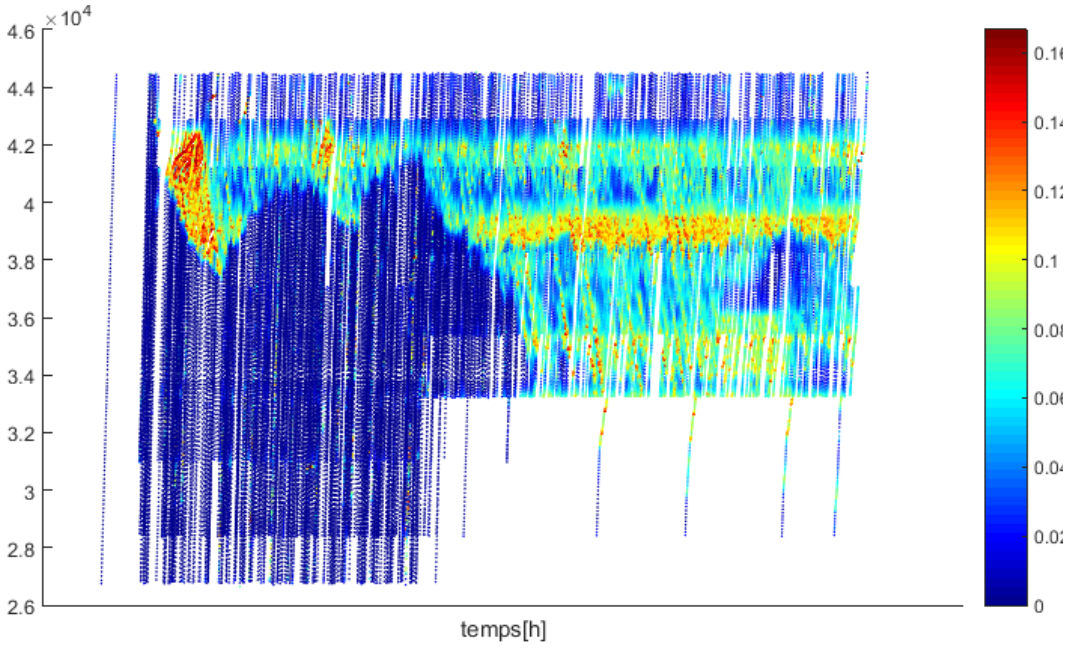

Figure 6.11 Trajectoires de voitures avec niveau de couleur correspondant à la densité calculée avec la formule (3.3)

Nous pouvons alors faire l'estimation différentiellement confidentielle de la densité (voir figure

6.12) à partir des trajectoires GPS en utilisant l'algorithme décrit section 5.2.3.

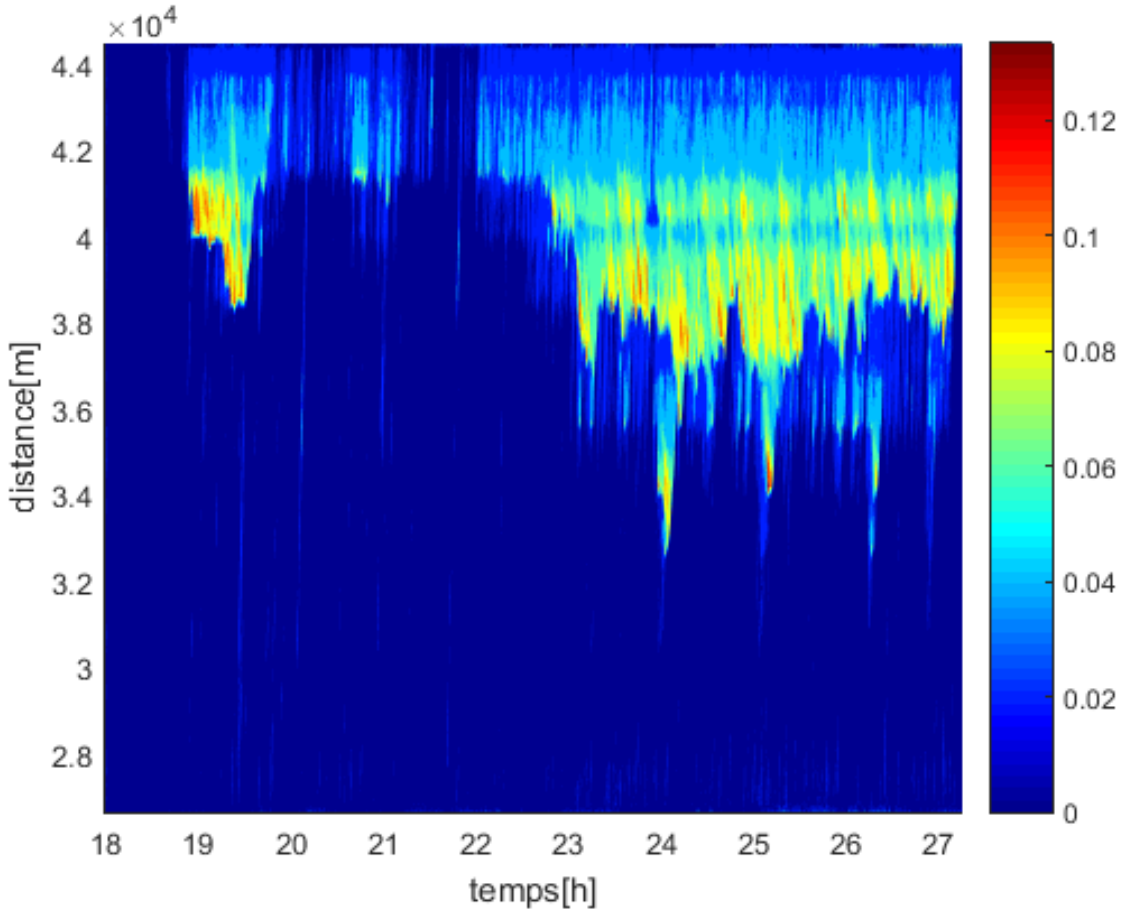

Figure 6.12 Estimation différentiellement confidentielle de la densité du trafic obtenue à partir des trajectoires GPS du jeu de données Mobile Century

Ces estimés de la densité sont assez précis pour être utilisables pour des applications liées au trafic et possèdent des garanties formelles de confidentialité protégeant les utilisateurs.

