Leveraging geolocalization technologies to model and estimate urban traffic

Aude Hofleitner

UNIVERSITÉ PARIS-EST

Thèse de doctorat d'Université Paris-Est

Villes, Transports et Territoires

Aude Hofleitner

Développement d'un modèle d'estimation des variables de trafic urbain basé sur l'utilisation des technologies de géolocalisation

Thèse dirigée par Alexandre Bayen et Habib Haj-Salem,
Co-encadrée par Jean-Patrick Lebacque

Soutenue le 4 décembre 2012

Rapporteurs:
Paola Goatin (INRIA - Opale)
Arnaud De La Fortelle (Mines ParisTech - CAOR)
Jury:
Alexandre Bayen (UC Berkeley)
Jean-Patrick Lebacque (IFSTTAR - GRETTIA)
Habib Haj-Salem (IFSTTAR - GRETTIA)
Ludovic Leclercq (IFSTTAR - LICIT)
François Peyret (IFSTTAR - MACS)

Synthèse

L'augmentation de la mobilité accroît continuellement la demande de transport, et en particulier la demande de transport routier. Face à cette augmentation de la demande, la première réaction des gouvernements a été d'augmenter l'offre de transport en augmentant la capacité des infrastructures existantes et en densifiant le réseau. Aujourd'hui, les politiques de développement durable cherchent à optimiser l'utilisation des infrastructures existantes. Afin d'améliorer l'utilisation des infrastructures, les usagers ont besoin d'information disponible en temps réel en fonction de leur position. Un système d'information du trafic à large échelle offre une vue générale de l'état du trafic et permet d'optimiser les choix d'itinéraires des utilisateurs et de faciliter la prise de décision des gestionaires du réseau.

Afin d'optimiser l'utilisation du réseau de transport, les usagers ont besoin d'information fiable, disponible en temps réel et qui dépend de la position géographique de l'usager. L'émergence des services internet sur les téléphones portables et la rapide prolifération des systèmes de géo-localisation permet le développement de nouveaux services visant à améliorer l'expérience des usagers et de modifier la façon dont les usagers interagissent avec les systèmes de transport.

Le développement de systèmes d'information du trafic a longtemps été synonyme de larges projets d'équipement du réseau avec des capteurs permettant de mesurer les conditions de circulation. Ces larges projets sont très couteux pour les collectivités publiques et sont généralement limités au réseau autoroutier. La densité du réseau urbain limite la possibilité de généraliser cette approche sur l'ensemble du réseau.

Le but de cette thèse est d'étudier la possibilité d'utiliser les technologies de géo-localisation ainsi que les capacités de communication des téléphones portables pour développer un système d'information du trafic sur le réseau urbain sans nécessiter l'installation de capteurs ou d'infrastructure additionnels.

Le Chapitre 1 souligne l'importance du développement de systèmes d'information du trafic. Le chapitre présente l'étendue des projets académiques et des projets industriels visant à améliorer l'information disponible en temps réel. Les différents types de données permettant d'obtenir des informations sur l'état du trafic sont détaillés dans la deuxième partie du chapitre. Cette analyse des données disponibles montre le potentiel des données de géo-localisation tout en soulignant le besoin de développer des algorithmes spécifiques afin d'extraire l'information de ces données qui sont souvent bruitées et ne sont envoyées, jusqu'à présent, que par une faible proportion des véhicules circulant sur le réseau.

Le Chapitre 2 étudie la littérature et analyse les modèles et algorithmes d'estimation du trafic urbain et autoroutier, en se concentrant sur l'estimation du trafic urbain. Il apparaît que les modèles existants ne permettent pas d'extraire l'information contenue dans les données de géo-localisation provenant de véhicules traceurs envoyant leur position de façon peu fréquente (au plus toutes les 20 ou 30 secondes). Les contraintes concernant la quantité et le type de données disponibles ainsi que la complexité des algorithmes existants limite les possibilités d'utiliser de tels modèles à large échelle. La littérature suggère que l'utilisation de modèles statistiques est adaptée au problème d'estimation du trafic urbain car ces modèles sont capables de représenter la variabilité des conditions de circulation. Cette thèse choisi cette approche et s'intéresse au développement de modèles capables d'incorporer les données provenant de véhicules traceurs qui envoient leur position périodiquement avec une fréquence faible, aussi bien pour développer des modèles d'apprentissage de la dynamique, de façon historique que pour estimer les conditions de circulation en temps réel.

Les données provenant de véhicules traceurs sont utilisées avec succès pour estimer les conditions de circulation sur le réseau autoroutier. Sur le réseau autoroutier, les algorithmes d'estimation sont souvent basés sur des modèles hydrodynamiques qui représentent l'écoulement des véhicules sur le réseau. Le Chapitre 3 présente un algorithme permettant d'utiliser les données provenant de véhicules traceurs pour estimer de façon fiable et robuste les temps de parcours sur le réseau autoroutier. Le succès de ce modèle motive le développement de modèles hydrodynamiques pour le réseau urbain adaptés à la faible fréquence des données envoyées par les véhicules traceurs.

Le Chapitre 4 présente un modèle de trafic urbain basé sur la modélisation du trafic comme un fluide hydrodynamique et utilisant les résultats existants dans ce domaine. Le modèle fait des hypothèses simplificatrices qui permettent d'obtenir des expressions analytiques et de contrôler la complexité du modèle tout en lui laissant des caractéristiques réalistes. Le modèle caractérise la formation de files d'attentes horizontales en amont des intersections.

Ce modèle de trafic urbain est utilisé dans le Chapitre 5 pour obtenir l'expression paramétrique de la probabilité de distribution de la position des véhicules sur un arc du réseau, délimité par des intersections. Les paramètres de la distribution ont une interprétation physique et sont appris à partir de données de position envoyées par les véhicules traceurs. En plus de l'information sur le niveau de trafic d'un arc, ces distributions paramétriques ont des applications intéressantes dans le domaine systèmes de transport intelligents. Les distributions peuvent être utilisées afin d'utiliser les temps de parcours des véhicules traceurs lorsque le temps de parcours est mesuré entre deux points qui ne coïncident pas avec les débuts et fins des arcs. La loi de probabilité peut également être utilisée afin de détecter automatiquement la présence de feux de signalisations afin de créer ou de mettre à jour les bases de données cartographiques à un coût plus faible et avec une meilleure efficacité et fiabilité que les techniques actuelles.

Le modèle de trafic urbain est également utilisé au sein du Chapitre 6 pour obtenir l'expression paramétrique de la loi de probabilité des temps de parcours entre deux points quelconques du réseau. Il est essentiel que ces distributions soient obtenues entre deux points arbitraires du réseau pour pouvoir les utiliser avec des données provenant de véhicules traceurs envoyant leur position avec une fréquence faible. Cette caractéristique du modèle est une des nouveautés les plus importantes développées dans ce travail.

Le Chapitre 7 continue le développement du modèle de trafic urbain présenté dans le Chapitre 4. Le modèle cherche à relaxer une des hypothèses formulées dans le Chapitre 4 concernant la structure des flux d'arrivée à l'amont de chaque arc du réseau. L'augmentation de la précision du modèle est utile lorsque les données contiennent suffisamment d'information pour capturer le nouveau phénomène pris en compte par le modèle. En particulier, ce modèle a des applications intéressantes dans les corridors urbains pour lesquels la synchronisation des feux est importante et peut être optimisée pour minimiser les temps d'attentes des véhicules. Le chapitre présente les dérivations analytiques permettant la synchronisation des feux et propose des stratégies de contrôle du trafic en corridor urbain. Les expressions analytiques ont l'avantage de permettre la mise en place de stratégies de contrôle en temps réel et de les adapter rapidement lorsque les conditions de circulations évoluent. Les expressions aident également à améliorer la compréhension de la dynamique du trafic urbain.

Mots clés

Estimation
Traffic urbain
Modèle statistique
Données de géo-localisation
Prédiction

Résumé

Face à l'augmentation de la mobilité, les politiques de développement durable cherchent à optimiser l'utilisation des infrastructures de transport existantes. En particulier, les systèmes d'information du trafic à large échelle ont le potentiel d'optimiser l'utilisation du réseau de transport. Ils doivent fournir aux usagers une information fiable en temps réel leur permettant d'optimiser leurs choix d'itinéraires. Ils peuvent également servir d'outils d'aide à la décision pour les gestionnaires du réseau. La thèse étudie comment l'émergence des services Internet sur les téléphones portables et la rapide prolifération des systèmes de géolocalisation permet le développement de nouveaux services d'estimation et d'information des conditions de trafic en réseau urbain.

L'utilisation des données provenant de véhicules traceurs nécessite le développement de modèles et d'algorithmes spécifiques, afin d'extraire l'information de ces données qui ne sont envoyées, jusqu'à présent, que par une faible proportion des véhicules circulant sur le réseau et avec une fréquence faible. La variabilité des conditions de circulations, due à la présence de feux de signalisation, motive une approche statistique de la dynamique du trafic, tout en intégrant les principes physiques hydrodynamiques (formation et dissolution de files d'attentes horizontales). Ce modèle statistique permet d'intégrer de façon robuste les données bruitées envoyées par les véhicules traceurs, d'estimer les paramètres physiques caractérisant la dynamique du trafic et d'obtenir l'expression paramétrique de la loi de probabilité des temps de parcours entre deux points quelconques du réseau.

La thèse s'appuie sur les données et les infrastructures développées par le projet *Mobile Millennium* à l'Université de Californie, Berkeley pour valider les modèles et algorithmes proposés. Les résultats soulignent l'importance du développement de modèles statistiques et d'algorithmes adaptés aux données disponibles pour développer un système opérationnel d'estimation du trafic à large échelle.

Title

Leveraging geolocalization technologies to model and estimate urban traffic

Keywords

Estimation
Arterial traffic
Statistical model
Geolocalization data
Forecast

Abstract

Sustainable mobility development requires the optimization of existing transportation infrastructure. In particular, ubiquitous traffic information systems have the potential to optimize the use of the transportation network. The system must provide accurate and reliable traffic information in real-time to optimize mobility choices. Successful implementations are also valuable tools for traffic management agencies. The thesis studies how the emergence of Internet services and location based services on mobile devices enable the development of novel Intelligent Transportation Systems which estimate and broadcast traffic conditions in arterial networks.

Sparsely sampled probe data is the main source of arterial traffic data with the prospect of broad coverage in the near future. The small number of vehicles that report their position at a given time and the low sampling frequency require specific models and algorithms to extract valuable information from the available data. On the one hand, the variability of traffic conditions in urban networks, caused mainly by the presence of traffic lights, motivates a statistical approach of arterial traffic dynamics. On the other hand, an accurate modeling of the physics of arterial traffic from hydrodynamic theory (formation and dissolution of horizontal queues) ensures the physical validity of the model. The thesis proposes to integrate the dynamical model of arterial traffic in a statistical framework to integrate noisy measurements from probe vehicle data and estimate physical parameters, which characterize the traffic dynamics. In particular, the thesis derives and estimates the probability distributions of vehicle location and of travel time between arbitrary locations.

The thesis leverages the data and the infrastructure developed by the *Mobile Millennium* project at the University of California, Berkeley to validate the models and algorithms. The results underline the importance to design statistical models for sparsely sampled probe vehicle data in order to develop the next generation of operation large-scale traffic information systems.

Laboratoires d'accueil

Partners for Advanced Transportation TecHnology (PATH)
University of California, Berkeley
2105 Bancroft Way, Suite 300
Berkeley, CA 94720-3830
United States of America

Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux (IFSTTAR)
Génie des Réseaux de Transport Terrestres et Informatique Avancée (GRETTIA)
2 rue de la Butte Verte
93166 Noisy le Grand
France

Contents

List of Figures

List of Tables

Acknowledgements

I would like to dedicate this PhD to my mom, Anne. She had a passion for travel, for working in international environment and socializing with people from various cultures which has strongly shaped my personality and my interests both personally and professionally. In particular, she has always encouraged me to seek international opportunities and to benefit from the amazing cultural experiences that they provide. One of these opportunities was to go to Berkeley to join the *Mobile Millennium* team for a one year internship, starting in July 2008. I seized the opportunity and immerse myself in an amazing journey of learning about the world and about myself! The end of 2008 brought dramatic emotions as my mom unexpectedly and so suddenly passed away. I am grateful for the support I received from my dad, Patrick and my sister, Céline as well as my family and friends. They encouraged me to continue my journey, helped me overcome this tragic life event and support me despite the distance.

I was also amazed by the attentive and personal support of my colleagues with whom I had only been working for a few months. In particular, Alex Bayen, my internship supervisor at the time, and my friend and colleague Saneesh Apte played an incredibly valuable role. The research environment had already exceeded my expectations in terms of dynamism, interest in the project and intellectual challenge; the realization that I could relate to my colleagues on the personal level as well convinced me that I wanted to pursue my graduate studies in this successful and fulfilling environment. With the help of Jean-Patrick Lebacque and Habib Haj-Salem, I demonstrated the interest of my research directions to the Corps des Ponts et Chaussées and started my PhD in August 2009.

During my internship and the beginning of my PhD, I benefited from the guidance of senior PhD students. Within the first month, Christian Claudel taught me about his work on viability theory. His excitement and interest for research and mathematics were contagious and lead to fruitful collaborations throughout my PhD. Saurabh Amin struck me with his mind overflowing with novel ideas and theoretical contributions. I really enjoyed our white board discussions and brainstorming sessions on traffic modeling, estimation, statistics, change detection, sampling, privacy and so much more. In many ways, these discussions have inspired a lot of the research which I have pursued throughout my PhD. In particular, Saurabh had a great vision on modeling arterial traffic, focusing on defining the appropriate level of model complexity given specific estimation, sensing or control goals, which I have followed and developed in my research.

Through his work on the *Path Inference Filter*, Timothy Hunter has made possible a lot of the numerical validation of my work. Timothy has spent countless hours developing, implementing, improving and running his algorithm on the millions of data points received in the *Mobile Millennium* system everyday, to turn noisy sparse GPS measurements into filtered path travel times which can be used for traffic estimation. Anyone who has worked, looked at or even imagined working with this type of data will understand the hard and valuable work of Timothy and realize the importance that it had in the completion of my work. Besides Timothy, the entire team of *Mobile Millennium* has developed the technical support and the infrastructure to develop the numerical applications of my

work. In particular, a lot of the results of this dissertation would not have been possible without the work of the team to set up and maintain the data feeds and the database, the development of the *Mobile Millennium* system and the road network abstraction.

I have found my best and dearest PhD collaborator in Ryan Herring. Ryan's view to research and scientific problems seems to perfectly complement mine. Working with him has been incredibly productive and enjoyable. Between developing models, designing code architecture, taking turns at writing papers or working on homework sets together, we have shared a lot of the PhD joys together. Ryan was also able to help me through the hardest times of the PhD and his trust in my potential to succeed and overcome obstacles has always amazed me. I feel privileged to have the chance to have someone so dear to me to celebrate the happy times of the PhD life and give me the courage and strength to overcome the deceptive times.

I am incredibly thankful to Pieter Abbeel for pointing out an idea which I had first developed as part of a class project. Pieter had started following my work halfway through my internship and was offering his machine learning and artificial intelligence expertise for the arterial estimation efforts that I was pursuing with Ryan, Saurabh, and later Timothy. My class project idea consisted in a simplification of the arterial traffic flow physics to model the distribution of vehicle location on a road segment. Pieter pushed me to develop this idea further and generalize it to travel times, in a hybrid approach of traffic theory and statistical modeling and inference. I am really happy for the vision of Pieter in this work, which turned into an essential part of my PhD. I really enjoyed his enthusiasm and appreciate the trust he had in this model. These exciting and dynamic meetings and research directions in the first six months of my PhD drove my research interest far beyond my expectations.

Since I first met him in a café in Paris in the winter of 2008, Alex Bayen has struck me with his constant level of energy and excitement about the work and research of his team. I had never met anyone so passionate about his research. The dynamism and high level of expectations easily convinced me to join the *Mobile Millennium* team and be part of an exciting adventure. I was not disappointed as I arrived in Berkeley and Alex entrusted me with more responsibilities than I had ever thought I would be able to manage. Alex keeps on surprising me with his trust and the confidence that he expresses for my work and career. He always encourages me to take on new and additional challenges, pushing me beyond what I think myself capable of. The completion of this PhD is largely due to his support and contagious dynamism and will to succeed. Beyond the career drive, Alex has been of incredible personal support. Throughout the years, Alex has been able to combine enjoying relaxing and social times with other fellow students from the lab, working on papers in the middle of the night, congratulating achievements, discussing personal matters and career choices for long hours, tirelessly editing, reshaping and improving presentation of research ideas, giving motivation and energy during low times of the PhD, and so much more.

The last thanks go to my friends, who I have met over the years, and accompany me everyday, giving advice, sharing thoughts, discussing both important and silly matters and enjoying life. I am thankful for my friends in France to always be around when I come back. Keeping friendships with thousands of miles of distance is not easy and I truly appreciate these lasting friendships which are not affected by distance and time. I would not be able to go through a long PhD journey without feeling so much positive emotions, happiness and social support around me.

Chapter 1

Societal need for large scale traffic estimation capabilities leveraging sparsely sampled probe vehicles

Reliable, large scale traffic estimation has become a societal necessity both on highways and on the secondary network (arterial roads). This thesis investigates the development of global traffic monitoring capabilities on arterial networks. The present chapter motivates the development of global traffic monitoring capabilities and presents different projects conducted by universities, government agencies and industry to solve this challenging problem. It presents the different sources of traffic data and analyzes their potential for arterial traffic estimation. The subsequent chapters propose adapted modeling approaches and estimation algorithms to reliably infer traffic conditions in real-time in large arterial networks.

1.1 Global scale traffic monitoring

To maximize the benefit for the users of the transportation networks and for public agencies responsible for their administration, it is important to understand the societal needs for arterial traffic estimation and have a synthetic representation of the different projects working on related problematics.

1.1.1 Societal needs for traffic monitoring

Over time, the demand for mobility has dramatically increased, leading to a $78 billion annual drain on the U.S. economy in the form of 4.2 billion lost hours in commute and 2.9 billion gallons of wasted fuel. The economical cost of congestion has doubled between 1997 and 2007 [141, 142]. Long term projections in 2009 by the *Energy Information Administration* (EIA) at the US Department of Energy and the *Federal Highway Administration* (FHWA) show a flattening - but not a significant decline - in this trend [142]. In Europe, the situation is similar. The European Commission has published a white paper entitled a *Road-map to a Single European Transport Area* [13]. This white paper summarizes the importance of managing and optimizing transportation systems. Other sources provide important insights regarding the societal costs of transportation. Eurostat states that, on average, 13.2% of every household's budget is spent on transport goods and services. The TREMOVE model, cited by the European commission white paper estimates that Congestion costs Europe about 1% of Gross Domestic Product (GDP) every year. The time spent in congestion is one of the important costs of congestion. For example, 20% of London commuters spend more than two hours a day traveling to and from work, which

adds up to one working day a week [12]. In Germany, 37% of commuters spend one hour a day commuting [63]. European Union drivers currently own one third of the world's 750 million cars. The IEA [153] projects that by 2050 car numbers worldwide will increase to more than 2.2 billion, with the sharpest growth in emerging economies.

Traditional, infrastructure oriented mitigation strategies, such as building more highways, widening existing roads, and expanding or starting new transit or rail routes, are unlikely to be successful by themselves. These approaches are not sustainable in the current economic and environmental climate and their popularity is rapidly decreasing. Rather, innovative paradigms are needed to marshal breakthroughs for operations of the transportation network, to transform the manner by which the traffic management issues are addressed. One missing component, key to addressing these issues, is an exhaustive *traffic information system* deployed at a global scale, providing an integrated view of the full transportation system. Such a system needs to be centered around information gathering, integration, and distribution. To participate in a change in the way the transportation system is used, the public needs *location based* information, available in real time, to make more efficient decisions. These decisions may include the choice of the mode of transportation (individual vehicle, ride sharing, public transit and so on), the choice of route, the choice of departure time. In particular, these decisions should improve the efficiency of the travels and increase mode shifts towards the public transit system. The emergence of the *mobile internet* on portable devices and the rapid proliferation of location based services allow for the development of initiatives that can result in paradigm shifts in the way the public interacts with the transportation network, in particular by giving incentives to the public to use new technology to improve its transportation experience.

Instrumenting a system at a global scale is a difficult and expensive problem for public agencies. Most developed countries have deployed significant amounts of instrumentation (in particular loop detectors). However, the maintenance and operations of the existing infrastructure represent a substantial cost. Today, this infrastructure is generally limited to the highway network (see Section 1.2 for a detailed review of sensing infrastructure characteristics) and the equipment of the secondary network (arterial roads) with such a dedicated sensing infrastructure has a prohibitive cost (because of the density of the network). In addition to its cost, the loop detector system suffers significant performance and reliability issues which can be solved using complementary information from other sensing sources.

1.1.2 Existing and developing large scale traffic monitoring platforms

Traffic information systems have witnessed an important revolution in the past years, leveraging the development of communication and sensing capabilities of mobile devices. Furthermore, the emergence of the Web 2.0 paradigm, crowd-sourcing and location based services offers unprecedented potential for the development of transportation cyberphysical systems, *i.e.* transportation platforms which merge information (data) and constitutive laws modeling the physics of the system [163].

In the past decades, traffic information was merely coming from dedicated fixed infrastructure such as loop detectors embedded in the pavement [98], RFID transponders, radars or cameras [118]. A processed form of this information was displayed to the public on variable message signs, on the internet [1, 5] and, more recently, on phones using cell phone versions of these websites. These services only provide information where the

transportation network is equipped with such sensors. To provide a global solution to this traffic information gathering problem, mobile probes provide an affordable alternative for this problem, leveraging existing communication and sensing capabilities. Industries and government agencies realized early the potential of crowd-sourcing to reduce the costs and increase the coverage and the reliability of traffic estimation.

Publicly funded projects

Government agencies have seen specific interests in the reduction of the installation and maintenance costs of sensing infrastructure. They have also acknowledged the importance of having reliable traffic estimates on the secondary network, too dense for a global deployment of fixed dedicated sensing infrastructure. The information provided by global traffic estimation platforms also allows for better management and operation of available infrastructure and decreases the economical costs of congestion. Numerous research projects have been funded by Department of Transportation or equivalent public agencies throughout the world. Some of these projects are successful illustrations of partnerships between public agencies, universities and sometimes industry partners. A non-exhaustive list of specific pilot projects and deployments includes:

- The *Urban Platform for Advanced and Sustainable Mobility* (Plateforme Urbaine de Mobilité Avancée et Soutenable, PUMAS) was launched in the region of Rouen, France in 2009. It is a partnership between several research institutions (ARMINES, ESIGELEC, INRIA, INSA de Rouen), public agencies (Agglomération de Rouen and funding from the French Department of Industry) and industries (Induct, Egis Mobilité, Intempora, Sodit). The project aims at developing a software platform for collecting and analyzing road traffic information in real-time. The data is obtained by equipping a significant sample of vehicles operating constantly in a given area with an on-board device with GSM and localization capabilities. The potential of this deployment (and its further development) will be assessed in late 2012.

- The *SINERGIT* project (Système d' INformation sur les d'Eplacements par véhicules tRaceurs avec Galileo pour l'agglomératIon Toulousaine) is another example of large scale partnership with the goal of optimizing traffic management and of providing real-time information on driving conditions to all drivers using up-to-date personal navigation devices in the Toulouse region, France [35]. The SINERGIT platform relies on data from fixed sensors, floating cars, tracking of personal navigation devices and cellular phones (GSM). The project produces maps of traffic conditions which are publicly available, the data is available to public agencies for further analysis. The project is a joint collaboration between private companies (Sodit, Thalès Alenia Space, France Telecom R&D, ASF, Polestar), research laboratories from IFSTTAR (ex-INRETS) and public agencies (Cete Sud-Ouest, City of Toulouse). Unfortunately, I have not found any assessment of the estimation accuracy publicly available.

- The *CarTel* project [95] collects, processes and visualizes data from sensors located on mobile units (mobile phones and in-car embedded devices). The project relies on devices which are frequently sampled but with a low spatial accuracy (WiFi hot-spots, GSM). An efficient map-matching algorithm [154] filters this data into accurate probe trajectories used for traffic estimation. This project is developed by the Massachusetts Institute of Technology (MIT) and is funded by both public and private entities (National Science Foundation, Google and Quanta Computer Inc.)

- The *Mobile Millennium* project [25] gathers data available from fixed dedicated sensing infrastructures (loop detectors, radars), vehicles equipped with geo-localization

and communication enabled devices (smart phones, fleet management devices) in California. The project receives millions of data points every day which are processed in real-time to provide traffic estimation to the public and real-time monitoring information to the public agencies (US and California Department of Transportation). The project started with a successful proof of concept that highway traffic could be accurately estimated with sparsely sampled GPS enabled probes only. This proof of concept took place on February 8th, 2008 [59] on a 10 mile stretch of I-880 in the East Bay Area. The success of this demonstration led to a larger deployment, started in November 2008: the *Mobile Millennium* project. The *Mobile Millennium* project includes the implementation of a real-time monitoring system for California. The project was launched as a partnership with Nokia, which had developed one of the first community sensing mobile application, providing users with real-time detailed traffic information as they shared privacy aware information about their journey [91]. The accuracy of the estimates have been carefully analyzed [25] on both the highway and the arterial networks.

- In Sweden, the *Mobile Millennium Stockholm* project [17] was initiated by the Swedish Transport Administration in order to address the need for a useful and timely traffic information system. The project builds upon the knowledge and experience from the *Mobile Millennium* project at University of California, Berkeley. The project focuses on algorithm for data fusion to leverage the potential of the different sources of data available today for traffic estimation and prediction. The project is developed as a partnership between Linköping University, the KTH Royal Institute of Technology and Sweco Infrastructure in Sweden, and UC Berkeley in the United States. In parallel to this project, KTH devotes an important research effort to data collection and management in partnership with IBM [131]. This partnership has lead to an experimental ITS laboratory which collects and analyzes loop detector data, incident information, congestion pricing data, travel time data from automatic license plate readers, public transportation efficiency metrics, weather data and pollutant emissions.

Navigation and traffic information providers

In parallel to publicly funded research projects, industries have developed their own estimation platforms. With the generalization of geo-localization devices on board of vehicles, real-time navigation has become standard over the past ten years. The development of this technology is naturally improved with real-time traffic information to efficiently route customers around traffic. Tom-Tom's HD Traffic was one of the first industrial development, offered as a service with a monthly charge to the owners of Tom-Tom devices [10]. Tom-Tom's deployment was followed by other tele-navigation companies such at TeleNav [9] and actors of the automotive industry, such as Renault [97]. Note that there are several other navigation and traffic information providers interested in this technology including Intellione [4], Octo-Tellematics [57] or NAVTEQ [6]. Pioneering work in social community building and social games for traffic has been developed by Waze [11]: through a game interface, this recent routing company crowdsources mapping, routes and traffic news, rather than relying on historical data or official information from transport authorities. The interface and the social gaming have drawn a lot of success to this mobile application, which has seen the number of its users increase dramatically in the past few years.

Large technology companies

A significant proportion of the leading US-based technology companies have also developed their own platforms. For these companies, the motivations are multiple. By providing quality services, they can develop their user base, increase user engagement and user retention for the variety of products that they offer. Another strong motivation is the collection of geo-tagged data from their users: when using the application on mobile devices, the GPS data collected for positioning is generally sent to the company providing the service. This data is very valuable for these large companies: they can improve location based services and develop targeted advertising. Both Apple, Microsoft and Google provide global traffic estimation and routing directions which take into account traffic conditions, although no assessment of the accuracy of the estimate is available, besides each users' feeling of using the system. Microsoft Research has also developed its own traffic estimation capabilities and pioneered research in crowd-sourcing applications [94, 101, 171]. IBM Research has a slightly different approach as it offers its Smart City and Smart Planet services to both government agencies, other businesses and universities [29, 131]. Other companies, such as INRIX, specialize in providing processed data which constitutes traffic information.

1.2 Sampling strategies

A number of sensors have been developed in the past 50 years designed to collect various types of traffic data. Traffic data comes with various specifications including the type of data (flow, occupancy, velocity, travel time, trajectories), the sampling frequency (how often measurements are taken), the level of deployment (density of fixed sensors, penetration rate of mobile sensors in the flow of vehicle) or the accuracy of the measurements. The section briefly describes the most important sensors and corresponding traffic data and analyzes their potential for arterial traffic estimation.

1.2.1 Fixed, dedicated sensing infrastructure

Loop Detectors

Inductive loop detectors are embedded into the roadway and detect vehicles as they pass over the detector. A properly calibrated loop detector provides high-accuracy flow and occupancy data [38] as well as velocity when two detectors are placed close together (double-loop detectors). The sensors suffer from important reliability issues requiring filtering to produce quality input data to traffic estimation algorithms [45]. Loop detectors are commonly found on most major highways throughout the United States and Europe where they have communication capabilities to transmit the data to a central server in real-time (that can subsequently be used in traffic information systems). In the United States, most loop detectors installed on arterial roads do not have internet connection, preventing their use for arterial estimation. Rather, this data is generally used locally for signal timing control.

Radar

Radars can be placed on poles along the side of the road enabling them to collect flow, occupancy and velocity data. In general, radars provide lower accuracy data than loop detectors, especially in urban environments [118] and their deployment remains limited. For these reasons radar data is not considered as a viable input for arterial estimation algorithms.

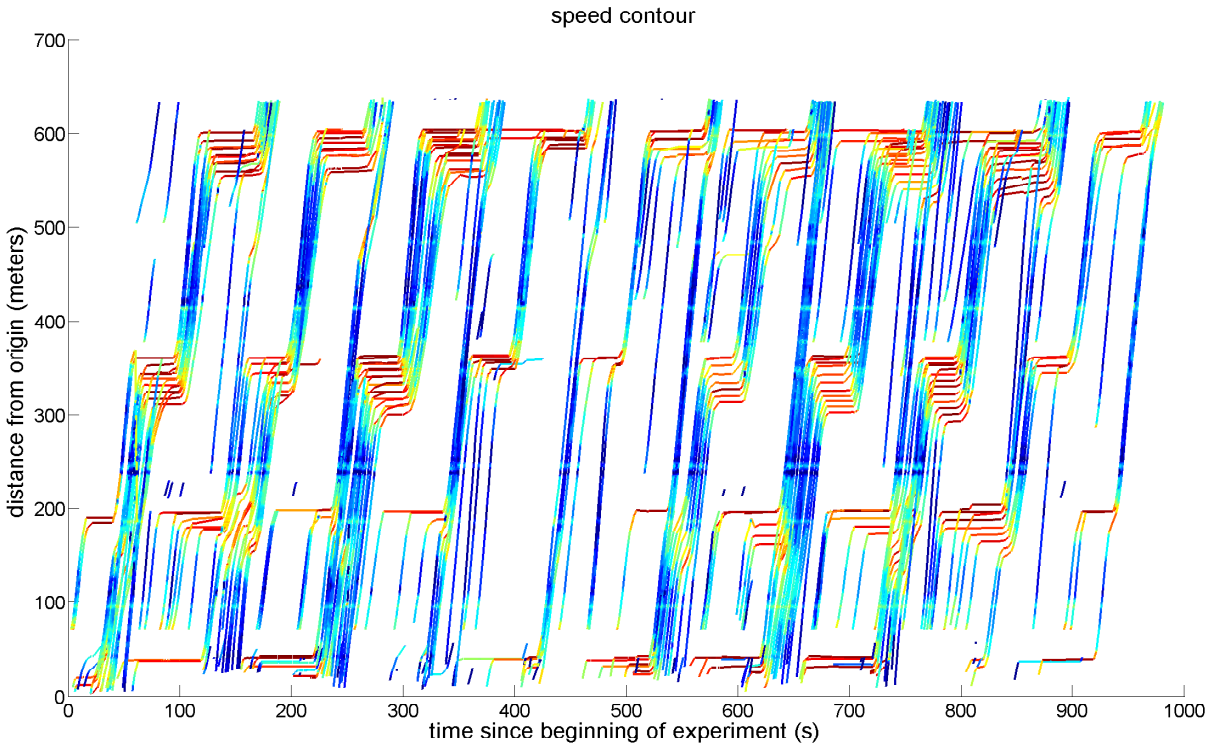


Figure 1.1: Trajectories of vehicles obtained after processing of high-resolution video camera data. The colors represent the speed of vehicles. The data was obtained by the NGSIM project. It corresponds to a fifteen minutes arterial dataset (4pm - 4:15pm) representing traffic patterns on a segment of Peachtree Street in the Midtown neighborhood of Atlanta, Georgia.

High-resolution video camera

High resolution cameras placed high above the roadway track all vehicles within the view of the camera. They cannot provide data in real-time due to the large amount of post-processing work that needs to be done on the images to turn them into actual vehicle trajectory data [7]. When properly processed, video provide high-resolution vehicle trajectories (vehicle positions every tenth of a second), as illustrated by the NGSIM project [7]. This technology is expensive to deploy and process, thus hard to deploy globally. It provides researchers with high-accuracy vehicle trajectories over a small spatio-temporal domain (in the order of one mile stretch for fifteen minutes) to validate modeling assumptions and estimation capabilities but cannot be used in real-time traffic information systems. The quality of the data obtained with high-resolution video camera is illustrated Figure 1.1. The figure represents the Peachtree dataset, a fifteen minutes arterial dataset (4pm - 4:15pm) representing traffic flows on a segment of Peachtree Street in the Midtown neighborhood of Atlanta, Georgia.

License Plate Readers

License plate readers automatically extract the license plate identification from passing vehicles. They are generally used in pairs along the road to extract high-accuracy travel times for vehicles passing both locations [3]. The deployment of these sensors require the identification of appropriate locations to place them and often remains limited to specific data collection studies.

RFID Transponders

Radio-Frequency Identification (RFID) is a ubiquitous technology used in numerous industry applications. Transit agencies make use of RFID in particular for toll or fare collection [23]. The RFID technology can be used for traffic data collection by placing readers at various points along the roadway. Travel times can be collected between pairs of points and processed similarly as license plate readers data. The accuracy of travel times provided by RFID transponders varies depending on the strength of the signal: stronger signals increase the chance of detection but increase the duration and area of detection, leading to a loss in accuracy especially for short distance travel times. RFID readers are generally placed far apart from each other in current deployments, making them useful for collecting long distance travel time information, but not for providing input data to detailed traffic estimation algorithms. They are placed almost exclusively on highways, making it uncommon to find this technology on arterial roads. The density of the arterial network and the high number of possible routes and itineraries decreases the probability to detect a specific vehicle at two distant readers, unless the entire network is equipped with such a technology.

Bluetooth

Bluetooth readers are capable of scanning the surrounding airwaves for Bluetooth enabled devices. They have received a lot of attention in recent years [161]. Pairs of Bluetooth readers provide travel time measurements for all vehicles carrying a Bluetooth device, passing both detectors successively. The Bluetooth readers suffer from similar limitations as the RFID readers: increasing the detection range decreases the accuracy, making it mostly applicable for long distance (a few miles) travel time measurements, with the risk of decreasing the re-identification rate if the network is not globally equipped. Bluetooth readers are an emerging technology and remain scarcely available on the network. The main deployments (such as SANEF in France) are on highway networks, where the limitation of the detection technology is not critical. It is likely that enhancement algorithms will be developed that will correct for the inaccuracies in the data provided by these devices, even though current state of the art does not provide the accuracy required for short distance travel time data.

Wireless Sensors

Wireless sensors are devices embedded into the roadway. They are similar to loop detectors but record the magnetic signature of vehicles passing them which is used for vehicle re-identification at downstream sensors with up to 80% accuracy [103]. Besides flow and occupancy, wireless sensors provide travel times between pairs of sensors for all the vehicles for which the algorithm found a signature match between the upstream and the downstream sensors. If there are intersections between the pair of sensors, vehicles can enter or exit the road segment delimited by two consecutive sensors. These vehicles are only detected by one of the sensors and will not lead to travel time measurements unless side roads are equipped with sensors as well. In this case, the re-identification algorithm also needs to be designed to allow for signature matches between all possible pairs of sensors, potentially decreasing the re-identification accuracy. The wireless sensors are cheaper to deploy and maintain than loop detectors. Moreover, they provide travel times for a larger portion of the flow and with higher accuracy than Bluetooth readers and RFID readers. These characteristics make them appealing for large-scale deployments on arterials even though they are only available in a small number of locations at the current time and monitor specific routes rather than portions of a network. *Sensys Networks* [8]

is currently one of the leading providers of these sensors.

1.2.2 Mobile sensing

Fixed dedicated infrastructure has been providing most of the traffic data available for highway traffic estimation in the past 50 years. The extension of this infrastructure to arterial networks is challenging, mainly because of the density of the network: (a) the installation cost of sensors on the entire network is often considered as prohibitive, (b) the number of intersections and possible routes reduces the probability to re-identify vehicles between long distance travel time measurement sensors (pairs of RFID, Bluetooth or License Plate Readers) and increases the needs for filtering. Mobile sensing offers a promising alternative with limited costs for global deployment, as it leverages the existing localization and communication capabilities. Experimental research on cell phone based traffic monitoring [24, 169, 149, 170, 140] has investigated the ability to locate the position of users using trilateration- or triangulation-based methods. It has shown limited success for estimation of travel times due to the position measurement inaccuracy, particulary on short distances and dense networks [110, 91]. For this reason, the use of probe data obtained from the cellular network is not considered for arterial traffic estimation, even though it may provide a valuable source of information for highway traffic estimation. For arterial traffic, GPS-based traffic information has a great potential and this thesis will focus on this source of data. There exists different sampling strategies with respective characteristics and potential for global deployment in arterial networks, as described below.

Virtual Trip Lines

A Virtual trip lines (VTL) is a *virtual* line drawn on the road. Its geometry is stored in a digital system [91]. A GPS enabled device using the VTL sensing paradigm downloads a list of the VTLs "deployed" in its vicinity. Whenever the device crosses a VTL, it sends an update to the central VTL server indicating its velocity, time of crossing, anonymized identifier and VTL identifier. The accuracy of the velocity data generated by frequent GPS sampling varies greatly with the type of GPS chip and the sensing environment. It is generally accurate enough for highway traffic estimation when properly filtered [165, 81]. For arterials, the velocity measurements are not reliable, both because of the variability of speeds and the lower accuracy of GPS devices in urban environments. For this reason, only travel time measurements (obtained between two consecutive VTL crossings) are considered for arterial traffic estimation. Figure 1.2 shows the variability of speeds obtained by GPS units, both in terms of magnitude and in terms of direction, underlining why this data is impractical to use.

The presence of a computing device on board of vehicles enables the design of other sensing paradigms which offer trade-offs between privacy guarantees, level of traffic information which can be reconstructed from the data and amount of data to transfer to the central server. Existing projects such as PUMAS have analyzed potential paradigms leveraging the computing power of certain devices. When considering mobile devices, which are not integrated to vehicles and may have battery limitations, the complexity of the sensing algorithm also needs to be taken into account. The complexity of the front end computations will have a direct implication on the battery life of the devices.

The VTL sensing paradigm was developed by Nokia in 2007 [92] and validated for highway traffic estimation during the *Mobile Century* experiment in February 2008 [81]. This traffic data collection system is a privacy aware "participatory sensing" system that allows individuals to download an application onto their GPS-enabled smartphone that both sends traffic data as well as receives traffic information and alerts. An experimental

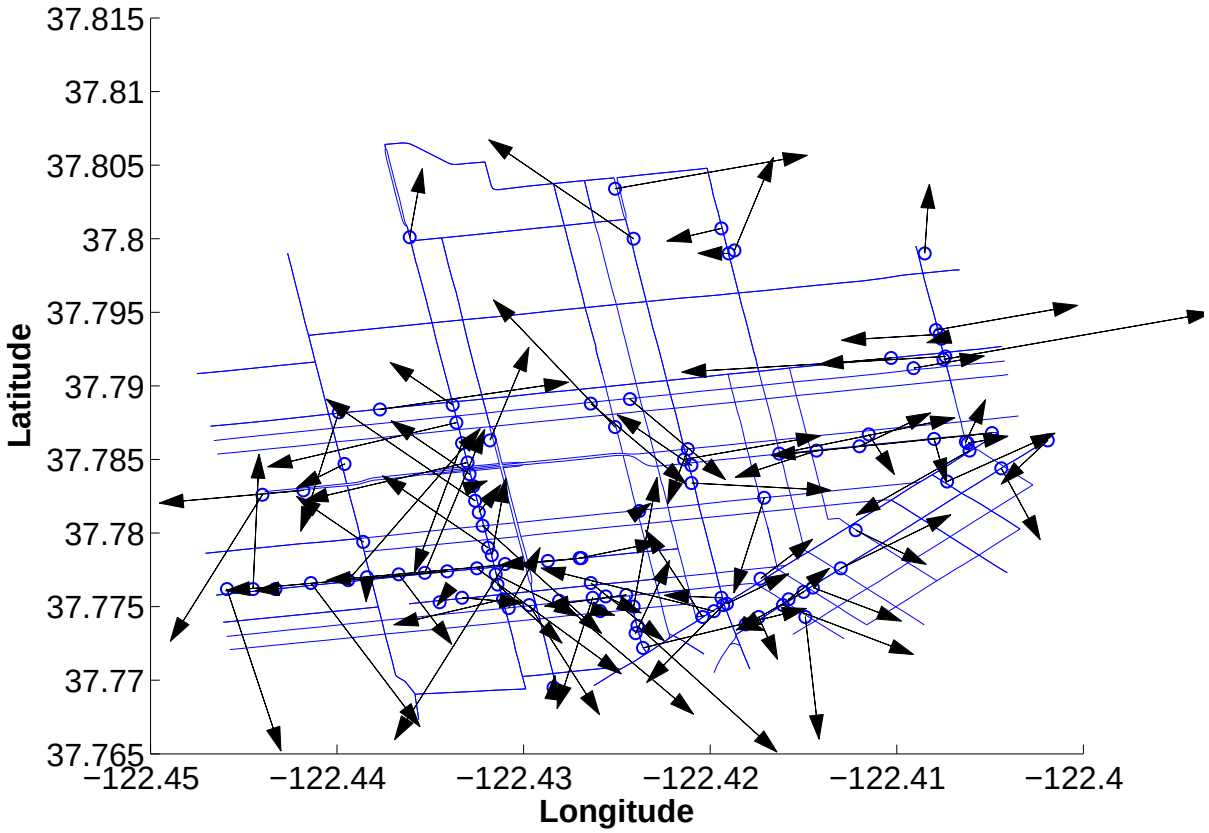


Figure 1.2: Speed measurements collected from the GPS units of probe vehicles in San Francisco. Each arrow starts at the location where the measurement was received and points towards the direction indicated by the vehicle. The length is proportional to the speed of the vehicle. The speed is obtained by the GPS units using standard finite differencing techniques of high frequency locations with correction algorithms. As can be seen, the urban canyoning effects and multipath effects results in significant errors in the computation of the speed, leading to misalignment with the road network.

Figure 1.3: Example of a Bay Area VTL deployment as part of the *Mobile Millennium* system.

deployment of VTLs was tested as part of the *Mobile Millennium* project, covering all of California and illustrated in the Bay Area Figure 1.3. The GPS devices reporting VTL data are limited to the ones which have downloaded a specific application and agreed to share pieces of their trajectories through the VTL sensing paradigm. For this reason, we also consider other sources of GPS data to complement the data gathered by the VTL deployment.