6.10 Modélisation dynamique d'ordre supérieur

Nous pouvons vérifier la robustesse de la qualité de notre estimateur en estimant avec ce même estimateur un trafic synthétique qui n'a pas été généré par un modèle hydrostatique. Nous pouvons utiliser des modèles microscopiques qui décrivent le trafic véhicule par véhicule et qui sont donc mieux adaptés pour décrire des comportement humains personnels.

6.10.1 Description mathématique

Les véhicules α sont décrits par l'abscisse curviligne $x_\alpha(t)$ du pare choc avant le long de la route (croissante dans la direction de la route), la vitesse $v_\alpha(t)$. Différents paramètres de conduite sont attribués à chaque conducteurs, comme le temps de réaction. Nous notons l_α la longueur du véhicule α. La distance entre le véhicule α et $\alpha - 1$ (voir schéma 6.13) est

$$s_\alpha = x_{\alpha-1} - l_{\alpha-1} - x_\alpha$$

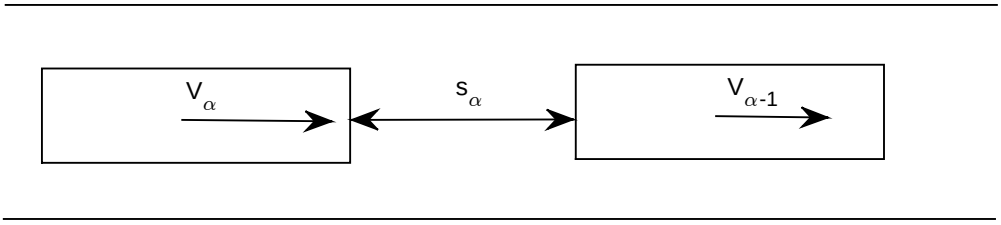

Figure 6.13 Schéma de modèle de suivi

Un modèle d'état simple correspondant à un contrôle de l'accélération par l'utilisateur en fonction de sa vitesse, de la vitesse du véhicule devant lui, et de l'espacement entre ces deux véhicules peut s'écrire

$$\dot{x}_\alpha(t) = \frac{\mathrm{d}x_\alpha(t)}{\mathrm{d}t} = v_\alpha(t)$$
$$\dot{v}_\alpha(t) = a(s_\alpha, v_\alpha, v_{\alpha-1})$$

Avec $\Delta v = v_\alpha - v_{\alpha-1}$. Nous utiliserons le modèle de différence de vitesses complet amélioré (Treiber and Kesting, 2013, p171)

$$\dot{v} = \frac{v_{\mathrm{opt}-v}}{\tau} - \frac{\gamma \Delta v}{\max\left[1, \frac{s}{v_0 T}\right]} \tag{6.1}$$

avec v_{opt} la vitesse optimale en régime permanent. Avec s l'espacement avec la voiture pré-

cédente nous avons les conditions sur $v_{opt}(s)$

$$v'_{opt}(s) \geq 0, v_{opt}(0) = 0, \lim_{s\to\infty} v_{opt}(s) = v_0$$

une proposition de fonction de vitesse optimale peut être trouvée dans Treiber and Kesting (2013, p169)

$$v_{opt}(s) = v_0 \frac{\tanh\left(\frac{s}{\Delta s} - \beta\right) + \tanh\beta}{1 + \tanh\beta}$$

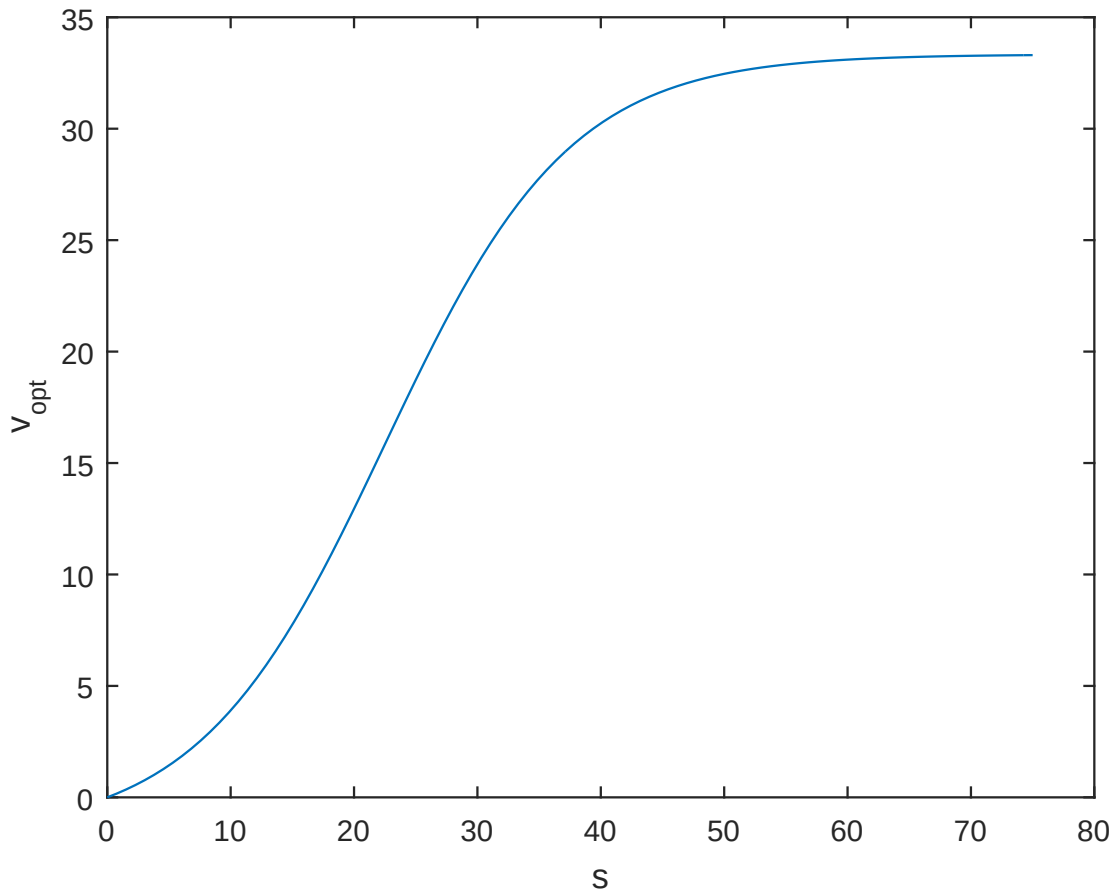

Figure 6.14 Vitesse optimale en fonction de l'espace inter véhicule

avec Δs une longueur de transition, β un facteur de forme, voir Fig 6.14. Le paramètre τ dans le modèle est une constante de temps pour l'accélération due à la différence de vitesse entre

la vitesse optimale est la vitesse réelle. Le paramètre γ contrôle la partie de l'accélération d'anticipation due à la variation de l'espace inter-véhicule. Le paramètre T contrôle la distance a partir de laquelle le conducteur débute l'anticipation.

6.10.2 Simulation de trafic et de données

Nous construisons un trafic synthétique sur le modèle de suivi (6.1). Nous pouvons voir le résultat de trajectoires figure 6.15. Les trajectoires représentée par un trait plus épais sont les trajectoires de véhicules équipés de GPS envoyant leurs données.

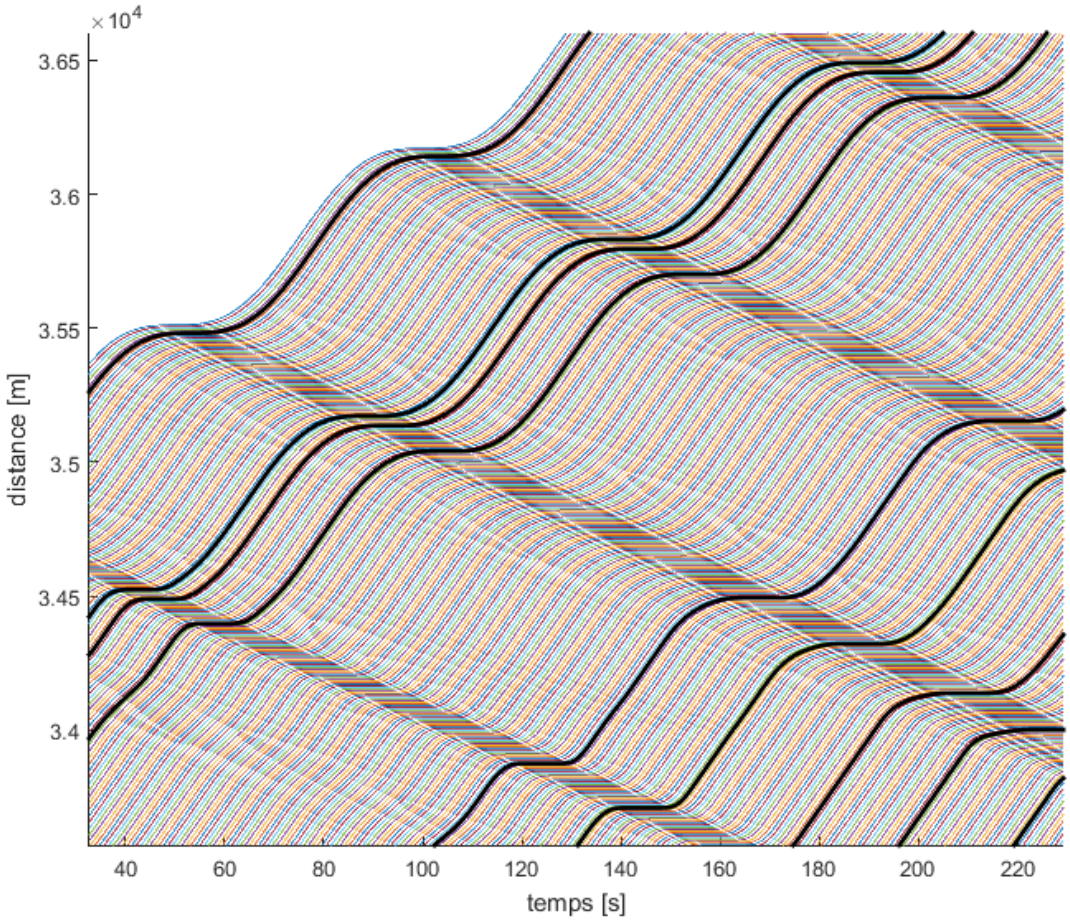

Figure 6.15 Trajectoires obtenues par modèle d'accélération de suivi

Nous pouvons appliquer notre algorithme d'estimation différentiellement confidentielle pour estimer ce trafic obtenu avec ce modèle supposé plus réaliste. Le résultat graphique est

présenté figure 6.16. Le trafic est globalement assez fluide, sauf sur certaines petites portions où il est congestionné.