Sparsely-sampled GPS

Sparsely-sampled probe GPS data refers to data received from probe vehicles reporting their current location at a low frequency (either temporally or spatially, *e.g.* every 60 seconds, every 400 meters and so on, depending on the sensing strategy). The frequency is not high enough to directly measure velocities or link travel times (*i.e.* duration between successive measurements greater than 20 seconds). Today, sparsely sampled GPS data is the main data source available in the US for arterial traffic with the prospect of global coverage in the near future. Sparsely-sampled probe data is a "massively available source of sparse data", making it both a challenging and a promising source of data. Sparsely-sampled probe data shares many of the characteristics used to define *Big Data*, which has received a lot of attention in the past few years: the amount of available measurements (received by all the vehicles) require specific data management and processing infrastructure and the sparsity of each of the measurement requires specific algorithms to extract

valuable information:

- A pre-processing step required to use sparsely-sampled probe data is to map GPS measurements to the road network representation used by the traffic information system and to reconstruct the path between successive measurements, as vehicles may travel several links between successive measurements. This process is known as map matching and path inference, and has been studied in parallel with the development of mobile sensing technologies [113, 154, 96]. The sparsely-sampled GPS data used in this thesis has been preprocessed by an algorithm using models of GPS accuracy and driving behavior in a Markov random field developed at UC Berkeley [96] and operational in the *Mobile Millennium* system.

- Since probe vehicles can often travel multiple links between measurements when the sampling frequency is low, one must infer the travel times on each link traversed by the vehicle. This process is known as *travel time allocation* or *travel time decomposition* [79, 84].

- Finally, the measurements being taken uniformly in time (or in space), measurements are received at any location on the network. In particular, the path of the vehicle does not necessarily cover entire portions of the links of the network, leading to information on *partial* link travel times which must be incorporated in the estimation framework.

Sparsely-sampled probe GPS data is currently the most ubiquitous data source on the arterial network. An example of this type of data comes from the Cabspotting project [2], which provides the positions of 500 taxis in the Bay Area approximately once per minute. Figure 1.4 shows one full day of raw data, which demonstrates that even just a single data source such as taxis can provide broad coverage of a city. This data clearly has some privacy issues as it is possible to track the general path of the vehicle. However, the majority of this data today comes from fleets of various sorts (such as UPS, FedEx, taxis, etc.). Most of this data is privately held among several companies, but between all sources there are millions of records per day in many major urban markets. Figure 1.5 shows the distribution of sampling frequencies for the data feeds available in the *Mobile Millennium* system and underlines the importance to design algorithms adapted to sparsely sampled probe data. The *Mobile Millennium* system receives several millions of data points per day in North California from vehicles sampled with a very low frequency (duration between successive measurements of most of the available data feeds is in the order of one minute). The research and development conducted as part of the *Mobile Millennium* project illustrate both the challenge and the potential impact of probe vehicle data for the future of Intelligent Transportation Systems.

High-frequency GPS

High-frequency probe GPS data refers to the case where probe vehicles send their current GPS location every few seconds, providing detailed trajectory data, especially after appropriate map-matching and filtering [96]. From this data, one can directly infer velocities and short distance travel times. Figure 1.6 depicts a sample of high-frequency data collected as part of the *Mobile Millennium* project during a field test experiment. This figure illustrates the level of detail that can be extracted from high-frequency data, but also shows the relatively low percentage of vehicles that were being tracked, even during a data gathering field experiment. Sampling a vehicle's position every few seconds is privacy invasive and also comes with large communication costs and battery drain issues. For these reasons, it is not common to receive this data with any kind of regularity or

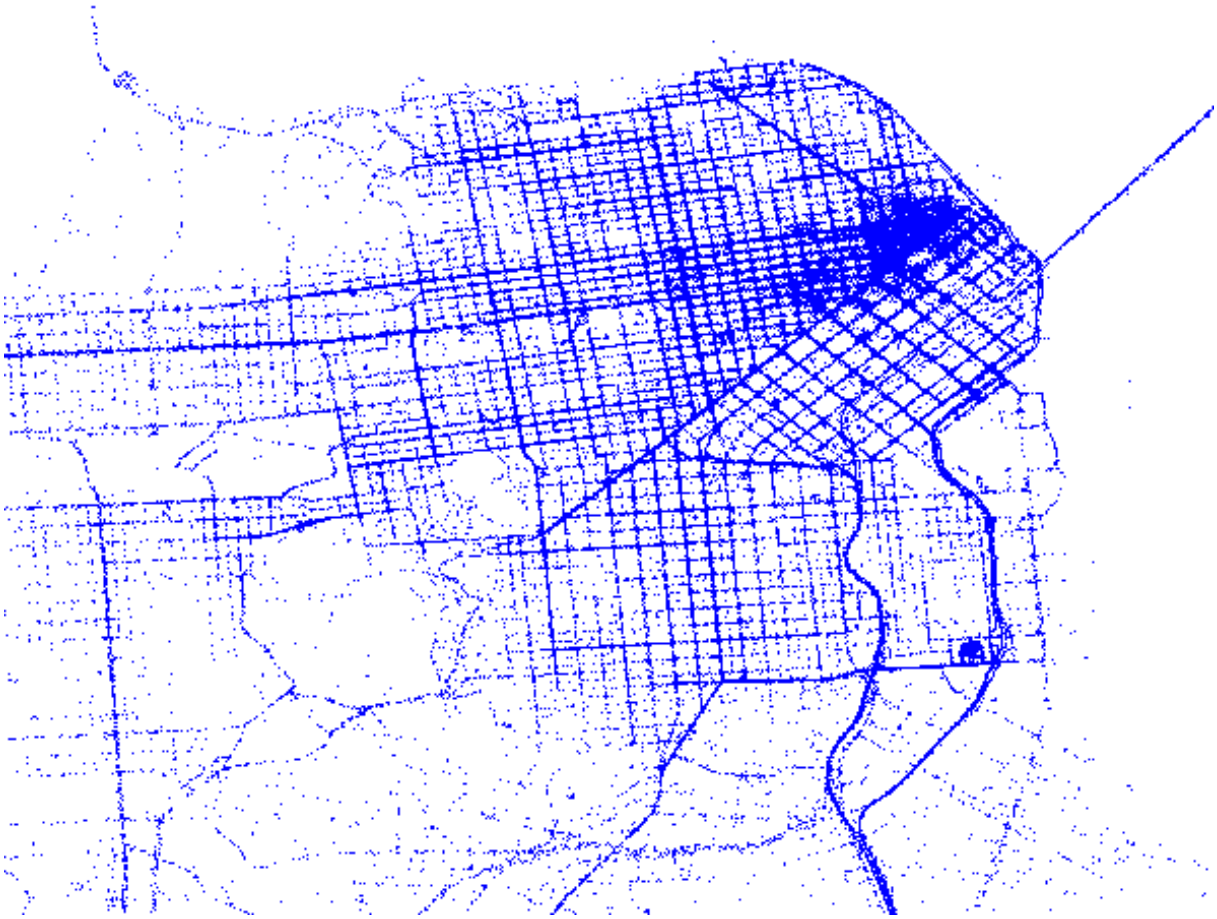


Figure 1.4: One day of sparsely-sampled GPS data from San Francisco taxi drivers as provided by the Cabspotting project. [Courtesy of T. Hunter [96]]

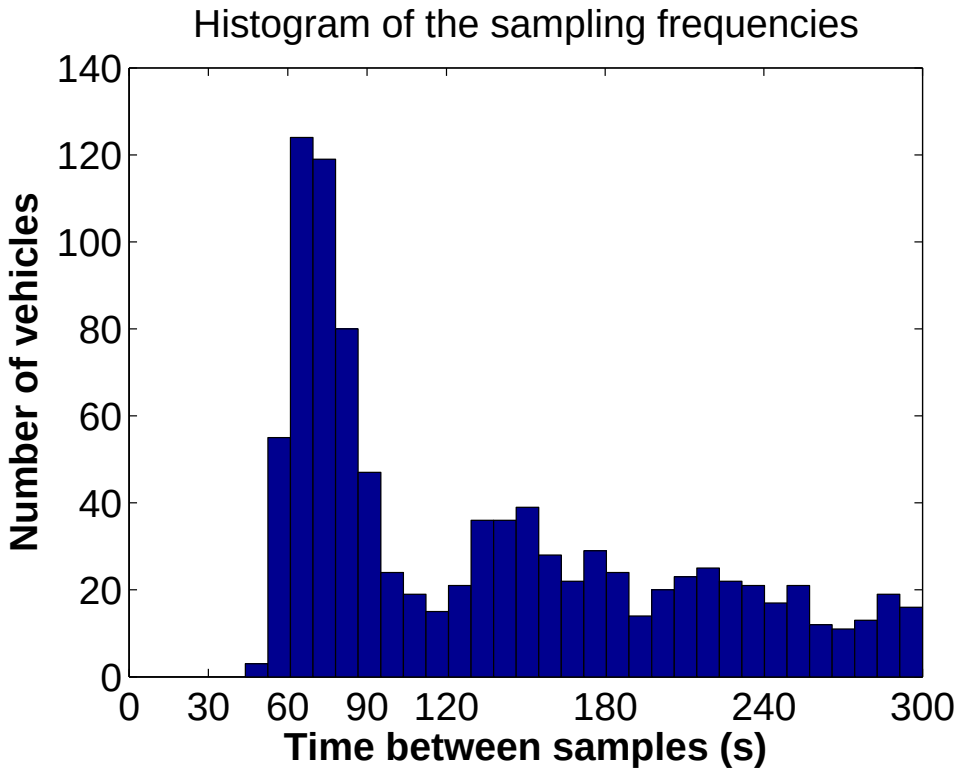


Figure 1.5: Example of distribution of sampling frequency for a commercial feed of probe data (vendor confidential and undisclosed). The histogram shows the number of probe vehicles as a function of their average sampling rate. For the clarity of the figure, only the probes which were sampled more frequently than once per five minutes are considered. In the United States, the majority of commercially available probe vehicle feeds is sampled at a rate close to once per minute.

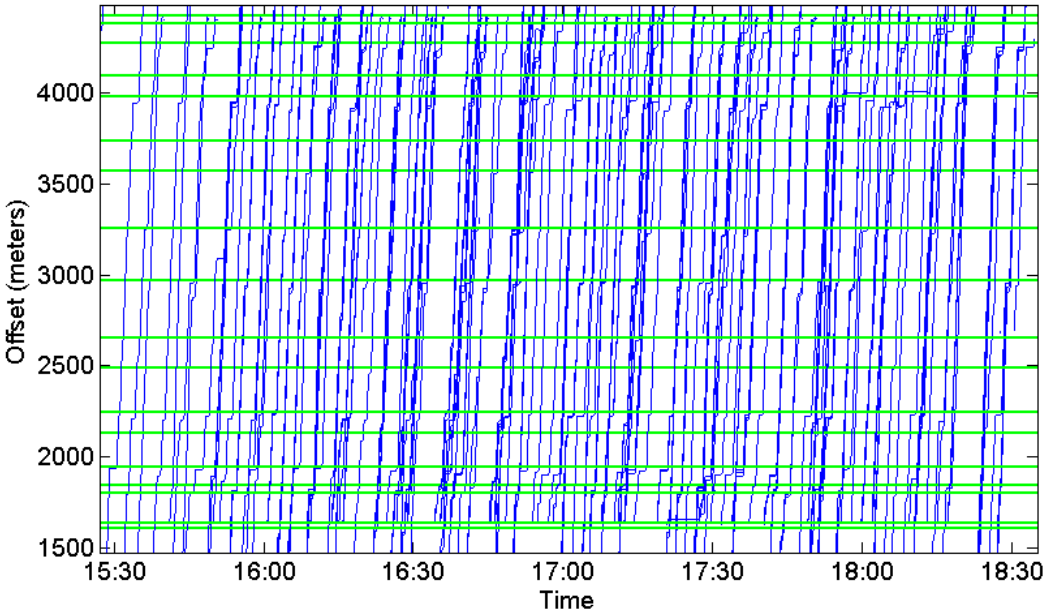


Figure 1.6: Vehicle trajectories from the *Mobile Millennium* evaluation experiment on San Pablo Avenue in Berkeley, Albany and El Cerrito, California. The high-frequency GPS data in this figure is represented as distance (meters) from an arbitrary start point upstream of the experiment location. The horizontal lines represent the locations of the traffic signals along the route.

global coverage. This data is often collected for specific experimental studies, but is not generally available for real-time traffic information systems.

The analysis of available traffic data for arterial traffic estimation demonstrates the potential of sparsely sampled probe vehicle data to provide the main input for estimation algorithms. Even though this data is massively available, its processing is challenging and specific modeling and estimation capabilities must be developed to turn the data into valuable information. Wherever additional sensing infrastructure is available and provides measurements with appropriate quality, it is possible to fuse the different sources of data into the estimation algorithms. However, these sources do not provide the coverage necessary for global traffic estimation. For this reason, this thesis focuses on sparsely sampled probe data. In order to leverage the existing work in the field of traffic modeling and estimation, a literature review is presented in the following chapter.

1.3 Problem statement

This introduction Chapter has presented the importance of real-time, large scale traffic estimation capabilities. Providing up-to-date information to the transportation network users allows them to make informed mobility choices including changes in mode of transportation, itinerary or departure time, leading to a more optimized use of available transportation infrastructures. The information is also valuable to traffic management center to deploy real-time control strategies such as ramp metering or variable speed limitations. This need is well understood by government agencies and universities which devote significant resources to solve this problem. The private sector has also acknowledged the potential of such systems. Most major Tech-Companies (Google, Apple, IBM, Microsoft)

have developed their own traffic estimation capabilities and work on the extension of these systems on the secondary network and on the improvement of the reliability of the estimates and predictions. The first building block for traffic estimation is traffic data, coming from a variety of sensors, presented and analyzed in the previous section.

Given this perspective, a natural question arises: "What can be done using probe data for traffic?" This thesis focuses principally on traffic information systems, and more specifically on the traffic estimation process in arterial networks. This thesis investigates the following fundamental question: "Is it possible to create traffic information systems in urban environments based principally on probe data?" A concurring question is the following: "Which models and estimation algorithms can be developed to create such a traffic information system?" These question are not only of interest to researchers, but also to the government and to industry, as demonstrated by the large amount of traffic applications that have emerged in the past few years. This thesis focuses on the technical feasibility of modeling and estimation of arterial traffic from probe vehicle data (floating car data). The thesis does not detail the data collection process, the privacy issues or the policy questions raised by this problematic. The thesis will take the example of the *Mobile Millennium* project to demonstrate the feasibility of real-time arterial traffic estimation from probe vehicle data only.

1.4 Organization of the thesis

Chapter 1 introduced the problem of interest developed by the present thesis. It underlined the need for large scale traffic monitoring systems and argues for the potential of sparsely sampled probe vehicle as the main source of traffic data in urban networks available at a large scale in the near future.

Chapter 2 reviews existing models and estimation algorithms in both highway and arterial networks, with a focus on arterial networks. From this literature review, it appears that the available models and algorithms are not able to leverage sparsely sampled probe vehicle data. The data requirements and/or computational complexity of these models limit their applicability at a large scale. The review of previous work argues for the development of statistical models and inference algorithms to represent the variability of arterial traffic dynamics, which is the approach chosen in this thesis. The chapter shows the necessity to develop large scale statistical models of arterial traffic. The models need to be able to incorporate sparsely sampled probe vehicles, both to calibrate the model (historical training) and to estimate and predict traffic conditions in real-time with streaming data.

Probe vehicle have been successfully used for highway traffic estimation based on flow models derived from hydrodynamic theory. Chapter 3 presents a possible use of probe vehicle data to robustly estimate travel times on sections of the highway network. The successful integration of probe vehicle data into highway traffic flow models motivates the development of flow models adapted to urban traffic which can integrate sparsely sampled probe vehicle data [44].

Chapter 4 presents a model of arterial traffic based on hydrodynamic theory. The model requires simplifying assumptions to keep the derivations tractable and increase the interpretability of the results while keeping it realistic. It characterizes the formation and dissolution of horizontal queues upstream of traffic signals [87].

The arterial traffic model is used in Chapter 5 to derive parametric probability distributions of the locations of vehicle on a (generic) link of the network, delimited by signalized intersections. The parameters of the probability distributions have physical interpretations and can be learned from location reports of probe vehicle data [87]. Besides

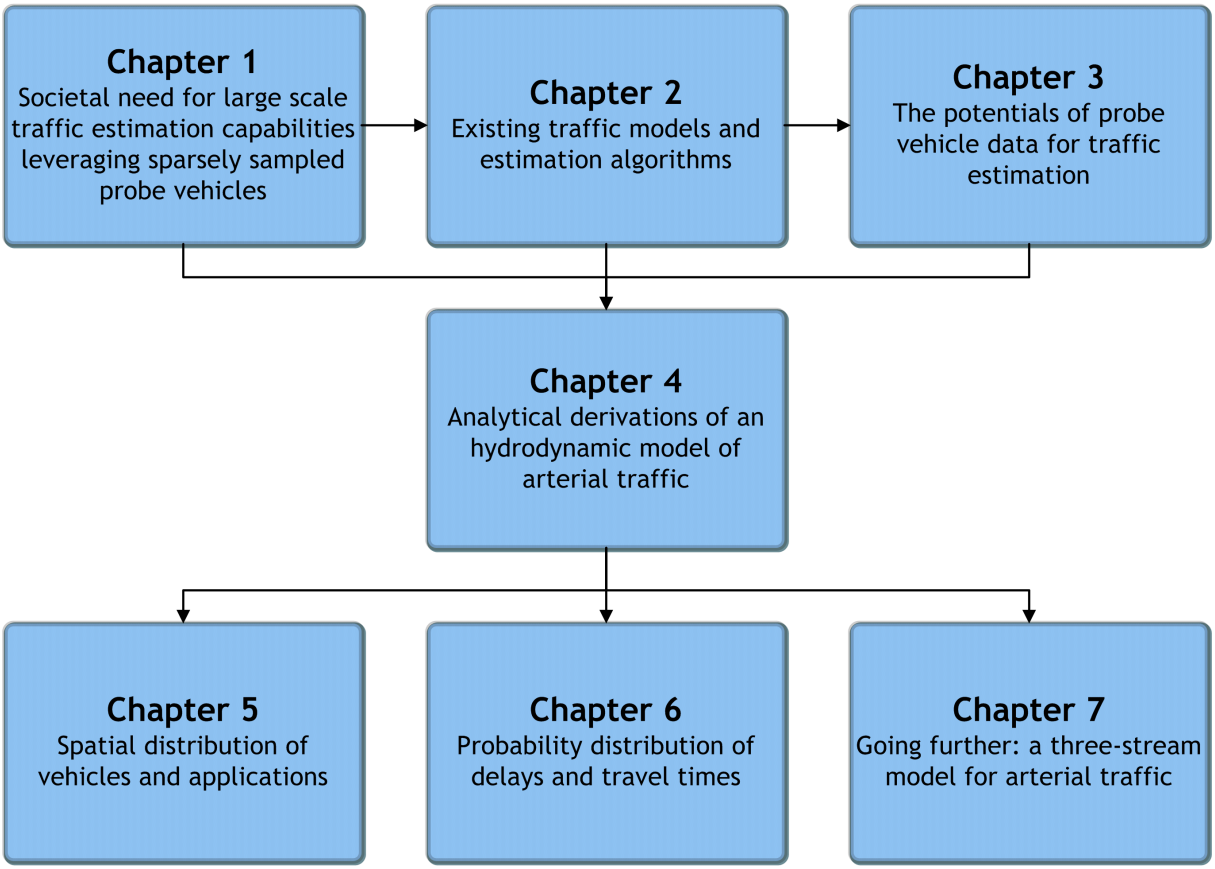


Figure 1.7: Structural organization of the dissertation

providing general information on the congestion level of a link, these parametric distributions have valuable applications in Intelligent Transportation Systems. They can be used to scale travel times of probe vehicles when the location reports do not coincide with the beginning or the end of the links of the network, providing an algorithm to use travel time information on *partial* links into travel time information on the entire link [82]. The parametric probability distributions can also be used to automatically detect the presence of traffic signals and create or update digital map databases at much lower cost and higher time efficiency than current techniques [86].

The arterial traffic model is also used in Chapter 6 to derive parametric probability distributions of travel times between arbitrary locations. The fact that these distributions are valid between arbitrary locations is crucial for the development of arterial traffic monitoring platforms based on sparsely sampled probe vehicle data and is one of the most important novelty of this work [87, 90].

Chapter 7 builds upon the arterial traffic model of Chapter 4. It relaxes one assumption made in Chapter 4 regarding the structure of the arrival flows on arterial links. The refinement of the model is valuable when the amount of data available provides enough information for the more precise modeling to be captured. It is of particular interest on arterial corridors in which the signal coordination is important. The model also has interesting applications for real-time signal control. The chapter derives analytical signal coordination control strategies on arterial corridors. The analytical derivations have the advantage to allow real-time control and provide interesting insight and understanding of the traffic dynamics [21].

Figure 1.7 shows the organization of the thesis and the logical relationship between the different chapters.

Chapter 2

Existing traffic models and estimation algorithms

A major contribution of arterial traffic research is the development of models and algorithms to estimate traffic conditions from the available measurements. A possible approach is to directly report traffic conditions as measured by the sensors: the sensor measurements or a simplistic processing of them are given as output to represent the traffic conditions. With such an approach, it is necessary to place sensors everywhere in the network where traffic information is needed. The density of the sensor deployment and the sampling frequency must be high enough to represent the spatio-temporal variability of arterial traffic. The quality of the representation of traffic conditions depends directly on the quality and reliability of the sensors. In arterial networks, researchers have tried to directly measure delays and travel times from probe vehicles [144, 137, 168] or through vehicle re-identification [125, 103]. This work analyzes the bias and required sample size to characterize delay and travel time [78]. These techniques require an ubiquitous fixed sensing infrastructure or high penetration rate of probe vehicles, both of which do not correspond to the available data for the entire arterial road network (or even close to that coverage), as analyzed in Chapter 1.2. For this reason, traffic models and data assimilation algorithms must be developed to efficiently transform the available measurements into reliable traffic information.

A pre-requisite for reliable estimation is to reconcile the features of the measurements (input), the physics of the underlying processes (system) and the estimates (output). The choice of model complexity has to find an appropriate compromise between the model complexity (a more complex model may be a more accurate representation of the physics) and the risk of overfitting. Overfitting generally occurs when the model complexity does not match the amount and precision of the available measurements, leading to poor predictive performance. This trade-off will be underlined in the analysis of existing research and the development of novel models and algorithms.

This chapter reviews the existing models and inference algorithms developed over the past fifty years to characterize traffic conditions from available measurements. The chapter starts with a review of highway traffic modeling and estimation, which is important for both historical reasons and for its influence in arterial traffic estimation. Arterial traffic estimation has seen the influence of two main fields: traffic flow theory and statistical models. The chapter reviews the state of the art of arterial traffic estimation and analyzes the additional research needed to develop reliable large scale arterial traffic monitoring platforms.

2.1 Highway traffic estimation

Traffic estimation algorithms have first been developed for the highway network. This network is not as dense as the arterial network and thus easier (and not as expensive) to equip with fixed dedicated sensing infrastructure. For this reason, the highway network has been well-instrumented with fixed-location sensors (mostly loop detectors) throughout most of the United States and Europe. For highway networks covered by such an infrastructure, it has become common practice to perform both system identification of highway parameters (free flow speed, traffic jam density and flow capacity) and estimation of traffic state (flow, density, length of queues, bulk speed and shockwave location) at a very fine spatio-temporal scale [165, 28]. These highway traffic monitoring approaches heavily rely upon both the ubiquity of data and highway traffic flow models developed over the last half century [108, 49, 152].

Highway networks were also the first to be studied when GPS probe data first started becoming available. The *Mobile Century* experiment [59] (a predecessor to *Mobile Millennium*) demonstrated the capability of estimating highway traffic conditions using GPS probe data only [81, 80, 165]. Several extensions to that original work have been developed in the past two years, leading to a thorough understanding of how to perform traffic estimation using a combination of fixed-location sensor data and GPS probe data [165, 42, 43]. The techniques used in these papers combine an underlying flow model of highway traffic with an *Ensemble Kalman Filtering* (EnFK) algorithm [61] for estimating real-time traffic conditions. These techniques rely on the assumption that highway traffic acts roughly as a continuous fluid. The important point to note here is that the problem of using GPS probe data on highways (along with fixed-location sensor data) has been solved with a high degree of accuracy.

For arterials (the secondary network), traffic monitoring is more challenging: probe vehicle data is the only significant data source with the prospect of global coverage in the future. Aside from less abundant sensing compared to existing highway traffic monitoring systems, the arterial network presents additional modeling and estimation challenges as the underlying flow physics which governs it is more complex. The main challenges arise from the presence of traffic lights (often with unknown cycles), intersections, stop signs, parallel queues, and others [105, 39]. Highway traffic can be considered as an *un*-interrupted flow, whereas urban traffic is interrupted at intersections, causing specific modeling and estimation challenges.

2.2 Traffic flow models for arterial traffic estimation

The review of arterial traffic estimation can be separated into two classes: model based and data driven. On the one hand, model based estimation relies on an abstraction (modeling) of physical principles representing the dynamics of traffic on a link, a route or a network. On the other hand, data driven estimation is based on models learned (trained) from measurements. They may be inspired from physical principles (spatio-temporal dependencies between variables representing how congestion spreads on a network) and a posteriori give information on the physics of the system, but this is not a requirement. This review of arterial traffic estimation starts with model based estimation (Sections 2.2-2.4) and presents data driven algorithms in Section 2.5.

The fundamental idea for estimation techniques based on flow models is to adapt traffic theory principles used for highway traffic estimation, in most cases the LWR model [108, 134], to arterial traffic and to estimate the traffic quantities of interest (flow, density, velocity, travel time). For arterial traffic, the goal is generally to estimate travel

times, because it is a metric with a high interpretability power, both for the network users and the traffic management center. Following classical traffic flow theory, most contributions in arterial traffic estimation from traffic flow models model vehicular flow as a continuum and represent it with macroscopic variables of *flow* $q(x,t)$ (veh/s), *density* $\rho(x,t)$ (veh/m) and *velocity* $v(x,t)$ (m/s). Researchers commonly make the assumption of a triangular fundamental diagram for both estimation and control applications [72, 173]. Moreover, the differences in fundamental diagram do not change significantly the dynamics of traffic on a large scale [30].

Arterial traffic estimation research has developed aggregate models to estimate average speeds or travel times from loop detectors measurements [69, 145, 146, 172]. This research defines general guidelines to relate travel times to flows in arterial networks. The models depend on the specific features of each road segment (*e.g.* volume to capacity ratio), the cycle time and the ratio of green to cycle time, as summarized in the Highway Capacity Manual [155]. These models have interesting long term planning potentials but provide limited accuracy for real-time traffic estimation. In particular, this research does not take into account the variability of travel times among the vehicles traveling on the network [124]. The literature review and this thesis focus on the importance of taking into account travel time variability for arterial traffic estimation.

More recently, several authors have studied the ability to define fundamental diagrams for arterial traffic, an empirical relationship relating the travel times to the density of vehicles on specific routes [70, 54] and arterial links [77]. These novel approaches have received a lot of attention and provide a macroscopic large scale representation of arterial traffic, even though they do not take into account variability for estimation purposes.

On the other end of aggregate models are simulation models, which reconstruct the movement of each individual vehicle (microscopic simulation), represent the evolution of macroscopic variables such as flow and density (macroscopic simulation) or simulate trajectories at an aggregate level from queuing theory or speed-density relationships (mesoscopic simulation). Simulation models have the capacity to represent very accurately complicated physical phenomena. They can be used for estimation and control of arterial networks [40, 58]. However, they require precise calibration of a large number of parameters, including *Origin-Destination* (OD) matrices, turn ratios, precise signal plans. This calibration and the computational costs of such models make them impractical for real-time large scale traffic estimation, even though they represent great tools for planning, control and validation purposes. The large number of parameters needed to fully characterize the road dynamics in these simulation models increases the risk of overfitting (too many degrees of freedom compared to the precision of the available data). Moreover, some of these parameters are hard to estimate accurately such as origin destination matrices or turn ratios.

2.3 Estimation of delays

Delays at intersections cause a major part of travel times on arterials [158]. Arterial travel times are commonly represented as the sum of the *free flow travel time* and the *delay*. The free flow travel time is the travel time that a vehicle would have experienced with no interaction with other vehicles nor with the signalization, *i.e.* the travel time experienced if the vehicle was able to choose its speed and maintain it on its entire path. This free flow travel time may depend on user behavior or other features (weather, type of road, visibility, and so on), and this discussion will be continued in the following chapters. Delays represent the additional time spent on the network by the vehicle. A major

source of delay is the necessity to stop at signalized intersections. As will be underlined further, the duration of the delay is strongly correlated with the time at which a vehicle enters a road segment (because of the traffic light cycle) and represents a major reason for travel time variability in arterials (Sections 2.4). If the number of vehicles present on a link at the initial time is known, delays can be reconstructed from precise vehicle counts upstream and downstream of the segments using the cumulative number of vehicles function [51]. However, such precise information is rarely available: the initial state of a network is rarely accessible, as it would correspond to taking an aerial photo of the network. Moreover, such a method is not robust to noise in the measurements and its accuracy will decrease over time because of the potential drift in the vehicle counts of the loop detectors. The conservation of vehicles may also be violated because of vehicles entering and exiting the network through parking lots and residential lanes.

As delays vary with the arrival time in an arterial link, previous research has characterized delays on arterial routes depending on the entrance time. In these models, it is common to assume that all the intersections share the same cycle length and that the cycle and the offsets between consecutive signals are fixed. From one intersection to the next, the fixed offset determines the arrival profile of vehicles, and thus the delays experienced by the vehicles [71, 147, 66]. This work relies on the characterization of platoons and arrival profiles along the links of an arterial [71, 148].

The requirements of these models are that loop detectors are installed along the route and the signal system is capable of providing data about the green and red light times back to the traffic estimation system. These techniques generally require precise calibration (fundamental diagram parameters) and have strong data availability assumptions (precise loop detector data for each signal at each intersection, real-time information about signal settings, etc.). The strong assumptions in terms of data availability allow for very precise estimation of traffic conditions: the models estimate the travel time of each arriving vehicle and characterize the queue length of each of the links in real-time. This line of work provides a very sound basis for estimating travel times on arterial roads and has inspired significantly the work presented in this thesis. The primary limitation of the work is the fact that loop detectors and traffic signals in urban networks are very rarely connected to a communication network that would allow for processing of the data in real-time. Moreover, this dedicated sensing infrastructure does not have global coverage nor does it have the prospect of extended coverage in the near future. These models also require precise calibration of the parameters of the arterial road network. Documenting the detailed parameters into an accessible electronic database would require the cooperation of numerous government agencies, making this information unreliable (most of this information is not stored in electronic format and is shared among different organizations) and tedious to obtain (each city, county or district has the information for the network it is responsible for managing), at least until the data volume significantly increases [128, 157]. The other limitation is that the model focuses on estimating travel times along a specific route and does not study travel times between any two locations in an arterial network. Other models have studied delay and travel time estimation in arterial networks from loop detector data [166, 111, 167], limiting their applicability for global monitoring purposes.

Flow models have also been developed to leverage probe vehicle data. Assuming that vehicles report their link travel times (through VTL sampling for example) it is possible to reconstruct travel time profiles, depending on the arrival time of vehicles [22]. This algorithm does not assume the availability of traffic signal settings. It requires the calibration of the fundamental diagram parameters and only provides estimates a posteriori (reconstruction of the state of traffic). Provided that the communication of

probe vehicle travel times to the central server is instantaneous, the model reconstructs at a given time the situation of the previous traffic signal, which is not a strong limitation in practice. In reality, communication of data is not instantaneous and the latency between real-time and estimates is closer to five minutes. As for the macroscopic models using loop detector data, this approach leverages traffic flow principles to provide accurate reconstruction of traffic conditions. The main limitation of this approach is that it relies on probe vehicle penetration rates which are far from standard today (about 20 to 30% penetration rate is required to obtain reliable results). Moreover, this model does not have forecast capabilities (only a posteriori estimation) and is not robust to missing data.

2.4 Characterizing the variability of arterial travel times

An important challenge in arterial traffic estimation is the characterization of travel time distributions. Travel time variability is typically categorized into three groups [124]:

- *vehicle-to-vehicle variability* represents the differences of travel times experienced by different vehicles departing within the same departure period,
- *over the course of the day variability* characterizes the variation of travel times through the different *Time Of the Day* (TOD *e.g.* morning rush hour, mid-day, evening rush hour, evening),
- *day-to-day variability* describes the variations due to weather, special events and so on, which are responsible for the variations between a specific *Day Of the Week* (DOW) and TOD from one week to the next.

The *over the course of the day* and *day-to-day* variability are due to differences in the traffic conditions (weather, demand, capacity of the network) leading to different level of congestion. The vehicle-to-vehicle variability does not characterize the evolution of congestion on the network. Rather, it models the variability due to the traffic signals: their presence causes queues and delays which are experienced differently depending on the entrance time in the network. Other factors add to the vehicle-to-vehicle variability, in particular differences in driving behavior and interactions with other flows of the transportation network (bus, pedestrians, parked vehicles, transit priorities). The characterization of this variability is crucial for arterial traffic estimation. The available data prevents a detailed temporal description of these variations, suggesting that a statistical approach is adapted. It is often acknowledged that arterial traffic has periodic dynamics, when studied over the duration of several traffic cycles, the periodicity being forced by the cycle of the traffic lights [70, 148]. Under such considerations, it is possible to study the probability distribution of delay or travel times.

The study of speeds and travel time distributions is part of ongoing research that started in the 1950s with the emergence of flow-based traffic engineering [27]. This research area has been closely related to queuing theory and delay estimation for fixed cycle traffic lights. A major early contribution was developped by Webster [162], modeling arrivals of vehicles as a Poisson Process and deriving the mean average delay and queue length at the end of the green time from analytical expressions and numerical simulation. This work has been used widely to set cycle timings at isolated intersections for which the demand and capacity have been previously estimated.

This work was followed by other statistical models generalizing the assumptions and the setting of Webster's work [117, 116, 76]. These results are generalized to arrival

processes for which the number of arrivals per time interval is a discrete random variable, not necessarily with a Poisson distribution [55, 156, 106]. The *probability generating function* (pgf) of overflow queue [55] was derived for general arrival distributions. The characterization of the stationary delay distribution was derived under simplification assumptions [16, 76] and with computational methods [126]. These articles model queues at traffic signals under stationary assumptions and numerically characterize the link delay distribution based on vertical queueing theory: vehicles are assumed to stack up upon one another at the point where congestion begins or at the stop line of a traffic signal. This modeling has the limitation to consider that vehicles incur no delay traveling to the point of congestion. The stationarity assumption is also limiting and has been addressed using time dependent vertical queueing models [100, 159] and discrete time systems.

Horizontal queueing theory has received recent interest [62] and has a lot of potential for arterial traffic estimation from probe vehicles, as this thesis will demonstrate.

2.5 Data driven models

By nature, arterial traffic has very high variability, which makes it challenging to use flow models for arterial networks. A statistical approach is suitable because sensing every vehicle is impractical and because this allows for the incorporation of other information types (such as human mobility patterns [73]). Data driven models may be easier to implement and show robust performances, even though they may not necessarily have a physical justification nor interpretation.

Previous research has studied the estimation and short term prediction of sensor readings using Dynamic Bayesian Networks [102, 129] and regression models [119]. These articles assume that sensors (such as loop detectors) provide measurements with a fixed frequency at fixed locations. Probe data on arterials is assumed to be available at random times and random locations, making this assumption limiting for arterial traffic estimation, as sparsely sampled probe data is available at random times and random locations. Other approaches [83, 65] assume that either a single measurement per time interval or aggregated measurements per time interval are available for each road segment of the network (according to the map discretization). This assumption limits the capacity to represent the variability of travel times among the vehicles traveling on the network. Moreover, such approaches are not adapted to missing data, when no information is available on some parts of the network.

A specific statistical approach inspired from the Ising model was developed in [65]. It relies on measurements of the level of congestion, as a value in the $[0, 1]$ interval. Transforming traffic data into binary congested/uncongested values is a difficult process by itself and has not been specifically addressed in the literature. Neural networks and pattern matching [57] have been used to estimate traffic from GPS data. The model presented in [57] makes the critical assumption that the velocity is spatially homogeneous and similar among drivers. This assumption does not take into account the variability of travel times due to the frequent stops at traffic signals. Some researchers have examined the processing of high-frequency probe data (one measurement approximately every 20 seconds or less) [95], which allows for reliable calculation of short distance speeds and travel times.

Markov chains have also been used to compute route travel times [109, 132], considering the travel time on a link as a mixture distribution (*e.g.* the components of the mixture represent different delay patterns such as delayed and non-delayed vehicles). A transition matrix models the probability to experience a certain type of delay, given the type of delay

experienced on the upstream link. This approach is very promising, its main limitation being the amount of data or prior knowledge currently needed to train the models as presented in [109, 132].

2.6 Conclusion

Existing research in the field of traffic modeling and estimation underlines the importance of two approaches: physical and statistical models.

- *Physical models* describe the dynamics of traffic flows on the network. They have been successfully applied for highway traffic estimation, using data from dedicated infrastructure, mobile sensors or a combination of both. The models developed for arterial traffic are also very promising. Their main limitation is often the type and/or amount of available data which is assumed to be available. The calibration of the models can also be a source of inaccuracy. It may also limit the possibility of large scale deployment of the models.

- *Statistical models* provide a very general framework to represent traffic dynamics. The models take advantage of the increasing amounts of data collected today. They have the prospect to estimate and forecast traffic conditions very accurately with appropriate training. However, these models may be hard to interpret and may produce results which are physically infeasible.

From these insights, a statistical model based on the physics of traffic flow seems to be a very promising approach for urban traffic modeling and estimation. This is the approach developed in the present dissertation.

Chapter 3

The potential of probe vehicle data for traffic estimation

Chapter 1 underlines the need for real-time reliable traffic information on both the major freeways and the secondary network. To provide a global solution to this traffic information gathering problem, one needs traffic information everywhere where there is congestion within the transportation network. Given the high costs of deploying a traffic monitoring system and the lack of public infrastructure, the thorough analysis of the potential of mobile probes to provide a feasible alternative for this problem is very valuable for both traffic researchers and practitioners. In this chapter, the attention focuses on highway traffic estimation to motivate the use of sparsely sampled probe data for traffic estimation. The case of arterial traffic is analyzed in the remaining chapters.

One of the major challenges in using mobile probe data for traffic estimation is the difficulty to incorporate this data into traffic models, which are traditionally used to describe highway traffic. Several types of models can be used: statistical models [101, 28], and flow models [108, 134]. When a flow model is used, this process is known as *inverse modeling* or *data assimilation*: it consists in incorporating data in the mathematical model of a physical system, in order to estimate the current state of the system and forecast its future state [107, 26]. In the field of inverse modeling, *Lagrangian* sensing specifically refers to measurements performed along a sensor's trajectory which it usually cannot control. Examples of this are smartphones traveling onboard cars following highway traffic flow. This is in contrast to *Eulerian* sensing, in which sensors are fixed (for example, video cameras or loop detectors along highways) and monitor a specific location or domain.

While inverse modeling using Lagrangian sensors is a somewhat established field in some areas such as oceanography [107, 26], it is still a relatively novel technique in the field of transportation engineering. Traditional approaches such as *Kalman Filtering* (KF) and its extensions have been applied to traffic models to perform estimation, in particular using first order models such as the *Cell Transmission Model* (CTM) [49, 50], see in particular [120, 122, 121, 151]. *Extended Kalman* (EKF) filtering has been used to handle second order models, when the discretization scheme used allows it, see for example [160]. *Ensemble Kalman Filtering* (EnKF) [61] arises as a technique capable of capturing shocks [30] and has been used for speed estimation on the highway [164, 165]. All the aforementioned methods produce a best estimate of traffic (in some sense, for instance in the least square sense), sometimes with statistics related to the produced results, such as confidence or probability associated with the prediction. The research in [164, 165] demonstrates the potential of probe data for highway traffic estimation. This chapter investigates a practical problem which goes beyond the specific data assimilation problem (i.e. production of an estimate), which we define as the *guaranteed range of travel times*

as follows.

Problem 3.1 (Guaranteed range of travel times)**.** *Given a set of loop detectors, and given a set of probe vehicles equipped with GPS, (i) how to reconstruct travel time, (ii) how to produce a guaranteed range for travel time, given the available data?*

This problem specifically addresses needs from the traveling public, since it is aimed at providing the public with the travel time information [46], and a range of validity of this information. The term "guaranteed range" refers to the possible range of travel times, taking into account the uncertainties in the model, and assuming that the loop detector and probe data are exact. In addition, the problem of sampling strategy is also analyzed for its importance in large scale deployments and its relationship to estimation accuracy and confidence:

Problem 3.2 (Sampling strategy)**.** *What is the influence of the penetration rate of equipped vehicles and of the spatial sampling strategy on the range of the travel time estimation?*

This second problem is helpful for cellular network operators and cellular phones manufacturers, who are currently in the process of mapping the transportation network with "virtual detectors", *i.e.* GPS data collection mechanisms which partly rely on the geometry of the transportation network.

This chapter is organized as follows: Section 3.1 summarizes the flow models used in this study and introduces the notation specific to the present chapter. Section 3.2 describes how sparsely sampled probe vehicle data, and in particular data gathered through the specific spatial sampling strategy developed by Nokia (Virtual Trip Lines, see Section 1.2.2) can complement existing loop detector data. The section summarizes the data assimilation procedure and the corresponding algorithm developed for the estimation of ranges of guaranteed travel times. In section 3.3, the method is implemented using the *Mobile Century* data set [81, 59, 164], which consists of loop detector and GPS-based smartphone data collected for 8 hours of traffic on I880 in Union Landing, CA, for 100 vehicles equipped with Nokia N95 phones. This section presents some conclusions of numerical results obtained for travel time estimation, as well as a study of the influence of penetration rate and spatial sampling strategy.

3.1 Background

3.1.1 Traffic flow models

Kinematic wave theory

Traffic flow on a highway segment can be described using both density and flow functions, which represent an aggregated number of vehicles per space (respectively time) unit. The *Lighthill-Whitham-Richards* (LWR) *partial differential equation* (PDE) [108, 134] is a first order model obtained from conservation of vehicles and an empirical relation between vehicle flow $q(t,x)$ and vehicle density $\rho(t,x)$ and commonly used in traffic engineering:

$$\frac{\partial \rho(t,x)}{\partial t} + \frac{\partial q(\rho(t,x))}{\partial x} \;=\; 0 \tag{3.1}$$

The flow-density relation $q(\rho)$ is known as *flux function* or *fundamental diagram*. The theory [20, 42] used in this chapter assumes that the flux function is concave. For simplicity, we choose a specific flux function, the triangular diagram [49, 50], as commonly done in the literature:

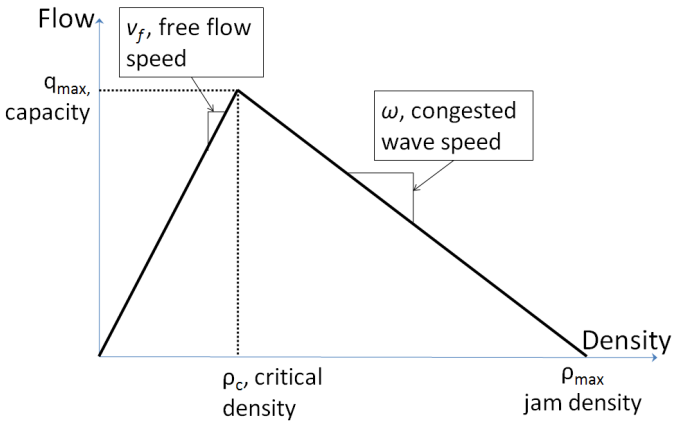


Figure 3.1: Representation of a triangular flux function.

$$q(\rho) = \begin{cases} v_f \rho & \text{if} \quad \rho \in [0, \rho_c] \\ w(\rho_c - \rho) + v_f \rho_c & \text{if} \quad \rho \in [\rho_c, \rho_{\max}] \end{cases}$$

In the diagram of Figure 3.1, v_f is the free flow speed, ρ_c is the critical density, and $\rho_{\max}$ is the maximal density. All these quantities are illustrated in Figure 3.1. The capacity of the road is the maximal flow $q_{\max} = \rho_c v_f$. The parameters v_f, w, ρ_c and $\rho_{\max}$ are related by $\rho_{\max} w = \rho_c(v_f + w)$, which means that the triangular fundamental diagram is fully characterized by three parameters.