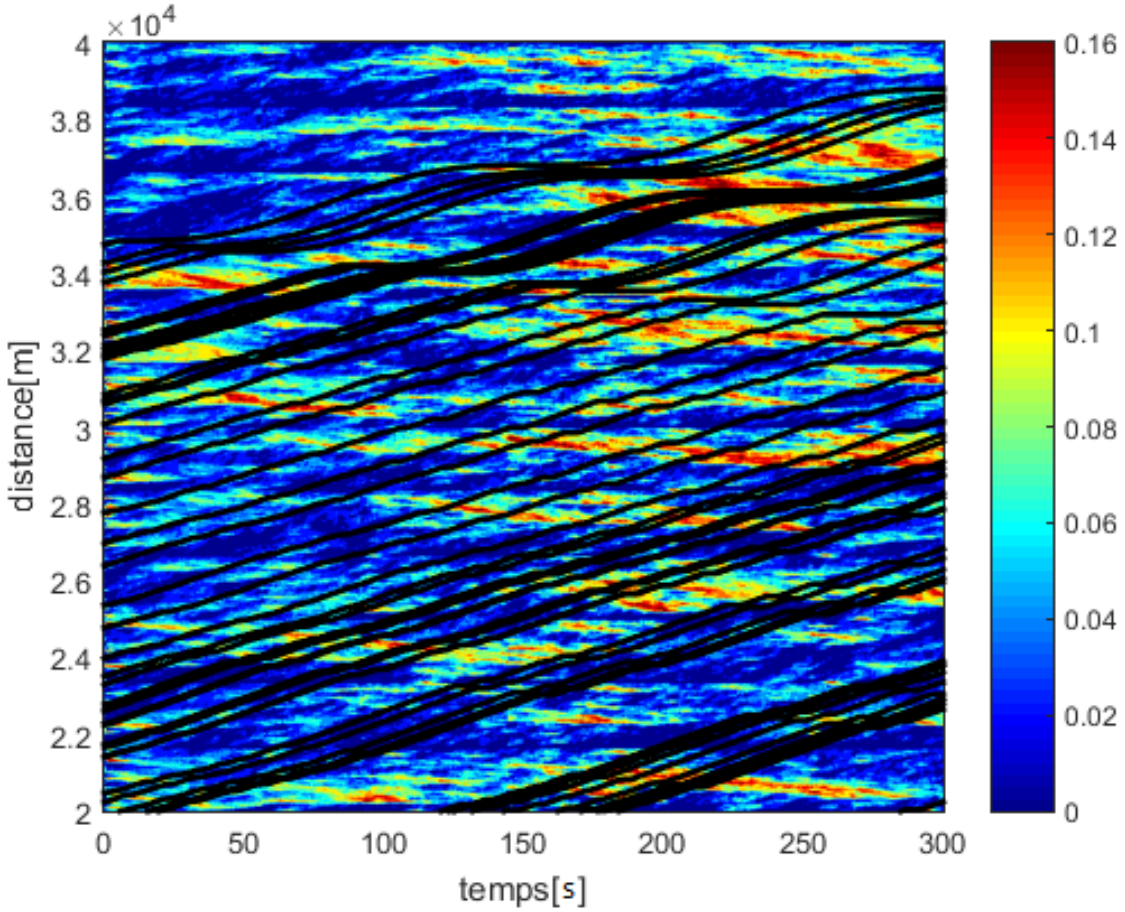

Figure 6.16 Estimation différentiellement confidentielle du trafic obtenue à partir de trajectoires simulées sur un modèle de suivi

Notre estimateur est donc également capable d'estimer des trafics générés avec d'autres modèles.

CHAPITRE 7 CONCLUSION

Cette étude nous a permis de concevoir et de tester un estimateur différentiellement confidentiel.

7.1 Synthèse des travaux

Pour concevoir ce nouvel estimateur, une étude bibliographique a été entreprise pour déterminer l'état de l'art en matière d'estimation de trafic et de confidentialité, et l'utilisation des différents filtres. Il en est ressorti un espace de conception pour un filtre de Kalman d'ensemble, ou un filtre à particules, différentiellement confidentiels. Nous avons présenté le modèle dynamique choisi et les variables d'états, ainsi que les modèles de mesures. Nous avons ensuite définit la confidentialité, la notion d'adjacence pour le trafic ainsi que les principaux résultats sur la confidentialité différentielle. Nous avons ensuite présenté les méthodes d'assainissement des mesures ainsi que les hypothèses faites pour restreindre la relation d'adjacence diminuer le bruit à introduire dans les mécanismes. Nous avons ensuite vérifié la plausibilité des vitesses faites à l'aide d'un trafic idéal et de données issues du jeu de données Mobile Century. Les différents mécanismes sont exposés sous la forme de théorèmes et les algorithmes d'implémentations sont fournis. Un module de changement dynamique de la position des capteurs est proposé. Les mécanismes développés sont testés et comparés sur des données synthétiques puis appliqués au jeu de données Mobile Century pour fournir un estimé différentiellement confidentiel de la densité.

7.2 Limitations de la solution proposée

Notre algorithme d'estimation différentiellement confidentiel, bien que donnant de bons résultats pour notre environnement de travail d'une route simple, fait face à des limitations lorsqu'on élargi son champ d'application. Nous énumérons les limitations ci-dessous.

1. La solution proposée est limitée au cas d'une route sans intersections. Trouver un moyen de conserver la confidentialité différentielle à travers les intersections permettrait d'étendre l'estimation à l'ensemble du réseau routier.
2. Nous n'avons pas fait d'hypothèses quand à la nature du type de trafic (voitures, camions,etc.).
3. Le mécanisme nécessite d'avoir un certain nombre de capteurs sur la route qu'on ne peut dépasser. Nous sommes donc limités à un nombre finis de points d'observation,

même si ceux-ci peuvent varier dans le temps.