Because density is an aggregated quantity, which cannot be measured by probe vehicles directly, the LWR PDE is difficult to use as such to incorporate vehicle trajectory data available from probe vehicles. To face this issue, the present model uses an alternate (equivalent) representation of traffic which was introduced by Newell and Daganzo, following the work of Moskowitz [123, 52, 53]. The modeling uses consecutive integer labels assigned to vehicles entering the highway at a user defined location $x = x_{\text{in}}$. The vehicles are counted from the reference point $(t = 0, x = x_{\text{in}})$. The first vehicle is assigned an arbitrary label, usually chosen to be 0. The choice of this arbitrary label at $t = 0$ and $x = x_{\text{in}}$ does not influence the results. Assuming that vehicles do not pass each other, one could imagine that an observer at location $x = x_{\text{in}}$ numbered the vehicles as they passed him, starting to count from the arbitrary first label (usually chosen to be zero). The *Moskowitz function* $\mathbf{M}(t, x)$ (also known as *cumulative number of vehicles function*) is a continuous representation of the label of the vehicle at location x and time t, and encodes the distribution of the vehicles on the highway at all times. The space and time derivatives of the Moskowitz function are related to the flow and density functions as follows [123, 52, 53]:

$$\frac{\partial \mathbf{M}(t,x)}{\partial t} = q(\rho(t,x)) \qquad \text{and} \qquad \frac{\partial \mathbf{M}(t,x)}{\partial x} = -\rho(t,x). \tag{3.2}$$

Using equation (3.2), one can transform equation (3.1) into the following Moskowitz *Hamilton-Jacobi* PDE [52, 53]:

$$\frac{\partial \mathbf{M}(t,x)}{\partial t} - q\left(-\frac{\partial \mathbf{M}(t,x)}{\partial x}\right) = 0. \tag{3.3}$$

Solution of the Moskowitz HJ PDE

Solutions to PDE (3.3) subject to initial and boundary conditions are known [48, 64] and can be computed with standard numerical analysis tools. Solving equation (3.3) requires

the knowledge of the initial state of the highway, *i.e.* the knowledge of an initial function $\mathbf{M}_0(x) := \mathbf{M}(0, x)$ at time $t = 0$, which would represent a distribution of labels of vehicles initially on the highway. Note that when this knowledge is not available, one can use the "flush" effect of the highway (i.e. waiting long enough until initial vehicles have disappeared from the section of interest) to avoid the need for this data. Assuming that loop detector data is available at locations $x = x_{\text{in}}$ (upstream) and $x = x_{\text{out}} > x_{\text{in}}$ (downstream), one can prescribe counts at these locations, i.e. $\mathbf{M}(t, x_{\text{in}}) = \gamma(t)$ and $\mathbf{M}(t, x_{\text{out}}) = \beta(t)$, where $\gamma(t)$ and $\beta(t)$ are the vehicle counts measured by the detectors. In other words, the label $\gamma(t)$ is incremented by one each time a car drives by the location $x = x_{\text{in}}$. A similar rule applies at $x = x_{\text{out}}$. Finally, given a vehicle with an integer label $\overline{\mathbf{M}}_i$ and a trajectory given by $x_i(t)$, the value of the function $\mathbf{M}(t, x_i(t))$ is constant and equal to $\overline{\mathbf{M}}_i$. This fact is true because the value of the label $\overline{\mathbf{M}}_i$ does not change along the trajectory $x_i(t)$ of vehicle $\overline{\mathbf{M}}_i$, thus $\mathbf{M}(t, x_i(t)) = \overline{\mathbf{M}}_i$ for all times t during which the vehicle is on the corresponding segment of highway. The value imposed to the soluton of 3.3 along the trajectory of a vehicle is an *internal condition*.

The mathematical properties of the solution of (3.3) require specific treatments to introduce internal boundary conditions. A specific control framework based on Lax-Hopf's formula and viability theory [20, 42] enables the introduction of internal boundary conditions. Using this framework, one can derive the unique Barron-Jensen/Frankowska solution of the Hamilton Jacobi Partial Differential Equation. The details of the derivations of the solution are out of the scope of this thesis and can be found in [42, 43]. It is possible to introduce passing rates in the model as done in [42, 43]. The corresponding *initial*, *boundary* and *internal* conditions on the function $\mathbf{M}(t, x)$ can thus be summarized by:

- Initial condition $\mathbf{M}(0, x) = \mathbf{M}_0(x)$ — Vehicle distribution at initial time
- Left boundary condition $\mathbf{M}(t, x_{\text{in}}) = \gamma(t)$ — Inflow of vehicles
- Right boundary condition $\mathbf{M}(t, x_{\text{out}}) = \beta(t)$ — Outflow of vehicles
- Internal conditions $\overline{\mathbf{M}}(t, x_i(t)) = \mathbf{M}_i$ labeled $\overline{\mathbf{M}}_i$ for all i — Trajectory measurement for the vehicle

3.1.2 Reconstruction of a posteriori travel time function

This work is focused on the computation of the *a posteriori* travel time $TT(t)$ at time t from the estimation of the Moskowitz function. The *a posteriori* travel time is defined as follows. If a vehicle $\overline{\mathbf{M}}_i$ crosses the upstream boundary x_{in} of the highway at time τ, and crosses the downstream boundary x_{out} at time t, the a posteriori travel time $TT(t)$ is defined by $TT(t) = t - \tau$, and represents the time necessary to cross the road section observed by the vehicle $\overline{\mathbf{M}}_i$ leaving the highway at time t. The a posteriori travel time can thus be obtained from the boundary condition functions $\gamma(\tau)$ and $\beta(\tau)$ using the following procedure:

Algorithm 1 Algorithm for computing the travel time function

Input t (time at which one wants to compute travel time)
Read $\overline{\mathbf{M}} := \beta(t)$ from downstream loop detector
Find τ such that $\gamma(\tau) = \overline{\mathbf{M}}$ from upstream loop detector (using backtracking search)
Compute $TT(t) := t - \tau$ (if τ exists)

3.2 Data assimilation using mixed Eulerian/Lagrangian Data

3.2.1 Model of label evolution and trajectories

Boundary conditions

Figure 3.2(a) illustrates the physical interpretation of the boundary conditions $\gamma(\cdot)$ and $\beta(\cdot)$ of the Moskowitz function in terms of the loop detector counts. As can be seen in this figure, data is prescribed at $x = x_{\text{in}}$: $\mathbf{M}(t, x_{\text{in}}) = \gamma(t)$ where $\gamma(t)$ is the label function at x_{in}, which is constructed from the loop detector measurements by summing the vehicles as they pass it. For example, if the label of the vehicle passing at $t = 17$ (min) is $\gamma(17) = 25$, and five vehicles pass between $t = 17$ (min) and $t = 18$ (min), then $\gamma(18) = 22$. Similarly, downstream, at $x = x_{\text{out}}$: $\mathbf{M}(t, x_{\text{out}}) = \beta(t)$ where $\beta(t)$ is the label function at x_{out}, which is constructed from the loop detector at $x = x_{\text{out}}$.

Initial conditions

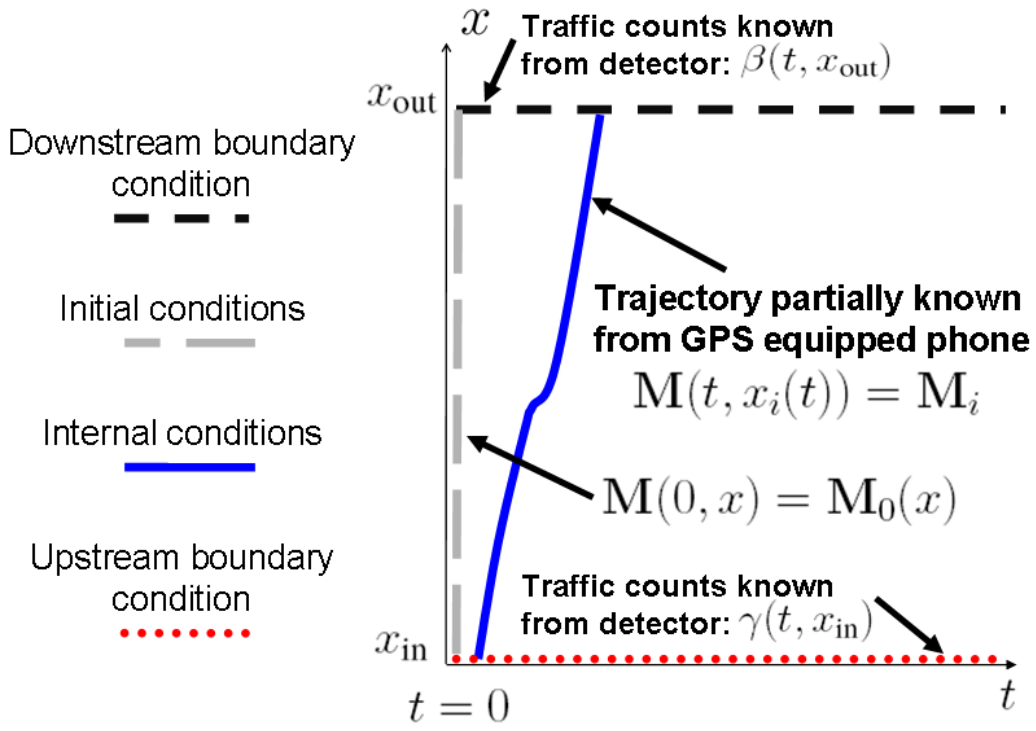
The knowledge of the initial vehicle distribution (similar to an aerial picture of the road at time $t = 0$), would give the initial condition represented with a dash line in Figure 3.2(a), in other words the labels of cars as initially positioned on the highway. The corresponding function is called $\mathbf{M}_0(x)$. For example, if the vehicle at $x = x_{\text{in}}$ and $t = 0$ is (arbitrarily) labeled vehicle zero ($\mathbf{M}(0, x_{\text{in}}) = 0$), and if there are 20 vehicles between $x = x_{\text{in}}$ and $x > x_{\text{in}}$, then $\mathbf{M}(0, x) = -20$. Note that the labels can all be arbitrary shifted by the same amount, which is prescribed at $x = x_{\text{in}}$ and $t = 0$: if one arbitrarily decided to call the vehicle at $x = x_{\text{in}}$ vehicle 28, *i.e.* $\mathbf{M}(0, x_{\text{in}}) = 28$, it would follow that $\mathbf{M}(0, x) = 8$. It can easily be seen that this shift does not impact equation (3.3) since the PDE only depends on derivatives of $\mathbf{M}(\cdot, \cdot)$. For practical reasons, initial condition data is not easily measurable and rarely available. For this reason, the method proposed in this chapter does not assume knowledge of $\mathbf{M}_0$.

Probe vehicle conditions

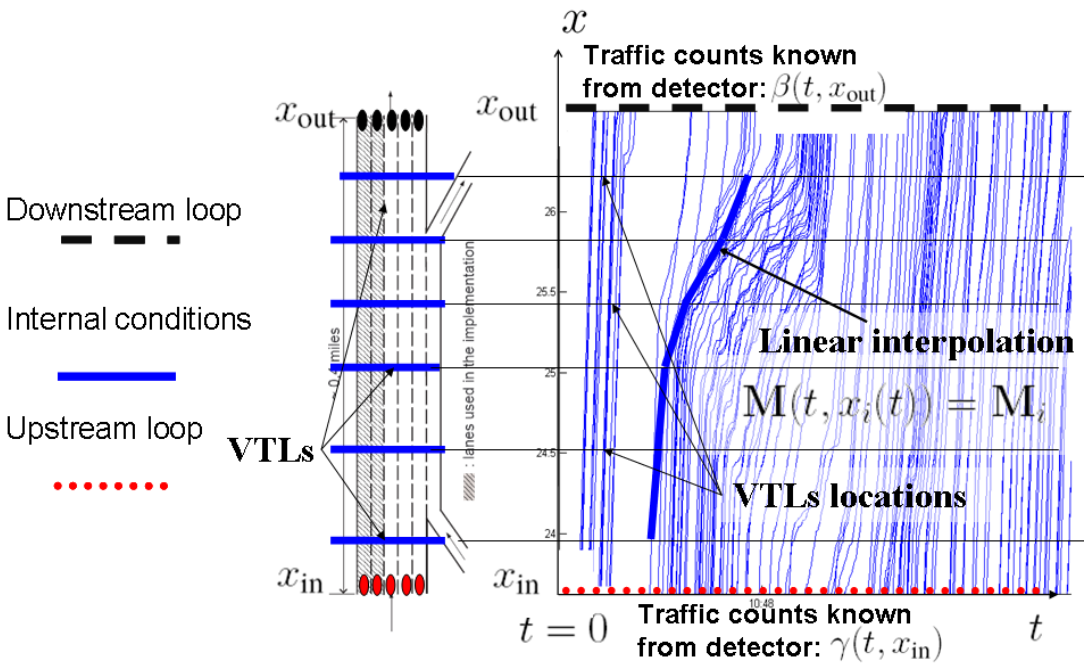
In Figure 3.2(a) the solid trajectory line on the time-space diagram represents the successive locations where the phone can be probed along a vehicle trajectory using the VTL sampling paradigm. The travel times between successive VTLs are then linearly interpolated to represent the vehicles trajectory. For privacy reasons, the system does not track entire trajectories of the vehicles (it is represented this way on the figure for illustration purposes – in practice only subsets of this trajectory would be transmitted to the system). The sparsely sampled trajectory of vehicle $\overline{\mathbf{M}}_i$ is characterized by the times t when measurements are received and corresponding locations $x_i(t)$. When the vehicle reports its position $x_i(t)$ at time t, it provides information on the Moskowitz function: $\mathbf{M}(t, x_i(t)) = \overline{\mathbf{M}}_i$. In practice, the label $\overline{\mathbf{M}}_i$ of the vehicle reporting its position is not known. The estimation algorithm takes into account this unknown.

3.2.2 Computation of the boundary condition functions using loop detector data

Figure 3.2(b) illustrates the procedure developed to collect the data used for the data assimilation procedure.



(a) Illustration of the notation used in the kinematic wave theory.



(b) Data collection procedure used by the algorithm. The data used for the picture comes from the *Mobile Century* experiment [81, 59].

Figure 3.2: Physical interpretation of the boundary conditions on the Moskowitz function.

Assimilation of the boundary conditions

The data assimilation procedure requires the incorporation of loop detector data (flow data only) into the flow model. For this, the knowledge of the functions γ and β is necessary. By definition of these loop detector flows, γ and β can be obtained (modulo a constant) by direct integration of the flows measured by the detectors. Thus, calling $q_{\text{inflow}}(t)$ the flow measured by the upstream detector and $q_{\text{outflow}}(t)$ the flow measured by the downstream detector, it follows that:

$$\gamma(t) = \int_0^t q_{\text{inflow}}(\theta)d\theta \qquad \text{and} \qquad \beta(t) = \int_0^t q_{\text{outflow}}(\theta)d\theta + \Delta \tag{3.4}$$

The parameter Δ represents the value[1] of $\beta(0, x_{\text{out}})$. The total number of vehicles present on the highway at time $t = 0$ is $-\Delta \geq 0$, which is an unknown of our problem. If the parameter Δ was known (for instance by taking a picture of the highway at initial time and counting the total number of vehicles on the highway), then the a posteriori travel time could be computed exactly, assuming that the loop detector flow data was errorless. However, since it is not the case, the parameter Δ must be estimated. Once Δ has been estimated, Algorithm 1 can then be used to compute the travel time. As will appear in the next sections of this algorithm, Δ is never known a priori (unless $\mathbf{M}_0$ is known, which is difficult in practice). It will therefore become one of the decision variables of the algorithm.

Assimilation of the probe data

Internal conditions of the problem are collected using *Virtual Trip lines* (VTLs), as described in Section 1.2. Recall that VTLs are geographical markers stored in the client (i.e. the mobile handset), which trigger a position and speed update whenever a probe vehicle crosses them. A VTL can thus be viewed as a virtual loop detector, which detects crossing of equipped vehicles. In essence, it provides readings of the $x_i(t)$ function at specific geographical locations on the highway according to a sampling procedure used by Nokia to guarantee privacy of the users. Figure 3.2(b) illustrates the data available through VTL sampling. As can be seen from Figure 3.2(b) the location of probe vehicles is known at sampling locations (VTLs) and corresponding crossing times. Each vehicle is uniquely identified at each crossing VTL and reports its travel time between successive VTL crossings. Between successive VTL crossings, we assume in this model that the vehicle has a constant speed, corresponding to linear trajectories in the space-time diagram. In practice, trajectories are not necessarily linear between sampling locations, in particular if the sampling locations are distant and cover junctions, ramps or merges.

3.2.3 Linear Program formulation of the data assimilation problem

Integration of the loop and probe data into the model

The problem of integrating initial, boundary and internal conditions into the Moskowitz equation (3.3) is in general extremely challenging. This complex problem can be decomposed into many simple problems associated to each of the value conditions (initial, boundary, and internal). For this, define a *component* function as the solution of the Moskowitz PDE associated to each individual value condition [42, 43]. The definition of the different components is done as follows:

[1] Note that $\gamma(0) = 0$ by assumption.

- *Initial condition component:* $\mathbf{M}_{\mathbf{M}_0}(t,x)$. This initial condition function can be computed directly from the initial condition $\mathbf{M}_0(x)$; it encodes the dependence of the solution on the initial condition.

- *Left boundary condition component:* $\mathbf{M}_\gamma(t,x)$. This left boundary condition function can be computed directly from the left boundary condition $\gamma(t)$; it encodes the dependence of the solution on the left boundary condition (upstream loop detectors).

- *Right boundary condition component:* $\mathbf{M}_\beta(t,x)$. This right boundary condition function can be computed directly from the right boundary condition $\beta(t)$; it encodes the dependance of the solution on the right boundary condition (downstream loop detectors).

- *Internal condition component:* $\mathbf{M}_{\overline{\mathbf{M}}_i}(t,x)$. This internal condition function can be computed directly from the trajectory $x_i(t)$ of vehicle $\overline{\mathbf{M}}_i$; it encodes the dependance of the solution on the internal condition.

The component functions can be computed individually using dynamic programming methods [52, 53], a Lax-Hopf formula [19], or using the minimization of closed-form expression functions [43]. Compatibility conditions [42, 43], which have a direct physical interpretation, ensure the well-posedness of the problem. In this chapter, we choose the last method to compute the component functions. The method has the advantage to compute exact value of the Moskowitz function at any time t and location x and does not require any discretization scheme. The details of the derivations are presented in [42, 43, 115] and are out of the scope of this thesis. As shown in [115], the assumption of a triangular fundamental diagram simplifies even more the computation, even though any concave diagram can be used in this setting. When physical compatibility conditions are met [42], the solution to the Moskowitz equation (3.3) can be simply computed as the minimum of the component functions [19, 42, 43]:

$$\mathbf{M}(t,x) = \min\left(\mathbf{M}_{\mathbf{M}_0}(t,x), \mathbf{M}_\gamma(t,x), \mathbf{M}_\beta(t,x), \mathbf{M}_{\overline{\mathbf{M}}_1}(t,x), \ldots, \mathbf{M}_{\overline{\mathbf{M}}_n}(t,x)\right) \tag{3.5}$$

This property is proved in detail in [19, 42]. In order to be able to simultaneously impose the initial, boundary and internal conditions, these conditions must satisfy necessary and sufficient conditions known as *compatibility conditions* [42, 43]. Indeed, any arbitrary initial, boundary and internal conditions cannot be simultaneously imposed on the Moskowitz function. For instance, if the initial condition consists in a completely congested highway, no positive inflow can be imposed at the entrance x_{in}, since there is no available space for entering vehicles. Similarly, if the highway is initially empty, no positive outflow can be imposed since no cars are present.

Linear program formulation

Let X be defined as $X := (\Delta, \overline{\mathbf{M}}_1, \overline{\mathbf{M}}_2, \ldots, \overline{\mathbf{M}}_n)$ where n is the total number of probe vehicles used for the internal conditions. The variable X represents the decision variable of the estimation problem. Recall that Δ represents the number of vehicles initially present on the road segment and that $\overline{\mathbf{M}}_i$ is the label of the i^{th} probe vehicle. The compatibility conditions can be shown to be equivalent to a set of k linear inequalities [2] in the variable X. The mathematical proof of this fact is cumbersome and algebraically involved. It is out of the scope of this thesis and available in [41]. The linear inequalities resulting

[2]The number k depends upon the type of conditions used for the reconstruction, the repartition of the car trajectories, and the time horizon.

from [41] can be written formally as $AX \leq b$, where $A \in \mathbb{R}^{k \times n+1}$ and $b \in \mathbb{R}^k$. The matrix A and the vector b depend upon the inflow, outflow, and trajectory data.

Let the vector c be defined as $c^T = (1, 0, 0, \ldots, 0)$. Using this definition, the analytical expression of Δ is given by $\Delta = c^T X$. To compute the value of Δ, consider the following *Linear Programs* (LPs):

$$\begin{array}{ll} \textbf{min:} & c^T X \\ \textbf{s.t.:} & AX \leq b \end{array} \quad \text{and} \quad \begin{array}{ll} \textbf{max:} & c^T X \\ \textbf{s.t.:} & AX \leq b \end{array} \tag{3.6}$$

The solution of the above LPs yield two objective values $\Delta_{\min}$ and $\Delta_{\max}$, which can be interpreted as follows: the value $-\Delta_{\min}$ represents the maximal number of vehicles that can possibly be present on the highway, and the value $-\Delta_{\max}$ represents the minimal number of vehicles that must be present on the highway (assuming our data is exact). To the possible range of values for Δ, $[\Delta_{\min}, \Delta_{\max}]$, corresponds a possible range of travel times defined by $TT_{\min}(t)$ and $TT_{\max}(t)$ and computed using Algorithm 1. The overall method is represented in Algorithm 2.

Algorithm 2 Process used to construct travel time ranges from Eulerian/Lagrangian measurements.

Input Loop detector data (flow data) and VTL data
Compute Boundary and internal conditions γ, β and μ_i
Compute A and b
Compute $\Delta_{\min}$ and $\Delta_{\max}$ (using LP (3.6))
Compute $TT(t)$

3.3 *Mobile Century* Implementation

This section analyzes the estimation capabilities of the previously described procedure using loop detector data, probe vehicle data, and video detector data collected during a field experiment known as *Mobile Century* [81, 59]. It presents an analysis of the sampling strategies and the number of equipped vehicles to implement this algorithm in practice.

3.3.1 Description of the experiment

The *Mobile Century* [81, 59] experiment took place on February 8th, 2008. It was conducted on Highway I-880, near Union City, CA, between Winton Ave. and Stevenson Blvd. (see Figure 3.3). This 10-mile long section of highway was selected specifically for its complex traffic properties, which include alternating periods of free-flowing, uncongested traffic, and slower moving traffic during periods of heavy congestion. The section also has a high density of loop-detectors (17 loops on the section of interest). The data from these sensors is collected by the *Freeway Performance Measurement System* (PeMS).

The experiment consisted in deploying 100 GPS- equipped Nokia N95 cell phones on a freeway during eight hours. For the duration of the experiment, 165 UC Berkeley students drove loops on the section of interest between 10am and 6pm. This period encompasses both free flow and congested traffic, and the transition between the two of them.

The data was collected in two ways during the experiment.

(i) A privacy aware architecture was developed, which uses the concept of *Virtual Trip Line* (VTL) explained earlier (Section 1.2.2). During the experiment, 40 VTLs were deployed in the section of interest (each VTL covers both travel directions). This

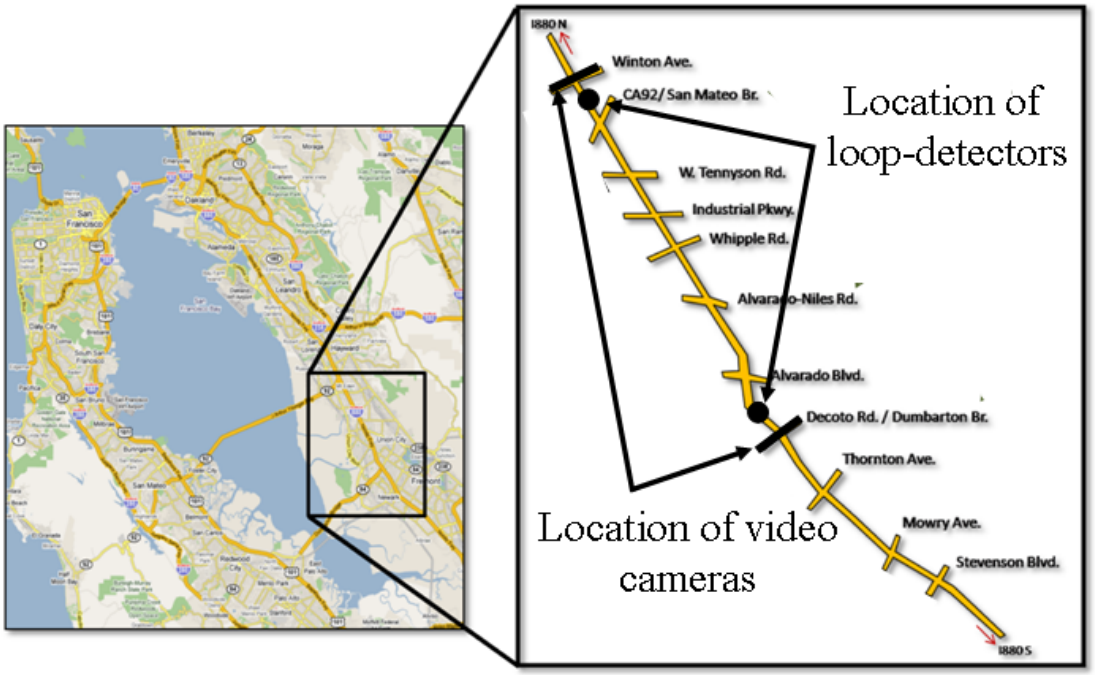


Figure 3.3: Set-up of the *Mobile Century* experiment. The figure represents the placement of the video camera and the loop detectors selected for the computation and the validation of our results.

data was used to produce real-time travel time and speed estimates on the section of interest and presented to the press in real-time. During the experiment, a data assimilation algorithm based on *Ensemble Kalman Filtering* (EnKF) [164] provided travel time estimates in real-time. On February 8^{th}, 2008, the goal of the algorithm was to show the possibility of reconstructing speeds and travel time in real-time from a set of probe vehicle measurements.

(ii) Each cell phone was storing its position and velocity every three seconds. This data (trajectory data) was stored for archival purposes in order to evaluate the quality of the data a posteriori. It was only generated for experimental validation, and would not be collected in an operational system.

In order to validate the estimation results of all algorithms developed during this field experiment, video data was collected from three bridges in the deployment area. The travel times were experimentally measured using a set of six HDV cameras deployed on the Mowry Ave., Decoto Rd. and Winton Ave. overpasses. The video data was sent to six laptops which were synchronized before the experiment and archived the data with timestamps. The camera footage was analyzed after the experiment, and the identifiable license plates (approximately 70% of the vehicles) were subsequently stored in a database. A matching algorithm was used to compute travel time through license plate reidentification. Figure 3.3 shows the deployment area of two of the three camera locations used to produce the data presented in this chapter. The travel times measured by the video camera are depicted by dots in the figures of the subsequent sections.

3.3.2 Analysis of probe vehicle penetration and VTL spacing

This section presents an analysis of probe vehicle penetration rate and VTL spacing, using the data presented in the previous section.

Penetration rate

The penetration rate achieved during the experiment was 3% to 5% depending on the time of the day [81, 59]. VTL data was collected at each of the 40 VTLs deployed on the section of highway. In an operational system, not all VTL measurements would be used by the system to probe all the vehicles systematically. In practice, a vehicle traveling across a VTL is randomly probed. In the present study, the number of measurements used by the algorithm is artificially decreased in order to simulate lower penetration rates. This procedure is a way to degrade the data set in order to assess the performance of the algorithm for low penetration rates and determine the penetration rate required for robust estimation.

VTL degradation

In order to study the influence of VTL spacing on the quality of the results, the available VTLs deployed for this study was reduced, *i.e.* some of the VTLs were artificially suppressed in the parametric study. The goal of this procedure is to study the performance of the algorithm when measurements are spatially sparse.

Parametric study of penetration rate and VTL sampling

Using the two decimation procedures outlined above, the two linear programs (3.6) for $\Delta_{\min}$ and $\Delta_{\max}$ are solved for each of the penetration rates and the VTL spacing chosen in the study. For each of these values, the length of the period considered for the assimilation is 1 hour and 25 minutes. The corresponding bounds on travel times (computed from the $\Delta_{\min}$ and $\Delta_{\max}$) are obtained for a vehicle initially entering the segment of interest. The results are summarized in Figure 3.6. This figure represents, for each pair (penetration rate - VTL spacing), the difference $TT_{\max}(t) - TT_{\min}(t)$. This range represents the guaranteed bound on average travel time under the assumption that overtaking and measurement noise are relatively small enough to be ignored. As expected, the range is the smallest for high penetration rates and low VTL spacing. The best corresponding range provided by the method is of the order of 100 seconds, for an average travel time of about 1200 seconds. This corresponds to a 8% error provided by the method. These results are very encouraging: they show that even with low penetration rates, and reasonable VTL spacing, the method is able to evaluate a guaranteed range for average travel time within less than 10% of its actual value. This type of information is very helpful for deployment studies, in order to determine the operational conditions for which this system would become valuable.

3.3.3 Results

Figures 3.4, 3.5 and 3.6 illustrate the numerical results obtained by solving problem (3.6).

- Figure 3.4 shows the range of travel times provided by the method if loop detectors only are used. Obviously, with two loop detectors separated by 6.06 miles, the range cannot be tight because of the extremely large uncertainty of traffic when only two measurement stations are available for such a long distance. The figure displays the validation data obtained from the video camera. Notice that all measurements

are within the guaranteed range of travel times. The figure also shows the estimate which would be obtained by computing the travel time from the PeMS speed data directly. This travel time falls outside of the guaranteed range and does not match the validation data accurately. As can be seen in the right subfigure, the estimate using the 16 PeMS speed detectors available in the section of interest falls inside the range and is closer to the cloud of points. However, this density of loop detector is rarely available on the highway network because of the important installation and maintenance costs.

- Figure 3.5 represents the guaranteed bounds on the travel time obtained from the algorithm throughout the experiment duration. As expected, this range gets tighter as the number of vehicles reporting their travel time through the VTL sensing paradigm increases. The addition of a single vehicle trajectory already produces drastic improvements in the predicted range. By construction of the optimization program, incorporating new trajectory information adds new constraints, which reduce the size of the feasible set $AX \leq b$, thus reducing $\Delta_{\max} - \Delta_{\min}$ and the range of guaranteed average travel times. Because the assumption that overtaking and measurement noise are negligible may not hold very accurately in practice, the ranges computed for the travel time function do not necessarily encompass exactly the actual travel times anymore. The next section provides explanations for this fact. Note however that, while there might be some relatively small error with these bounds, one can clearly see from these four figures that the estimates are not only tight, but also reproduce the trends of the validation data (in the present case the progressive decrease of travel time as morning congestion dissipates).

3.3.4 Comments on the results

The quality of the results depends on the validity of the assumptions of the model. The following analyzes some of the sources of inaccuracy of the results.

- The model assumes that cars do not overtake each other (no shearing) and computes an associated possible range of travel times for all the vehicles driving at the same time. This is a common assumption in numerous transportation engineering articles. However, vehicles overtake each other, their label is not constant along their trajectory. As the distribution of the travel times (in Figure 3.5) do not satisfy the non-shearing assumption, some travel times may not satisfy the guaranteed bounds (which are computed for a non-shearing situation only). The mathematical framework [42, 43] used to compute the components of the Moskowitz function allows for the introduction of a passing rate for each probe vehicle. Note however that the LWR model does not take overtaking into account and the potential of this refinement of the model is not analyzed in the present example.

- Flow from and to on- and off-ramps are not taken into consideration by this study, because of the lack of available data. The conservation of vehicles assumption is therefore violated. Data shows that the cumulative number of vehicles that exits the road section by the loop at point B is increasing faster than the cumulative number of vehicles that enter at point A. Over the course of eight hours, the difference reaches 9247 vehicles, which exceeds the road section capacity (5041 vehicles) and consequently invalidate this assumption. To produce the results, the model integrates this additional flow of vehicles as a constant throughout the day. A temporal refinement of this flow correction or measurements of inflows and outflows at ramps would improve the accuracy of the estimation.

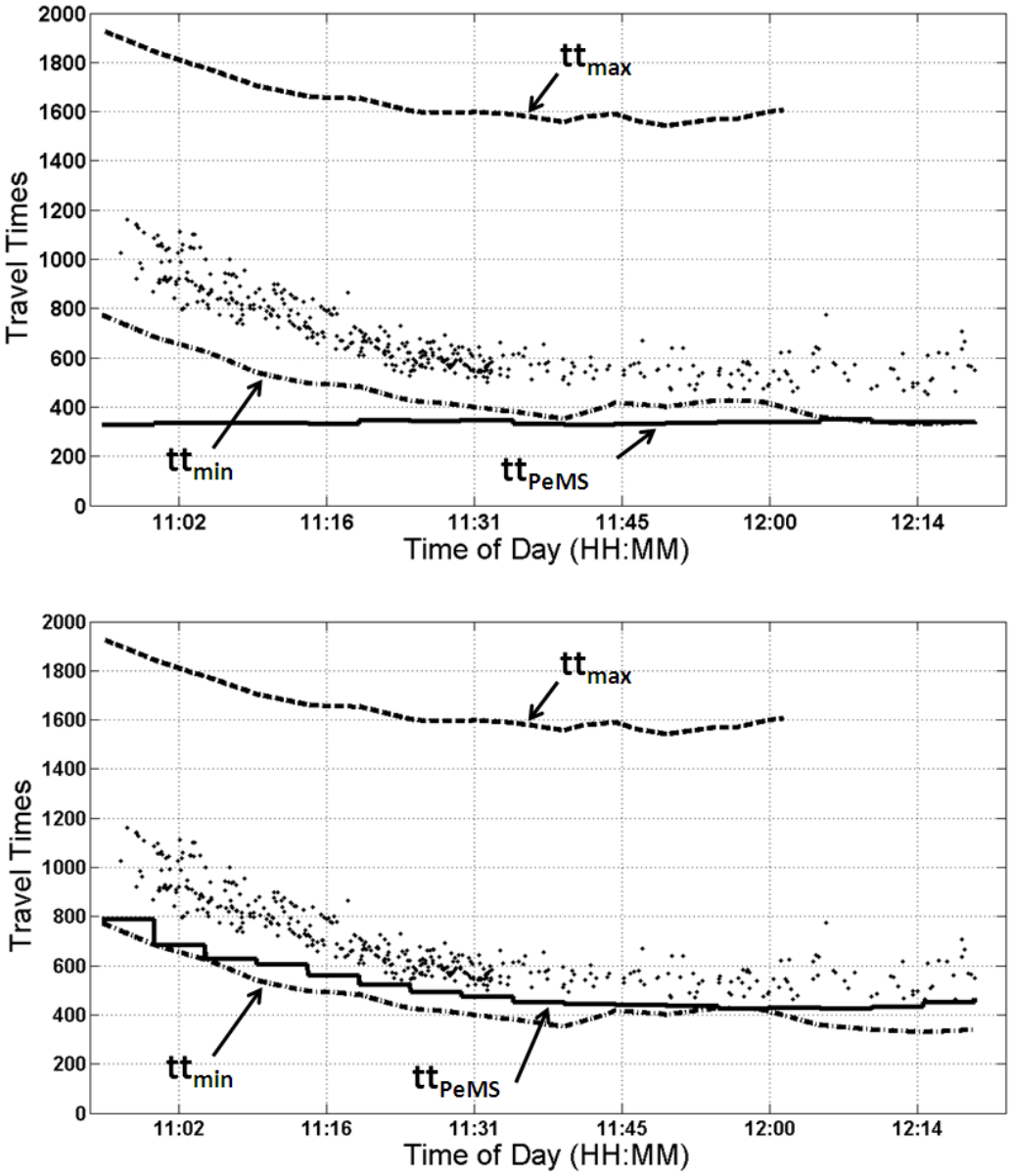
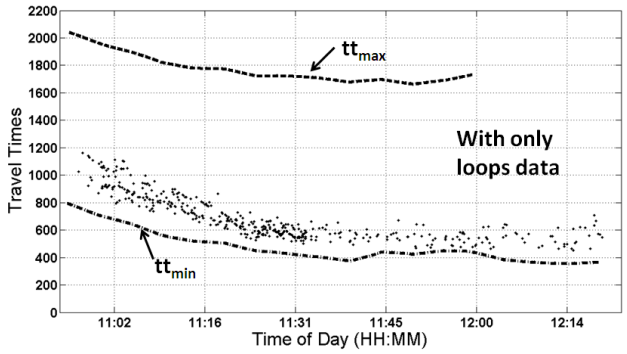
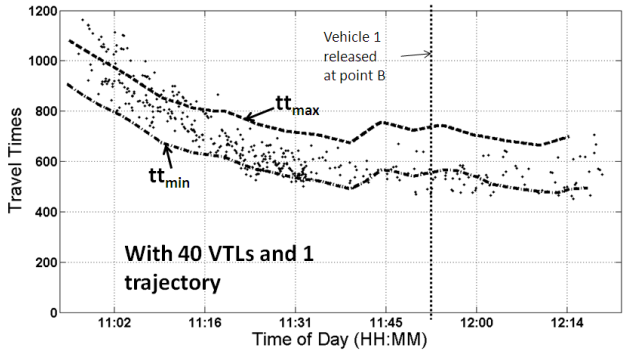


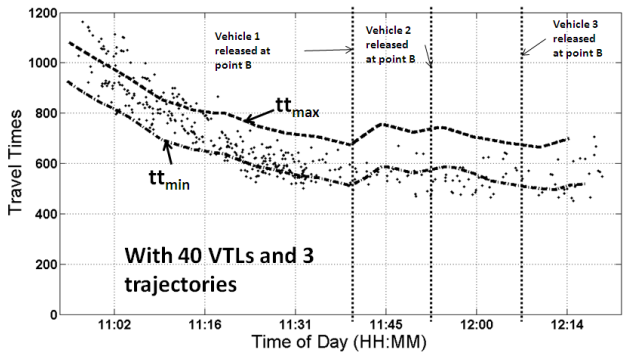
Figure 3.4: **Comparison of the computation of the travel time using our algorithm and a naive method based on integration of the speed data given by the loop detectors**. Loop detector data only is used in this example (no VTL data). The horizontal axis represents the time t (unit: hours and minutes). The vertical axis represents the travel time (unit: sec). The upper and lower bounds on travel time, computed by the algorithm (using flow data) described before, are represented by dashed lines. The actual travel times obtained from the video data are represented by dots. On both figures, they are computed using flow data from the first and last loop detectors only. The travel time computed by the commonly used PeMS algorithm using velocity data is represented by a solid line. It uses only the first and last loop in the top figure, whereas all the intermediate loop are used in the bottom figure.



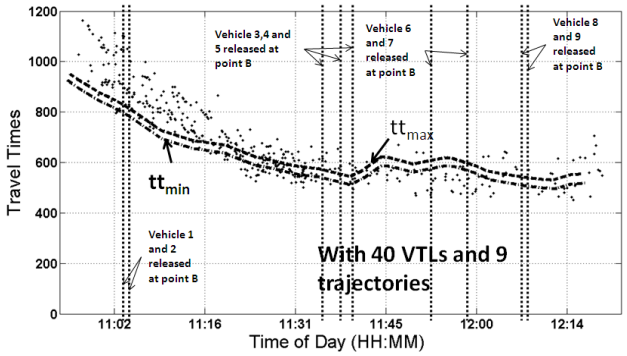
(a) Loop detectors data only.



(b) Loop detectors data+1 trajectory.



(c) Loop detectors data+3 trajectories.



(d) Loop detectors data+9 trajectories.

Figure 3.5: **Estimation of the travel time function.** The horizontal axes represent the time t (unit: hours and minutes). The vertical axes represent the travel time TT. Subfigure legends are the same as in the previous figure. The influence of the penetration rate (respectively 0%, 0.01%, 0.03% and 0.1%) is illustrated in the subfigures (respectively (a), (b), (c) and (d)).

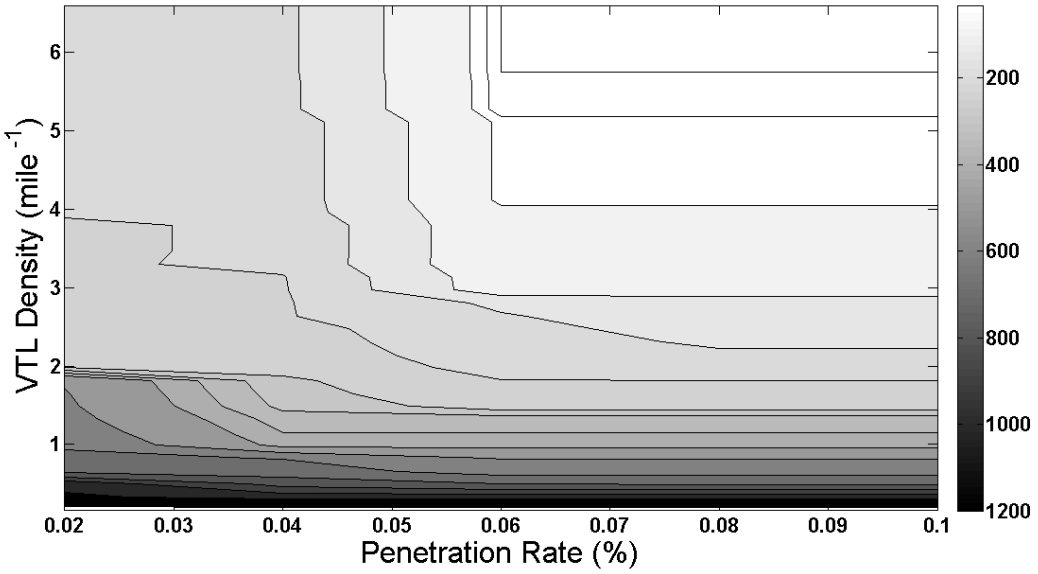


Figure 3.6: **Guaranteed range (unit: seconds) for the estimation of the travel time $TT(t)$ for a randomly chosen time t.** The horizontal axis represents the penetration rate (unit: percent) of the GPS-equipped phones in the traffic. The vertical axis represents VTL density (unit: number of VTL per mile). The total error $TT_{\max}(t) - TT_{\min}(t)$ on the parameter $TT(t)$ is indicated using a grey scale, a darker shade representing a larger error.

- Loop detector data is known to be noisy and biased. Loop detectors report their raw data to the PeMS system every 30 seconds. This raw data is hardly directly useable. PeMS filters this data and averages it on a five minutes time period. Despite the filtering and the data processing, this data still has significant reliability issues which are not specifically addressed in this study. The bounds are guaranteed only when the VTL and loop detector data are exact. Since this is never the case in practice, the guarantee is lost.

- The time sampling of the loop detectors is low. The average five minute update gives a low time resolution for the boundary conditions. Using a confidence interval or a probability distribution for the upstream and downstream boundary conditions improves the robustness of the model by making the bounds depend on this uncertainty [85].

- In the experimental set-up, the location of the first and last loop detectors does not match perfectly the location of the video cameras, as illustrated in Figure 3.3. The travel times computed by the algorithm are linearly interpolated to match the distance between the two video cameras located on the bridges. As the road section considered for the computations (delimited by the loop detectors) represents 88% of the road section delimited by the cameras, the computed travel time is assumed to be equal to 88% of the travel time between the two video cameras. This is obviously an approximation, and may induce additional numerical error.

3.4 Conclusion

This chapter investigates a specific data assimilation technique used to incorporate probe vehicle data into flow models. The study uses a traditional traffic flow model (the Lighthill-

Whitham-Richards theory), formalized using the Moskowitz framework. This framework enables the use of a function, whose isolines correspond to vehicle trajectories. The problem of including loop detector data (flow data) in this framework is done following the traditional kinematic wave approach. The problem of fusing it with probe data is formulated using the Moskowitz function. Based on this formulation, two linear programs are created to compute upper and lower bounds of the estimate of initial numbers of vehicles on the highway segment. These bounds are taking into account the uncertainty in the model, and are used to find bounds on travel time through the corresponding section of highway. The method is implemented on data collected during the *Mobile Century* experiment, using 100 Nokia N95 phones traveling onboard vehicles driving loops on I880 in California. A sampling and penetration rate study shows that the method provides accurate travel time estimates at low penetration rates and reasonable, privacy aware, sampling strategies. The numerical results show the importance of using probe vehicle data to improve travel time estimates without the need for a dense deployment of costly dedicated infrastructure. The results also underline that the data gathered by a few vehicles is sufficient to significantly improve results obtained by sparse loop detectors. It is possible to take into account uncertainty of the data and to relax the non-overtaking assumption to improve the robustness and the reliability of the estimation.

Chapter 4

Analytical derivations of an hydrodynamic model of arterial traffic

An important challenge in arterial traffic estimation is the characterization of travel time distributions. The literature review (Chapter 2) indicates that statistical models representing the dynamics of traffic flows are well-suited towards designing a robust, scalable arterial traffic monitoring system. The integration of statistical models and traffic theory is a very promising approach to capture the variability of arterial traffic and leverage the increasing amount of traffic data collected from probe vehicles. The mathematical abstraction of physical phenomena by traffic flow models is coupled with a statistical model to fully take advantage of the traffic data collected today.

This chapter presents an arterial traffic model based on hydrodynamic theory to represent the physics of traffic flows. The model characterizes the formation and dissolution of horizontal queues upstream of capacity reductions caused by traffic signals. Recall that the main difference between horizontal queues and vertical queues is as follows: vertical queues consider that vehicle stack up vertically, incurring no delay to travel to the point of congestion. On the other hand, horizontal queues model the physical location where vehicles queue, modeling the time necessary to travel to the point of congestion when the queue releases. The model is based on specific assumptions on the physics of traffic flow which make the problem tractable, while keeping it realistic.

In traffic monitoring systems relying on probe data, probe vehicles send periodic location measurements, which provide two sources of indirect information about the arterial traffic link parameters: (i) as the location measurements are taken uniformly over time, more densely populated areas *of the link* will have more location measurements and (ii) the time spent between two consecutive location measurements provides information on the speed at which the vehicle drove through the corresponding arterial link(s). We use these properties and the arterial model presented in the present chapter to develop a statistical model of arterial traffic, which can be trained with probe data and used for inference in arterial networks:

- Chapter 5 derives parametric probability distributions of vehicle locations on arterial links, delimited by signalized intersections which has valuable applications in *Geographic Information Systems* and traffic monitoring. Because of the presence of traffic signals, vehicles spend more time downstream of the links, where they experience delay. The probability distribution of vehicle locations enables the estimation of queue lengths, which is a measure of congestion.

- Chapter 6 derives parametric probability distributions of delays and travel times

between arbitrary points of the network. These distributions are characterized by a small set of parameters with direct physical interpretation (signal timing, queue length). These parameters represent valuable information for traffic management entities.

The present chapter presents the traffic theory results derived from hydrodynamic models and queuing theory which are used in the following chapters to derive probability distributions of vehicle locations on a link and probability distributions of travel times between arbitrary location on a link from horizontal queuing theory. Section 4.1 introduces the macroscopic traffic flow model. Section 4.2 details the specific assumptions on the physics of traffic flow which make the problem tractable, while keeping it realistic. Section 4.3 derives the horizontal queuing theory results. The notation is summarized in Section 4.4 and the potential applications of the horizontal queuing theory are presented in Section 4.5, before being expanded in the following chapters.

4.1 Hydrodynamic theory

In traffic flow theory, it is common to model vehicular flow as a continuum and represent it with macroscopic variables of *flow* $q(x,t)$ (veh/s), *density* $\rho(x,t)$ (veh/m) and *velocity* $v(x,t)$ (m/s). The definition of flow gives the following relation between these three variables [108, 134]:

$$q(x,t) = \rho(x,t)\, v(x,t). \tag{4.1}$$

This property is used throughout the derivations of traffic models.

For low values of density, experimental data shows that the velocity of traffic is relatively insensitive to the density; and all vehicles travel close to the so called free flow velocity of the corresponding road segment v_f. As density increases, there is a critical density ρ_c at which the flow of vehicles reaches the capacity $q_{\max}$ of the road. As the density of vehicles increases beyond ρ_c, the velocity decreases monotonically until it reaches zero at the maximal density $\rho_{\max}$. The maximal density can be thought of as the maximum number of vehicles that can physically fit per unit length, and at this density, vehicles are unable to move without additional space between vehicles. Experimental data indicates a decreasing linear relationship between flow and density, as the density increases beyond ρ_c. The slope of this line is referred to as the congested wave speed, noted w. This leads to the common assumption of a triangular *fundamental diagram* (FD) to model traffic flow dynamics [54, 72].

A triangular fundamental diagram was used in Chapter 3 to estimate guaranteed travel time bounds on highway segments. It is illustrated in Figure 3.1. The triangular fundamental diagram is fully characterized by three parameters: v_f, the free flow speed (m/s); $\rho_{\max}$, the jam (or maximum) density (veh/m); and $q_{\max}$, the capacity (veh/m). Its expression is reminded for convenience and reads:

$$q(\rho) = \begin{cases} v_f \rho & \text{if} \quad \rho \in [0, \rho_c] \\ w(\rho_c - \rho) + v_f \rho_c & \text{if} \quad \rho \in [\rho_c, \rho_{\max}] \end{cases},$$

where $w = \frac{\rho_c v_f}{\rho_{\max} - \rho_c}$. Note that ρ_c represents the boundary density value between (i) free flowing conditions for which cars have the same velocity and do not interact and (ii) saturated conditions for which the density of vehicles forces them to slow down and the flow to decrease. When a queue dissipates, vehicles are released from the queue with the maximum flow—capacity $q_{\max}$—which corresponds to the critical density $\rho_c = q_{\max}/v_f$.

For a given road segment of interest, the arrival rate at time t, i.e. the flow of vehicles entering the link at t, is denoted by $q_a(t)$. Conservation of flow relates it to the arrival density $\rho_a(t) = q_a(t)/v_f$.

4.2 Modeling assumptions

The model is based on the following assumptions on the dynamics of traffic flow:

1. *Triangular fundamental diagram*: this is a standard assumption in transportation engineering, both for highway and arterial traffic modeling and estimation. Chapter 3 also uses a triangular fundamental diagram to successfully derive robust travel time bounds on highway segments. Other arterial models commonly make this assumption [54, 72].

2. *Stationarity of traffic*: during each estimation interval, the parameters of the light cycles (red and cycle time) and the arrival density ρ_a are constant. Moreover,there is not a consistent increase or decrease in the length of the queue, nor instability. With these assumptions, the traffic dynamics are periodic with period C (length of the light cycle). The work is mainly focused on deriving travel time distributions for cases in which measurements are sparse. Constant quantities (for a limited period of time) do not limit the derivations of the model since the aim of this model is to study trends rather than fluctuations.

3. *First In First Out* (FIFO) model: overtaking on the road network is neglected. When traffic is congested, it is generally difficult or impossible to pass other vehicles. In undersaturated conditions, vehicles can choose their own free flow speed, but as done in the LWR model, the derivations are based on the assumption that the free flow speeds are similar enough that the "no overtaking" assumption is a good approximation.

4. *Model for differences in driving behavior*: the free flow pace (inverse of the free flow speed) is not the same for all vehicles: it is modeled as a random variable with vector of parameter θ_p—*e.g.* the free flow pace has a Gaussian or Gamma distribution with parameter vector $\theta_p = (\bar{p}_f, \sigma_p)^T$ where $\bar{p}_f$ and σ_p are respectively the mean and the standard deviation of the random variable.

4.3 Horizontal queuing theory

In arterial networks, traffic is driven by the formation and the dissipation of queues at intersections. The dynamics of queues are characterized by shocks, which are formed at the interface of traffic flows with different densities.

There exists two discrete traffic regimes: *undersaturated* and *congested*, which represent different dynamics of the arterial link depending on the presence (respectively the absence) of a remaining queue when the light switches from green to red. Figure 4.1 illustrates these two regimes under the assumptions made in Section 4.2. The speed of formation and dissolution of the queue are respectively called v_a and w. Their expression is derived from the Rankine-Hugoniot [60] jump conditions and given by

$$v_a = \frac{\rho_a v_f}{\rho_{\max} - \rho_a} \quad \text{and} \quad w = \frac{\rho_c v_f}{\rho_{\max} - \rho_c}. \tag{4.2}$$

Undersaturated regime

In this regime, the queue fully dissipates within the green time. This queue is called the *triangular queue* (from its triangular shape on the space-time diagram of trajectories). It is defined as the spatio-temporal region where vehicles are stopped on the link. Its length is called the maximum queue length, denoted $l_{\max}$, which can also be computed from traffic theory:

$$l_{\max} = R\frac{wv_a}{w - v_a} = R\frac{v_f}{\rho_{\max}}\frac{\rho_c\rho_a}{\rho_c - \rho_a}. \tag{4.3}$$

The duration between the time when the light turns green and the time when the queue fully dissipates is the *clearing time* or *queue dissipation time* denoted τ, derived as follows

$$\tau = l_{\max}\left(\frac{1}{w} + \frac{1}{v_f}\right). \tag{4.4}$$

Replacing the $l_{\max}$ and w by their expressions derived in equations(4.3) and (4.2), the expression of τ reads

$$\tau = R\frac{\rho_a}{\rho_c - \rho_a}. \tag{4.5}$$

Congested regime

In this regime, there exists a part of the queue downstream of the triangular queue called *remaining queue* with length l_r corresponding to vehicles which have to stop multiple times before going through the intersection.

All notations introduced up to here are illustrated for both regimes in Figure 4.1.

Stationarity of the two regimes

Assumption 2 made earlier implies the periodicity of these queue evolutions (see Figure 4.1). As indicated by the slopes of the trajectories in the figure, when vehicles enter the link, they travel at the free flow speed v_f. The distance between two vehicles is the inverse of the arrival density $1/\rho_a$. The time during which vehicles are stopped in the queue is represented by the horizontal line in the queue. The length of this line represents the delay experienced at the corresponding location. The assumption of a triangular fundamental diagram implies that the distance between vehicles stopped in the queue is the inverse of the maximum density $1/\rho_{\max}$. When the queue dissipates, vehicles are released with a speed v_f and a density ρ_c. The trajectory is represented by a line with slope v_f, the distance between two vehicles is $1/\rho_c$.

For the two distinct traffic regimes, the following chapters present the derivations of the *probability distribution function* (pdf) for the location of vehicles on a link and for the travel time along a link (Chapter 5) and of the pdf of delays and travel times between any two arbitrary locations on the network (Chapter 6).

4.4 Notation

The list below summarizes the notation introduced earlier and to be used to fully characterize the arterial traffic model. The parameters are specific for each network link j. The index j is omitted for notational simplicity.

- Model parameters:

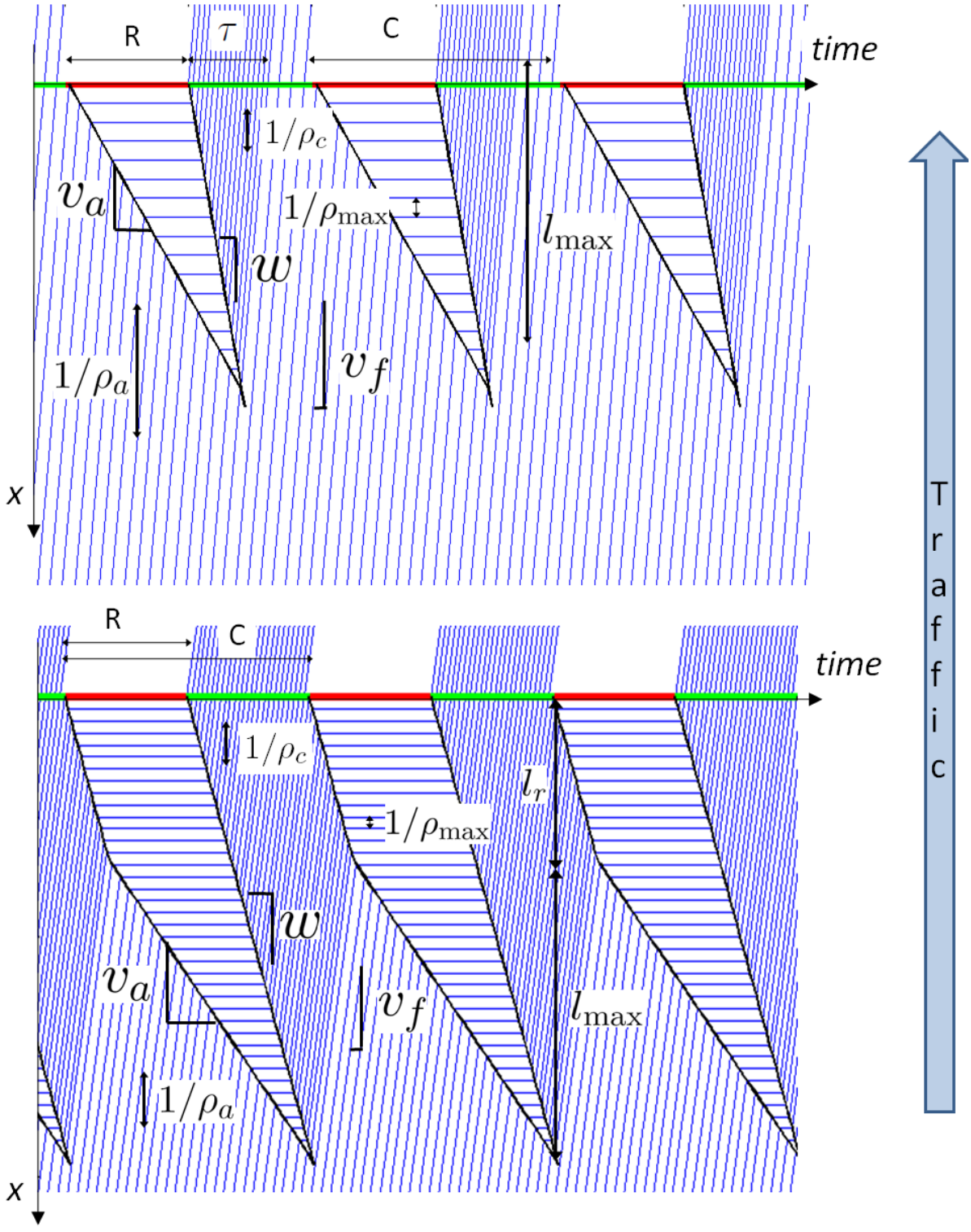


Figure 4.1: Space time diagram of vehicle trajectories with uniform arrivals under an undersaturated traffic regime (top) and a congested traffic regime (bottom).

  - Free flow pace, p_f (seconds/meter), inverse of the free flow velocity v_f. The free flow pace is a random variable. Its *probability distribution function* (p.d.f.) is denoted $\varphi^p(p)$; it models the different driving behavior by assuming a distribution of the free flow pace among the different drivers,
  - Cycle time, C (seconds),
  - Red time, R (seconds),
  - Length of the link, L (meters).

- Traffic state variables:
  - Clearing time τ,
  - Triangular queue length $l_{\max}$,
  - Remaining queue length l_r.

This set of variables is sufficient to characterize the model and the time evolution of the state of traffic. The location x on a link corresponds to the distance from the location to the downstream intersection. From these variables, it is possible to compute the other traffic variables, including velocity, flow, and density of vehicles at any x and time t and queue length. The remaining queue length l_r is specific to the congested regime ($l_r = 0$ in the undersaturated regime). Similarly, the existence of a clearing time is specific to the undersaturated regime (the clearing time is null and thus $\tau = C - R$ in the congested regime). The undersaturated and congested regimes are labeled u and c respectively.

4.5 Applications of the arterial traffic model

This chapter reviewed an analytical model of arterial traffic dynamics on a link delimited by signalized intersections. The model builds upon traditional first order traffic flow model (the LWR model) and makes specific assumptions to maintain the tractability of the model while keeping it realistic. The goal of these derivations is to characterize the formation and dissolution of queues upstream of traffic signals. Characterizing the formation and dissolution of queues is key to enable estimation algorithms specific to arterial traffic. Indeed the periodic formation and dissolution of queues is the main specificity of arterial traffic when compared to highway traffic. It is important to accurately capture this main feature of arterial traffic as it is the main source of variability in arterial traffic, both temporally and spatially.

The importance of the variability of traffic variables, both in space and in time motivates the derivation of statistical models of arterial traffic dynamics. In particular, there is specific interest in developing probability distribution of vehicle location on a link and probability distribution of travel times between arbitrary locations on the network. The probability distribution of vehicle location represent the spatial variability of traffic within an arterial link, delimited by signalized intersections. The probability distribution of travel times illustrate the differences in travel time experienced by vehicles traveling during the same time interval but which enter the network at different time instants.

The following chapters present the derivation of these probability distributions. The probability distribution of vehicle locations is denoted $f^s(x)$, $s \in \{u, c\}$. The variable x indicates the distance *to* the intersection, so the location of the intersection is at $x = 0$ and the start of the link is at $x = L$. The function f^s encodes the probability of a vehicle to be at location x, which depends on x because of the spatial heterogeneity of the density, due

to the formation of queues at intersections, as can be seen in Figure 4.1. The probability distributions for the delay δ_{x_1,x_2} and travel time y_{x_1,x_2} between two locations x_1 and x_2 on a link of the network are denoted respectively $h(\delta_{x_1,x_2})$ and $g(y_{x_1,x_2})$. The derivations are based on the stationarity assumption, which allows for the definition of temporal averages of the traffic variables. These averages are then taken over a light cycle C. For example, chapter 5 defines the variable Z as the average number of vehicles present on a link, with index u (resp. c) for the undersaturated (resp. the congested) regime.

Chapter 5

Spatial distribution of vehicles and applications

Leveraging horizontal queuing theory, this chapter presents the mathematical characterization of the probability distribution of vehicle location on a link. The distributions are mathematical abstractions of the following fact: the (temporal) average density of vehicles within a link increases with the average delay experienced at this location. The arterial model of Chapter 4 provides the basis for the derivation of parametric probability distributions of vehicle location. As illustrated in the data exploration, the parameters of these distributions can be learned with probe vehicle data: probes which send periodic location measurements are more likely to report their position where they experience delay. The parameters of the probability distributions can be learned from the probe measurements collected over large periods of time. There are interesting applications of this work for arterial traffic:

- It provides a convenient way to *scale* partial travel times, *i.e.* to infer the travel time on an arterial link from the travel time on a portion of this link. This method is more accurate than scaling with respect to the distance traveled because of the important differences in average speeds within an arterial link caused by the presence of signals.
- The distribution of vehicle location can be used to *automatically detect the presence of traffic signals* (stop signs or traffic lights), which is a valuable application for the creation and update of accurate Geographic Information Systems.
- The training of the model and estimation of the parameters provide an estimation of the queue length which is an *indicator of congestion levels*.

The present chapter is organized as follows. Section 5.1 presents the derivations of the statistical model representing the spatial statistical distribution of vehicles. The derivation relies on the computation of the average vehicle density over a traffic cycle based on the arterial traffic model of Chapter 4. Section 5.2 illustrates the potential of the model to scale partial link travel times and validates it with probe data collected by the *Mobile Millennium* system. Besides the interest for arterial traffic estimation, the analysis of the results motivate the application of the model for automatic signal detection. The challenges of the creation of accurate Geographic Information Systems are presented in Section 5.3 as well as the potentials of probe data to make this process automatic and more reliable. The automatic detection algorithm of traffic signals is presented in Section 5.4 and the potential of the algorithm are analyzed in Section 5.5.

5.1 Modeling the spatial distribution of vehicles on an arterial link

This Section builds upon the arterial traffic model presented in Chapter 4 to derive parametric probability distribution for the location of vehicles on an arterial link.

5.1.1 Statistical modeling

Using the stationarity assumption, the density at location x is time periodic with period C. Define the average density $d(x)$ at location x as the temporal average of the density $\rho(x,t)$ at location x and time t:

$$d(x) = \frac{1}{C}\int_0^C \rho(x,t)\,dt.$$

In practice, flow is never perfectly periodic of period C (even in stationary conditions), but the model assumes that the above averaging over a duration C is a good proxy of a longer average.

According to the model assumptions, the density at location x and time t takes one of the three following values, numbered 1 to 3 for convenience: (1) $\rho_1 = \rho_{\max}$, when vehicles are stopped, (2) $\rho_2 = \rho_c$ when vehicles are dissipating from a queue, (3) $\rho_3 = \rho_a$ when vehicles arrive at the link and have not stopped in the queue.

The average density at location x is thus

$$d(x) = \sum_{i=1}^{3} \alpha_i(x)\rho_i$$

where $\alpha_i(x)$ represents the fraction of time that the density is equal to ρ_i at location x.

The probability distribution $f(x)$ of vehicle location at location x is obtained from the average density $d(x)$ at location x by normalization. The normalization constant Z is given by $Z = \int_0^L d(x)\,dx$ and the expression of the distribution f reads

$$f(x) = \frac{d(x)}{\int_0^L d(x)\,dx}.$$

In the undersaturated and the congested regime, the computation of the $\alpha_i(\cdot)$, $i = 1\ldots3$ enables the derivation of the probability distribution of vehicle locations. The following section presents how the analytical expressions of $\alpha_i(\cdot)$, $i = 1\ldots3$ are derived from the arterial traffic model presented in Chapter 4.

5.1.2 Undersaturated regime

Upstream of the maximum queue length, the density remains constant at ρ_a throughout the entire light cycle. Downstream of the maximum queue length, the value of the density varies over time during the light cycle and takes one of the three density values ρ_1, ρ_2 or ρ_3.

Using the assumption that the fundamental diagram is triangular and that the arrival density is constant, the average density increases linearly from ρ_a to the value it takes at the intersection, where $x = 0$. At the intersection, the density is $\rho_{\max}$ during the red time R. The density is ρ_c when the queue dissipates, *i.e.* during the clearing time $\tau = l_{\max}(\frac{1}{w} + \frac{1}{v_f})$. Replacing w and $l_{\max}$ by their expressions, the time during which the queue dissipates is $R\frac{\rho_a}{\rho_c - \rho_a}$. The rest of the cycle has a duration $C - R\frac{\rho_c}{\rho_c - \rho_a}$ and it has density ρ_a. The average density at the intersection is the sum of the arrival, maximum and

critical densities, weighted by the fraction of the cycle during which each of the density is experienced. The average density at the intersection is:

$$\begin{aligned} d(0) &= \frac{1}{C}\left(\underbrace{R\rho_{\max}}_{\substack{\text{Red time } R \\ \text{at density } \rho_{\max}}} + \underbrace{R\frac{\rho_a}{\rho_c-\rho_a}\rho_c}_{\substack{\text{Clearing time } \tau \\ \text{at density } \rho_c}} + \underbrace{\left(C-\left(R+R\frac{\rho_a}{\rho_c-\rho_a}\right)\right)\rho_a}_{\substack{\text{Extra green-time } C-(R+\tau) \\ \text{at density } \rho_a}} \right) \\ &= \frac{R}{C}\rho_{\max}+\rho_a \end{aligned}$$

Given that the density grows linearly between the end of the queue and the intersection, the density at location x is given by

$$\begin{aligned} d(x) &= \rho_a & \text{if } x \geq l_{\max} \\ d(x) &= \rho_a + \tfrac{R}{C}\rho_{\max}\tfrac{l_{\max}-x}{l_{\max}} & \text{if } x \leq l_{\max}, \end{aligned}$$

which can be summarized as

$$d(x) = \rho_a + \frac{R}{C}\rho_{\max}\frac{\max(l_{\max}-x,0)}{l_{\max}}.$$

The normalization constant Z_u is defined by $Z_u = \int_0^L d(x)\,dx$ and represents the temporal average of the number of vehicles on the link. Its explicit value is given by $Z_u = L\rho_a + \frac{l_{\max}}{2}\frac{R}{C}\rho_{\max}$. The normalized density of vehicles as a function of the position on the link, defined by $f^u(x) = d(x)/Z_u$ is thus equal to

$$\begin{aligned} f^u(x) &= \tfrac{\rho_a}{Z_u} & \text{if } x \geq l_{\max} \\ f^u(x) &= \tfrac{\rho_a}{Z_u} + \tfrac{R}{C}\rho_{\max}\tfrac{l_{\max}-x}{l_{\max}Z_u} & \text{if } x \leq l_{\max} \end{aligned}.$$

When vehicles report their location arbitrarily in time, this function represents the probability of receiving a measurement at location x.

5.1.3 Congested regime

In the congested regime, the average density is constant upstream of the maximum queue length—equal to ρ_a—and increases linearly until the remaining queue. In the remaining queue, it is constant and equal to $\frac{R}{C}\rho_{\max}+(1-\frac{R}{C})\rho_c$. The different spatio-temporal domains of the density regions are illustrated Figure 5.1 (left). The probability distribution of vehicle locations is:

$$\begin{aligned} f^c(x) &= \tfrac{\rho_a}{Z_c} & \text{if } x \geq l_{\max}+l_r \\ f^c(x) &= \tfrac{\rho_a}{Z_c} + \left(\tfrac{R}{C}\rho_{\max} + \left(1-\tfrac{R}{C}\right)\rho_c - \rho_a\right)\tfrac{-x+l_{\max}+l_r}{l_{\max}Z_c} & \text{if } x \in [l_r,\, l_{\max}+l_r] \\ f^c(x) &= \tfrac{R}{C}\tfrac{\rho_{\max}}{Z_c} + \left(1-\tfrac{R}{C}\right)\tfrac{\rho_c}{Z_c} & \text{if } x \leq l_r \end{aligned}. \quad (5.1)$$

where Z_c is the normalizing constant that ensures that the integral of the function on $[0, L]$ equals 1. We have

$$Z_c = L\rho_a + \left(\frac{l_{\max}}{2}+l_r\right)\left(\frac{R}{C}\rho_{\max} + \left(1-\frac{R}{C}\right)\rho_c - \rho_a\right).$$

Remark 5.1 (General case)**.** *The undersaturated regime is a special case of the congested regime, in which the remaining queue length l_r is equal to zero. In the remainder of this report, we consider the congested regime as the general case for the spatial distribution of vehicle location.*

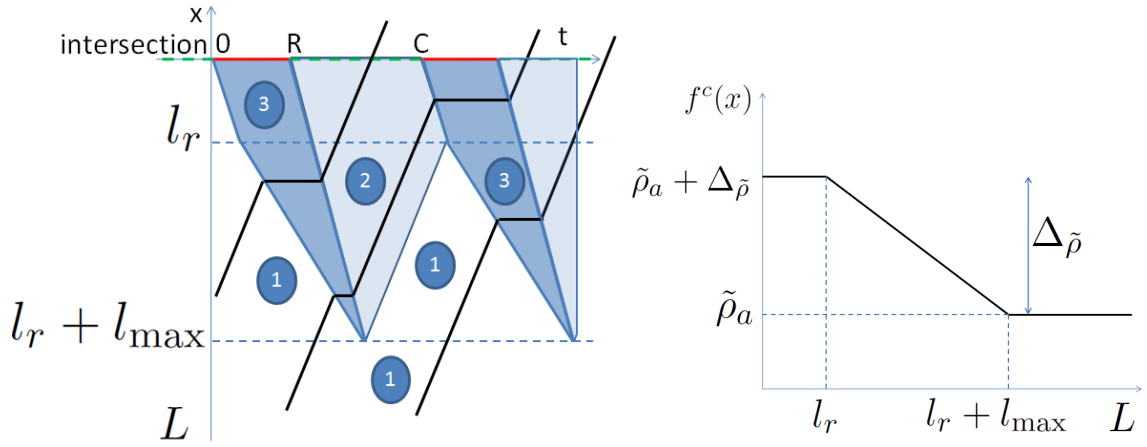


Figure 5.1: **Left**: The estimation of vehicle spatial distribution on a link is derived from the queue dynamics of the traffic flow model. The space-time plane represents the space-time domain in which density of vehicles is constant. Domain 1 represents the arrival density ρ_a, domain 2 represents the critical density ρ_c and domain 3 represents the maximum density $\rho_{\max}$. **Right**: Using the stationarity assumption, we compute the average density at location x and normalize to derive the probability distribution of vehicle locations on the link.

The probability distribution of vehicle location is fully determined by three independent parameters. The following parameters are used to non-ambiguously specify the probability distribution f^c: the remaining queue length l_r, the triangular queue length $l_{\max}$ and the normalized arrival density $\tilde{\rho}_a = \rho_a / Z_c$. Using this parameterization, the probability distribution of vehicle location is illustrated in Figure 5.1 (right) and reads:

$$\begin{array}{lcll} f^c(x) & = & \tilde{\rho}_a & \text{if } x \geq l_{\max} + l_r \\ f^c(x) & = & \tilde{\rho}_a + \frac{(l_r + l_{\max}) - x}{l_{\max}} \Delta_{\tilde{\rho}} & \text{if } x \in [l_r,\, l_{\max} + l_r] \\ f^c(x) & = & \tilde{\rho}_a + \Delta_{\tilde{\rho}} & \text{if } x \leq l_r \end{array} ,$$

$$\text{with } \Delta_{\tilde{\rho}} = \frac{1 - \tilde{\rho}_a L}{l_{\max}/2 + l_r}. \tag{5.2}$$

The expression of $\Delta_{\tilde{\rho}}$ above, can be obtained easily by noticing that $\int_0^L f^c(x)\, dx$ must be equal to 1 as f^c is a probability distribution function. The expression of $\Delta_{\tilde{\rho}}$ can also be obtained by direct computation from Equation (5.1), by replacing Z_c by its expression (Equation (5.2)) and ρ_a by $Z_c \tilde{\rho}_a$. The explicit dependency of f^c on the parameters l_r, $l_{\max}$ and $\tilde{\rho}_a$ is omitted for the sake of notational simplicity, as long as it does not impair the understandability and readability of the derivations. When the explicit dependency is needed, $f^c(x; \tilde{\rho}_a, l_{max}, l_r)$ denotes the density of probability of receiving a measurement at location x for the specific value of the parameters.

5.1.4 Learning the parameters of the model

The parameters $\tilde{\rho}_a$, l_{max} and l_r of the distribution of vehicle location on a link are learned by maximizing the likelihood of the set of location observations. The location observations are position reports communicated by probe vehicles and sampled uniformly in time. The set of observations is denoted $(x_o)_{o \in O}$ and may be collected over large periods of time (historical data) for specific Time Of the Day (TOD) and Day of the Week (DOW). Because the logarithm is an increasing function, it is equivalent to maximize the logarithm of the likelihood (referred to as *log-likelihood*). For numerical reasons, it is commonly

preferred to maximize the log-likelihood rather than the likelihood. The optimization problem is given by

$$\begin{aligned} \underset{\tilde{\rho}_a, l_r, l_{\max}}{\textbf{maximize}} & \sum_{o \in O} \ln(f^c(x_o)) \\ \textbf{s.t.} \quad & 0 \leq \tilde{\rho}_a \leq \tfrac{1}{L}, 0 \leq l_r,\ 0 \leq l_{\max}, \\ & l_r + l_{\max} \leq L \end{aligned} \tag{5.3}$$

The constraints come from the physics of the problem. The first constraint illustrates the fact that the arrival density is inferior to the average density on the link. The other constraints illustrate that the total queue cannot extend beyond the length of the link and that the triangular queue and the remaining queue must be non-negative. Note that the constraints on the queue lengths do not limit the generality of the model. Under spill-over conditions (queue length extending beyond the upstream intersection), we consider that the queue length extends up to the upstream intersection, the rest of the queue is accounted for in the upstream links. Bounds on the parameters can be added based on physical a priori knowledge about the parameters to limit the feasible set and ensure the quality of the results when little data is available.

Remark 5.2. *The objective function is not concave in the optimization variables. However, the search space is limited (three bounded parameters) and we perform a grid search followed by a local gradient ascent for the B best solutions of the grid-search. The definition of the grid and the value of B are left to the user. We found that a very fine grid was not necessary to provide good results, leading to efficient computation times.*

5.2 Scaling of partial link travel times

Since probe vehicles can send their positions at any location on the network, the path between successive location reports can start and end at any location and the first and last links of the corresponding path are not fully traversed by the vehicle. They are called *partial* links of the path. Besides the probability distribution of travel times on each link of the network, the model has to define the pdf of travel times on partial links, *i.e* the pdf of travel times on link i between offsets x_1 and x_2 (where x_m, $m = 1, 2$ represents the distance to the downstream intersection). Let $Y^i_{x_1,x_2}$ define the random variable representing the travel time on *partial* link i between offsets x_1 and x_2 ($x_1 \geq x_2$). Following this notation, $Y^i_{L^i,0}$ represents the travel time on link i (between offsets L^i, length of link i, and 0). Numerous traffic estimation algorithms do not rely on an arterial flow model. When using these models to estimate arterial traffic with sparsely sampled probe vehicles, the estimation algorithm requires a specific module to take into account *partial* link travel times. There are two main solutions to integrate partial link travel times in such models:

- Refine the spatial discretization, by dividing each link into cells. This approach will increase the adequation of the location reports with the beginning and end of the cells of the network. However, it increases significantly the number of parameters which need to be estimated leading to risks of overfitting and unreliable estimation and prediction capabilities.

- Define a *scaling* function. The scaling function estimates the duration of travel on an entire link based on the travel time on a partial link. More specifically, a scaling function may be defined as a function α^i such that the travel time $Y^i_{x_1,x_2}$ on a portion of link i, delimited by x_1 and x_2 is related to the travel time $Y^i_{L^i,0}$ on link i by $Y^i_{x_1,x_2} = \alpha^i(x_1, x_2)Y^i_{L^i,0}$. This approach includes the possibility to refine the spatial discretization but provides more flexibility. The choice of the function α^i

determines the complexity of the model and can be chosen appropriately depending on the amount and the quality of available data.

This section focuses on the derivation of such a function α^i adapted to the amount of probe vehicle data available today. The function α^i is defined on the domain $\mathcal{D}_{\alpha^i} \subset [0, L^i]^2$. The domain $\mathcal{D}_{\alpha^i}$ is defined as follows:

$$\mathcal{D}_{\alpha^i} = \{(x_1, x_2) \in [0, L^i]^2 \ : x_1 \geq x_2\}.$$

5.2.1 Modeling the scaling function

For the consistency of the model, the function α^i must satisfy the following conditions:

- The travel time on a partial link is a fraction of the link travel time: $\forall (x_1, x_2) \in D_{\alpha^i}$, $\alpha^i(x_1, x_2) \in [0, 1]$. In particular, if the partial link spans the entire link, the partial travel time has the same distribution as the link travel time *i.e.* $\alpha^i(L^i, 0) = 1$,
- If a partial link is included in another partial link, its travel time should be smaller: $\forall x_1$, the function $x_2 \mapsto \alpha^i(x_1, x_2)$ is a decreasing function of x_2 and $\forall x_2$, the function $x_1 \mapsto \alpha^i(x_1, x_2)$ is an increasing function of x_1.
- The probability for a vehicle to experience delay increases as the location gets closer to the downstream intersection. For the same distance traveled, travel times are on average longer close to the downstream intersection because of the presence of traffic signals. To model this intuitive fact, another condition is imposed on α^i: $\forall x_1$, the function $x_2 \mapsto \alpha^i(x_1, x_2)$ is a convex function of x_2. Similarly, $\forall x_2$, the function $x_1 \mapsto \alpha^i(x_1, x_2)$ is a concave function of x_1.

For example, the function defined by $\alpha^i(x_1, x_2) = (x_1 - x_2)/L^i$ would satisfy these conditions. However it assumes that the travel time on a partial link is proportional to the distance traveled on the link, but does not take into account the presence of traffic signals. The idea developed here is to leverage the hydrodynamic model of traffic flow to derive a parametric model for the function α^i. The parameters of the functions α^i can be learned from the sparse measurements of probe vehicle locations. More specifically, define α^i as the *cumulative distribution function* (cdf) of a specific random variable. For a probe vehicle sampled uniformly in time and reporting its position while traveling on link i, the random variable represents the position of the vehicle on the link as it reports its location. Its probability density function is precisely the probability distribution of vehicle location derived in Section 5.1 from the hydrodynamic model of traffic flow. Its parameters are learned from past measurements of probe vehicle locations.Because of the presence of traffic signals, the distribution is a decreasing function of the distance to the downstream intersection (increasing function of the distance from the upstream intersection). The function α^i is chosen as follows:

$$\alpha^i(x_1, x_2) = \int_{x_2}^{x_1} f^c(x)\, dx,$$

which satisfies all the above requirements.

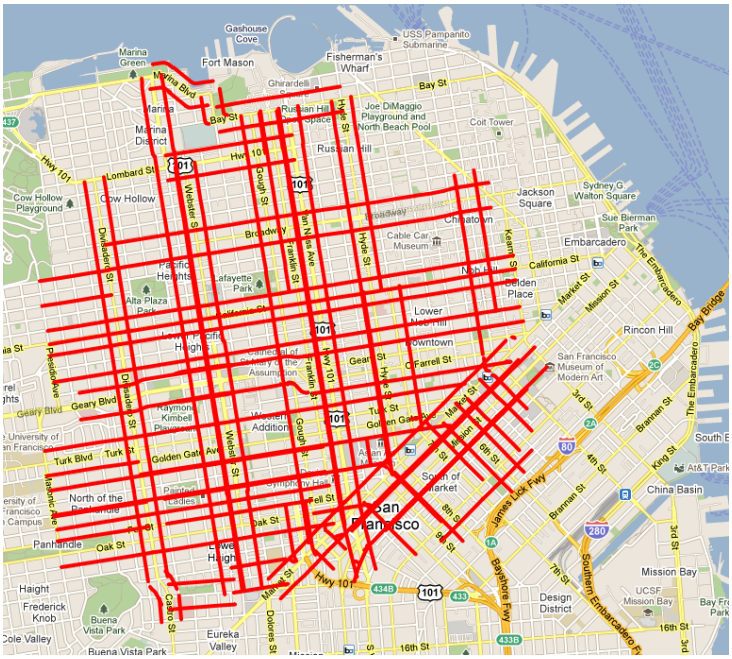


Figure 5.2: The subnetwork of San Francisco, CA used for the validation of the travel time scaling.

Table 5.1: Outcome of statistical tests.

		True hypothesis	
		H_0 is true	H_0 is false
Decision	Accepts H_0	Right decision	Wrong decision, Type II error, rate β
	Rejects H_0	Wrong decision, Type I error, rate α	Right decision

5.2.2 Numerical validation of the travel time scaling capabilities

Experimental setup

The numerical validation is based on data collected by the *Mobile Millennium* system from a fleet of 500 taxis in San Francisco, CA. The vehicles report their location every minute. The study focuses on a subnetwork of San Francisco (Figure 5.2) with 815 links and 527 intersections (more than 12.6 km. of roadway). For the application of interest, a *historical time interval* is a tuple consisting of a day of the week, a start time, and an end time. The locations reported by the vehicles on each link of the network are aggregated for each historical time interval. The parameters of the density model are learned from these observations. The numerical results analyze the data on historical intervals representing Tuesdays from 4 to 8pm with 15 minutes duration each (*i.e.* (Tuesday, 4pm, 4:15pm), ..., (Tuesday, 7:45pm, 8pm)).

Description of the statistical test of the model

For each link of the network and each historical interval, the *Kolmogorov-Smirnov* (K-S) statistics tests whether the locations of the probe vehicles are distributed according to the density model [114] or not. The K-S statistic is computed as the maximum difference between the empirical and the hypothetical cdf (density model). In contrast to other tests (*e.g.* T-test that tests uniquely the mean, or chi-squared test that assumes that the data is normally distributed), the K-S test is a standard *non-parametric* test to

Table 5.2: Percentage of positive K-S tests for different values of α and the two hypothesis (density model or uniform distribution).

Distribution of vehicles	α			Mean p-value
	0.1	0.05	0.01	
Density model	0.75	0.80	0.89	0.35
Uniform	0.46	0.55	0.67	0.15

state whether samples are distributed according to a hypothetical distribution. The K-S statistics is the basis for a test which decides to accept (or to reject) the null hypothesis. The null hypothesis H_0 is stated as follows, "H_0*: The measurements of probe vehicles are distributed according to the density model*". When performing a statistical test, four situations described in Table 5.1 arise. The performance of a statistical test is defined by its *statistical significance* (1 - α, where α is the probability to reject H_0 when it is actually true) and *statistical power* ($1 - \beta$, where β is the probability to accept H_0 when it is actually false). The p-value is used to decide to accept or reject the null hypothesis H_0. Low p-values indicate that the data does not follow the proposed distribution. The hypothesis H_0 is rejected at the α significance level if the p-value is smaller than α and accept it otherwise. The parameter α is commonly set to values ranging from 0.001 to 0.1, and often conventionally set to $\alpha = 0.05$. Smaller levels of α increase confidence in the determination of significance, but increase the risk of Type II errors, and so have less *statistical power*. The K-S test has a probability of Type II error β that tends to zero as the number of samples tends to infinity. Since the number of samples is finite, the *power* of the statistical test is maintained by (i) not testing links that do not have enough measurements, (ii) experimenting with different levels of significance, and (iii) reporting the p-value for each decision.

The goal of the numerical analysis is to validate the capability of the density model to properly scale travel time on portions of arterial links. In particular, it is interesting to compare the results of the density model to a simpler scaling function consisting in scaling travel times proportionally to the distance traveled. Such a scaling function arises from the assumption that vehicles are uniformly distributed on the length of the link. The expectation is that the scaling with the function α^i will perform better than a scaling proportional to the distance traveled because the density model takes into account the fact that vehicles are more likely to experience delay close to the downstream intersection. To illustrate this reasoning, the outcome of the K-S test for the density model is compared to the outcome of a K-S test for which the null hypothesis would consider that vehicles are uniformly distributed on the link. The comparison of the test outcomes on both hypotheses is presented in Table 5.2.