4. Notre modèle dynamique ne prend pas en compte la typologie de la route (intersections, relations entre les voies, etc.).
5. La faible densité des données mobiles est une limitation à la qualité de l'estimation. Nous souhaitons par exemple posséder plusieurs mesures de vitesse, assez proches dans le temps, au passage d'une LPV pour en déduire la vitesse géométrique moyenne.

7.3 Améliorations futures

Nous n'avons pas de filtre à particules différentiellement confidentiels convergents. Le bruit à introduire avec les deux mécanismes étudiés (modification des poids et celui utilisant l'incertitude de la mesure) divergent avec le nombre de particules. Un travail futur serait de trouver les bonnes hypothèses pour faire fonctionner un tel filtre. Avoir un filtre à particules serait d'une grande utilité pour conserver plusieurs hypothèses de trafic. Nous pouvons reprendre points par points l'énumération de la sous-section des limitations (section 7.2) et envisager les améliorations possibles.

1. Trouver les bonnes hypothèses pour conserver la confidentialité différentielle après une intersection. Intuitivement, le degré de confidentialité va être amélioré au passage d'une intersection (incertitude sur la route empruntée par les véhicules).
2. Une analyse des temps d'occupations individuel $o_{p,t}^{(i)}$ véhicule par véhicule pourrait révéler le type de véhicule si nous avons une bonne idée de la vitesse. Le type de véhicule influe sur la longueur effective du véhicule dans le modèle de mesure.
3. La nature discrète du problème (éléments d'observations étant les voitures unitaires, présence ou absence pour les comptes et la relation d'adjacence), entraîne une contrainte sur le nombre discret de capteurs fixes (ou de LPV) à utiliser. Résoudre le problème précédent sur les intersections résoudrait en partie ce problème, puisque les routes présente fréquemment des intersections, laissant une densité de capteurs suffisante entre deux intersections pour rendre compte de l'état du trafic sur la portion de route.
4. Notre modèle pourrait être enrichi pour rendre compte de la typologie de la route, de la présence des intersections et de la signalisation pour intégrer les notions de code de la route, en prenant aussi en compte la présence de camions. Notre système pourrait aussi être connecté à une source d'information météorologique (le brouillard réduirait la vitesse libre v_0 par exemple, etc.).
5. La diffusion massive de smartphones, de voitures connectées et une gestion commune des données permet d'accroître la densité de données accessibles. Il n'est cependant

pas exclu qu'un meilleur algorithme de collecte des vitesses soit possible.

L'amélioration de l'utilité de l'estimé s'est faite en restreignant la relation d'adjacence. Il n'est pas exclu non plus de trouver une meilleure relation d'adjacence que celle proposée.

RÉFÉRENCES

C. Alecsandru, A. Quddus, K. C. Huang, B. Rouhieh, A. R. Khan, et Q. Zeng, "An assessment of the cell-transmission traffic flow paradigm : Development and applications", dans *Transportation Research Board 90th Annual Meeting*, no. 11-1152, 2011.

R. Bassily, A. Smith, et A. Thakurta, "Differentially private empirical risk minimization : Efficient algorithms and tight error bounds", *arXiv preprint arXiv :1405.7085*, 2014.

J. Bi, C. Chang, et Y. Fan, "Particle filter for estimating freeway traffic state in beijing", *Mathematical Problems in Engineering*, vol. 2013, 2013.

R. Billot, N.-E. El Faouzi, J. Sau, et F. De Vuyst, "Integrating the impact of rain into traffic management : online traffic state estimation using sequential monte carlo techniques", *Transportation Research Record : Journal of the Transportation Research Board*, no. 2169, pp. 141–149, 2010.

R. Boel et L. Mihaylova, "Modelling freeway networks by hybrid stochastic models", dans *Intelligent Vehicles Symposium, 2004 IEEE*. IEEE, 2004, pp. 182–187.

——, "A compositional stochastic model for real time freeway traffic simulation", *Transportation Research Part B : Methodological*, vol. 40, no. 4, pp. 319–334, 2006.

M. Burger, M. Van Den Berg, A. Hegyi, B. De Schutter, et J. Hellendoorn, "Considerations for model-based traffic control", *Transportation Research Part C : Emerging Technologies*, vol. 35, pp. 1–19, 2013.

E. S. Canepa et C. G. Claudel, "A framework for privacy and security analysis of probe-based traffic information systems", dans *Proceedings of the 2nd ACM international conference on High confidence networked systems (HiCoNS)*, 2013, pp. 25–32.

O. Cappé, S. J. Godsill, et E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo", *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

J. Carpenter, P. Clifford, et P. Fearnhead, "Improved particle filter for nonlinear problems", dans *Radar, Sonar and Navigation, IEE Proceedings-*, vol. 146, no. 1. IET, 1999, pp. 2–7.

H. B. Celikoglu, “An approach to dynamic classification of traffic flow patterns”, *Computer-Aided Civil and Infrastructure Engineering*, vol. 28, no. 4, pp. 273–288, 2013.

B. Y. Chen, H. Yuan, Q. Li, W. H. Lam, S.-L. Shaw, et K. Yan, “Map-matching algorithm for large-scale low-frequency floating car data”, *International Journal of Geographical Information Science*, vol. 28, no. 1, pp. 22–38, 2014.

H. Chen et H. A. Rakha, “Real-time travel time prediction using particle filtering with a non-explicit state-transition model”, *Transportation Research Part C : Emerging Technologies*, vol. 43, pp. 112–126, 2014.

H. Chen, H. A. Rakha, et S. Sadek, “Real-time freeway traffic state prediction : A particle filter approach”, dans *Intelligent Transportation Systems (ITSC), 2011 14th International IEEE Conference on.* IEEE, 2011, pp. 626–631.