The results indicate that for a majority of arterial links, the average location of vehicles is a random variable that follows the density model. The spatial distribution of vehicle location is better represented by the density model than by a uniform distribution. A graphical representation of the data for the different links of the network provides valuable qualitative information:

- Cumulative locations reported by the probe vehicles. For each link of the network, the cumulative distribution of measurements is displayed as follows

    1. sort the locations reported by the probe vehicles in ascending order,
    2. plot the points $(x_i, i/N)$ for $i = 1 \dots N$, where N is the number of locations collected for the link and historical interval and x_i is the i^{th} location on the

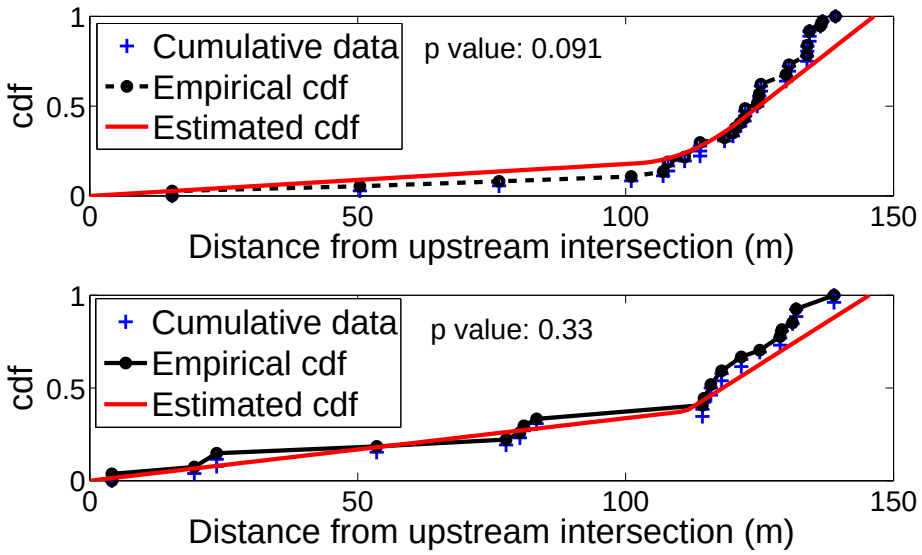


Figure 5.3: Empirical and proposal cdf of the average vehicle locations. **(Top):** Link with p-value equal to 0.09. The model predicts a sharp increase in the density of measurements towards the downstream extremity of the link but no measurements are received in the last 15 meters of the link and the hypothesis H_0 is rejected for $\alpha = 0.1$. Note that the digital map does not model the width of the road or the intersection, which might be the reason for the absence of measurements on the last 15 meters. **(Bottom):** Link with p-value equal to 0.33. The model learns the characteristics of the distribution of vehicle locations. In particular, it estimates the historical queue length (around 30 meters) which provides information on the average congestion of the link.

link (in meters from the upstream intersection).

- Empirical (Kaplan-Meier) cumulative distribution function [99] of the measurements
- Proposed distribution: the distribution of vehicle locations derived from the hydrodynamic model of arterial traffic and calibrated with the training data.

Figure 5.3 illustrates the cdfs obtained for two links of the network during the first historical interval. The first link shows a good qualitative fit. However, the p-value is only 0.091. The relatively poor result may be due to the map discretization: it does not take into account the width of intersections. For this reason, the length of the link stored in the database represents the length of the link between the middle of the upstream and downstream intersections, which is larger than the length of the link up to the signal. The second link has an average p-value and shows a good fit to the model. In both cases, the data follows the sharp increase in the density of measurements close to the downstream intersection, as predicted by the model because of the presence of a queue caused by the traffic signal. The model also provides an estimate of the historical queue length on each link of the network which can be used for planning and network congestion analysis.

Additional application for traffic signal detection

The thorough investigation of the results on the links with very poor p-values helps to understand the limitations of the model and provides valuable insight for additional applications. Figure 5.4 presents the result for a link with p-value equal to 6.8×10^{-4}. The model predicts sharp increases in the density of measurements to occur upstream of traffic signals. This work relies on a NAVTEQ digital map database. The database contains attributes of the transportation network, such as road characteristics, presence of traffic

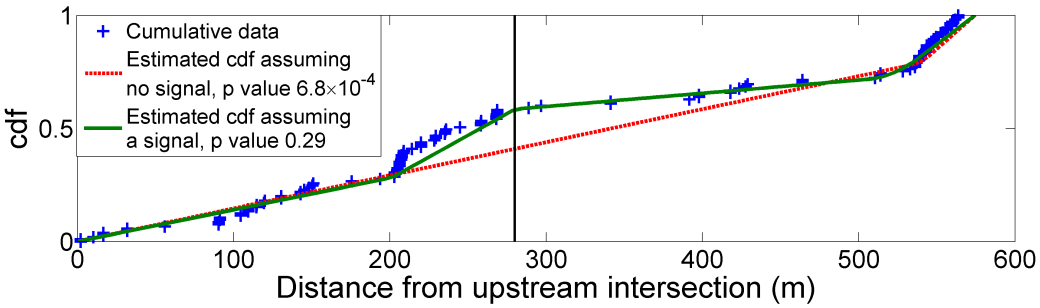


Figure 5.4: Detecting signal locations using the average spatial distribution of vehicles. The figure illustrates an example of very low p-value for a link of the network. Analyzing the results with Google Street View confirmed the intuiting that a signal was missing in the database, explaining the poor fit of the model.

lights, and so on. On this link, the cumulative distribution of vehicle location exhibits two important increases, whereas only one signal was present in the map database.

The analysis of the location of the link in Google Street View confirms that there was a signal which was not in the database. With the corrected information and updated proposal distribution, the p-value is equal to 0.29. This realization motivates the use of probe vehicle data and of the model of probability distributions of vehicle location to automatically detect the presence of traffic signals. Such a capability is very valuable for digital map companies as it has the potential to speed up and reduce the costs of map creation and updates while improving their reliability and accuracy. Another potential interest is for the development of autonomous driving or assisted driving services. Other sources of poor fitting of the model are related to specific behaviors of the taxi, such as waiting in front of major hotels, which can be filtered, when considering successive locations of the vehicle.

5.3 Creating large scale accurate Geographic Information Systems

Geographic Information Systems (GIS) are designed to capture, store and present all types of geographically referenced data. In the context of Intelligent Transportation Systems, accurate and reliable GIS (such as NAVTEQ or OpenStreetMap) are needed for a large number of applications. For example, an accurate description of the geometry and the features of the road network is necessary to develop real-time traffic information systems [25, 95], routing [139] and driver assistance technologies. The collection and processing of data for GIS is an expensive and time consuming process, making automated techniques desirable. Besides the cost and time inefficiencies, the developments and updates of digital maps are traditionally based on surveying methods and digitizing of satellite images which lead to inaccuracies and systematic errors. The emergence of new technologies opens new possibilities to increase the efficiency in developing and maintaining reliable GIS, in particular in developing countries where the infrastructure is evolving rapidly.

The use of GIS for Intelligent Transportation Systems applications has attracted a lot of interest in the computer vision and robotics community to improve driver assistance technologies. For example, the use of real-time, video processing, detection algorithms combined with accurate GIS significantly improves the capabilities to infer speed limitations [130]. Similarly, object recognition algorithms based on image processing allow for

the real-time detection of traffic signals by intelligent vehicles [56]. These algorithms can update existing map databases as intelligent vehicles travel the network. The distributed nature of the problem makes crowdsourcing approaches appealing as they leverage information collected by a large number of vehicles traveling on the road network.

GPS traces have attracted the interest of the machine learning community to lower the costs of producing and updating digital maps while improving their accuracy. In particular, GPS traces have been used to learn the map network geometry using clustering and graph inference algorithms [36]. In [136], data-mining approaches are used to process GPS traces and refine existing digital maps to enable safety applications, such as lane-keeping, and convenience applications, such as lane-changing advice. However, GPS traces requiring high frequency sampling remain scarcely available, mainly because of privacy concerns, communication costs and limitation of the battery life of portable GPS devices. Sparsely sampled probe data is the main source of geo-location data with the prospect of global coverage in the near future. For these reasons, the present section focuses on the use of sparsely sampled data for digital map learning.

The potential of sparsely sampled probe vehicle data for transportation engineering applications has been demonstrated through the successful implementations of reliable real-time traffic information systems on both the highway and the arterial networks [25, 95]. The following sections investigates the use of sparsely sampled probe data collected by the *Mobile Millennium* system [25] to improve and update existing digital map databases. More specifically, the chapter studies how to use this data, in addition to the road network geometry, to automatically detect the presence of traffic signals (traffic lights or stop signs) at each intersection.

The algorithm relies on the derivations of Section 5.1 to develop a model which characterizes the probability distribution of the location of a vehicle on an arterial road segment. It is motivated by the numerical results of Section 5.2 to learn the parameters of the probability distribution of the location of vehicles and present an unsupervised classification algorithm which identifies whether there is a signal at the upstream end of each road segment in the network. In this chapter, a road segment (link) is defined as the stretch of road between consecutive intersections.

The following section (Section 5.4) describes the unsupervised classification algorithm based on model selection information criterion. Section 5.5 analyzes the signal detection potentials of the algorithm using data collected by the *Mobile Millennium* system in an arterial network of San Francisco, CA of over 1,000 links.

5.4 Automatic signal detection

The present section presents two algorithms based on the statistical model derived in Section 5.1 to automatically identify the presence (resp. absence) of traffic signals at intersections with model selection criteria. The distribution of vehicle locations is fully determined by three independent parameters: the remaining queue length l_r, the triangular queue length $l_{\max}$ and the normalized arrival density $\tilde{\rho}_a = \rho_a / Z$. The dependency on these parameters is underlined by denoting $\phi_L(x; \tilde{\rho}_a, l_{max}, l_r)$ the probability density of measurement at location x for the specific value of the parameters. These parameters are learned for each link of the network independently.

5.4.1 Detection based on a single link

The algorithm independently classifies each link of the network as having a traffic signal (stop sign or traffic light) at the downstream intersection or not. Under the hypothesis that there is a traffic signal, the parameters of the probability distribution are the optimizers of Equation (5.3), computed from location measurements x_o sent by sparsely sampled probe vehicles. For link i with length L_i and learned parameters $\tilde{\rho}_a^i, l_{max}^i, l_r^i$, the learned distribution of measurements is denoted

$$\varphi_i^{\text{sig}}(x) = f^c(x; \tilde{\rho}_a^i, l_{max}^i, l_r^i).$$

Under the hypothesis that there is no traffic signal, the distribution of measurements $\varphi_i^{\text{no sig}}$ is expected to be uniform on the link:

$$\varphi_i^{\text{no sig}}(x) = \frac{1}{L_i}\mathbf{1}_{[0,L_i]}(x),$$

where $\mathbf{1}_{[0,L_i]}$ is the indicator function of interval $[0, L_i]$.

5.4.2 Detection based on two consecutive links

Another possible approach to the signal detection is the following. Consider two consecutive links i and j (with link i upstream of link j), connected by an intersection. The decision algorithm detects the presence (or absence) of a signal at the downstream intersection of link i. According to the density modeling of Section 5.1, the decision is based on the results of two models trained on the location measurements: the first one assumes that there is a signal at the end of link i, the second one assumes that there is no signal at the end of link i.

Under the assumption that there is a signal at the downstream intersection of link i, the parameters of the distribution of vehicle location on link i and on link j are learned independently, using the measurements received on each link respectively. These models are combined to represent the distribution of measurements on the stretch of road representing link i and j. For this purpose, x denotes the distance from the upstream intersection of link i on the stretch of road (i, j). The distribution of measurements is denoted $\psi_{i,j}^{\text{sig}}(x)$ and is given by

$$\begin{aligned}\psi_{i,j}^{\text{sig}}(x) = {} & \alpha^{i,j} f^c(x; \tilde{\rho}_a^i, l_{max}^i, l_r^i)\mathbf{1}_{[0,L_i]}(x) \\ & + (1-\alpha^{i,j}) f^c(x - L_i; \tilde{\rho}_a^j, l_{max}^j, l_r^j)\mathbf{1}_{[L_i, L_i+L_j]}(x)\end{aligned}$$

The maximization of the likelihood with respect to $\alpha^{i,j}$ leads to a closed form formula: the number of measurements received on link i over the number of measurements received on links i and j combined.

In the United States, links which do not have a signal at the downstream intersection have the priority at the downstream intersection. If traffic conditions are not congested, this limits the possibilities for delays at the downstream intersection to left turn movements. Under the assumption that there is no signal at the downstream intersection of link i, links i and j are aggregated as a single link of length $L_i + L_j$. The parameters of the distribution of vehicle on this aggregated link are learned using the measurements received on both link i and link j. The parameters are denoted $\tilde{\rho}_a^{i,j}$, $l_{max}^{i,j}$ and $l_r^{i,j}$ and the corresponding distribution of vehicle location is denoted $\psi_{i,j}^{\text{no sig}}(x)$ and given by

$$\psi_{i,j}^{\text{no sig}}(x) = f^c(x; \tilde{\rho}_a^{i,j}, l_{max}^{i,j}, l_r^{i,j})$$

5.4.3 Model selection

Whether the modeling considers the distribution of vehicles on a single link i or on two consecutive links i and j, the signal detection problem amounts to a model selection problem. For the single link approach, the algorithm decides whether φ_i^{sig} or $\varphi_i^{\text{no sig}}$ represents the distribution of measurements most accurately. The model φ_i^{sig} is fit using the probe vehicle data received on link i. The model $\varphi_i^{\text{no sig}}$ does not require fitting as it corresponds to a uniform distribution of the measurements over the length of the link. Model selection criterion are used to select the model which explains the best the data. The model with the higher prediction power, *i.e.* the model that minimizes the selection criterion is chosen for the signal detection decision.

The different models, representing the presence or the absence of a traffic sign, have different complexities. Here, model complexity refers to the number of parameters required to specify the model. Various criteria have been developed to trade-off between model fit and model complexity. Increasing the model complexity (*i.e*, the number of model parameters) leads to a better fit, and thus a higher likelihood but may overfit the available data. Information criteria *penalize* the score of the models depending on their complexity, *i.e.* models with a large number of parameters will be penalized more than models with a small number of parameters. Information criteria are used to compare the fit of models with different number of parameters, as is the case here.

In the present case, the uniform distribution is a special case of the parametric distribution function $f^c(x; \tilde{\rho}_a^i, l_{max}^i, l_r^i)$ for which $\tilde{\rho}_a^i = 1/L_i$. The other parameters, l_r^i and $l_{\max}^i$ can be chosen arbitrarily in their admissible range. The likelihood of the training data under the distribution φ_i^{sig} will therefore necessarily be higher than under the distribution $\varphi_i^{\text{no sig}}$. For the two links modeling approach, both models require fitting based on the available probe measurements. As for the one link approach, the model $\psi_{i,j}^{\text{sig}}$ has a higher number of parameters. The following results shows the additional complexity of the model $\psi_{i,j}^{\text{sig}}$.

Proposition 5.1 (Comparison of fitting potentials)**.** *If $\alpha^{i,j}$ is considered as a free parameter, the following result holds. For all set of parameter $(\tilde{\rho}_a^{i,j}, l_r^{i,j}, l_{\max}^{i,j})$ there exists a set of parameters $(\alpha^{i,j}, \tilde{\rho}^i, l_r^i, l_{\max}^i, \tilde{\rho}^j, l_r^j, l_{\max}^j)$ such that $\forall\, x \in [0, L_i + L_j]$ $\psi_{i,j}^{sig}(x) = \psi_{i,j}^{no\ sig}(x)$. This property shows that the model representing the assumption that there is a signal at the downstream end of link i can also represent the assumption that there is no signal at the downstream end of link i.*

Proof. For a given set of parameters for the model without signal, the following choice of parameters provides the same distribution of vehicles on the aggregated links i and j:

$$\tilde{\rho}_a^i = 1/L_i \quad \begin{matrix} \alpha^{i,j}\tilde{\rho}_a^i = \tilde{\rho}_a^{i,j} \\ l_r^j = l_r^{i,j} \end{matrix} \quad \begin{matrix} (1-\alpha^{i,j})\tilde{\rho}_a^j = \tilde{\rho}_a^{i,j} \\ l_{\max}^j = l_{\max}^{i,j} \end{matrix}$$

□

Note that the choice of $\alpha^{i,j}$ is driven by the number of measurements received on links i and j rather than on the respective distribution of the measurements on the link. For this reason, this parameter may not be considered as a free parameter which is set during the fitting.

Property 5.1 indicates that the model $\psi_{i,j}^{\text{sig}}$ should provide a higher likelihood score on training data than the model $\psi_{i,j}^{\text{no sig}}$. Comparing the likelihood of the models is thus not

informative as the model with a signal will always have a better score, whether there is actually a signal or not, underlining the importance to define appropriate model selection criteria.

The selection capabilities of three model selection criteria are compared in the experiments: the Aikaine Information Criterion (AIC), its correction for finite sample sizes (AICc) and the Bayesian Information Criterion. The model selection capabilities of the AIC and the AICc have theoretical motivations from Information theory [15, 37], whereas the derivations of the BIC arise from Bayesian statistics [143]. The different criteria have an analytical expression given by:

$$\begin{aligned} \text{AIC} &= -2\ln(\Lambda) + 2p, \\ \text{AICc} &= -2\ln(\Lambda) + 2p\frac{n}{n-p-1}, \\ \text{BIC} &= -2\ln(\Lambda) + p\ln(n), \end{aligned}$$

where Λ is the likelihood of the estimated model, p is the number of model parameters and n is the data size. All the criteria consist of the sum of the opposite of the log-likelihood and a penalization term that depends on the complexity of the model (number of parameters p) and, for some of them, on the size n of the dataset used to train the model. The dependency on n in the information criteria takes into account the finite sample sizes.

For the one link approach, the parameters of the model with signal are $\tilde{\rho}_a^i$, $l_{\max}^i$ and l_r^i and thus $p = 3$. The model without signal does not have free parameters: the two parameters of the uniform distribution are 0 and the length of the link and are not set based on the data. As for the two link approach, the parameters of the model with signal are $\tilde{\rho}_a^i$, $l_{\max}^i$, l_r^i, $\tilde{\rho}_a^j$, $l_{\max}^j$, l_r^j and $\alpha^{i,j}$, and thus $p = 7$. The model without signal is parameterized by $\tilde{\rho}_a^{i,j}$, $l_{\max}^{i,j}$ and $l_r^{i,j}$ and thus $p = 3$

5.5 Validation of the automatic signal detection algorithm

5.5.1 Experimental setup

The signal detection algorithm is tested on data collected by the *Mobile Millennium* system in the Bay Area of San Francisco, CA. The system collects several millions of GPS data points per day from probe vehicles reporting their location at a given sampling frequency (typically about every minute). The data used for the specific study presented below comes from a sub-fleet of around 500 probe vehicles within the *Mobile Millennium* system [25] collected on Tuesdays from 6 am to 10 am. In practice, numerical experiments have shown that the assumption of uniform arrival rates and periodicity (Section 5.1) does not limit the decision capabilities of the algorithm. For signal detection, the most important feature of the model is the queue length which specifically characterizes the presence of a traffic signal. The numerical analysis presented in the present article was also performed on data collected during 15 consecutive days (all times of day from January 1st, 2011 to January 15th, 2011) with very similar conclusions.

5.5.2 Automatic signal detection

Before being used for signal detection, the GPS measurements are filtered and mapped on the road network using a map-matching and path inference algorithm [96] which combines models of GPS measurements and drivers' behavior into a conditional random field. The road network is given by the NAVTEQ Inc. digital map which includes the geometry of

each of the links along with numerous attributes such as the presence of a traffic light or a stop sign. The present model cannot distinguish between traffic lights and stop signs who. Both types of signalization induce the formation of queues and the existence of delay experienced by the probe vehicles at the downstream end of the link. For this reason, each link is classified according to whether or not the algorithm determines that it has a signal (traffic light or stop sign) at the downstream intersection. The sub-network of San Francisco, CA of interest is composed of 1,172 links which have either signal or no traffic signal at the downstream intersection. The percentage of signalized links, as indicated in the NAVTEQ database, is 54%. For each link of the network, the probe data collected by the *Mobile Millennium* system is integrated in the classification algorithm to identify the presence of a traffic signal. The decision of the algorithm is then compared with the information available in the NAVTEQ database (Quarter 3, 2008).

Section 5.4 investigates two different approaches to identify the presence of a signal using data collected on the link or on the link and the downstream consecutive link. The section also suggestes different model selection criteria to classify each link as being signalized or not. The results of all the proposed approaches are summarized in Table 5.3.

The confusion matrices obtained seem to be at first glance a little deceptive with only around 70% of matches with the digital map database for all the criterions (AIC, AICc, BIC) and all the methods (one link, two links). Furthermore, the results show an unexpectedly high false-positive rate (e.g. 30.7% for BIC using two links), *i.e.* a significantly large number of links for which the algorithm detects the presence of a signal while the database does not indicate the presence of a signal. The differences between the one link models and the two links models are quite small, and the same apply for AIC and AICc which give very similar results. We expected the results of the AIC and the AICc to be similar as the difference between the criteria tends to zero as the sample size increases and most links received several hundreds of measurements. The main differences are observed between the AIC type criterion (AIC and AICc) and the BIC criterion. AIC approaches give higher false positive rates comparing to BIC which is more balanced between the two kinds of errors. This observation is natural since, BIC penalizes more the complex models than AIC does.

However, the performance analysis reported in Table 5.3 is based on a GIS which is also prone to errors, even though the quality of the NAVTEQ maps is internationally acknowledged. The existence of inaccuracies in the digital map database was first noticed in Section 5.2 and motivated the derivation of an algorithm for automatic signal detection. The analysis of the empirical distribution of the measurements on the links that were identified as signalized by the algorithm whereas they were recorded as not signalized in the GIS database (i.e. the false positive) is very informative. It underlined the idea that the map database might have erroneous information. The confusion matrices presented in Table 5.3 do not directly indicate the performance of the algorithm: they represent a quantitative comparison between the GIS labels (which contain some label noise) and the labels provided by the algorithm for each of the proposed approach (model with one or two links and different model selection criteria). To improve the conclusion regarding the importance of the approach, it was important to check the accuracy of the GIS used for validation. The analysis of the database accuracy consisted in a manual labeling of a subset of the map database using *Google Street View*. The attention was focused on the false positive of the BIC-two links decision algorithm. This choice was made to limit the number of manual checks. The false positives were considered to be the most likely to arise from errors in the database. The results of the human labeling are reported in the next section and underline the importance of the automatic labeling approach.

Table 5.3: Confusion matrix between prediction and GIS information for the two approaches (one link or two links) and the different model selection criteria.

AIC - one link

		Prediction	
		Signal	No signal
Actual	Signal	508	125
	No signal	241	297

AICc - one link

Actual	Signal	506	127
	No signal	237	301

BIC - one link

Actual	Signal	384	249
	No signal	149	389

AIC - two links

Actual	Signal	551	82
	No signal	276	262

ACCc - two links

Actual	Signal	548	85
	No signal	269	269

BIC - two links

Actual	Signal	429	204
	No signal	165	373

5.5.3 GIS manual labeling and update

Besides producing the confusion matrices, the code developed for the experiment produces a plot of the empirical distribution of the measurements and the models corresponding to the two possible decisions regarding the presence of a signal at the downstream end of the link or not. This qualitative analysis of the results seemed to indicate that, for a large number of the false-positive decisions, the model which was estimated under the assumption that the link was signalized fit the empirical data very accurately, whereas the model corresponding to the assumption of a non-signalized intersection did not capture the shape of the distribution. Figures 5.5 and 5.6 illustrate the fitting results on such a link: the algorithm classified the link as signalized whereas the database did not indicate the presence of a signal. The figures represent the empirical distribution of measurements along with the fitted distribution of measurements under the hypothesis that the link is signalized (red dashed line) or not signalized (green dotted line). Figure 5.5 represents the results obtained with the one-link approach and Figure 5.6 corresponds to the two-link approach. The intuition that a signal was actually present was validated by obtaining the geometry of the link (latitude and longitude points which characterize the location of the link) and checking for the presence of a traffic signal using *Google Street View*.

The manual verification lead to the realization that some of the signalized intersections were not present in the map database. In order to analyze the importance of the missing labeling, the links which were classified (using BIC on two links) as being signalized when the database indicated no signal (false-positives) were manually checked using *Google Street View*. The analysis conducted has important implications for the results of the algorithm and underlines the importance of automatic signal detection. Among the 165

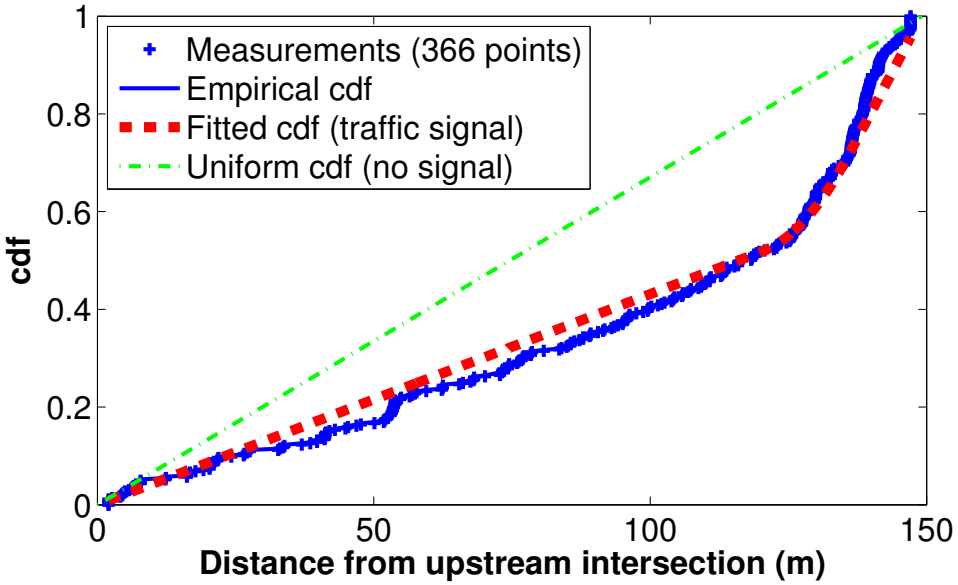


Figure 5.5: Qualitative comparison of the one-link models. The model with traffic signal fits the empirical data accurately, whereas the assumption of uniform distribution does not fit the observations. All information criteria accurately classify the link as having a traffic signal at the downstream intersection.

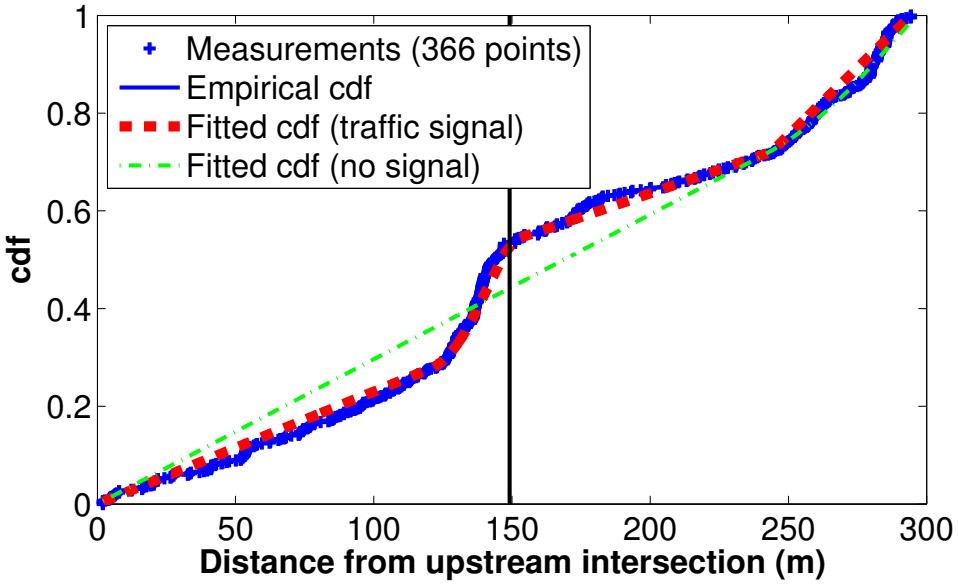


Figure 5.6: Qualitative comparison of the two-links models. The position of the intersection of the two links is depicted by the vertical black line. The two models independently learn the parameter under the assumption that there is a signal at the end of the first link or that there is no signal at the end of that link. The second fitting cannot capture the sharp increase in the empirical cdf and results in a poor fitting of the data. The information criteria accurately classify the link as having a traffic signal at the downstream intersection of the first link.

false-positives, 40% actually had traffic signals, 17% had stop signs. Therefore more than half of the false positives are in reality true positives. Among the 71 remaining links, 17% had specific features that explain the false-detection. Such features include the presence of a pedestrian crossing at the downstream end of the link even though there is no *actual* signal. Other factors include tunnels in which GPS reception and cellular communications are nonexistent or inaccurate or complicated intersections for which the downstream end of the link does not coincide with the the location of the signalization, even though queuing occurs on a regular basis on these links. Some of these cases (pedestrian crossing and complicated intersection) are illustrated in Figure 5.7.

5.6 Conclusion

From the hydrodynamic theory, the present chapter proposes a statistical model to represent the probability distribution of the locations reported by the probe vehicles. The distribution characterizes the mathematical model of traffic flow dynamics and illustrates the fact that vehicles spend, on average, more time close to downstream signalized intersections because it is the location where they are the most likely to experience delay.

The mathematical characterization of this empirical fact is important to properly scale the travel time of probe vehicles when they only travel a fraction of the link (partial link). A scaling which takes into account the spatial inhomogeneity of travel times significantly improve estimation capabilities based on machine learning algorithms [89].

The automatic signal detection algorithm leverages the physics of traffic to derive a statistical model representing the distribution of vehicles on a link, depending on the presence (resp. the absence) of signalization. The model parameters can be estimated using sparsely sampled probe vehicle data which makes it very promising given the emergence of this data at a large scale. The method is a first step towards automated GIS updates regarding signal location. The algorithm produces interesting result (more than half of the automatic signal detections that were not recorded in the GIS database correspond to actual stop signs or traffic lights). Furthermore, experiments with one week coverage data, lead to similar results as experiments based on data collected for a longer period of time, for a specific Day Of the Week and Time Of the Day. This realization motivates a periodic use of the algorithm to update and correct a GIS, in particular in areas where the infrastructure evolves rapidly such as developing countries.

This first step towards automated GIS updates and cleaning has the prospect to be improved and generalized by taking into account additional features. For example, the graph structure of the network structure can be leveraged to improve the decision results *e.g.* at an intersection with a light, all links are signalized and therefore a global decision by intersection (and not by link) should improve the robustness of the decision. Another possible extension of the methodology regards the discrimination of stop signs from traffic lights. Other information derived from sparsely sampled probe vehicle data such as travel times should be a mean to perform this discrimination by looking at the distribution of delay (and maybe average speed as well). However, no specific algorithm with promising results has been derived so far.

Figure 5.7: Illustration of the features of the downstream intersection of some links detected as being signalized by the algorithm (false-positives).

Chapter 6

Probability distribution of delays and travel times

The location of probe vehicle measurements provide important information regarding the relative density of measurements on a link which has valuable application for scaling travel times and automatic digital maps creation and update. When the vehicle can be re-identified between two measurements, the data can be used to infer travel times on the network.

The literature review presented in Chapter 2 motivates the development of statistical models for arterial traffic. In particular, the use of horizontal queuing theory to derive probability distribution of delays shows promising results [174]. Previous work has focused on the derivations of link travel time distributions. Since probe vehicles can send their positions at any location on the network, the path can start and end at any location and the first and last links of the corresponding path are not fully traversed by the vehicle (partial links). The derivations of link travel time distributions is not adapted to leverage the travel time information contained in the *partial* links of the path. One possibility is to use the travel time scaling algorithm presented in Section 5.2. However, the hydrodynamic based arterial traffic model (Chapter 4) develops the necessary framework to derive a model which specifically characterizes the probability distribution of delays and travel times between arbitrary locations. This approach has the prospect to yield more accurate results as it models the changes in the shape of the distribution depending on the measurement locations rather than only providing a scaling factor of the distribution.

From the hydrodynamic theory presented in Chapter 4, the chapter derives a statistical model for the probability distribution of travel times on an arterial network. The travel time between two locations is considered as the sum of two independent variables: the delay (stopping time) and the free flow travel time. For the delay, the origin of variability is the entrance time in the link (compared to the beginning of the cycle), which is considered as a random variable. For the free flow travel time, the randomness represents differences in driving behavior. Section 6.1 derives the probability distribution of delays between arbitrary locations. Section 6.2 derives the probability distribution of travel times and prove the quasi concavity of link travel time distributions. This feature is key to enable machine learning algorithms used in large scale traffic estimation algorithms [88].

6.1 Probability distribution of delay among the vehicles entering the link in a cycle

The travel time experienced by vehicles traveling on arterial networks is conditioned on two factors. First, the traffic conditions, given by the parameters of the network, dictate the state of traffic experienced by all the vehicles entering the link. Second, the time (after the beginning of a cycle) at which each vehicle arrives at the link determines how much delay will be experienced in the queue due to the presence of a traffic signal and the presence of other vehicles. Under similar traffic conditions, drivers experience different travel times depending on their arrival time. Using the assumption that the arrival density (and thus the arrival rate) is constant, arrival times are uniformly distributed on the duration of the light cycle. This allows for the derivation of the pdf of delay, which depends on the characteristics of the traffic light and the traffic conditions as defined in Section 4.4.

In this work, the data available for traffic estimation comes from probe vehicles which report their location periodically in time with a low frequency (on average once per minute). The vehicles send tuples of the form (x_1, t_1, x_2, t_2) where x_1 is the location of the vehicle at t_1 and x_2 is the position of the vehicle at t_2. This is representative of the most widely available data source with the prospect of global coverage in the near future (see Section 1.2 for a review of the different data types and their specificities). It includes taxi fleets or delivery fleets (FedEx, UPS and so on) which typically send data every minute in urban networks for fleet management purposes. Each tuple sent by the vehicles is considered independently. The assimilation of the measurements into the model has to be done independently for each data points if the sampling strategy does not track vehicle for more than two successive location reports.

6.1.1 *Total delay* and *measured delay* between locations x_1 and x_2

Let us consider a vehicle traveling from location x_1 to location x_2 and sending its location x_1 at time t_1 and its location x_2 at time t_2.

Definition 6.1 (Measured delay from x_1 to x_2)**.** *The* measured delay from x_1 *to* x_2*,* experienced in the time interval $[t_1, t_2]$*, in short "measured delay from* x_1 *to* x_2*" is defined as the difference between the travel time of the vehicle* $(t_2 - t_1)$ *and the travel time that the vehicle would experience between* x_1 *and* x_2 *without the presence of other vehicles nor signals.*

Definition 6.2 (Free flow travel time between x_1 and x_2)**.** *For a vehicle with free flow pace* p_f*, we call free flow travel time between* x_1 *and* x_2*, the quantity* $y_{f;x_1,x_2} = p_f(x_1 - x_2)$*, representing the travel time between* x_1 *and* x_2 *if the vehicle is not slowed down or stopped on its trajectory.*

The delay experienced between x_1 and x_2 is the difference between the travel time y_{x_1,x_2} of the vehicle between x_1 and x_2—not necessarily at free flow speed—and the free flow travel time $y_{f;x_1,x_2}$. In this model, vehicles are either stopped or driving at the free flow speed. The measured delay from x_1 to x_2, experienced in the time interval $[t_1, t_2]$ is the cumulative stopping time between t_1 and t_2.

Definition 6.3 (Total delay from x_1 to x_2)**.** *The* total delay *from* x_1 *to* x_2 *is defined as the cumulative stopping time of the vehicle on its trajectory from* x_1 *(from the first time it joined the queue, if the vehicle was in the queue at* x_1*) to* x_2 *(until the time it left the queue, if it was in the queue at* x_2*).*

In particular, if the vehicle stops at x_1 or at x_2 the total delay from x_1 to x_2 covers the full delay experienced during the stop, without taking into account the sampling scheme. Note that for vehicles sampled at x_1 and x_2 that do not stop at x_1 nor at x_2 the total delay is equal to the measured delay. For vehicles stopping in x_1 or in x_2, the measured delay is less than or equal to the total delay experienced by the vehicle (Figure 6.1 (right)).

A simple case illustrates the difference between *measured* and *total* delay and improves the understanding of the definitions. Let a vehicle be sampled every 30 seconds. Assume that the vehicle stops at the traffic signal ($x = 0$) and that the duration of the red time is 40 seconds. The vehicle sends its locations x_1 at t_1 and x_2 at $t_2 = t_1 + 30$. We do not receive additional information on the trajectory prior to t_1 or past t_2. The measured delay is at most 30 seconds (sampling rate); the total delay is 40 seconds. As a general remark, a vehicle reporting its delay during a stop reports a delay that is less than or equal to the total delay experienced on the trajectory, it represents the delay experienced between the two sampling times.

The following section derives the pdf of the *measured* and the *total* delay between any two locations x_1 and x_2 from the modeling assumptions defined in Section 4.1 and Definitions 6.1 and 6.3 . Given two sampling locations x_1 and x_2, the probability distribution of the total (resp. measured) delay δ_{x_1,x_2} is denoted $h^t(\delta_{x_1,x_2})$ (resp. $h^m(\delta_{x_1,x_2})$). The stationarity and constant arrival assumptions lead to analytical expressions for the speed of formation and dissolution of the queue, respectively denoted v_a and w (4.2). Under the stationarity assumption, the traffic variables are periodic with period C. To each arrival time corresponds a specific delay experienced by the vehicles on its trajectory from x_1 to x_2, traveled during the time interval $[t_1, t_2]$, as illustrated in Figure 4.1. The uniform arrival assumptions allows analytical derivations of the probability distribution of (measured or total) delays between any location x_1 and x_2.

6.1.2 Probability distribution of the total and measured delay between x_1 and x_2 in the undersaturated regime

Pdf of the *total* delay between x_1 and x_2

In the undersaturated regime, let $\eta^u_{x_1,x_2}$ denote the fraction of the vehicles entering the link during a cycle that experience a delay between x_1 and x_2. The remainder of the vehicles entering the link in a cycle travels from x_1 to x_2 without experiencing any delay. The proportion $\eta^u_{x_1,x_2}$ of vehicles delayed between x_1 and x_2 in a cycle, is computed as the ratio of vehicles joining the queue between x_1 and x_2 over the total number of vehicles entering the link in one cycle (Figure 6.1, left). The number of vehicles joining the queue between x_1 and x_2 is the number of vehicles stopped between x_1 and x_2: $(\min(l_{\max}, x_1) - \min(l_{\max}, x_2))\,\rho_{\max}$. The number of vehicles entering the link is $v_f C \rho_a$. The proportion of vehicles delayed between x_1 and x_2 is thus:

$$\eta^u_{x_1,x_2} = (\min(x_1, l_{\max}) - \min(x_2, l_{\max}))\frac{\rho_{\max}}{v_f C \rho_a}.$$

The total stopping time experienced when stopping at x is denoted by $\delta^u(x)$ for the undersaturated regime. Because the arrival of vehicles is constant, the delay $\delta^u(x)$ increases linearly with x. At the intersection ($x = 0$), the delay is maximal and equals the duration of the red light R. At the end of the queue ($x = l_{\max}$) and upstream of the queue ($x \geq l_{\max}$), the delay is null. Thus the expression of $\delta^u(x)$:

$$\delta^u(x) = R\left(1 - \frac{\min(x, l_{\max})}{l_{\max}}\right).$$

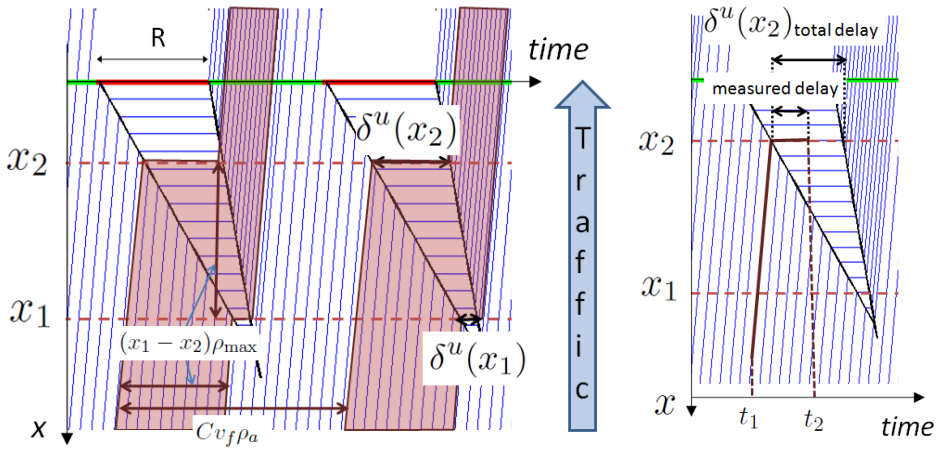


Figure 6.1: (**Left**) The proportion of delayed vehicles $\eta^u_{x_1,x_2}$ is the ratio between the number of vehicles joining the queue between x_1 and x_2 over the total number of vehicles entering the link in one cycle. The trajectories highlighted in purple represent the trajectories of vehicles delayed between x_1 and x_2. (**Right**) The vehicles reporting their location during a stop at x_2 experience a delay $\delta \in [0, \delta^u(x_2)]$ in the time interval $[t_1, t_2]$. This delay is less than or equal to the total delay ($\delta^u(x_2)$) experienced on the trajectory.

Given that the arrival of vehicles is uniform in time, the distribution of the location where the vehicles reach the queue between x_1 and x_2 is uniform in space. For vehicles reaching the queue between x_1 and x_2, the probability to experience a delay between locations x_1 and x_2 is uniform. The uniform distribution has support $[\delta^u(x_1), \delta^u(x_2)]$, corresponding to the minimum and maximum delay between x_1 and x_2.

The *total* delay experienced between x_1 and x_2 is a random variable with a mixture distribution with two components. The first component represents the vehicles that do not experience any stopping time between x_1 and x_2 (mass distribution in 0), the second component represents the vehicles reaching the queue between x_1 and x_2 (uniform distribution on $[\delta^u(x_1), \delta^u(x_2)]$). We note $\mathbf{1}_A$ the indicator function of set A,

$$\mathbf{1}_A(x) = \begin{cases} 1 \text{ if } x \in A \\ 0 \text{ if } x \notin A \end{cases}$$

Let $\text{Dir}_{\{a\}}(\cdot)$ denote the Dirac distribution centered in a, used to represent the mass probability. The pdf of total delay between x_1 and x_2 (Figure 6.3, left) reads:

$$h^t(\delta_{x_1,x_2}) = (1 - \eta^u_{x_1,x_2})\text{Dir}_{\{0\}}(\delta_{x_1,x_2}) + \frac{\eta^u_{x_1,x_2}}{\delta^u(x_2) - \delta^u(x_1)}\mathbf{1}_{[\delta^u(x_1),\delta^u(x_2)]}(\delta_{x_1,x_2})$$

The cumulative distribution function of total delay $H^t(\cdot)$ reads:

$$H^t(\delta_{x_1,x_2}) = \begin{cases} 0 & \text{if } \delta_{x_1,x_2} < 0 \\ (1 - \eta^u_{x_1,x_2}) & \text{if } \delta_{x_1,x_2} \in [0, \delta^u(x_1)] \\ (1 - \eta^u_{x_1,x_2}) + \eta^u_{x_1,x_2}\frac{\delta_{x_1,x_2} - \delta^u(x_1)}{\delta^u(x_2) - \delta^u(x_1)} & \text{if } \delta_{x_1,x_2} \in [\delta^u(x_1), \delta^u(x_2)] \\ 1 & \text{if } \delta_{x_1,x_2} > \delta^u(x_2) \end{cases}$$

Pdf of the *measured* delay between x_1 and x_2

Because of the sampling scheme, the measured delay differs from the total delay experienced by the vehicles.

In the following, i refers to the upstream or the downstream measurement locations ($i \in \{1, 2\}$). When sending their location x_i, some vehicles are stopped at this location.