H. Chen, H. Rakha, S. Sadek, et B. Katz, “Particle filter approach for real-yime freeway traffic state prediction”, dans *Transportation Research Board 91st Annual Meeting*, no. 12-2605, 2012.

J. Chen, K. H. Low, C. K.-Y. Tan, A. Oran, P. Jaillet, J. M. Dolan, et G. S. Sukhatme, “Decentralized data fusion and active sensing with mobile sensors for modeling and predicting spatiotemporal traffic phenomena”, *arXiv preprint arXiv :1206.6230*, 2012.

P. Cheng, Z. Qiu, et B. Ran, “Particle filter based traffic state estimation using cell phone network data”, dans *Intelligent Transportation Systems Conference, 2006. ITSC’06. IEEE.* IEEE, 2006, pp. 1047–1052.

——, “Traffic estimation based on particle filtering with stochastic state reconstruction using mobile network data”, dans *Transportation Research Board Annual Meeting*, 2006.

C. F. Daganzo, “The cell transmission model : A dynamic representation of highway traffic consistent with the hydrodynamic theory”, *Transportation Research Part B : Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, et V. D. Blondel, “Unique in the crowd : The privacy bounds of human mobility”, *Scientific Reports*, vol. 3, 2013.

G. Dervisoglu, G. Gomes, J. Kwon, R. Horowitz, et P. Varaiya, “Automatic calibration of the fundamental diagram and empirical observations on capacity”, dans *Transportation Research Board 88th Annual Meeting*, vol. 15, 2009.

R. Dong, A. A. Cárdenas, L. J. Ratliff, H. Ohlsson, et S. S. Sastry, "Quantifying the utility-privacy tradeoff in the smart grid", *arXiv preprint arXiv :1406.2568*, 2014.

C. Dwork et A. Roth, "The algorithmic foundations of differential privacy", *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2014.

C. Dwork, F. McSherry, K. Nissim, et A. Smith, "Calibrating noise to sensitivity in private data analysis", dans *Proceedings of the Third Theory of Cryptography Conference*, 2006, pp. 265–284.

C. Dwork, "Differential privacy", dans *Automata, languages and programming*. Springer, 2006, pp. 1–12.

C. Dwork et A. Roth, "The algorithmic foundations of differential privacy", *Theoretical Computer Science*, vol. 9, no. 3-4, pp. 211–407, 2013.

G. Evensen, "The ensemble Kalman filter : Theoretical formulation and practical implementation", *Ocean dynamics*, vol. 53, no. 4, pp. 343–367, 2003.

L. Fan, L. Xiong, et V. Sunderam, "Fast : differentially private real-time aggregate monitor with filtering and adaptive sampling", dans *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. ACM, 2013, pp. 1065–1068.

S. Fan et B. Seibold, "A comparison of data-fitted first order traffic models and their second order generalizations via trajectory and sensor data", *arXiv preprint arXiv :1208.0382*, 2012.

X. Feng, S. Xu, et X. Yan, "Traffic flow signal based traffic event reconstruction using sequential monte carlo methods", 2015.

S. Gisdakis, V. Manolopoulos, S. Tao, A. Rusu, et P. Papadimitratos, "Secure and privacy-preserving smartphone-based traffic information systems", *Intelligent Transportation Systems, IEEE Transactions on*, vol. 16, no. 3, pp. 1428–1438, 2015.

N. J. Gordon, D. J. Salmond, et A. F. Smith, "Novel approach to nonlinear/non-Gaussian Bayesian state estimation", dans *Radar and Signal Processing, IEE Proceedings F*, vol. 140, no. 2. IET, 1993, pp. 107–113.

P. Handel, J. Ohlsson, M. Ohlsson, I. Skog, et E. Nygren, "Smartphone-based measurement systems for road vehicle traffic monitoring and usage-based insurance", *Systems Journal, IEEE*, vol. 8, no. 4, pp. 1238–1248, 2014.

B. Hardjono, A. Wibowo, M. F. Rachmadi, et W. Jatmiko, “Mobile phones as traffic sensors with map matching and privacy considerations”, dans *Micro-NanoMechatronics and Human Science (MHS), 2012 International Symposium on.* IEEE, 2012, pp. 450–455.

B. Hardjono, R. Akbar, A. Wibisono, P. Mursanto, W. Jatmiko, et A. M. Arymurthy, “Fundamental diagram estimation using virtual detection zone in smart phones’ application and cctv data”, dans *Consumer Electronics (GCCE), 2014 IEEE 3rd Global Conference on.* IEEE, 2014, pp. 465–469.

A. C. Harvey, *Forecasting, structural time series models and the Kalman filter.* Cambridge university press, 1990.

A. Hegyi, L. Mihaylova, R. Boel, et Z. Lendek, “Parallelized particle filtering for freeway traffic state tracking”, dans *Control Conference (ECC), 2007 European.* IEEE, 2007, pp. 2442–2449.

B. Hoh, T. Iwuchukwu, Q. Jacobson, M. Gruteser, A. Bayen, J.-C. Herrera, R. Herring, D. Work, M. Annavaram, et J. Ban, “Enhancing privacy and accuracy in probe vehicle based traffic monitoring via virtual trip lines”, *IEEE Transactions on Mobile Computing*, vol. 11, no. 5, May 2012.

B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J.-C. Herrera, A. M. Bayen, M. Annavaram, et Q. Jacobson, “Virtual trip lines for distributed privacy-preserving traffic monitoring”, dans *Proceedings of the 6th international conference on Mobile systems, applications, and services.* ACM, 2008, pp. 15–28.

B. Hoh, T. Iwuchukwu, Q. Jacobson, D. Work, A. M. Bayen, R. Herring, J. C. Herrera, M. Gruteser, M. Annavaram, et J. Ban, “Enhancing privacy and accuracy in probe vehicle-based traffic monitoring via virtual trip lines”, *Mobile Computing, IEEE Transactions on*, vol. 11, no. 5, pp. 849–864, 2012.

Z. Huang, Y. Wang, S. Mitra, et G. E. Dullerud, “On the cost of differential privacy in distributed control systems”, dans *Proceedings of the 3rd international conference on High confidence networked systems.* ACM, 2014, pp. 105–114.

S. E. Jabari et H. X. Liu, “A stochastic model of traffic flow : Theoretical foundations”, *Transportation Research Part B : Methodological*, vol. 46, no. 1, pp. 156–174, 2012.

Z. Jia, C. Chen, B. Coifman, et P. Varaiya, “The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors”, dans *Proceedings of the 4th*

IEEE Conference on Intelligent Transportation Systems, 2001.

S. J. Julier et J. K. Uhlmann, "New extension of the Kalman filter to nonlinear systems", dans *AeroSense'97*. International Society for Optics and Photonics, 1997, pp. 182–193.

B. Jun, G. Wei, et Q. Long-Tao, "A genetic resampling particle filter for freeway traffic-state estimation", *Chinese Physics B*, vol. 21, no. 6, p. 068901, 2012.

Z. Khan, T. Balch, et F. Dellaert, "An MCMC-based particle filter for tracking multiple interacting targets", dans *Computer Vision-ECCV 2004*. Springer, 2004, pp. 279–290.

D. Kifer et A. Machanavajjhala, "A rigorous and customizable framework for privacy", dans *Proceedings of the 31st symposium on Principles of Database Systems*. ACM, 2012, pp. 77–88.

Q.-J. Kong, Z. Li, Y. Chen, et Y. Liu, "An approach to urban traffic state estimation by fusing multisource information", *Intelligent Transportation Systems, IEEE Transactions on*, vol. 10, no. 3, pp. 499–511, 2009.

Q.-J. Kong, Q. Zhao, C. Wei, et Y. Liu, "Efficient traffic state estimation for large-scale urban road networks", *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 1, pp. 398–407, 2013.

F. Koufogiannis et G. Pappas, "Diffusing private data over networks", *arXiv preprint arXiv :1511.06253*, 2015.

F. Koufogiannis, S. Han, et G. J. Pappas, "Optimality of the Laplace mechanism in differential privacy", *arXiv preprint arXiv :1504.00065*, 2015.

J. Le Ny, A. Touati, et G. J. Pappas, "Real-time privacy-preserving model-based estimation of traffic flows", dans *Proceedings of the Fifth International Conference on Cyber-Physical Systems (ICCPS)*, April 2014.

J. Le Ny, "Privacy-preserving nonlinear observer design using contraction analysis", *arXiv preprint arXiv :1507.02250*, 2015.

J. Le Ny et G. J. Pappas, "Differentially private filtering", *Automatic Control, IEEE Transactions on*, vol. 59, no. 2, pp. 341–354, 2014.

N. Li, T. Li, et S. Venkatasubramanian, "t-closeness : privacy beyond k-anonymity and l-diversity", dans *Proceedings of the 23rd IEEE International Conference on Data Enginee-*

ring, 2007.

T. Liebig, “Privacy preserving centralized counting of moving objects”, dans *AGILE 2015*. Springer, 2015, pp. 91–103.

L. Mihaylova, R. Boel, et A. Hegyi, “Freeway traffic estimation within particle filtering framework”, *Automatica*, vol. 43, no. 2, pp. 290–300, 2007.

L. Mihaylova, A. Hegyi, A. Gning, et R. K. Boel, “Parallelized particle and Gaussian sum particle filters for large-scale freeway traffic systems”, *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 1, pp. 36–48, 2012.

L. Mihaylova, A. Y. Carmi, F. Septier, A. Gning, S. K. Pang, et S. Godsill, “Overview of Bayesian sequential monte carlo methods for group and extended object tracking”, *Digital Signal Processing*, vol. 25, pp. 1–16, 2014.

A. Narayanan et V. Shmatikov, “Robust de-anonymization of large sparse datasets (how to break anonymity of the Netflix Prize dataset)”, dans *Proceedings of the IEEE Symposium on Security and Privacy*, 2008.

K. Nummiaro, E. Koller-Meier, et L. Van Gool, “An adaptive color-based particle filter”, *Image and vision computing*, vol. 21, no. 1, pp. 99–110, 2003.

K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, et D. G. Lowe, “A boosted particle filter : Multitarget detection and tracking”, dans *Computer Vision-ECCV 2004*. Springer, 2004, pp. 28–39.

A. Pascale, M. Nicoli, et U. Spagnolini, “Cooperative Bayesian estimation of vehicular traffic in large-scale networks”, *Intelligent Transportation Systems, IEEE Transactions on*, vol. 15, no. 5, pp. 2074–2088, 2014.

C. Quek, M. Pasquier, et B. Lim, “A novel self-organizing fuzzy rule-based system for modelling traffic flow behaviour”, *Expert Systems with applications*, vol. 36, no. 10, pp. 12 167–12 178, 2009.

L. J. Ratliff, R. Dong, H. Ohlsson, A. A. Cárdenas, et S. S. Sastry, “Privacy and customer segmentation in the smart grid”, dans *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 2136–2141.

S. Särkkä, *Bayesian Filtering and Smoothing*. Cambridge University Press, 2013.

C. Snyder, T. Bengtsson, P. Bickel, et J. Anderson, “Obstacles to high-dimensional particle filtering”, *Monthly Weather Review*, vol. 136, no. 12, pp. 4629–4640, 2008.