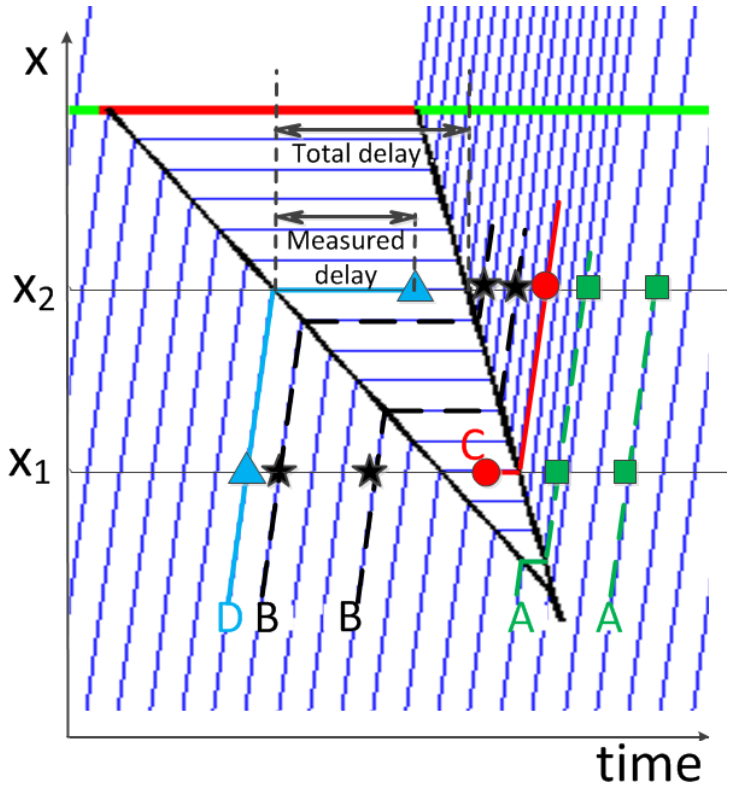


Figure 6.2: Classification of the trajectories depending on the stopping location. Vehicles travel between the two measurement points x_1 and x_2. They are sampled uniformly in time.

These vehicles may not report the full delay associated with location x_i (Figure 6.1, right). In particular, a vehicle stopped at x_1 when sending its location at time t_1 will only report the delay experienced after t_1. Similarly, a vehicle stopped at x_2 when sending its location at time t_2 will only report the delay experienced before t_2.

- For a measurement received at x_i, the probability that it comes from a vehicle stopped in the queue is $\frac{\delta^u(x_i)}{C}$, which is the ratio of the time spent by the stopped vehicle at x_i over the duration of the cycle.
- For a vehicle stopped at x_2, observed at t_2 coming from a previous observation point x_1 (at t_1), the probability of a stopping time experienced by the vehicle until t_2 is uniform between 0 and $\delta^u(x_2)$, since the vehicle is sampled arbitrarily during its stopping phase.

The vehicles traveling from x_1 to x_2 (where x_1 and x_2 are measurement points) are classified depending on the locations of their stop with respect to x_1 and x_2. This classification of the trajectories is also illustrated in Figure 6.2:

A) Vehicles do not experience any delay between the measurement points x_1 and x_2.

B) Vehicles reach the queue at x with $x \in (x_2, x_1)$. These vehicles are not stopped when they send their location at x_1 and x_2.

C) Vehicles reach the queue at x_1, where they report their location at time t_1. At t_1, the vehicle was already stopped (or was just stopping) and the measurement only accounts for the delay occurring after t_1, which is less than or equal to $\delta^u(x_1)$. Because of the uniform sampling in time, the reported delay has a uniform distribution on $[0, \delta^u(x_1)]$

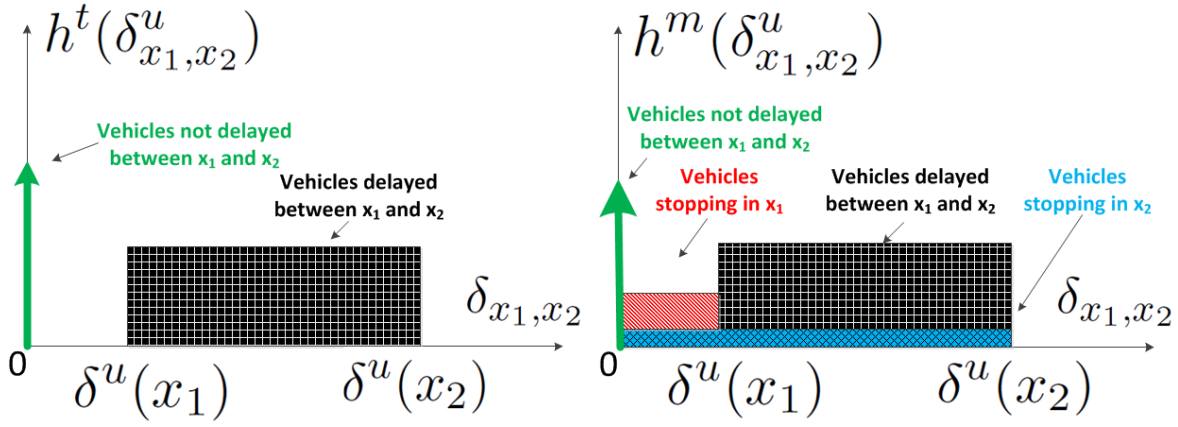


Figure 6.3: (**Left**) Probability distribution of the total delay between x_1 and x_2 in the undersaturated regime. (**Right**) Probability distribution of the measured delay between x_1 and x_2 in the undersaturated regime. Vehicles are assumed to be sampled uniformly in time.

D) Vehicles reach the queue at x_2, where they report their location at time t_2. At t_2, the vehicle is still stopped and the measurement only represents the delay occurring up to t_2, which is less than or equal to $\delta^u(x_2)$. Because of the uniform sampling in time, the reported delay has a uniform distribution on $[0, \delta^u(x_2)]$

Let s_{x_i} denote the event "*vehicle stops at location* x_i". Denoting by $\mathcal{P}(A)$ the probability of event A, we have $\mathcal{P}(s_{x_i}) = \frac{\delta^u(x_i)}{C}$. The notation $\bar{s}_{x_i}$ represents the event "*vehicle does not stop at location* x_i". The notation $(\bar{s}_{x_1}, \bar{s}_{x_2})$ represents the event "*vehicles do not stop at location* x_1 *nor* x_2". We assume that the events $\bar{s}_{x_1}$ and $\bar{s}_{x_2}$ are independent. The probability of event $(\bar{s}_{x_1}, \bar{s}_{x_2})$ reads:

$$\begin{aligned} \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) &= \mathcal{P}(\bar{s}_{x_1})\mathcal{P}(\bar{s}_{x_2}) && \text{Independence assumption} \\ &= (1 - \mathcal{P}(s_{x_1}))(1 - \mathcal{P}(s_{x_2})) && \text{Complementary events} \end{aligned}$$

The event $(\bar{s}_{x_1}, \bar{s}_{x_2})$ corresponds to trajectories of type A (vehicles do not stop between x_1 and x_2) and trajectories of type B (vehicles stop strictly between x_1 and x_2 but neither in x_1 nor in x_2). Among the vehicles stopping at none of the measurement points, a fraction $\eta^u_{x_1,x_2}$ is delayed between x_1 and x_2 (trajectories of type B) and a fraction $1-\eta^u_{x_1,x_2}$ does not experience delay between x_1 and x_2 (trajectories of type A). Given a delay measurement between locations x_1 and x_2, the probability that it was sent by a vehicle with a trajectory of type A is $\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})(1-\eta^u_{x_1,x_2})$. Similarly, the probability that it was sent by a vehicle with a trajectory of type B is $\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})\eta^u_{x_1,x_2}$.

Given that a measurement is received at location x_i, the probability that this measurement is sent by a vehicle that joined the queue at x_i is proportional to the delay experienced at location x_i. Given successive measurements at locations x_1 and x_2, the probability that a vehicle reports its location x_i ($i \in \{1, 2\}$) while being stopped at this location is denoted ζ_{x_i}. From this definition and given that the delay measurement is received between locations x_1 and x_2, the probability that it was sent by a vehicle with a trajectory of type C is ζ_{x_1}. The probability that it was sent by a vehicle with a trajectory of type D is ζ_{x_2}. Note that vehicles cannot stop both at x_1 and x_2 (they stop only once in the queue); thus $\mathcal{P}(s_{x_1}, s_{x_2}) = 0$. Given that a vehicle was sampled at x_1 and x_2, the following equation holds:

$$\underbrace{\zeta_{x_1}}_{\substack{\text{Prob. that the veh} \\ \text{stopped at } x_1 \text{ only}}} + \underbrace{\zeta_{x_2}}_{\substack{\text{Prob. that the veh} \\ \text{stopped at } x_2 \text{ only}}} + \underbrace{\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})}_{\substack{\text{Prob. that the veh} \\ \text{stopped neither at } x_1 \text{ nor at } x_2}} + \underbrace{\mathcal{P}(s_{x_1}, s_{x_2})}_{\substack{\text{Prob. that the veh} \\ \text{stopped both at } x_1 \text{ and at } x_2 \\ (=0)}} = 1$$

The probability of stopping either at x_1 or at x_2 is $1 - \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$ (complementary of stopping neither at x_1 nor at x_2). Among these vehicles, the proportion that stops in x_i is proportional to the delay experienced in x_i for $i \in \{1, 2\}$. It follows that

$$\left\{ \begin{array}{rcl} \zeta_{x_i} & \propto & \delta^u(x_i),\ i \in \{1,2\} \\ \zeta_{x_1} + \zeta_{x_2} & = & 1 - \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \end{array} \right. \Rightarrow\ \zeta_{x_i} = (1 - \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})) \frac{\delta^u(x_i)}{\delta^u(x_1) + \delta^u(x_2)} \quad i \in \{1,2\}$$

The probability distribution of measured delay is a finite mixture distribution, in which each component is a mass probability or a uniform distribution. The theoretical probability distribution function is illustrated Figure 6.3, right. It is the sum of the following terms that also refer to Figure 6.2:

(A) a mass probability in 0 with weight $(1 - \eta^u_{x_1,x_2})\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$, representing the vehicles that do not reach the queue between x_1 and x_2,

(B) a uniform distribution on $(\delta^u(x_1), \delta^u(x_2))$ with weight $\eta^u_{x_1,x_2}\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$, representing the vehicles that reach the queue strictly between x_1 and x_2,

(C) a uniform distribution on $[0, \delta^u(x_1)]$ with weight ζ_{x_1}, representing the vehicles that stop in x_1,

(D) a uniform distribution on $[0, \delta^u(x_2)]$ with weight ζ_{x_2}, representing the vehicles that stop in x_2.

The pdf of the measured delay is related to the pdf of the total delay as:

$$h^m(\delta_{x_1,x_2}) = \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) h^t(\delta_{x_1,x_2}) + \frac{\zeta_{x_1}}{\delta^u(x_1)} \mathbf{1}_{[0,\delta^u(x_1)]}(\delta_{x_1,x_2}) + \frac{\zeta_{x_2}}{\delta^u(x_2)} \mathbf{1}_{[0,\delta^u(x_2)]}(\delta_{x_1,x_2})$$

6.1.3 Probability distribution of the measured delay between x_1 and x_2 in the congested regime

In the congested regime, the delay distribution can be computed using a similar methodology as for the undersaturated regime, by deriving the delay experienced between x_1 and x_2 for each arrival time. Let n_s be the maximum number of stops experienced by the vehicles in the remaining queue between the locations x_1 and x_2. The delay experienced at location x when reaching the triangular queue at x is readily derived from the expression of the delay in the undersaturated regime. The delay experienced when reaching the remaining queue is the duration of the red time R. The expression of the delay at location x is then

$$\delta^c(x) = \left\{ \begin{array}{cl} R & \text{if } x \leq l_r \\ R\,\frac{l_r + l_{\max} - x}{l_{\max}} & \text{if } x \in [l_r, l_r + l_{\max}] \\ 0 & \text{if } x \geq l_r + l_{\max} \end{array} \right.$$

The details of the derivation are given in Appendix A and illustrated in Figures A.1–A.4. Note that to satisfy the stationarity assumption, the distance traveled by vehicles in the queue in the duration of a light cycle is $l_{\max}$.

The section summarizes the derivations, classified depending on the location of the positions x_1 and x_2 with respect to the remaining and triangular queue lengths:

1. x_1 *Upstream* – x_2 *Remaining* (Figure A.1): The location x_1 is upstream of the queue and the location x_2 is in the triangular queue. Let the critical location x_c be defined by $x_c = x_2 + n_s l_{\max}$. Vehicles reaching the triangular queue upstream of x_c stop n_s times

in the remaining queue. On the road segment $[x_1, x_2]$, vehicles reaching the triangular queue downstream of x_c stop $n_s - 1$ times in the remaining queue. The vehicles experience a delay uniformly distributed on $[\delta_{\min}, \delta_{\max}]$ with $\delta_{\min} = (n_s - 1)R + \delta^c(x_c)$ and $\delta_{\max} = n_s R + \delta^c(x_c) = \delta_{\min} + R$. The probability distribution of total delay reads:

$$h^t(\delta_{x_1,x_2}) = \frac{1}{\delta_{\max} - \delta_{\min}} \mathbf{1}_{[\delta_{\min},\delta_{\max}]}(\delta_{x_1,x_2}), \quad \begin{array}{l} \delta_{\min} = \delta^c(x_c) + (n_s - 1)R \\ \delta_{\max} = \delta^c(x_c) + n_s R \end{array}$$

2. x_1 *Triangular* – x_2 *Triangular* (Figure A.2): Both locations x_1 and x_2 are upstream of the remaining queue (in the triangular queue or upstream of the queue). Given that the path is upstream of the remaining queue, this case is similar to the undersaturated regime, where derivations are updated to account for the fact that the triangular queue starts at $x = l_r$. Adapting the notation from Section 6.1.2, $\eta^c_{x_1,x_2}$ denotes the fraction of the vehicles entering the link in a cycle that experience delay between locations x_1 and x_2.

$$\eta^c_{x_1,x_2} = \frac{\min(x_1 - l_r, l_{\max}) - \min(x_2 - l_r, l_{\max})}{l_{\max}}$$

This delay is uniformly distributed on $[\delta^c(x_1), \delta^c(x_2)]$. The reminder do not stop between x_1 and x_2. The probability distribution of total delay reads:

$$h^t(\delta_{x_1,x_2}) = (1 - \eta^c_{x_1,x_2})\mathrm{Dir}_{\{0\}}(\delta_{x_1,x_2}) + \frac{\eta^c_{x_1,x_2}}{\delta^c(x_2) - \delta^c(x_1)} \mathbf{1}_{[\delta^c(x_1),\delta^c(x_2)]}(\delta_{x_1,x_2})$$

3. x_1 *Remaining* – x_2 *Remaining* (Figure A.3): Both locations x_1 and x_2 are in the remaining queue. Let the critical location x_c be defined by $x_c = x_2 + (n_s - 1)l_{\max}$. The vehicles reaching the queue between x_1 and x_c stop n_s times in the remaining queue between x_1 and x_2, their stopping time is $n_s R$. The reminder of the vehicles stop $n_s - 1$ times in the remaining queue and their stopping time is $(n_s - 1)R$. The probability distribution of total delay reads:

$$h^t(\delta_{x_1,x_2}) = \frac{x_1 - x_c}{l_{\max}} \mathrm{Dir}_{\{n_s R\}}(\delta_{x_1,x_2}) + \left(1 - \frac{x_1 - x_c}{l_{\max}}\right) \mathrm{Dir}_{\{(n_s-1)R\}}(\delta_{x_1,x_2})$$

4. x_1 *Triangular* – x_2 *Remaining* (Figure A.4): The upstream location x_1 is in the triangular queue and the downstream location x_2 is in the remaining queue. Let the critical location x_c be defined by $x_c = x_2 + n_s l_{\max}$.

   - If $x_1 \geq x_c$, a fraction $(x_1 - x_c)/l_{\max}$ of the vehicles entering the link in a cycle join the triangular queue between x_1 and x_c. They stop once in the triangular queue and n_s times in the remaining queue. Among these vehicles, the stopping time is uniformly distributed on $[\delta^c(x_1) + n_s R,\ \delta^c(x_c) + n_s R]$. A fraction $(x_c - l_r)/l_{\max}$ of the vehicles entering the link in a cycle join the triangular queue between x_c and $l_{\max}$. Among these vehicles, the stopping time is uniformly distributed on $[\delta^c(x_c) + (n_s - 1)R,\ n_s R]$. The remainder of the vehicles reach the remaining queue between l_r and $x_1 - l_{\max}$ and their stopping time is $n_s R$. The probability distribution of total delay reads:

$$\begin{array}{rll} h^t(\delta_{x_1,x_2}) = & \frac{x_1 - x_c}{l_{\max}} \frac{\mathbf{1}_{[\delta^c(x_1) + n_s R,\ \delta^c(x_c) + n_s R]}(\delta_{x_1,x_2})}{\delta^c(x_c) - \delta^c(x_1)} & \text{Vehicles stopping between } x_1 \text{ and } x_c \\ + & \frac{x_c - l_r}{l_{\max}} \frac{\mathbf{1}_{[\delta^c(x_c) + (n_s-1)R,\ n_s R]}(\delta_{x_1,x_2})}{R - \delta^c(x_c)} & \text{Vehicles stopping between } x_c \text{ and } l_r \\ + & \left(1 - \frac{x_1 - l_r}{l_{\max}}\right) \mathrm{Dir}_{\{n_s R\}}(\delta_{x_1,x_2}) & \text{Vehicles stopping between } l_r \text{ and } x_1 - l_{\max} \end{array}$$

- If $x_1 \leq x_c$, a fraction $(x_1 - l_r)/l_{\max}$ of the vehicles entering the link in a cycle join the triangular queue between x_1 and l_r. They stop once in the triangular queue and $n_s - 1$ times in the remaining queue. Among these vehicles, the stopping time is uniformly distributed on $[\delta^c(x_1) + (n_s - 1)R,\ n_s R]$. A fraction $1 - (x_c - l_r)/l_{\max}$ of the vehicles entering the link in a cycle join the remaining queue between l_r and $x_c - l_{\max}$. The stopping time of these vehicles is $n_s R$. The remainder of the vehicles experiences a stopping time of $(n_s - 1)R$. The probability distribution of total delay reads:

$$
\begin{array}{rlll}
h^t(\delta_{x_1,x_2}) = & \frac{x_1 - l_r}{l_{\max}} \frac{\mathbf{1}_{[\delta^c(x_1) + (n_s-1)R,\ n_s R]}(\delta_{x_1,x_2})}{R - \delta^c(x_1)} & \text{Vehicles stopping between } x_1 \text{ and } l_r \\
+ & \left(1 - \frac{x_c - l_r}{l_{\max}}\right) \mathrm{Dir}_{\{n_s R\}}(\delta_{x_1,x_2}) & \text{Vehicles stopping between } l_r \text{ and } x_c - l_{\max} \\
+ & \frac{x_c - x_1}{l_{\max}} \mathrm{Dir}_{\{(n_s-1)R\}}(\delta_{x_1,x_2}) & \text{Vehicles stopping between } x_c - l_{\max} \text{ and } x_1 - l_{\max}
\end{array}
$$

These cases represent the pdf of total delay. From the results derived in Section 6.1.2, we derive the pdf of measured delay. From the previous derivations, we have:

$$
\begin{array}{rcl}
\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_1}) & = & (1 - \mathcal{P}(\bar{s}_{x_1}))(1 - \mathcal{P}(\bar{s}_{x_2})) \\
\zeta_{x_i} & = & (1 - \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})) \frac{\delta^u(x_i)}{\delta^u(x_1) + \delta^u(x_2)} \quad i \in \{1, 2\}
\end{array}
$$

It is the sum of the following terms:

(i) the delay distribution given that the vehicles stop neither in x_1 nor in x_2, with weight $\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$,

(ii) the delay probability distribution given a stop in x_1, with weight ζ_{x_1},

(iii) the delay probability distribution given a stop in x_2, with weight ζ_{x_2}.

We summarize the different components of the delay distribution, described as a mixture distribution for all the different cases in Tables 6.1 and 6.2.

6.2 Probability distributions of travel times

On a path between x_1 and x_2, the travel time y_{x_1,x_2} is a random variable. It is the sum of two random variables: the delay δ_{x_1,x_2} experienced between x_1 and x_2 and the free flow travel time of the vehicles $y_{f;\,x_1,x_2}$. The free flow travel time is proportional to the distance of the path and the free flow pace p_f such that $y_{f;\,x_1,x_2} = p_f(x_1 - x_2)$. The travel time y_{x_1,x_2} is given by $y_{x_1,x_2} = \delta_{x_1,x_2} + y_{f;\,x_1,x_2}$.

In the following, the delay and the free flow pace are assumed to be independent random variables, thus so are the delay and the free flow travel time.

The differences in traffic behavior are modeled by considering the free flow pace p_f as a random variable with distribution φ^p and support $\mathcal{D}_{\varphi^p}$. For convenience, for a probability distribution φ with support $\mathcal{D}_\varphi$, let us define its prolongation by zero of out of $\mathcal{D}_\varphi$. With a slight abuse of notation, this new function is still denoted φ.

Using a linear change of variables, the probability distribution $\varphi^y_{x_1,x_2}$ of free flow travel time $y_{f;\,x_1,x_2}$ between x_1 and x_2 is derived as follows:

$$
p_f \sim \varphi^p(p_f) \Rightarrow \varphi^y_{x_1,x_2}(y_{f;\,x_1,x_2}) = \varphi^p\left(\frac{y_{f;\,x_1,x_2}}{x_1 - x_2}\right) \frac{1}{x_1 - x_2}
$$

The pdf of travel times is derived from the following fact:

Table 6.1: The pdf of measured delay is a mixture distribution. The different components and their associated weights depend on the location of stops of the vehicles with respect to the queue length and sampling locations (Table 1 of 2).

Case	Trajectories	Weight	Dist.	Support
Case 1 $x_1 \geq l_r + l_{\max}$, $x_2 \leq l_r$, $x_c = x_2 + n_s l_{\max}$	Does not stop at x_2	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$	Unif.	$[(n_s-1)R + \delta^c(x_c),$ $n_s R + \delta^c(x_c)]$
	Stop at x_2	$\zeta_{x_2} = 1 - \mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2})$	Unif.	$[(n_s-1)R + \delta^c(x_c),$ $n_s R + \delta^c(x_c)]$
Case 2 $x_1 \geq l_r$, $x_2 \geq l_r$	No stop between x_1 and x_2	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times$ $(1 - \eta^c_{x_1,x_2})$	Mass	$\{0\}$
	Reach the (triangular) queue between x_1 and x_2	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times$ $\eta^c_{x_1,x_2}$	Unif.	$[\delta^c(x_2), \delta^c(x_1)]$
	Stop at x_1	ζ_{x_1}	Unif.	$[0,\ \delta^c(x_1)]$
	Stop at x_2	ζ_{x_2}	Unif.	$[0,\ \delta^c(x_2)]$
Case 3 $x_1 \leq l_r$, $x_2 \leq l_r$, $x_c = x_2 + (n_s-1) l_{\max}$	Reach the (remaining) queue between x_1 and x_c	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times$ $\frac{x_1 - x_c}{l_{\max}}$	Mass	$\{n_s R\}$
	Reach the (remaining) queue between x_c and $x_1 - l_{\max}$	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times$ $\frac{x_c - x_1 + l_{\max}}{l_{\max}}$	Mass	$\{(n_s-1)R\}$
	Stop at x_1	ζ_{x_1}	Unif.	$[(n_s-1)R,\ n_s R]$
	Stop at x_2	ζ_{x_2}	Unif.	$[(n_s-1)R,\ n_s R]$

Fact 6.1 (Sum of independent random variables)**.** *If X and Y are two independent random variables with respective pdf f_X and f_Y, then the pdf f_Z of the random variable $Z = X + Y$ is given by $f_Z(z) = f_X * f_Y(z)$*

This classical result in probability is derived by computing the conditional pdf of Z given X and then integrating over the values of X according to the total probability law.

For each regime s, the probability distribution of travel times reads:

$$g^s(y_{x_1,x_2}) = \left(h^s * \varphi^y_{x_1,x_2}\right)(y_{x_1,x_2})$$

Notice that the delay distributions are mixtures of mass probabilities and uniform distributions. The importance of these classes of distribution motivates the derivation of the general expression of the travel time distributions when vehicles experience a delay with mass probability in Δ and when vehicles experience a delay with uniform distribution on $[\delta_{\min}, \delta_{\max}]$.

6.2.1 Travel time distributions

Travel time distribution when the delay has a mass probability in Δ

The stopping time is Δ. This corresponds to trajectories with n_s stops ($n_s \geq 0$) in the remaining queue. This includes the non stopping vehicle in the undersaturated regime, when the remaining queue has length zero. The travel time distribution is derived as

$$\begin{aligned} g(y_{x_1,x_2}) &= \mathrm{Dir}_{\{\Delta\}} * \varphi^y_{x_1,x_2}(y_{x_1,x_2}) \\ &= \varphi^y_{x_1,x_2}(y_{x_1,x_2} - \Delta). \end{aligned} \tag{6.1}$$

Table 6.2: The pdf of measured delay is a mixture distribution. The different components and their associated weights depend on the location of stops of the vehicles with respect to the queue length and sampling locations (Table 2 of 2).

Case	Trajectories	Weight	Dist.	Support
Case 4a $x_1 \in [l_r, l_r + l_{\max}]$, $x_2 \leq l_r$, $x_c = x_2 + n_s l_{\max}$, $x_c \leq x_1$	Reach the (triangular) queue between x_1 and x_c	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{x_1 - x_c}{l_{\max}}$	Unif.	$[n_s R + \delta^c(x_1), n_s R + \delta^c(x_c)]$
	Reach the (triangular) queue between x_c and l_r	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{x_c - l_r}{l_{\max}}$	Unif.	$[(n_s - 1)R + \delta^c(x_c), n_s R]$
	Reach the (remaining) queue between l_r and $x_1 - l_{\max}$	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{l_r - x_1 + l_{\max}}{l_{\max}}$	Mass	$\{n_s R\}$
	Stop at x_1	ζ_{x_1}	Unif.	$[n_s R, n_s R + \delta^c(x_1)]$
	Stop at x_2	ζ_{x_2}	Unif.	$[(n_s - 1)R + \delta^c(x_c), n_s R + \delta^c(x_c)]$
Case 4b $x_1 \in [l_r, l_r + l_{\max}]$, $x_2 \leq l_r$, $x_c = x_2 + n_s l_{\max}$, $x_c \geq x_1$	Reach the (triangular) queue between x_1 and l_r	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{x_1 - l_r}{l_{\max}}$	Unif.	$[(n_s - 1)R + \delta^c(x_1), n_s R]$
	Reach the (remaining) queue between l_r and $x_c - l_{\max}$	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{l_r - x_c + l_{\max}}{l_{\max}}$	Mass	$\{n_s R\}$
	Reach the (remaining) queue between $x_c - l_{\max}$ and $x_1 - l_{\max}$	$\mathcal{P}(\bar{s}_{x_1}, \bar{s}_{x_2}) \times \frac{x_c - x_1}{l_{\max}}$	Mass	$\{(n_s - 1)R\}$
	Stop at x_1	ζ_{x_1}	Unif.	$[(n_s - 1)R, (n_s - 1)R + \delta^c(x_1)]$
	Stop at x_2	ζ_{x_2}	Unif.	$[(n_s - 1)R, n_s R]$

Travel time distribution when the delay is uniformly distributed on $[\delta_{\min}, \delta_{\max}]$

Vehicles experience a uniform delay between a minimum and maximum delay respectively denoted $\delta_{\min}$ and $\delta_{\max}$. The probability of observing a travel time y_{x_1,x_2} is given by

$$g(y_{x_1,x_2}) \quad = \quad \frac{1}{\delta_{\max} - \delta_{\min}} \int_{-\infty}^{+\infty} \mathbf{1}_{[\delta_{\min},\delta_{\max}]}(y_{x_1,x_2} - z)\ \varphi^y_{x_1,x_2}(z)\, dz. \tag{6.2}$$

The integrand is not null if and only if $y_{x_1,x_2} - z \in [\delta_{\min}, \delta_{\max}]$, *i.e.* if and only if $z \in [y_{x_1,x_2} - \delta_{\max},\ y_{x_1,x_2} - \delta_{\min}]$. Since $\varphi^y_{x_1,x_2}(z)$ is equal to zero for $z \in \mathbb{R} \setminus \mathcal{D}_\varphi$, the integrand is not null if and only if $z \in [y_{x_1,x_2} - \delta_{\max},\ y_{x_1,x_2} - \delta_{\min}] \bigcap \mathcal{D}_\varphi$.

As an illustration, the derivation of the probability distribution of travel times on an entire link in the undersaturated regime, for a pace distribution with support on $\mathbb{R}^+$ is detailed below and illustrated in Figure 6.4 (left). The length of the link is denoted L. A fraction $1-\eta^u_{L,0}$ of the vehicles entering the link in a cycle has a delay with mass probability in 0 (vehicles do not stop on the link). The probability distribution of travel times of these vehicles is computed via Equation (6.1) with $\Delta = 0$. The reminder of the vehicles (fraction $\eta^u_{L,0}$) experiences a delay that is uniformly distributed on $[0, R]$. The probability distribution of travel times of these vehicles is computed via Equation (6.2) with $\delta_{\min} = 0$ and $\delta_{\max} = R$. The probability distribution of travel times on an undersaturated arterial link reads:

$$g^u(y_{L,0}) = \begin{cases} 0 & \text{if} \quad y_{L,0} \leq 0 \\ (1-\eta^u_{L,0})\varphi^y_{L,0}(y_{L,0}) + \dfrac{\eta^u_{L,0}}{R} \displaystyle\int_0^{y_{L,0}} \varphi^y_{L,0}(z)\, dz & \text{if} \quad y_{L,0} \in [0, R] \\ (1-\eta^u_{L,0})\varphi^y_{L,0}(y_{L,0}) + \dfrac{\eta^u_{L,0}}{R} \displaystyle\int_{y_{L,0}-R}^{y_{L,0}} \varphi^y_{L,0}(z)\, dz & \text{if} \quad y_{L,0} \geq R \end{cases}, \tag{6.3}$$

In the more general case of a travel time distribution on an undersaturated partial link between locations x_1 and x_2, the delay distribution is a mixture of mass probabilities and uniform distributions. Using the linearity of the convolution it is possible to treat each component of the mixture separately and sum them with their respective weights to derive the probability distribution of travel times.

The derivations are similar in the congested regime. For the different cases described in Section 6.1.3, the delay is a mixture of mass probabilities and uniform distributions. For example, the probability distribution of link travel times (Case 1) is illustrated in Figure 6.4 (right). When the delay is uniformly distributed on $[\delta_{\min}, \delta_{\max}]$, the probability distribution of travel times is computed via Equation (6.2) and reads

$$g^c(y_{L,0}) = \begin{cases} 0 & \text{if} \quad y_{L,0} \leq \delta_{\min} \\ \frac{1}{\delta_{\max}-\delta_{\min}} \int_0^{y_{L,0}-\delta_{\min}} \varphi^y_{L,0}(z) dz & \text{if} \quad y_{L,0} \in [\delta_{\min}, \delta_{\max}] \\ \frac{1}{\delta_{\max}-\delta_{\min}} \int_{y_{L,0}-\delta_{\max}}^{y_{L,0}-\delta_{\min}} \varphi^y_{L,0}(z)\, dz & \text{if} \quad y_{L,0} \geq \delta_{\max} \end{cases} \tag{6.4}$$

6.2.2 Quasi-concavity of the probability distributions of link travel times

The probability distributions of travel times depend on a set of parameters that must be estimated to fully determine the statistical distribution of the travel times. A parameter with true value θ_0 is estimated via an estimator $\hat{\theta}$. The estimator is in general chosen to have some optimality properties—extremum point based on an objective function, *e.g.*

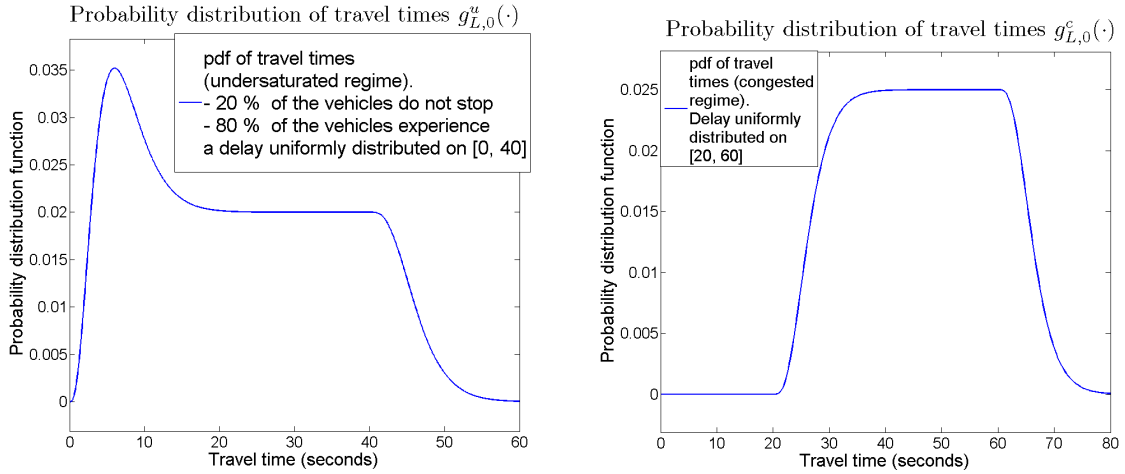


Figure 6.4: Probability distributions of link travel times. **Left:** Undersaturated regime. The figure represents the probability distribution function of travel times for the following choice of parameters: the traffic light has duration 40 seconds, 80% of the vehicles stop at the light ($\eta^u_{L,0} = .8$). **Right:** Congested regime. The figure represents the probability distribution function of travel times with the following choice of parameters: the traffic light has duration 40 seconds, all the vehicles stop in the triangular queue and 50% of the vehicles stop once in the remaining queue. Both figures are produced for a link of length 100 meters. The free flow pace is a random variable with Gamma distribution. The mean free flow pace is 1/15 s/m and the standard deviation is 1/30 s/m. The probability distribution function γ of a Gamma random variable $x \in \mathbb{R}^+$ with shape α and inverse scale parameter β is given by $\gamma(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, where Γ is the Gamma function defined on $\mathbb{R}^+$ and with integral expression $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t}\, dt$.

least square estimator, maximum likelihood estimator. In particular, the maximum likelihood estimator is widely used in statistics. Its computation requires the maximization of the likelihood (or log-likelihood) function. The likelihood function is the probability (probability density in the continuous case) of observing a set of data points as a function of the parameters of the distribution. The log-likelihood is the logarithm of the likelihood function and is commonly used to simplify numerical expressions and improve the numerical stability of the optimization algorithms.

Here, the function to maximize is the probability distribution of travel times. Properties on the concavity of this function are important for designing efficient maximization algorithms with guaranty global optimality. This section presents the proof of the quasi-concavity of the link travel time distributions in both the undersaturated and the congested regimes. We also prove the log-concavity of the different components of the distribution of travel times, considered as mixture distributions.

Definition 6.4 (Quasi-concavity (Boyd)). *[34] A function $f : \mathbb{R}^n \to \mathbb{R}$ is called quasi-concave if its domain is convex and if $\forall \alpha \in \mathbb{R}$, the superlevel set Sf_α ($Sf_\alpha = \{x \in \mathcal{D}_f | f(x) \geq \alpha\}$) is convex.*

From this definition, one can derive equivalent characterization of quasi-concavity, when f has first (and second) order derivatives. Further references on quasi-concavity are detailed in [34]. In particular, the proofs of quasi-concavity of the probability distributions functions use the characterization of continuous quasi-concave functions on $\mathbb{R}$ (Lemma 6.1) and the second order characterization (Lemma 6.2):

Lemma 6.1 (Characterization of continuous quasi-concave functions on $\mathbb{R}$). *[34] A*

continuous function $f : \mathcal{D}_f \to \mathbb{R}$ is quasi-concave if and only if at least one of the following conditions holds:

- *f is nondecreasing*
- *f is nonincreasing*
- *there is a point $x_c \in \mathcal{D}_f$ such that for $x \leq x_c$ (and $x \in \mathcal{D}_f$), f is nondecreasing, and for $x \geq x_c$ (and $x \in \mathcal{D}_f$), f is nonincreasing.*

Lemma 6.2 (Second order characterization of quasi-concave functions)**.** *[34] $f \in \mathcal{C}^2$ is quasi-concave if and only if $\forall (x, y) \in {\mathcal{D}_f}^2$, $y^T \nabla f(x) = 0 \Rightarrow y^T \nabla^2 f(x) y \leq 0$. If f is unidimensional, f is quasi-concave if and only if $f'(x) = 0 \Rightarrow f''(x) \leq 0$.*

Note that for probability distributions, the focus is often on the properties of the logarithm of the probability distribution function.

Definition 6.5 (Log-concavity (Boyd))**.** *[34] A function $f : \mathbb{R}^n \to \mathbb{R}_*^+$ is log-concave if and only if its logarithm $\ln(f)$ is concave. The second order characterization is as follows: $f \in \mathcal{C}^2$ is log-concave is and only if $\forall x \, f(x) \nabla^2 f(x) \preceq \nabla f(x) \nabla f(x)^T 0$.*

Fact 6.2. *For f a twice differentiable function taking values in $\mathbb{R}_*^+$, f is quasi-concave $\Leftrightarrow \ln(f)$ is quasi-concave.*

Proof. We have $\nabla \ln f(x) = \dfrac{\nabla f(x)}{f(x)}$ and $\nabla^2 \ln f(x) = \dfrac{f(x)\nabla^2 f(x) - \nabla f(x) \nabla f(x)^T}{f(x)^2}$.

- If f is quasi-concave then $ln(f)$ is quasi-concave:
  Since f is quasi-concave and has its image included in $(0, +\infty)$, it follows from Lemma 6.2 that for all $(x, y) \in \mathcal{D}_f^2$, $y^T \nabla f(x) = 0$ implies $y^T \nabla^2 f(x) y \leq 0$.
  Let x and y be such that $y^T \nabla \ln(f(x)) = 0$, *i.e.* $y^T \nabla f(x) = 0$. From the quasi-concavity of f, it follows that $y^T \nabla^2 f(x) y \leq 0$

$$\begin{aligned} y^T \nabla^2 \ln(f(x)) y &= \frac{f(x)\, y^T \nabla^2 f(x) y - y^T \nabla f(x) \nabla f(x)^T y}{f(x)^2} \\ &= \frac{f(x)\, y^T \nabla^2 f(x) y}{f(x)^2} \quad \text{since } y^T \nabla f(x) = 0 \\ &\leq 0 \qquad\qquad\qquad\quad \text{using the quasi-concavity of } f \end{aligned}$$

  So $y^T \nabla \ln(f(x)) = 0 \Rightarrow y^T \nabla^2 \ln(f(x)) y \leq 0$ and $\ln(f)$ is quasi-concave.

- If $\ln(f)$ is quasi-concave then f is quasi-concave:
  Since $\ln(f)$ is quasi-concave, it follows from Lemma 6.2 that for all $(x, y) \in \mathcal{D}_f^2$, $y^T \nabla \ln(f(x)) = 0$ implies that $y^T \nabla^2 \ln(f(x)) y \leq 0$. Using the expression of $\nabla \ln(f(x))$ and $\nabla^2 \ln(f(x))$, this condition can be rewritten as follows:

$$\forall (x, y), \; y^T \nabla f(x) = 0 \Rightarrow y^T \nabla^2 f(x) y \leq 0$$

  And this proves that f is quasi-concave.

□

In the following, we assume that the pdf φ^p of the free flow pace is strictly log-concave, and thus so is the pdf $\varphi^y_{x_1,x_2}$ of the free flow travel time between location x_1 and x_2. Note that most common probability distributions (*e.g.* Gaussian or Gamma with shape greater than 1) are log-concave.

Lemma 6.3. *In one dimension, a strictly log-concave probability distribution function φ defined on $\mathcal{D}_\varphi \subset \mathbb{R}$ has a unique critical point $y_c \in \mathcal{D}_\varphi$. If φ is defined on a finite or semi-finite interval, the critical point y_c may be one of the finite boundaries of $\mathcal{D}_\varphi$. On the domain $\mathcal{D}_\varphi$, the function φ is strictly increasing for $y \leq y_c$ and strictly decreasing for $y \geq y_c$.*

Proof. The function φ is strictly log-concave. Since the logarithm is strictly increasing, it follows that φ is either (i) strictly increasing or (ii) strictly decreasing or (iii) that there exists y_c such that φ is strictly increasing for $y \leq y_c$ and strictly decreasing for $y \geq y_c$. The case (iii) corresponds to the conclusion of the Lemma. It remains to prove that for case (i), there exists a such that $\mathcal{D}_\varphi \subset (-\infty, a)$ and that for case (ii), there exists b such that $\mathcal{D}_\varphi \subset (b, +\infty)$. For both cases, reasoning by contradiction and using the fact that φ is integrable on $\mathcal{D}_\varphi$ concludes the proof. □

Proposition 6.1 (Quasi-concavity of the probability distribution of link travel times in the undersaturated regime)**.** *The probability distribution function of travel times on an undersaturated link, denoted g^u and with analytical expression given in Equation* (6.3)*, is quasi-concave.*

Proof. Let Δ denote the maximum delay experienced (i.e. the red time R) and η the fraction of delayed vehicles (previously denoted $\eta^u_{L,0}$). The length of the link L is a scale parameter that does not change the concavity properties of the function. For notational simplicity, let φ denote the probability distribution function of travel times and omit the locations $x_1 = L$ and $x_2 = 0$ in this section. Recall the travel time probability distribution on an undersaturated link (Equation (6.3)):

$$g^u(y) = (1-\eta)\varphi(y) + \frac{\eta}{\Delta}\int_{y-\Delta}^{y} \varphi(z)\, dz$$

with the convention $\varphi(z) = 0$ for $z \leq 0$.

The function g^u is continuously differentiable on $\mathbb{R}^+$ and $\forall\, y \in \mathbb{R}^+$ its derivative reads:

$$(g^u)'(y) = (1-\eta)\varphi'(y) + \frac{\eta}{\Delta}(\varphi(y) - \varphi(y-\Delta)). \tag{6.5}$$

The function $(g^u)'$ is continuously differentiable on $\mathbb{R}^+$ and $\forall\, y \in \mathbb{R}^+$ the second derivative of g^u is given by:

$$(g^u)''(y) = (1-\eta)\varphi''(y) + \frac{\eta}{\Delta}(\varphi'(y) - \varphi'(y-\Delta)) \tag{6.6}$$

Using the expression of $(g^u)'(y)$, it follows that:

$$(g^u)'(y) = 0 \Leftrightarrow (1-\eta)\varphi'(y) = \frac{\eta}{\Delta}(\varphi(y-\Delta) - \varphi(y)). \tag{6.7}$$

The goal of the present proof is to show that $(g^u)'(y) = 0 \Rightarrow (g^u)''(y) \leq 0$, so let y be such that $(g^u)'(y) = 0$.

- Case 1: $\varphi'(y) > 0$
  Lemma 6.3 implies that φ is strictly increasing on $(-\infty, y]$. Thus $\varphi(y-\Delta) < \varphi(y)$. Plugging back into (6.5), it follows that $\varphi'(y) > 0 \Rightarrow (g^u)'(y) > 0$ which contradicts the hypothesis $(g^u)'(y) = 0$.