K. Staňková et B. De Schutter, “On freeway traffic density estimation for a jump Markov linear model based on daganzo’s cell transmission model”, dans *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on*. IEEE, 2010, pp. 13–18.

A. Sumalee, R. Zhong, T. Pan, et W. Szeto, “Stochastic cell transmission model (SCTM) : a stochastic dynamic traffic model for traffic state surveillance and assignment”, *Transportation Research Part B : Methodological*, vol. 45, no. 3, pp. 507–533, 2011.

L. Sweeney, “k-anonymity : A model for protecting privacy”, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 10, no. 05, pp. 557–570, 2002.

——, “Weaving technology and policy together to maintain confidentiality”, *Journal of Law, Medicine and Ethics*, vol. 25, pp. 98–110, 1997.

M. Treiber et A. Kesting, “Traffic flow dynamics”, *Traffic Flow Dynamics : Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*, 2013.

R. Van Der Merwe, A. Doucet, N. De Freitas, et E. Wan, “The unscented particle filter”, dans *NIPS*, vol. 2000, 2000, pp. 584–590.

Y. Wang, M. Papageorgiou, A. Messmer, P. Coppola, A. Tzimitsi, et A. Nuzzolo, “An adaptive freeway traffic state estimator”, *Automatica*, vol. 45, no. 1, pp. 10–24, 2009.

Y. Wang, P. Coppola, A. Tzimitsi, A. Messmer, M. Papageorgiou, et A. Nuzzolo, “Real-time freeway network traffic surveillance : Large-scale field-testing results in southern Italy”, *Intelligent Transportation Systems, IEEE Transactions on*, vol. 12, no. 2, pp. 548–562, 2011.

Y. Wang, Z. Huang, S. Mitra, et G. E. Dullerud, “Entropy-minimizing mechanism for differential privacy of discrete-time linear feedback systems”, dans *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014, pp. 2130–2135.

Y.-X. Wang, S. E. Fienberg, et A. Smola, “Privacy for free : Posterior sampling and stochastic gradient monte carlo”, *arXiv preprint arXiv :1502.07645*, 2015.

D. B. Work, O.-P. Tossavainen, S. Blandin, A. M. Bayen, T. Iwuchukwu, et K. Tracton, “An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices”, dans *Proceedings of the 47th IEEE Conference on Decision and Control*, December 2008, pp. 5062–5068.

H. Wu, F. Sun, et H. Liu, “Fuzzy particle filtering for uncertain systems”, *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 5, pp. 1114–1129, 2008.

P. Wu, H. Xue, et X. Hu, “Particle filter based traffic data assimilation with sensor informed proposal distribution”, dans *Proceedings of the 48th Annual Simulation Symposium*. Society for Computer Simulation International, 2015, pp. 173–180.

L. Xiong, “Adaptive differentially private data release for data sharing and data mining”, dans *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 891–891.

W. Yi, M. R. Morelande, L. Kong, et J. Yang, “A computationally efficient particle filter for multitarget tracking using an independence approximation”, *Signal Processing, IEEE Transactions on*, vol. 61, no. 4, pp. 843–856, 2013.

H. Zhang, Y. Shu, P. Cheng, et J. Chen, “Privacy and performance trade-off in cyber-physical systems”, *IEEE Networks*.

J.-D. Zhang, J. Xu, et S. S. Liao, “Aggregating and sampling methods for processing GPS data streams for traffic state estimation”, *Intelligent Transportation Systems, IEEE Transactions on*, vol. 14, no. 4, pp. 1629–1641, 2013.

ANNEXE A DESCRIPTION DU JEU DE DONNÉES MOBILE CENTURY

Données des détecteurs à boucle d'induction magnétique La description des données des boucles à induction est détaillée dans le tableau A.

`"pems_id"` :	Identifiant du capteur.
`"unixtime"` :	Temps unix (secondes).
`"flows"` :	Comptes des véhicules sur 30 secondes.
`"occs"` :	Taux d'occupation du capteur sur 30 secondes.

Propriétés des capteurs par boucle d'induction La description des propriétés des boucles à induction est détaillée dans le tableau A.

`"pems_id"` :	Identifiant du capteur.
`"abs_pm"` :	Position du capteur sur la route en miles.
`"lat","lon"` :	réfèrent à la latitude et la longitude de la position du capteur.

Lignes de Passage Virtuelles (LPV) La description des LPV est détaillée dans le tableau A. Les téléphones mobiles envoient leurs données GPS lorsqu'ils traversent une LPV.

`"vtl_id"` :	Identifiant du véhicule.
`"unixtime"` :	Temps unix (secondes).
`"coordinate"` :	Réfèrent à la latitude et la longitude de la position de la LPV.
`"vel_mph"` :	Réfèrent à la vitesse du capteur mobile à la traversée de la LPV.

Temps de trajets Les temps de trajets entre le boulevard Stevenson et la route Decoto sont détaillés dans le tableau A.

`"departure_time"` :	Temps unix du départ du véhicule du boulevard Stevenson.
`"travel_time"` :	Temps du trajet du véhicule jusqu'à la route Decoto.

Trajectoires individuelles Les trajectoires brutes non traitées issues de GPS de véhicules individuels. Voir tableau A.

`"unixtime"` :	Temps unix (secondes).
`"latitude","Longitude"` :	réfèrent à la latitude et la longitude de la position.
`"postmile"` :	Réfèrent à la position en mile du capteur mobile sur la route.
`"speed"` :	Vitesse (miles/h) du véhicule

Trajectoires individuelles Nokia N95 Les trajectoires brutes non traitées issues des téléphones Nokia N95. Voir tableau A.

`"unixtime"` :	Temps unix (secondes).
`"latitude","Longitude"` :	réfèrent à la latitude et la longitude de la position.
`"speed"` :	Vitesse (miles/h) du véhicule