- Case 2: $\varphi'(y) \leq 0$
  From (6.6) and the log concavity of φ, the following inequality holds

$$\begin{aligned}
(g^u)''(y) &\leq (1-\eta)\frac{(\varphi'(y))^2}{\varphi(y)} + \frac{\eta}{\Delta}(\varphi'(y) - \varphi'(y-\Delta)) \\
&\quad \text{Using (6.7), replace } (1-\eta)\varphi'(y) \text{ by } \frac{\eta}{\Delta}(\varphi(y-\Delta) - \varphi(y)) \\
&= \frac{\eta}{\Delta}\left(\frac{(\varphi'(y))}{\varphi(y)}\left(\varphi(y-\Delta) - \varphi(y)\right) + \varphi'(y) - \varphi'(y-\Delta)\right) \\
&= \frac{\eta}{\Delta}\left(\frac{(\varphi'(y))}{\varphi(y)}\varphi(y-\Delta) - \varphi'(y-\Delta)\right)
\end{aligned}$$

  Moreover, equation (6.7) and the condition $\varphi'(y) \leq 0$ imply that $\varphi(y-\Delta) \leq \varphi(y)$. Reasoning by contradiction, assume that $\varphi'(y-\Delta) \leq 0$. From Lemma 6.3, it follows that $\varphi'(y-\Delta) \leq 0$ implies $\varphi(y-\Delta) > \varphi(y)$, which contradicts the assumption of Case 2. Thus necessarily, $\varphi'(y-\Delta) \geq 0$ and plugging into (6.6), $(g^u)''(y) \leq 0$.

The derivations have proven that $(g^u)'(y) = 0 \Rightarrow (g^u)''(y) \leq 0$. From the definition of quasi-concavity (Definition 6.4), the proof concludes that $g^u(y)$ is quasi-concave. □

Proposition 6.2 (Quasi-concavity of the probability distribution of link travel times in the congested regime)**.** *The probability distribution function of travel times on a congested link, denoted g^c and with analytical expression given in Equation* (6.4)*, is quasi-concave.*

Proof. Let $\delta_{\min}$ (resp. $\delta_{\max}$) denote the minimum (resp. maximum) delay experienced. Recall the travel time probability distribution on a congested link:

$$g_c(y) = \frac{1}{\delta_{\max} - \delta_{\min}} \int_{y-\delta_{\max}}^{y-\delta_{\min}} \varphi(y)\, dy$$

The function g_c is continuously differentiable on $\mathbb{R}^+$ and $\forall\, y \in \mathbb{R}^+$ its derivative g_c' reads:

$$g_c'(y) = \frac{1}{\delta_{\max} - \delta_{\min}} \left(\varphi(y-\delta_{\min}) - \varphi(y-\delta_{\max})\right).$$

Let us prove that there exists an interval I such that $y \notin I \Rightarrow g_c'(y) \neq 0$. From the characterization of quasi-concave function given in Lemma 6.1, it follows that g_c is quasi-concave.

Referring to Lemma 6.3, let y_c denote the critical point of the pace distribution φ. The following holds:

- For $y \in [0,\, y_c + \delta_{\min}]$, the following inequalities hold $y - \delta_{\max} < y - \delta_{\min} \leq y_c$. Thus $\varphi(y-\delta_{\max}) < \varphi(y-\delta_{\min})$ and $g_c'(y) > 0$.

- For $y \in [y_c + \delta_{\max},\, +\infty]$, the following inequalities hold $y_c \leq y - \delta_{\max} < y - \delta_{\min}$. Thus $\varphi(y-\delta_{\max}) > \varphi(y-\delta_{\min})$ and $g_c'(y) < 0$.

- For $y \in [y_c + \delta_{\min},\, y_c + \delta_{\max}]$, the following inequalities hold $y - \delta_{\max} \leq y_c \leq y - \delta_{\min}$. For all $y \in [y_c + \delta_{\min},\, y_c + \delta_{\max}]$, $y - \delta_{\max} \leq y_c$ and thus the function $y \mapsto \varphi(y - \delta_{\max})$ is strictly increasing on $[y_c + \delta_{\min},\, y_c + \delta_{\max}]$. Similarly, for all $y \in [y_c + \delta_{\min},\, y_c + \delta_{\max}]$, $y - \delta_{\max} \geq y_c$ and the function $y \mapsto \varphi(y - \delta_{\min})$ is strictly decreasing on $[y_c + \delta_{\min},\, y_c + \delta_{\max}]$. The function g'^c is strictly decreasing on $[y_c + \delta_{\min},\, y_c + \delta_{\max}]$. Moreover $g'^c(y_c + \delta_{\min}) > 0$ and $g'^c(y_c + \delta_{\max}) < 0$. Using the monotonicity of g'^c on $[y_c + \delta_{\min},\, y_c + \delta_{\max}]$ and the theorem of intermediate values, it follows that g_c' is equal to zero in a unique point on $[y_c + \delta_{\min},\, y_c + \delta_{\max}]$.

The function g_c has a unique critical point (unique point where $g^{c\prime}$ equals zero). From the characterization of quasi-concavity given in Lemma 6.1, we conclude that $g_c(y)$ is quasi-concave. Note that we have also proven *strict* quasi-concavity since we showed that the critical point of g_c is unique. □

6.2.3 Learning the probability distribution of travel times

From traffic flow theory, we derived a probability distribution of travel times between arbitrary locations on an arterial link. These distributions are parameterized by the network parameters (average red time R, average cycle time C, driving behavior θ_p and saturation queue length $l^s_{\max}$) and the level of congestion represented by the queue length $l_{\max}$. As probe vehicles report their location periodically in time, the duration between two successive location reports x_1 and x_2 represents a measurement of the travel time of the vehicle on its path from x_1 to x_2. We use these travel time observations from probe vehicles to learn the parameters of the travel time distributions.

Common sampling rates for probe vehicles are around one minute and probe vehicles typically traverse several links between successive location reports. It is possible to optimally decompose the path travel time to estimate the travel time spent on each link of the path [84]. In this section, we assume that this decomposition has already been achieved and we focus on the estimation of the pdf of travel times. Since probe vehicles may report their location at any point x_1 and x_2, they provide partial link travel time measurements that allow for the estimation of the independent parameters of each link: the red time R, the queue length $l_{\max}$, the fraction of stopping vehicles on the link among the vehicles entering the link in one cycle $\eta^u_{L,0}$ and the driving behavior θ_p. The estimation of the parameters of link i is done by maximizing the likelihood (or more conveniently the log-likelihood) of the (partial) link travel times of this link with respect to these parameters. Note that the parameters of the travel time distribution (R^i, $\eta^u_{L,0}$, l^i and θ^i_p) do not depend on the locations x_1 and x_2 of measurement j. In particular, we learn the travel time distributions using travel time measurements which span different portions of the link, *i.e.* the locations x_1 and x_2 depend on the index of the measurement (j), even though we do not explicit this dependency for notational simplicity. Let $(y^j_{x_1,x_2})_{j=1:J^i}$ represent the set of (partial link) travel times allocated to link i. The estimation problem is given by:

$$\begin{aligned} \underset{R^i,\eta^u_{L,0},l^i,\theta^i_p}{\textbf{minimize}} \ & \sum_{j=1}^{J^i} -\ln(g^i(y^j_{x_1,x_2})) \\ \textbf{s.t. } & \eta^u_{L,0} \in [0,1],\ l^i \in [0, L^i]. \end{aligned} \tag{6.8}$$

Additional constraints and bounds may be added to limit the feasible set to physically acceptable values of the parameters and improve the estimation when little data is available. The optimization problem (6.8) is not convex but it is a small scale optimization problem (feasible set of dimension five). Numerous optimization techniques can be used to solve this problem including global optimization algorithms [93, 175]. Moreover, since the parameters represent physical parameters, they can be bounded to limit the feasible set to a compact set (of dimension five). It is thus possible to do a *grid search*. The grid search algorithm defines a grid on the bounded feasible set and evaluates the objective function for each set of parameters defined by the grid. We keep the B best set of parameters, associated with the lowest values of the objective function and perform a first or a second order optimization algorithm [34] from this best set of parameters. In the implementation of the algorithm used to produce the results of Section 7.3, we set $B = 4$ and used the `active-set` algorithm in the Matlab optimization toolbox, which is a second order optimization algorithm based on *Sequential Quadratic Programming* [31].

Figure 6.5: Arterial corridor in Chula Vista, CA (Telegraph Canyon Road) where the data was collected using *Sensys Networks* technology.

6.3 Numerical experiments and results

6.3.1 Experimental setup

The numerical experiments are based on data collected from wireless vehicle detection systems deployed by *Sensys Networks* [75]. The sensors record the magnetic signature of the vehicles and can re-identify a vehicle between successive sensors, thus providing travel time measurements of individual vehicles. In this article, we focus on data collected on February 10th, 2011 on an arterial corridor in Chula Vista, California (Telegraph Canyon Road). This network consists of 18 sensors in both directions located at the 9 signalized intersections, as illustrated Figure 6.5. *Sensys Networks* reports that approximately 70% of the vehicles traveling on the corridor are properly re-identified. Note that the unmatched vehicles include vehicles changing direction (no successive crossings) or vehicles with a magnetic signature that was not properly re-identified by the algorithm.

6.3.2 Model selection for travel time distributions

The chapter presents a statistical model of arterial traffic to characterize probability distribution of travel times on arterial networks. To validate the approach, the goodness of fit of the *traffic distribution*, derived from hydrodynamic theory, is compared to other standard distributions. The present section also analyzes the potential to improve the estimation capabilities by learning some of the traffic parameters from historical data and considering them fixed in the estimation procedure.

As for the numerical analysis of Chapter 5, information criterion are useful for model selection. The *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* measure the relative goodness of fit of statistical models and define trade-offs between the accuracy and the complexity of each model. Because the sample size is finite, AIC is typically corrected and denoted AICc. For a model with k parameters, trained with n samples and resulting likelihood L, the information criterion are defined as follows,

$$\mathrm{AIC} = 2k - 2\ln(L), \quad \mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1}, \quad \mathrm{BIC} = k\ln(n) - 2\ln(L).$$

The information criterion are computed for different classes of probability distributions. The class of probability distributions which represents the best trade-off between model complexity and accuracy is chosen to characterize the probability distribution of travel times. The traffic distribution, derived from hydrodynamic theory, is compared to standard distributions (normal, Gamma and log-normal) and to a Gaussian mixture

model with two components. The model allows to input prior knowledge on some of the parameters. Note that the prior information can also be incorporated in a *Bayesian* framework by considering that the parameters are r.v. with a specified distribution. Among the standard distributions, the log-normal distribution consistently performs better than both the normal and the Gamma distribution. For this reason, only the results for the log-normal distribution report are reported and compared to the results obtained for the traffic distribution. Regarding the possibility to fix some of the parameters of the traffic distribution, the results do not improve if the mean of the free flow pace is fixed. However, significant improvement is obtained when both the mean and the variance of the free flow pace are fixed. The analysis reports the results obtained when fixing none of the parameters (labeled *traffic*) or both the mean and variance of the free flow pace (labeled *traffic with fixed pace*).

Figure 6.6 presents the AICc and BIC computed for the different models for varying sizes of the training dataset. When only a few measurements are available (less than 7 or 10), the results are unstable and the risk of over-fitting is large. With more than 10 or 15 observations, the size of the training dataset does not influence the results significantly. The Gaussian mixture model has the largest values of both the AICc and the BIC. Even though the model provides a good fit on the training data, the larger number of parameters (compared to the other models) leads to a higher risk of overfitting. The traffic model and the log-normal model have similar results for the AICc and the log-normal model has a slightly better value for BIC, which penalizes the number of parameters more strongly than AICc does. The traffic model has the potential to fit the data more accurately but the complexity of the model is penalized. This motivates the investigation of reducing the traffic model estimation complexity by fixing some of the parameters. Fixing the free flow pace does not improve the AICc or BIC scores. However, fixing both the mean and the variance of the free flow pace provides much better results and is the model chosen in the remainder of the numerical analysis.

6.3.3 Validation of the fitting results

To further validate the choice of the traffic model, with fixed free flow pace distribution, 30 points are randomly sampled from the data set to form the *training set*. The parameters of the *traffic with fixed pace* model and the parameters of a log-normal distribution are learned on this training set. The learned distributions are compared to the remaining data (*validation set*) in Figure 6.7.

The large validation set underlines further the capacity of the traffic model to characterize the probability distribution of travel times. It captures specific aspects of traffic dynamics resulting from the presence of a traffic signal, in particular the differences between the stopping and non-stopping vehicles and the long tail of the distribution illustrating the variability of delays experienced depending on the arrival time in the link.

6.4 Conclusion on the probability distribution of delays and travel times

The present chapter proposes a parametric probability distribution of travel times between arbitrary locations on an arterial link. The probability distribution is derived from the hydrodynamic theory presented in Chapter 4 and represents the dynamics of traffic flow on arterial links. In particular, it captures the delay of vehicles due to the presence of a queue that forms and dissipates periodically because of the presence of a traffic signal at the downstream intersection. These distributions are parameterized by physical parameters

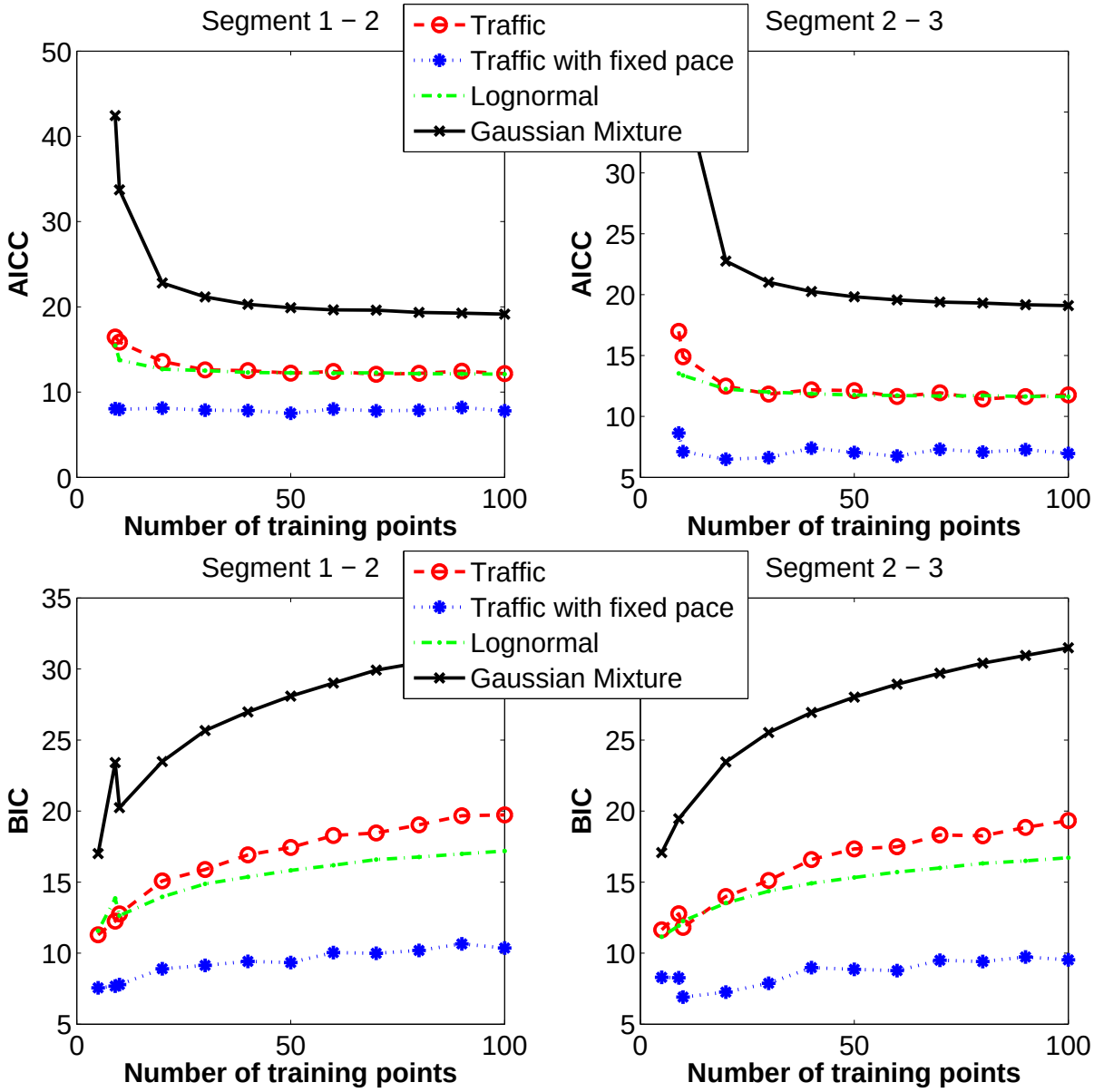


Figure 6.6: Information criterion computed on the different potential models to represent probability distribution of travel times on arterial networks. The two information metrics (AICc and BIC) lead to the same conclusions in terms of model selection. **Top:** Corrected Aikaike Information Criterion (AICc). **Bottom** Bayesian Information Criterion.

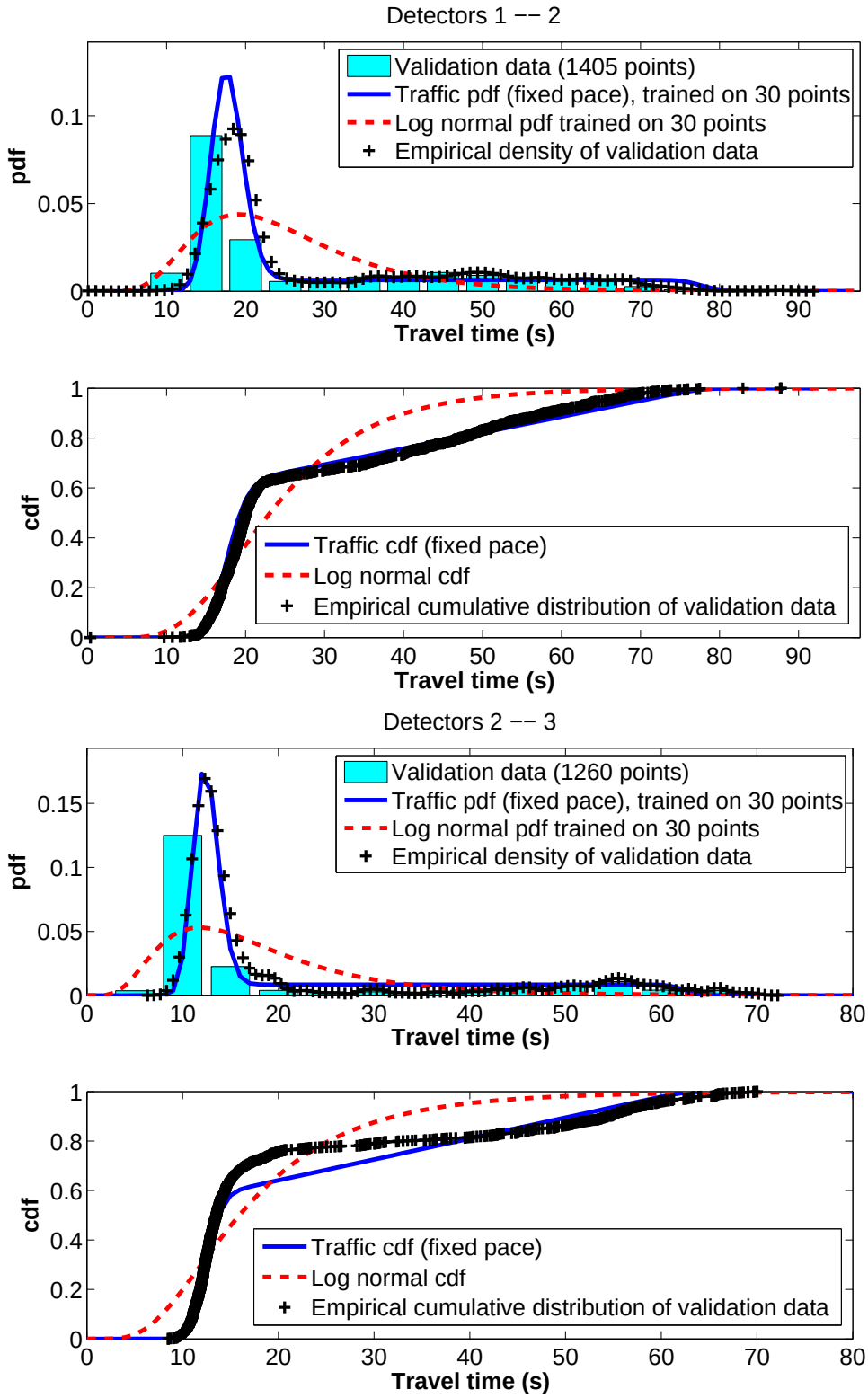


Figure 6.7: Validation of the pdf of travel time derived from horizontal queuing theory. The parameters of the log-normal distributions and the traffic distribution with fixed pace are learned using 30 travel time observations sampled randomly from the dataset. The histogram of the data and the empirical distribution illustrate the distribution of the validation data. The empirical pdf is computed using a normal kernel function [33] and the empirical cdf represent the Kaplan-Meier estimator [47]

which characterize the arterial network, the driving behavior and the level of congestion.

The travel time distributions are adapted to real-time traffic estimation from sparsely sampled probe data. The following specificities make the model particularly adapted to probe vehicle data:

- The pdf of travel times between any two locations on a link are parameterized by the same set of parameters. This makes the model particularly interesting for learning with sparsely sampled probe data. As seen in Section 5.2, travel speeds vary significantly on arterial links (the probability of delay is higher close to intersections). Scaling partial travel times takes into account the spatial inhomogeneity of travel times. The main limitation of scaling partial travel times is that it only scales the distribution without changing the shape of the distribution. This does not accurately reflect the evolution of the probability distribution function when the sampling locations change. Moreover, a finer discretization of the road network would imply the learning of a larger number of parameters which increase the risk of over-fitting given the amount of data available today at a large scale.

- The probability distributions of link travel times are quasi-concave distributions. The proof of this property is important to develop estimation capabilities from sparsely sampled probe vehicles and guarantee convergence of machine learning algorithms.

The goodness of fit of the distributions was tested on data collected by *Sensys Networks* in Chula Vista, CA. The numerical results show the superiority of the traffic distribution to represent the distribution of travel times compared to "classic" distributions (normal, log-normal, Gamma distributions, Gaussian mixture model), commonly used to represent the distribution of travel times. We also show how prior estimates of certain parameters (distribution of free flow pace) improves the estimation capabilities. The model represents the best trade-off between model complexity (in terms of number of parameters) and goodness of fit and performs well when a limited amount of data is available.

The assumption of uniform arrivals is the most restrictive assumption on which this work is based. It may limit the accuracy of the model when signal synchronization is important. The following chapter, Chapter 7 proposes a model to relieve this assumption while keeping a small number of parameters and controlling the computational complexity of the model. Improvements on the model are mostly important for control strategies which study light synchronization. The model presented in this chapter is sufficient to characterize the probability distribution given the information contained in the data available today.

Another possible refinement of the model is to consider a lane-dependent model which takes into account differences in the way queue forms depending on the turn movements. When arterial links have protected turn movements, the queuing dynamics on these lanes may differ from the queuing dynamics for the through movements. As probe vehicles report their location periodically in time, it is possible to differentiate the measurements depending on the turn movements. As the number of parameters per link increases, so do the risks of overfitting and a trade-off analysis is necessary to decide which level of model detail is adapted to the data available.

The model presented in this chapter characterizes the formation and dissolution of queues on each link of the network independently. The model can be extended to take into account network dynamics and model how congestion spreads and dissipates throughout the network. A model enforcing the conservation of vehicles at intersections and parameterized by the turn ratios provides a network-wide statistical dynamical model of urban traffic which can be used to estimate travel time distributions [88].

Chapter 7

Going further: a three-stream model for arterial traffic

The present chapter focuses on an extension of the hydrodynamic model of arterial traffic presented in Chapter 4. The extension proposes analytic derivations for arterial traffic modeling which release the most potentially limiting assumption of Chapter 4 while controlling the number of parameters required to characterize the model. This extension has important applications to model arterial traffic on a route, in particular when signal coordination invalidates the uniform arrival assumption of Chapter 4. After arterial traffic modeling and estimation, a natural continuation in traffic engineering is traffic control. Control is closely related to both modeling and estimation: in classical control theory, an estimation of the state of the system is fed into the controller to adapt the control strategy according to the dynamics of the system.

The majority of the studies on arterial traffic use numerical algorithms for signal optimization [127, 112, 77, 67], or rely on simulation. These methods can handle scenario analysis of complex systems and can generate the desired signal control numerically. However, the complexity of the solution process grows rapidly with the size of the problem [32], in addition to the fact that the amount of information needed for the optimization is large and tedious to obtain for large networks. In addition, numerical solutions might not provide physical insight on the traffic patterns controlled by such schemes. Analytical solutions provide a deeper understanding of traffic flow dynamics. The purpose of analytical methods is generally not to provide detailed solutions to specific problems, but to generate general principles to solve the problem, by making specific assumptions to reduce the number of parameters and the complexity of the problem. For example, [150] derives expressions for delays at signalized intersections assuming platoon inflow. The present article considers platoon traffic and ignores secondary traffic. This is complemented by [138] which considers both platoon traffic and secondary traffic.

This chapter focuses on analytical methods and proposes a novel model relying on hydrodynamic traffic models [18]. In the present model, the arrival and departure of traffic flows at each signalized intersection are represented by three streams of traffic during each cycle. Each traffic stream is characterized by its flow and duration (the time it takes for all the traffic within the stream to go through a point in space). This is realistic if one inspects the downstream of an intersection, where there are mainly three streams of traffic: no traffic during the red time, saturation flow during the beginning of the green time (as the queue dissipates), and less than saturation flow (if undersaturated) during the end of the green time, once the queue is fully dissipated. The present model approximates the third traffic stream with a constant flow. When both the arrival and the

departure traffic are modeled in this way, the results from a single signalized intersection are automatically applicable to a corridor including multiple signalized intersections. In addition, the number of parameters is limited and only grows linearly with the size of the network, facilitating analytical solutions.

The chapter is organized as follows. Section 7.1 presents the three stream model, which characterizes the departure traffic streams based on the arrival traffic streams. Section 7.2 presents how the model can be used for signal control and corridor optimization. The section proves that the total delay is a quasi-convex function in the traffic light offset. This property enables the derivation of the optimal offset under different scenarios for the input streams. The section also characterizes how the streams evolve through the corridor. The model is compared against microsimulation data in Section 7.3. Section 7.4 discusses the generality of the model and provides conclusion about the benefits of the method for both arterial traffic modeling, estimation and control.

7.1 Modeling traffic flows through a single signalized intersection

The present section develops a model of traffic dynamics through a single intersection. The model treats the arrival traffic as inputs and the departure traffic as outputs. The model describes traffic flow at each intersection with a limited number of parameters, which does not grow with the complexity of the network. This property, referred to as *parameter efficiency*, facilitates analytical solutions which are leveraged to derive tractable estimation and control algorithms. The model is structured so that results from a single intersection can easily be extended to a series of intersections.

7.1.1 Three stream model

The three-stream model builds upon the hydrodynamic arterial flow model of Chapter 4. As done in Chapter 4, vehicular flow is modeled as a continuum and represented with macroscopic variables of *flow* $q(x,t)$ (veh/s), *density* $\rho(x,t)$ (veh/m) and *velocity* $v(x,t)$ (m/s). Flow and density are empirically related by the *fundamental diagram*, as commonly done in traffic modeling [108, 134]. For arterial traffic, it is common to assume that this dependency is piecewise linear, leading to the assumption of a triangular fundamental diagram [54, 112]. The triangular fundamental diagram is illustrated Figure 3.1. It is fully characterized by three parameters: v_f, the free flow speed (m/s); $\rho_{\max}$, the jam (or maximum) density (veh/m); and $q_{\max}$, the capacity (veh/s). Its analytical expression is given by

$$q(\rho) = \begin{cases} v_f \rho & \text{if} \quad \rho \in [0, \rho_c] \\ w(\rho_c - \rho) + v_f \rho_c & \text{if} \quad \rho \in [\rho_c, \rho_{\max}] \end{cases}, \tag{7.1}$$

where $w = \frac{\rho_c v_f}{\rho_{\max} - \rho_c}$. Let ρ_c denote the *critical density*. It is the boundary density value between (i) free flowing conditions for which cars have the same velocity and do not interact and (ii) congested conditions for which the density of vehicles forces them to slow down and the flow to decrease.

Definition 7.1 (**Stream of vehicles of density ρ and duration T**). *A stream of vehicle of density ρ and duration T is a group of vehicles characterized by a uniform density ρ. As the arrival or departure streams always travel at free flow speed v_f, the flow within the stream is also uniform. The duration T of the stream is the time it takes for all vehicles within the stream to go through a point in space.*

Definition 7.2 (**Undersaturated/congested regime**). *The presence of traffic signals leads to the formation of queues during the red time which start to dissipate as the signal turns green. If the queue fully dissipates before the end of the green time, the traffic conditions are* undersaturated. *Otherwise, the regime is* congested.

Definition 7.3 (**Residual green time**). *In an undersaturated arterial link, the residual green time is the period of time between the end of the queue dissipation and the beginning of the red time.*

The solution of the Lighthill-Whitham-Richards partial differential equation when the fundamental diagram is triangular (as defined by (7.1)) and the arrival rate is constant is computed analytically. The analytical solution is represented in a space-time diagram, as shown in Chapter 4. For convenience, the space-time diagram is also displayed in Figure 7.1. Note the three distinct streams of the departure flow in this figure: (1) the red time with flow zero and duration R, (2) the queue dissipation time with flow at capacity $q_{\max} = \rho_c v_f$ and duration τ, and (3) the *residual green time* with flow equal to the arrival flow and duration $C - R - \tau$, where C is the cycle length. Note that in the congested regime, the duration of the third state is zero since there is no residual green time. Also note that the speed of the back propagating wave for queue dissipation is denoted by w, and that for queue formation is denoted by v_a.

As the departure streams of a link correspond to the arrival streams of its downstream link, model the arrival traffic as three streams, characterized by their density ρ_i and their duration T_i, $i \in \{1, 2, 3\}$. If the arrival traffic includes three streams, the departure traffic is not necessarily three streams, as shown in the Figure 7.2(d). The density of traffic during the residual green time may come from different streams and may not be uniform. This fact makes an exact modeling of the flow intractable. The number of streams increases in an unbounded fashion as the flow traverses the successive intersections of an arterial route. To reduce the number of parameters required to describe the system and make the model tractable, the output flow is modeled as a three stream flow. The three different streams correspond to the three stages of the output flow: the red time, the queue dissipation time and the residual green time. The density of traffic during the residual green time is approximated as a stream with uniform density. The average density during the extra green time is denoted ρ_f and computed analytically in the remainder of this section. This assumption is all the more appropriate if street segments are long and vehicle streams of different densities merge into one stream with uniform density [71].

The conservation of vehicles yields the following equation for the average density ρ_f of the last departure stream (of duration $C - R - \tau$):

$$\underbrace{\sum_{i=1}^{3} v_f \rho_i T_i}_{\text{Arrival streams}} = \underbrace{0 \cdot R}_{\text{Red time}} + \underbrace{v_f \rho_c \tau}_{\text{Queue dissipation time}} + \underbrace{v_f \rho_f (C - R - \tau)}_{\text{Residual green time}}.$$

Note that the triangular fundamental diagram yields a simple relation between the flow q and the density ρ as $q = v_f \rho$ whenever $\rho \leq \rho_c$. It yields the following expression for ρ_f

$$\rho_f = \frac{\sum_{i=1}^{3} \rho_i T_i - \rho_c \tau}{C - R - \tau}. \tag{7.2}$$

Note that, the density ρ_f depends on the duration of the queue dissipation τ. The expression of τ is derived analytically as a function of the arrival streams $(\rho_i, T_i)_{i=1:3}$ in Section 7.1.2.

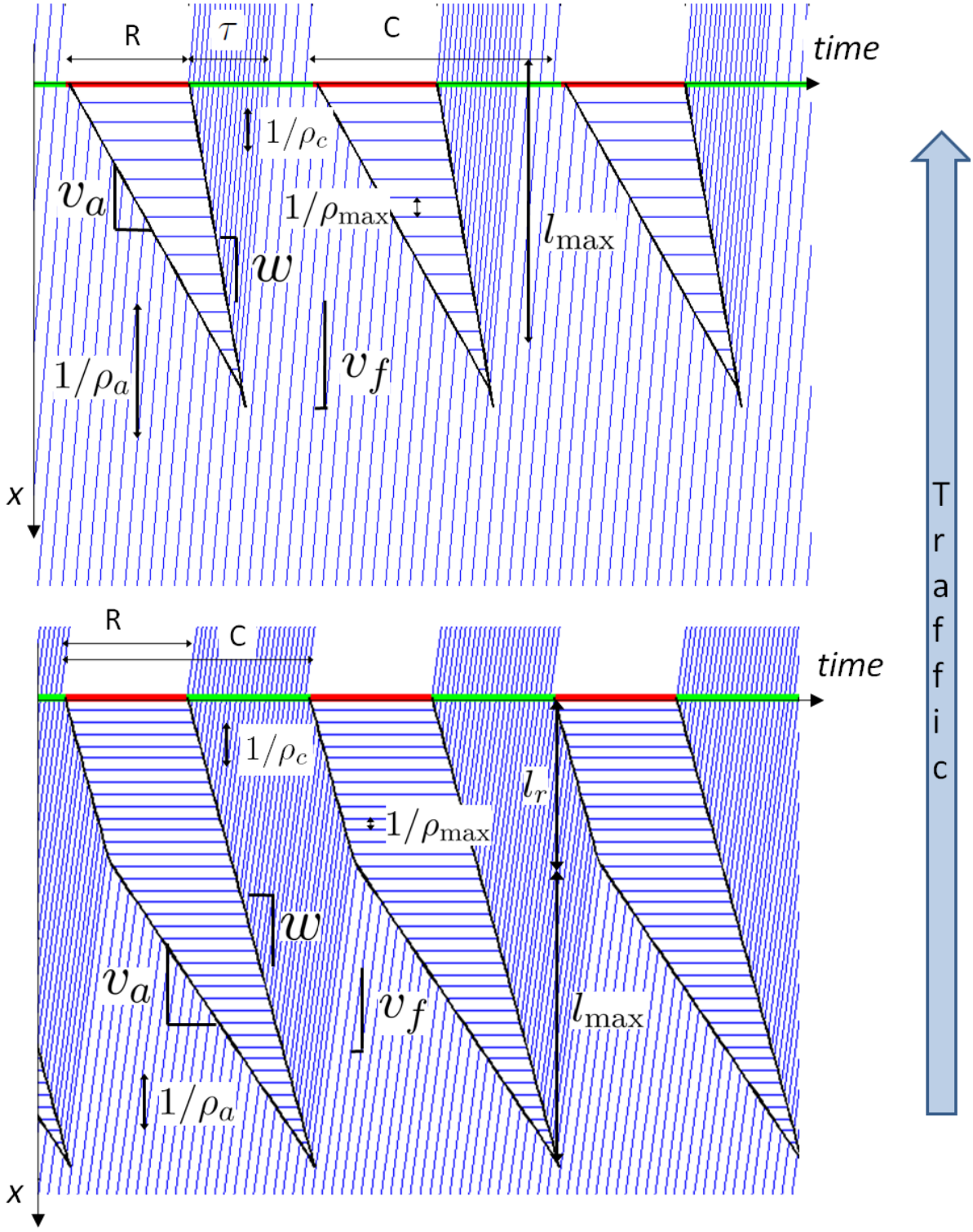
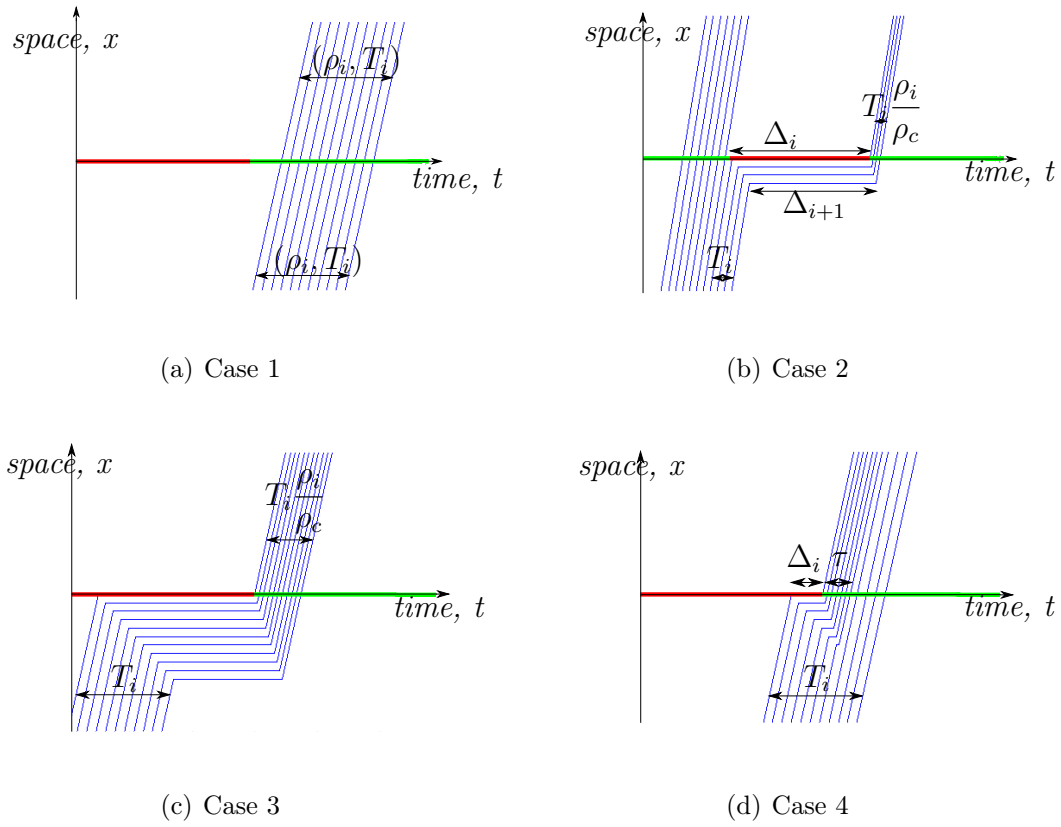


Figure 7.1: Space time diagram of vehicles trajectories under uniform arrivals of density ρ_a. **Top:** undersaturated regime. **Bottom:** Congested regime.

7.1.2 Dynamics of a stream through an intersection

Given an arrival stream of density ρ_i and duration T_i, its dynamics through the intersection follows one of the four cases:

Case 1. No vehicle of the stream stops in the queue. There is one departure stream with the same characteristics as the arrival stream, (ρ_i, T_i).



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 7.2: Dynamic of streams of vehicles through an intersection. Figure 7.2(a): All the vehicles of the stream go through the intersection without stopping (Case 1). Figure 7.2(b): The first few vehicles of the stream do not stop at the intersection, they represent a fraction α of the vehicles of the stream (Case 2). Figure 7.2(c): All the vehicles of the stream stop at the intersection (Case 3). Figure 7.2(d): The last few vehicles of the stream do not stop at the intersection, they represent a fraction α of the vehicles of the stream (Case 4).

Case 2. The first vehicles of the stream go through the intersection without stopping but some vehicles at the end of the stream stop in the queue. We denote by α the fraction of vehicles arriving in stream i that go through the link without stopping. Note that at most one arrival stream follows this case during a cycle. Downstream of the traffic signal, there are three departure streams: the non-stopping vehicles $(\rho_i, \alpha T_i)$, the red time stream $(0, R)$ and the stopping vehicles released at capacity during the queue dissipation $(\rho_c, (1-\alpha)T_i\frac{\rho_i}{\rho_c})$. This case is illustrated in Figure 7.2(b).

Case 3. All the vehicles of the stream stop at the red light. There is one departure stream corresponding to the queue dissipation of these vehicles. It has characteristics $(\rho_c, T_i\frac{\rho_i}{\rho_c})$.

Case 4. The first vehicles of the stream stop in the queue but the last ones go through the intersection without stopping. As for Case 2, we denote by α the fraction of vehicles of the stream that do not stop in the queue. The derivation of the departure streams is similar to Case 2: the stopping vehicles released at capacity during the queue dissipation $(\rho_c, (1-\alpha)T_i\frac{\rho_i}{\rho_c})$ and the non stopping stream $(\rho_i, \alpha T_i)$. This case is illustrated in Figure 7.2(d).

Note that there is a fifth possible case, corresponding to a combination of Cases 2 and 4. In this last case, the first vehicles of the stream go through the intersection without stopping. The middle of the stream stop in the queue. The last vehicles of the stream arrive after the queue dissipation and do not stop. This additional case will be incorporated in the modeling framework of Case 4 in the following.

Let Δ_i denote the delay experienced by the first vehicle of stream i. If the arrival flow is uniform, the speed of queue formation is constant and is denoted v_i. The speed of queue dissipation, w, is also constant. They can be derived from the Rankine-Hugoniot [133] jump conditions as

$$v_i = \frac{\rho_i v_f}{\rho_{\max} - \rho_i} \text{ and } w = \frac{\rho_c v_f}{\rho_{\max} - \rho_c} \tag{7.3}$$

Remark that the inequality $w \geq v_i$ necessarily hold. It follows that the delay decreases linearly among the vehicles of the stream. If the queue does not fully dissipate as the last vehicle in stream i arrives (Cases 2 and 3), this last vehicle will experience a delay $\Delta_{i+1} = \Delta_i - T_i(1 - \frac{\rho_i}{\rho_c})$ (see Figure 7.2(b)). This expression is valid if and only if $\Delta_i \geq T_i(1 - \frac{\rho_i}{\rho_c})$. If this condition is not satisfied (Case 4, Figure 7.2(d)), the queue dissipates before the end of stream i and the last vehicles of the stream do not experience delay. The general expression for Δ_{i+1} is

$$\Delta_{i+1} = \max\left(0, \Delta_i - T_i\left(1 - \frac{\rho_i}{\rho_c}\right)\right). \tag{7.4}$$

We introduce θ_i such that θ_i/T_i represents the fraction of stream i which stops at the intersection and have

$$\theta_i = \min\left(\Delta_i \frac{\rho_c}{\rho_c - \rho_i}, T_i\right). \tag{7.5}$$

7.1.3 Characterization of the departure streams

This section derives analytical expressions for the densities and durations of the departure streams, parameterized by the characteristics of the arrival streams. Without loss of generality, the sections assumes that the signal turns red at $t = 0$, and that stream 1 hits the red light at the beginning of the cycle.

A fraction $1 - \alpha$ of the vehicles of stream 1 reaches the intersection after the signal turns red whereas the remaining vehicles reach the intersection before the signal turns red. In this model, the arterial traffic dynamics are considered as periodic (as done in Chapter 4). With this consideration, it is natural to consider that the remaining vehicles reach the intersection at the end of the cycle. This second representation simplifies the notation in the derivations. The arrival streams are thus modeled as four streams with densities ρ_i and duration $\widetilde{T}_i$ with $\widetilde{T}_1 = (1-\alpha)T_1$, $\widetilde{T}_2 = T_2$, $\widetilde{T}_3 = T_3$, $\widetilde{T}_4 = \alpha T_1$ and $\rho_4 = \rho_1$.

In a corridor with several signalized intersections, α is determined by the offset between consecutive signals. The delay experienced by the first vehicle that stops at the signal is $\Delta_1 = R$.

The expressions of $(\Delta_i)_{i=1:5}$ and $(\theta_i)_{i=1:4}$ are computed for the four streams according to equations (7.4) and (7.5), with the initialization $\Delta_1 = R$ (see Figure 7.3). It follows that:

$$
\begin{aligned}
\Delta_1 &= R \\
\Delta_2 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c})\right) \\
\Delta_3 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c}) - \widetilde{T}_2(1-\tfrac{\rho_2}{\rho_c})\right) \\
\Delta_4 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c}) - \widetilde{T}_2(1-\tfrac{\rho_2}{\rho_c}) - \widetilde{T}_3(1-\tfrac{\rho_3}{\rho_c})\right) \\
\Delta_5 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c}) - \widetilde{T}_2(1-\tfrac{\rho_2}{\rho_c}) - \widetilde{T}_3(1-\tfrac{\rho_3}{\rho_c}) - \widetilde{T}_4(1-\tfrac{\rho_4}{\rho_c})\right)
\end{aligned}
\tag{7.6}
$$

and

$$
\begin{aligned}
\theta_1 &= R\frac{\rho_c}{\rho_c - \rho_1} \\
\theta_2 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c})\right)\frac{\rho_c}{\rho_c - \rho_2} \\
\theta_3 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c}) - \widetilde{T}_2(1-\tfrac{\rho_2}{\rho_c})\right)\frac{\rho_c}{\rho_c - \rho_3} \\
\theta_4 &= \max\left(0, R - \widetilde{T}_1(1-\tfrac{\rho_1}{\rho_c}) - \widetilde{T}_2(1-\tfrac{\rho_2}{\rho_c}) - \widetilde{T}_3(1-\tfrac{\rho_3}{\rho_c})\right)\frac{\rho_c}{\rho_c - \rho_1}
\end{aligned}
\tag{7.7}
$$

The intersection modifies the structure of the three arrival streams into several departure streams as follows:

$$
\overbrace{\left\{\begin{array}{c}(\rho_1, T_1)\\(\rho_2, T_2)\\(\rho_3, T_3)\end{array}\right\}}^{\text{Arrival streams}} \longmapsto \overbrace{\left\{\begin{array}{c}\left(0, R\right)\\\left(\rho_c, \min(\widetilde{T}_1, \theta_1)\frac{\rho_1}{\rho_c}\right)\\\left(\rho_1, \max(0, \widetilde{T}_1 - \theta_1)\right)\\\left(\rho_c, \min(\widetilde{T}_2, \theta_2)\frac{\rho_2}{\rho_c}\right)\\\left(\rho_2, \max(0, \widetilde{T}_2 - \theta_2)\right)\\\left(\rho_c, \min(\widetilde{T}_3, \theta_3)\frac{\rho_3}{\rho_c}\right)\\\left(\rho_3, \max(0, \widetilde{T}_3 - \theta_3)\right)\\\left(\rho_c, \min(\widetilde{T}_4, \theta_4)\frac{\rho_1}{\rho_c}\right)\\\left(\rho_1, \max(0, \widetilde{T}_4 - \theta_4)\right)\end{array}\right\}}^{\text{Departure streams}}
\tag{7.8}
$$

For simplicity, the presentation of the three stream model does not take into account side street traffic, *i.e.* vehicles exiting or entering the route at intersections. The derivations presented in this chapter can be generalized to take into account side street traffic. For example, it may be natural to add two parameters: one representing the average (potentially negative) flow corresponding to the balance of exiting and arriving vehicles during the red time and another one representing the average (potentially negative) flow corresponding to the balance of exiting and arriving vehicles during the green time. Another modeling approach might consider that there is a fixed turning ratio at each intersection and take these parameters into account in a network setting. Other characterization of side street traffic are possible and not detailed here. Section 7.2 analyzes how side street traffic may affect the model and its perturbation on signal optimization plans.

As seen in (7.8), the number of departure streams can be more than three. Indeed, each stream i, $i \in \{1, \ldots, 4\}$ leads to up to two streams: a stream representing the queue discharge if $\Delta_i > 0$ (otherwise this stream has duration zero) and a stream representing the vehicles which do not stop if $\Delta_{i+1} = 0$ (otherwise this stream has duration zero). This leads to up to eight streams to which we add the red phase of the signal which

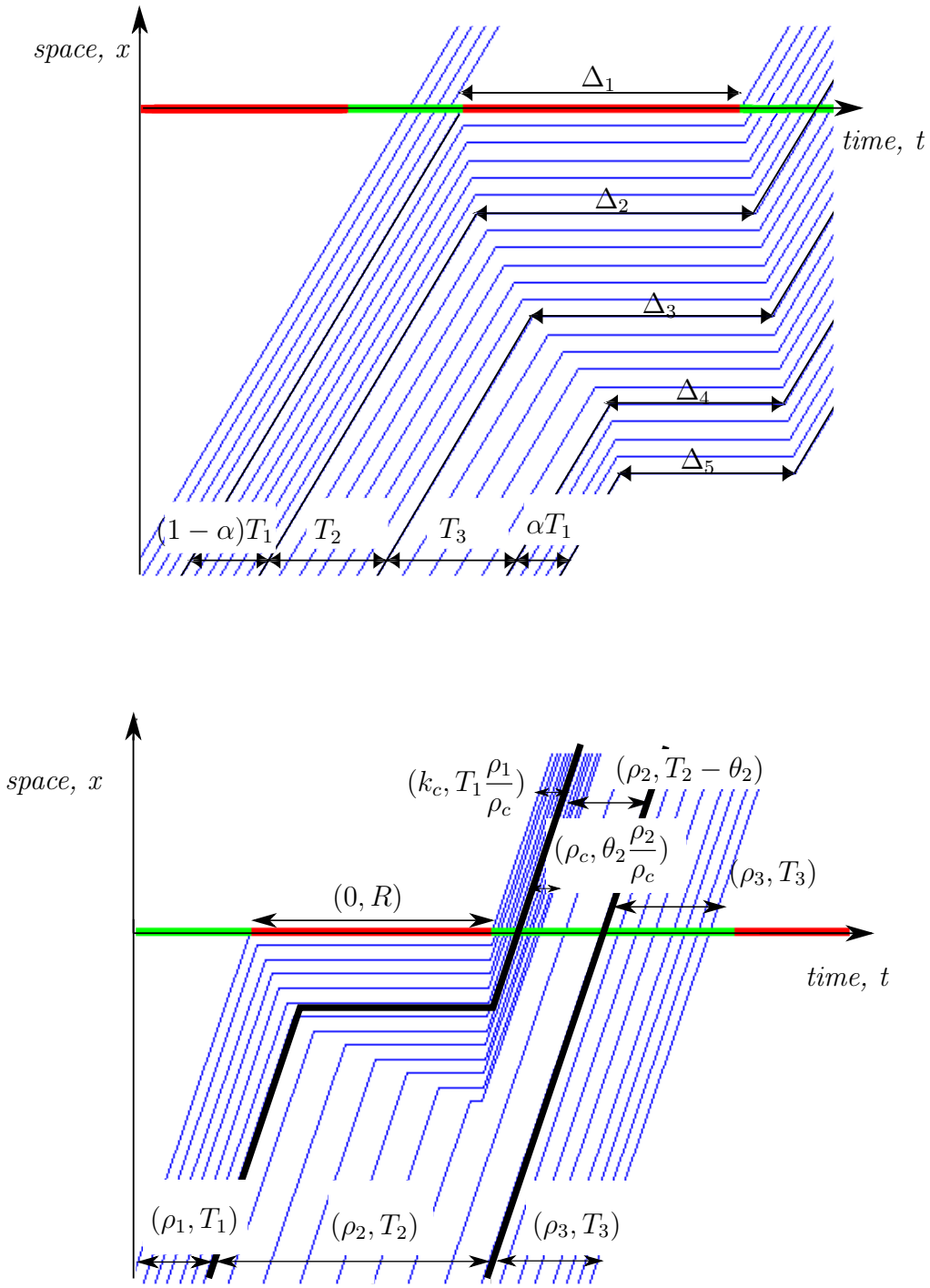


Figure 7.3: **Top:** Arrival streams of vehicles. The stream that reaches the signal as the traffic light turns red is split between two streams denoted stream 1 and stream 4. Stream 1 has duration $(1-\alpha)T_1 = \widetilde{T}_1$. It reaches the intersection as the signal turns red. Stream 4 has duration $\alpha T_1 = \widetilde{T}_4$. It reaches the intersection at the end of the cycle. The waiting times of the first and last vehicles of stream i are denoted Δ_i and Δ_{i+1}. Note that the Δ_i can be null. In particular, in an undersaturated regime, we have $\Delta_5 = 0$ since the queue fully dissipates as the signal turns red. **Bottom:** Dynamic of three arrival streams through a signalized intersection, illustrating equation (7.8)

creates a ninth stream of density 0 and duration R. Modeling the exact progression of the characteristics of streams through a corridor rapidly leads to an over-parameterized model. The level of details of the modeling does not match the precision of available data. It also assumes a perfect, non-noisy model of traffic which is also unrealistic. To limit the number of parameters and control the complexity of the model, the departure streams listed above are approximated by three departure streams, corresponding to the red time, the queue dissipation time, and the residual green time. The red time leads to a stream of density 0 and duration R. The queue dissipation leads to a stream of density ρ_c and duration τ and the multiple streams of the residual green time are aggregated into a single stream of density ρ_f and duration $C-(R+\tau)$, as derived in (7.2). The densities and durations of the three departure streams are given by

$$\overbrace{\left\{\begin{array}{c}(\rho_1, T_1)\\(\rho_2, T_2)\\(\rho_3, T_3)\end{array}\right\}}^{\text{Arrival streams}} \longmapsto \overbrace{\left\{\begin{array}{c}\Big(0, R\Big)\\\Big(\rho_c, \tau\Big)\\\Big(\rho_f, C-R-\tau\Big)\end{array}\right\}}^{\text{Averaged departure streams}} \tag{7.9}$$

with

- $\tau = \min(\alpha T_1, \theta_1)\frac{\rho_1}{\rho_c} + \min(T_2, \theta_2)\frac{\rho_2}{\rho_c} + \min(T_3, \theta_3)\frac{\rho_3}{\rho_c} + \min((1-\alpha)T_1, \theta_4)\frac{\rho_1}{\rho_c}$ the duration of the queue dissipation,
- ρ_f the merging density which only depends on τ and the parameters of the intersection as computed in (7.2).

7.2 Application to the optimization of traffic signals

The model described in Section 7.1 provides a framework to analyze the dynamics of traffic flows through an arterial corridor. The assumptions lead to analytical derivations and a better understanding of the dynamics, providing insight for the control of arterial networks. This section leverages this framework to analyze the well studied problem of one way corridor signal optimization. The derivations lead to analytical optimal control strategies for different scenarios of the arrival streams. This allows for timely adjustments of the control strategy in real time as congestion changes throughout the day.

7.2.1 Problem setting

The derivations focus on the minimization of the total delay D experienced at an intersection. The total delay is defined as

$$D = \int_0^C W(t)q(t)dt = \int_0^C W(t)v_f\rho(t)dt, \tag{7.10}$$

where $W(t)$ is the delay experienced by the flow entering at time t, $q(t)$ and $\rho(t)$ are the flow and the density of the stream that enters at time t. C is the cycle length assumed to have the same value for all signals.

The algorithm optimizes the *total delay* because it finds a compromise between the duration of the delay experienced by the stopping vehicles and the proportion of vehicles that go through the intersection without stopping. Other choices of optimization problems are possible such as the maximization of the number of vehicles going through the intersection without stopping or the minimization of the maximal delay.

The objective function has an analytical expression, which is derived under the assumption that vehicles arrive from an upstream intersection with a three-stream structure. The cost function is additive, in the sense that the contribution of each stream to the total delay can be computed independently. The individual contributions of each stream are summed to get the total delay.

As derived in Section 7.1.2, the delay decreases linearly among the stopping vehicles of a stream i (from the first stopping vehicle with delay Δ_i to the last stopping vehicle with delay Δ_{i+1}). The total delay experienced by the vehicles of a stream is the average delay of the stopping vehicles times the number of stopping vehicles. According to the definition of θ_i, the number of vehicles stopping in the queue is $\rho_i v_f \theta_i$ and the minimum and maximum delays of the stopping vehicles of stream i are given by Δ_{i+1} and Δ_i respectively (see Figure 7.3).

Remark 7.1 (**Control variables**). *In traffic signal optimization, the control variables are the duration of the red light and the offset between the two traffic signals. In a one way corridor, it is not relevant to minimize according to the duration of the red time because, without any constraints, the optimal value of the objective function is zero, corresponding to a red time equal to zero. For this reason, only the actual offset Θ between the two traffic signals is considered as a control variable. The standardized offset t_0 is defined as $t_0 = \Theta - \frac{L}{v_f}$. The standardized offset takes into account the free flow travel time of vehicles along the link. Here, L represents the length of the link between the two intersections.*

The standardized offset t_0 is related to α by $t_0 = (1-\alpha)T_1$. This gives the explicit expression of the total delay as a function of t_0, denoted $D(t_0)$. Moreover, the offset t_0 determines which stream hits the signal first. This leads to an implicit dependence represented by the cyclic permutation between the streams, so that the stream that reaches the intersection as the signal turns red is denoted 1. The analytical expression of the total delay $D(t_0)$ is the sum of the contributions of the three arrival streams. The previous derivations yield to:

$$D = v_f \Bigg[\rho_1 \min(\theta_1, T_1 - t_0)\frac{\Delta_1 + \Delta_2}{2} + \rho_2 \min(\theta_2, T_2)\frac{\Delta_2 + \Delta_3}{2} + \rho_3 \min(\theta_3, T_3)\frac{\Delta_3 + \Delta_4}{2} + \rho_1 \min(\theta_4, t_0)\frac{\Delta_4 + \Delta_5}{2}\Bigg] \quad (7.11)$$

In the case of a congested regime, all vehicles experience some delay. Let $\Delta_{\min}$ represent the minimum delay experienced by the vehicles on the link, then the total delay is given by $D_{\text{sat}} = D + \Delta_{\min} v_f \sum_{i=1}^{3} \rho_i T_i$. Only the first term of the sum, D, depends on t_0, it is therefore equivalent to minimize D or D_{sat}. Without loss of generality equation (7.11) is used to minimize the total delay in both the undersaturated and the congested regime.

7.2.2 Convexity of the cost function

Equation (7.7) implies that for all i, $\theta_i \leqslant \theta_{i-1}$. In particular, if there exists j such that $\theta_j = 0$, then $\theta_m = 0$ for any m greater than or equal to j. Moreover, $\Delta_m = 0$ for $m \geq j$ since $\theta_m = \dfrac{\rho_c}{\rho_c - \rho_m}\Delta_m$.

Proposition 7.1 (**Analytical expression of** D). *In an undersaturated regime, for all t_0, there exists a unique $m \in \{1, \ldots, 4\}$ such that $0 < \theta_m \leq \widetilde{T}_m$. The expression of the cost function is simplified as follows:*

$$D = v_f \sum_{i=1}^{m-1} \rho_i \widetilde{T}_i \frac{\Delta_i + \Delta_{i+1}}{2} + \rho_m \frac{\rho_c}{\rho_c - \rho_m}\frac{\Delta_m^2}{2} \quad (7.12)$$

Proof. Intuitively, the index m represents the stream of vehicle from Case 4, for which the last vehicles of the stream do not stop on the queue. The proof of existence and uniqueness of this index m is based on two intermediate results (Lemma 7.1 and Lemma 7.2).

Lemma 7.1. *For all i greater than or equal to 2, $\tau_i > 0$ if and only if $\theta_{i-1} > \widetilde{T}_{i-1}$.*

Proof. Replacing θ_i by its expression (Equation (7.5)), multiplying the strict inequality, $\theta_i > 0$, by the positive term $\frac{\rho_c - \rho_i}{\rho_c}$ and rearranging the sum yield the following inequality

$$R - \sum_{n=1}^{i-2} \widetilde{T}_n(1 - \frac{\rho_n}{\rho_c}) > \widetilde{T}_{i-1}(1 - \frac{\rho_{i-1}}{\rho_c}).$$

Multiplying this inequality by $\frac{\rho_c}{\rho_c - \rho_{i-1}}$, it follows that

$$\left(R - \sum_{n=1}^{i-2} \widetilde{T}_n(1 - \frac{\rho_n}{\rho_c})\right) \frac{\rho_c}{\rho_c - \rho_{i-1}} > \widetilde{T}_{i-1}$$

and in particular $\theta_{i-1} > \widetilde{T}_{i-1} > 0$. □

Lemma 7.2. *For all i less than or equal to 3, $\theta_i \leq \widetilde{T}_i$ if and only if $\theta_{i+1} \leq 0$.*

Proof. Replacing θ_i by its expression (Equation (7.5)) and multiplying the inequality, $\theta_i \leq \widetilde{T}_i$, by the positive term $\frac{\rho_c - \rho_i}{\rho_c}$ yield the following inequality

$$R - \sum_{n=1}^{i} \widetilde{T}_n(1 - \frac{\rho_n}{\rho_c}) \leq 0$$

By multiplying the inequality by the positive term $\frac{\rho_c}{\rho_c - \rho_{i+1}}$, the expression of θ_{i+1} from Equation (7.5) appears. Since, θ_{i+1} is non negative by definition, θ_{i+1} is necessarily equal to zero. It follows that $\theta_{i+1} \leq \widetilde{T}_{i+1}$. □

Proof of the existence and uniqueness of m: the proof is in two steps. The first step proves that if such an m exists, it is necessarily unique. The second step proves the existence of such an m.

- *Uniqueness.* Let m be an index such that $0 < \theta_m \leq \widetilde{T}_m$. By induction, Lemma 7.1 and 7.2 imply that $\forall\, j < m,\ \theta_j > \widetilde{T}_j > 0$ and $\forall\, j > m,\ \theta_j = 0 \leq \widetilde{T}_j$. This proves that if m exists, it is unique.

- *Existence.* Let j be defined by $j = \max\{n \in \{0, \ldots, 4\} | \theta_n > \widetilde{T}_n\}$, where θ_0 and $\widetilde{T}_0$ are chosen arbitrarily such that $\theta_0 > \widetilde{T}_0$. The proof shows that $m = j + 1$.

  In an undersaturated regime, $\theta_4 \leqslant t_0$, so $j \leq 3$. The condition $\theta_0 > \widetilde{T}_0$ implies that $j \geq 0$ and thus the definition of j is proper (j is not infinite). The maximality of j implies that $\theta_{j+1} \leq \widetilde{T}_{j+1}$. Lemma 7.2 implies that $\forall\, i \geq j + 2,\ \theta_i = 0$. It remains to prove that $\theta_{j+1} > 0$. Reasoning by contradiction, assume that $\theta_{j+1} = 0$.

  - If $j = 0$, this implies that $\forall\, n \in \{1, \ldots 4\},\ \theta_n = 0$ which means that no vehicle experiences delay and contradicts the assumption $\theta_{j+1} = 0$ as long as the red time is positive. and thus $\forall\, i \geq j + 2,\ \theta_i = 0$.
  - If $j \geq 1$, then Lemma 7.2 implies that $\theta_j \leq \widetilde{T}_j$, which contradicts the maximality of j.

It is now possible to conclude that $\theta_{j+1} > 0$ and thus $m = j+1$ is the unique index such that $0 \leq \theta_m \leq \widetilde{T}_m$.

□

Remark 7.2. *The index m is piecewise constant in t_0 and thus the expression of D holds on each of these intervals. Physically, m represents the index of the first stream in which some vehicles go through the intersection without stopping. Moreover, the expression holds in the case of a congested regime, with $m = 5$ and the convention $\rho_5 = 0$.*

Proposition 7.2 (**Property of** D)**.** *The function $t_0 \mapsto D(t_0)$ is piecewise quadratic.*

Proof. The analytical expression of the cost function $D(\cdot)$ over an interval in which m is constant is given by Proposition 7.1. Both the Δ_is and $\widetilde{T}_1$ are linear in t_0. All the terms of the sum from $i = 2$ to $m-1$ are linear in t_0. The first term of the sum is quadratic in t_0. Therefore, D is the sum of a quadratic term and of linear terms and is quadratic on each interval in which m is constant. On each of these intervals, the expression of $D(t_0)$ is written as a quadratic function: $D(t_0) = at_0^2 + bt_0 + c$ with

$$a = \frac{(\rho_c - \rho_1)(\rho_m - \rho_1)}{2(\rho_c - \rho_m)} \tag{7.13}$$

$$b = -\frac{R\rho_c(\rho_1 - \rho_m) - \sum_{i=1}^{\rho-1} T_i(\rho_c - \rho_1)(\rho_i - \rho_m)}{\rho_c - \rho_m} \tag{7.14}$$

and the optimum (either a minimum or a maximum according to the sign of a) is reached in:

$$-\frac{b}{2a} = \sum_{i=1}^{m-1} T_i \frac{\rho_m - \rho_i}{\rho_m - \rho_1} - R\frac{\rho_c}{\rho_c - \rho_1} \tag{7.15}$$

□

The function $D(\cdot)$ is piecewise quadratic. The study of its monotonicity on each interval in which m is constant leads to the characterization of the global optimum.

Definition 7.4 (**Quasi-convex function** [34])**.** *A function $f : D_f \to \mathbb{R}$ is called quasi-convex if its domain D_f and all its sublevel sets $S_{f_\alpha} = \{x \in D_f : f(x) \leq \alpha\}$ for $\alpha \in \mathbb{R}$ are convex. In particular, a function is quasi-convex if one of the following conditions hold: (1) f is non decreasing, (2) f is non increasing, (3) $\exists c \in D_f$ such that for $t \leq c$ (and $t \in D_f$), f is nonincreasing, and for $t \geq c$ (and $t \in D_f$), f is nondecreasing.*

Proposition 7.3 (**Quasi-convexity property**)**.** *If the time initialization is chosen such that $t = 0$ as the beginning of the stream with the highest density enters the link, $t_0 \mapsto D(t_0)$ is a quasi-convex function on $[0, C]$.*

Proof. Sketch of the proof, the full proof is available in [21].
The proof studies the monotonicity of D over each interval corresponding to the three arrival streams –*i.e.* over $[0, T_1], [T_1, T_1 + T_2], [T_1 + T_2, C]$– and proves that there exists t_c such that the function $t_0 \mapsto D(t_0)$ is non increasing for $t_0 \in [0, t_c]$ and non decreasing for $t_0 \in [t_c, C]$. The cost function is nonincreasing over the interval corresponding to the arrival stream with the highest density, it is nondecreasing over the interval corresponding to arrival stream with the lowest density and the behavior over the last interval is such that the minimum is either reached over this interval or at the bounds of this interval. It may be reached outside of this interval if the interval over which the cost function is nonincreasing and the interval over which the cost function is nondecreasing are consecutive. The enumeration of all possible cases proves the quasi-convexity of the function. □

7.2.3 Optimization of a one-way corridor

The variations of $D(\cdot)$ on $[0, C]$, derived in the proof of Proposition 7.3 provide analytical solutions for the optimal control (choice of the offset t_0) analytically. The following derivations are based on the definition of two families of control solutions: (i) the *corner* solutions in which t_0 corresponds to the beginning of a stream and (ii) the solutions in which t_0 lies inside the arrival time of a stream. The latter solutions only exist if the optimal t_0 is such that the first stream which stops at the signal is the one with the intermediate density, (see [21] for details). The intermediate density is indexed by 1. In the following, the convention $\rho_2 \leq \rho_1 \leq \rho_3$ is adopted. The optimal t_0 is denoted t_0^*.

In corridor optimization, the offset of traffic signals is optimized over several consecutive intersections. Optimizing the sum of the total delays at each intersection over each offset is a difficult problem to solve analytically. Instead, the algorithm solves an optimization problem for each intersection. Given the departure streams resulting from the optimal control at intersection i (arrival streams of the downstream intersection $i+1$), the algorithm computes the optimal control to be applied at intersection $i+1$. A *scenario* is defined as a class of arrival streams leading to a specific choice of t_0^*, denoted *control strategy*. A scenario s is *unstable* if it leads to a different scenario at the downstream intersection. The scenario of intersection i is unstable if either the structure of the arrival streams or the optimal control strategy of intersection $i+1$ is different from the structure of the arrival streams or the optimal control of intersection i. On the contrary, a scenario is *stationary* if, once this scenario occurs at an intersection, it will occur at all the downstream intersections. The derivations identify specific conditions on the arrival streams for each control strategy to be the optimal one. The derivations also identify conditions for scenario to be stationary or unstable. The details of the derivations can be found in [21]. The most insightful part of the derivation is the interpretation of the results.

The Optimal Control Is a Corner Solution

The optimal control t_0^* is either $0, T_1$, or $T_1 + T_2$. One of the three streams is coordinated such that its first car reaches the signal at the beginning of the red time. Intuitively, this stream should have the lowest density. However, the following stream, which may have a density close to ρ_c, can join the queue before it fully dissipates, causing a rapid increase in the queue length and thus in the total delay. Depending on the densities of the streams and on how many streams join the queue, the corner solution can be either of the three possibilities. Figure 7.4 summarizes the different scenarios representing an optimal control strategy associated with a class of arrival streams.

The Optimal Control Is Not a Corner Solution

There is only one scenario in which the optimal control is not a corner solution, then $t_0^* = T_1 + T_2\dfrac{\rho_3 - \rho_2}{\rho_3 - \rho_1} - R\dfrac{\rho_c}{\rho_c - \rho_1}$. In this scenario, the first stream, with the intermediate density, is split into: a stream which does not stop in the queue (stream 4) and a stream which reaches the intersection as the signal turns red (stream 1). As the offset increases, additional vehicles from the first stream experience long delays. These long delays are not compensated by the smaller number of vehicles from the third stream (with the highest density) which experience short delays. As the offset decreases, fewer vehicles from the first stream (intermediate density) experience delay. This reduction in the total delay for the first stream is overcompensated by the significant increase in the total delay experienced by the vehicles from the third stream (with the highest density). This illustrates a trade-off between having a few cars with long delays and a lot of cars with short delays.

7.2.4 Relations between the scenarios and convergence towards a unique stationary optimal control

Each scenario corresponds to a specific structure of the arrival streams. Each intersection modifies this structure, potentially leading to a different control strategy at the downstream intersection. The relationships and transitions from one control strategy to another as a vehicle traverses a corridor can be represented using a directed graph. Each node of the graph represents a control strategy for a class of arrival streams. The arrows indicate the possible transitions from a control strategy to the control strategy at the downstream intersection. The directed graph is represented Figure 7.4.

The green dotted arrows indicate that different evolutions of the control strategies are possible: the scenario at the downstream intersection depends on the parameters of the arrival streams. The solid red arrows indicate that there exists a unique scenario for the optimal control at the downstream intersection. This means that once the scenario occurs, there is a unique scenario possible at the downstream intersection. We notice that all the scenarios converge after a finite number of iterations towards the unique stationary scenario (bottom left of the figure).

Physically, the limit scenario corresponds to what is called a green wave [68]. A green wave is a flow of vehicles going through a series of intersections without stopping at any red light. This result is intuitive. Indeed, at each intersection, one of the departure stream has no vehicles, corresponding to the red light. Because of the conservation of vehicles, the two other streams have a higher density after each intersection, until it reaches the critical density ρ_c.

In a green wave, vehicles are clustered in a single stream of critical density. They arrive at the intersection during the green time and do not experience any delay. This is possible as long as the regime is undersaturated, since the duration of the single stream, at critical density, must be inferior to the duration of the green time. This minimum, expected to be local because we only optimize each intersection individually and not the entire set of intersections at once, is actually a global minimum because the cost function is null, it is not possible to do better.

However, a green wave has limitations when it comes to synchronizing traffic signals because it is very sensitive to external factors. At critical density, the traffic dynamics may be unstable (showing the limits of the modeling of traffic flow with a fundamental diagram). A single incident on the network (pedestrian, parking, accident and so on) or small calibration errors may cause significant delays and the formation of queues.

Real-time dynamic control improves the robustness of the optimization. At each intersection, the optimal control is chosen given the traffic conditions at the upstream intersection (from sensors for instance). This dynamic control strategy anticipates an incident which would have disrupted the green wave. This idea of real-time control traffic has already been studied with real-time computations [104, 14, 135]. Here, all computations can be done off-line and analytically, reducing the online computations to comparisons between parameters, which are quasi-instantaneous.

The presence of side traffic changes the values of the densities of the streams as they traverse each intersection. The structure of the stream (duration of each stream) may remain unchanged. The transitions between the different control strategies presented in Figure 7.4 no longer holds. However, if turning flows or turning ratios can be measured or estimated, the density of the streams can be updated and the derivations of the control strategies remain valid. In control theory, it is important to analyze the stability of the controller to avoid undesirable oscillations of the system. To limit this risk, it is possible to consider piecewise constant controls which average the information of the upstream links for a given interval before applying the control, leading to a smoother feedback which

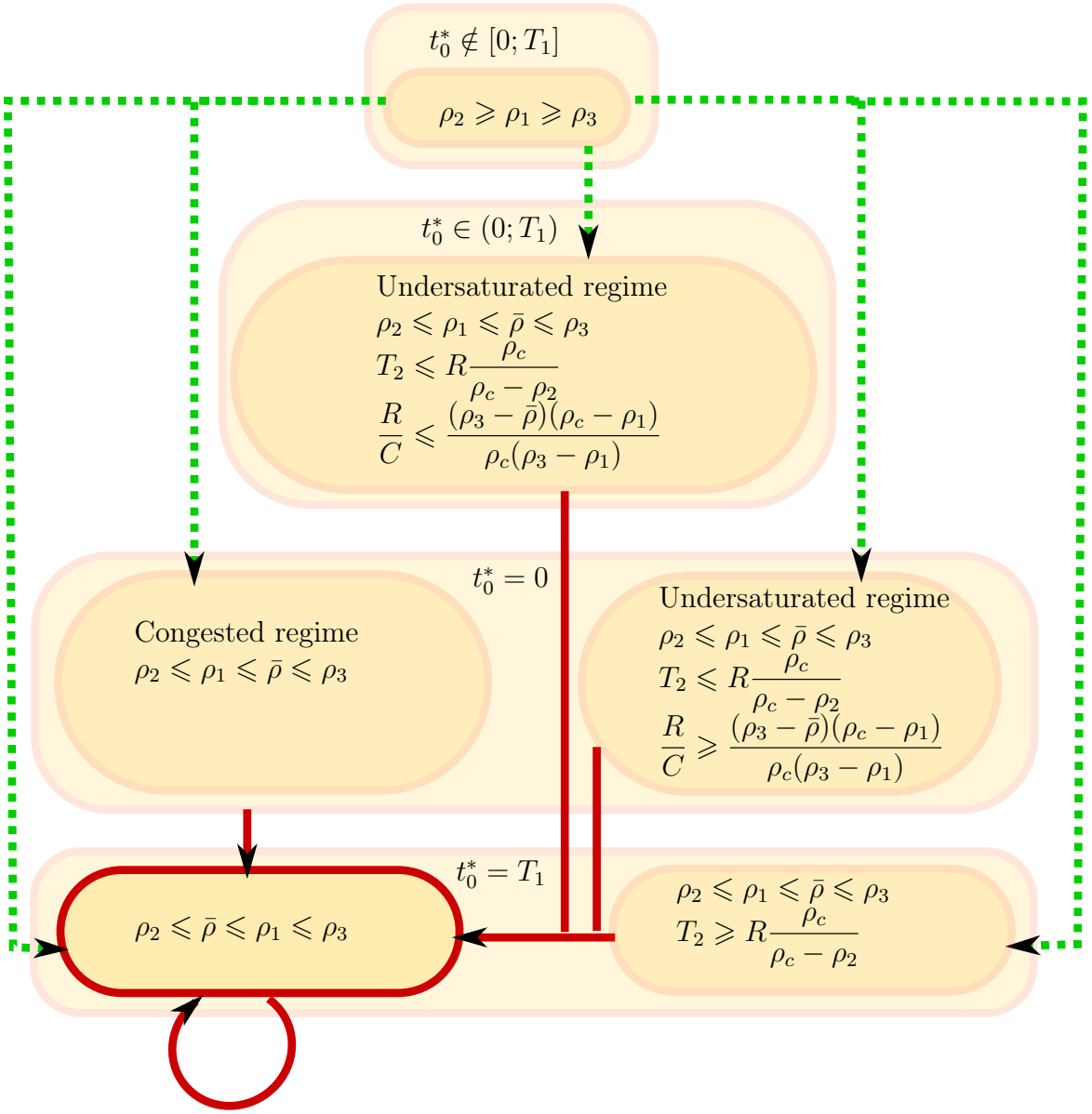


Figure 7.4: The figure represents the different control scenarios (optimal control strategy and corresponding class of arrival streams). It also shows the dynamics of the scenario in a corridor leading to a unique stationary scenario which corresponds to a green wave.

integrates the past dynamics.

7.3 Numerical analysis and validation

This section validates the model with microsimulation data. The flow dynamics predicted by the model are compared with the results from CORSIM [74].

CORSIM simulates an arterial corridor equipped with four signalized intersections. The comparison focuses on one direction of traffic. As the derivations considered traffic from/to side streets to be negligible, the simulation is set such that all the flow movements at the intersections correspond to through traffic. The intersections are numbered from 1 to 4 in the direction of traffic. The settings of the simulation are the following:

- The distance between two consecutive intersections is 500 feet (152.4 meters)
- The cycle has the same duration for every signal and lasts 60 seconds
- Every link is assumed to have one lane only
- Arrival flow upstream of the first intersection is 300 vehicles/hour
- Saturation flow is 2000 vehicles/hour

The arterial corridor is simulated for a range of values of the red time and the offset. For every simulation, the red time is common to every signal and the offset between two consecutive traffic lights is the same on each link. Each simulation is run 10 times for every set of values of the parameters and each simulation lasts 20 cycles. The comparison variable between the simulation and the model is the total delay of all the vehicles, experienced at an intersection, during a cycle. To avoid the effects of initialization, the total delays are averaged over the last 10 cycles of each simulation.

The model considers that the arrival flow upstream of intersection 1 is uniform. Departure streams of each intersection are computed according to Equation (7.9). The departure streams of intersection i are the arrival streams of intersection $i+1$. At each intersection, the total delay per cycle is computed using Equation (7.11).

Figure 7.5 compares the total delays per cycle from the simulation and from the model. The left column represents the results computed between intersections 1 and 2. From top to bottom, the figure represents the total delay per cycle computed by the model, the microsimulation and the difference between the microsimulation and the model. The results are presented as functions of the red time R and the offset t_0. The model underestimates the total delay by about 20% on average. The two surfaces have extremely similar shapes. The total delays computed by the simulation and by the model exhibit a similar dependency on the parameters (red time and offset), which implies that the assumptions of the models are reasonable for signal control.

The model is relevant to obtain better understanding of traffic flow dynamics and study problems for which the value of the objective function is not as important as gaining intuition on the response of a corridor to a change in the values of the parameters. The traffic signal optimization problem is a good application of the model because the key point of this problem is to obtain the value of the optimal control and not the one of the minimal total delay. Even though the minimal value of the total delay is underestimated by about 20% by our model, the optimal control derived by the model and by the simulation are close due to the similar shapes of the surfaces representing the total delay.

From a hydrodynamical theory point of view, if we consider an intersection with uniform arrivals (a single stream of density ρ and duration C), there are exactly three

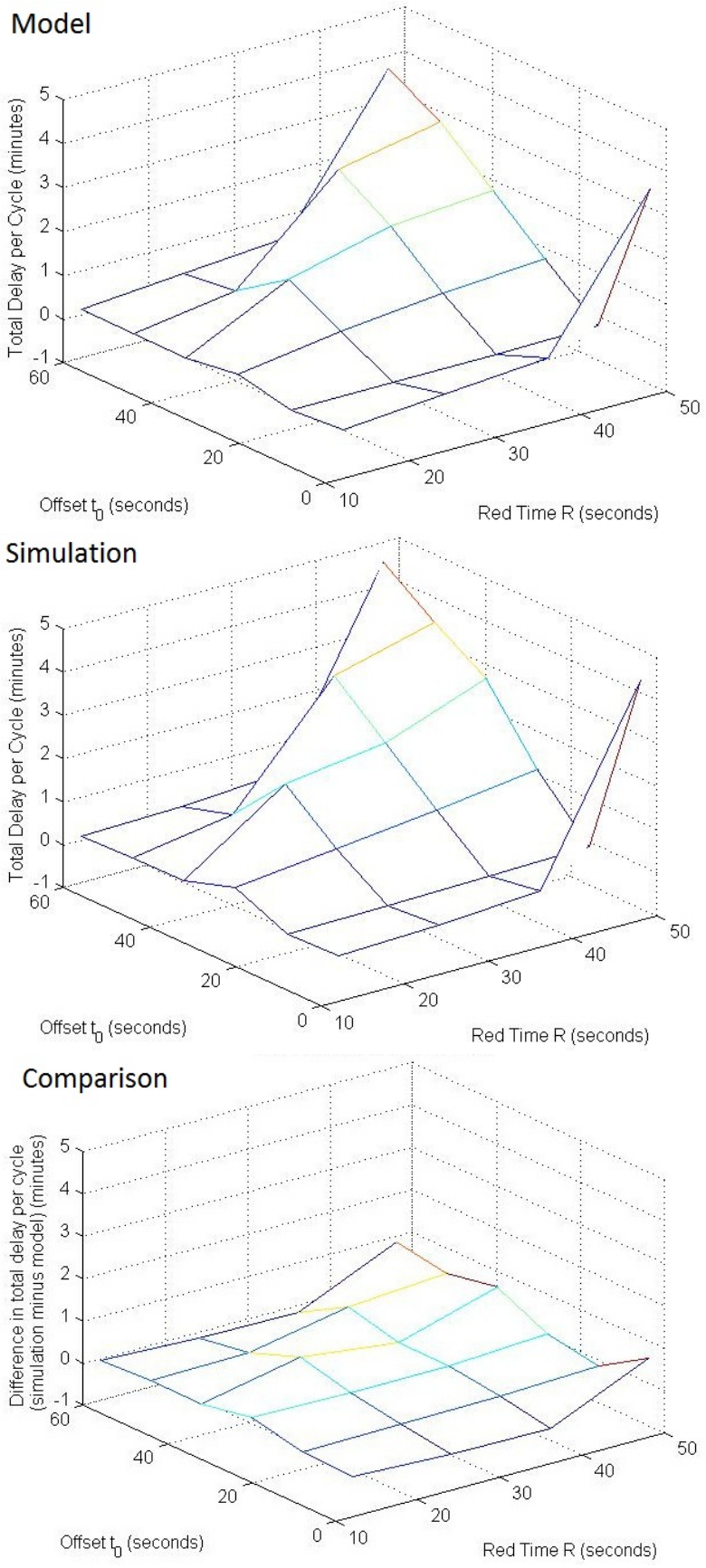


Figure 7.5: Comparison of the total delay computed by the microsimulation and by the model. **Top:** Total delay per cycle computed by the model. **Center:** Total delay per cycle computed by the microsimulation. **Bottom:** Difference between the total delay per cycle computed by the microsimulation and by the model.

streams downstream of the intersection (red time with density zero and duration R, queue discharge with density ρ_c and duration τ and residual green time with density ρ and duration $C-(R+\tau)$). The differences in the computation of the total delay between the model and the microsimulation do not result from the three-stream approximation. The 20% error may be due to the triangular shape of the fundamental diagram and to the deterministic trajectories of the vehicles. This 20% difference between the model and the simulation may be considered as a baseline error.

7.4 Discussion and conclusion

This work presents the derivations of a model of arterial traffic flow through signalized intersections. This model allows the traffic flow to be characterized by a small number of parameters. Moreover, the study of a corridor is made easier and analytical by the similar structure of the arrival and the departure flows at each intersection.

The model provides an analytical solution to the classic problem of traffic light coordination. The total waiting time of the vehicles during a cycle is a quasi-convex function of the offset between successive traffic signals. The proof of this quasi-convexity property is used to derive the optimal control analytically. For a corridor with multiple intersections, this analysis provides optimal control for the traffic signal at an intersection as a function of the departure streams of the upstream intersection.

After a few intersections, the analysis shows that the choice of the optimal offset leads to a green wave, an intuitive optimization of the offset on a corridor. The results go beyond recalling that the formation of a green wave is the optimal control strategy on a corridor. They provide analytical optimal control strategies for the choice of the offsets, as a function of the arrival streams. This provides valuable information for a real-time implementation with timely adaptation of the control strategies as traffic conditions change, since it does not require additional computation. Given flow measurements from sensors, the traffic signals can compute the optimal offset from the analytical expressions derived in this article. In particular, no online optimization is necessary which is crucial to implement real-time control strategies. The implementation of such algorithms have become a realistic approach to real-time traffic signal control in the recent years, with the emergence of novel sensing technologies available for online control [8].

This model is not limited to the one-way synchronization problem and could be applied to model the flow in numerous arterial traffic situations. The two-way corridor can be studied with the same method and preliminary results are available in [21]. In a similar fashion, the modeling approach presented in the present chapter can be generalized to a network. This network generalization requires to take into account additional specifities of traffic flow such as conflicting flows which may cause delays in the intersection. In a congested regime, it is optimal to optimize the direction of traffic with the longer red light duration. In addition to traffic lights synchronization, this model has potential applications to model the probability distribution of travel times on arterial corridor. In particular, we are interested in the additional accuracy provided by the three-stream approach compared to models which do not take into account light synchronization and assume constant arrival rates as done in Chapter 4 and in [90, 87].

Chapter 8

Conclusion

This thesis presents novel large scale modeling and estimation algorithms of urban traffic dynamics. The thesis focuses on the integration of hydrodynamic theory and statistical modeling in order to accurately represent the physical characteristics of traffic flow and capture the intrinsic variability of traffic dynamics. From hydrodynamic theory, the thesis derives parametric, analytical expressions for the probability distributions over the location of vehicles between intersections as well as for the probability distributions of delays and travel times between any two points on the road network.

Sparsely sampled probe vehicle data appears to be the main source of available data for arterial networks with the prospect of coverage in the near future, at a global scale. For this reason, it is essential that modeling and estimation techniques be designed to leverage data from sparsely sampled probe vehicles. The specific design of algorithms to extract valuable information from sparse and noisy probe vehicle data is the main contribution of this thesis and is essential to develop the next generation of traffic information and management systems. The main features of the approach which makes it particularly adapted to probe vehicle data are the following:

- The statistical modeling approach provides a framework to learn the dynamics historically and estimate traffic conditions in real-time from limited amounts of data. The dynamics are learned from large amounts of historical data. In real time, the statistical model leverages the historical learning to provide estimates of the traffic conditions. The historical learning increases the robustness of the real-time estimates when little data is available. This feature is important as the penetration rate of probe vehicles remains limited, in particular because of the diversity of data providers. Besides the robustness to missing data, statistical models are more robust to noise and are able to filter outliers and to detect deviations from the learned dynamics.

- The underlying physical model of hydrodynamic theory leads to a parametric model in which the parameters have physical interpretations. This property of the model increases the robustness to missing data. When the parameters are learned from historical data or updated in real-time, bounds on their minimum and maximum values and constraints between the variables can be added to represent the physical phenomenon. These bounds and constraints restrict the feasible set of solutions and ensure the quality of the solution when little data is available or when measurements are noisy.

- The statistical modeling approach captures the intrinsic variability of traffic dynamics. The current penetration rate of probe vehicles and the low sampling rates

do not provide sufficient information to fully characterize the macroscopic traffic state (flow, velocity, density) at each point of the network and at each time. As this precise characterization cannot be achieved, it is necessary to have a statistical approach to represent the various traffic conditions experienced by the vehicles traveling on the network at similar times. This variability of traffic conditions is mainly due to the presence of traffic lights which induce queues which form and dissipate periodically, leading to important differences in the delay experienced by the vehicles. This variability is not taken into account when only a mean travel-time is provided. The thesis also argues that common distributions (Gaussian, Log-Normal, Gamma) do not fully capture the variability phenomenon.

- The travel time distributions are derived between *arbitrary locations* on an arterial link. This feature is essential as probe vehicles report their position at a low frequency (on average once per minute) and the location reports do not coincide with the network discretization and a finer discretization would require learning additional parameters with important risks of over-fitting.

- Besides the modeling and estimation of traffic conditions, the thesis presents other applications of probe vehicle data for Intelligent Transportation Systems. The derivations of the density of probe vehicle measurements can be used to automatically detect the presence of traffic signals (either traffic lights or stop signs) which is a key component of digital map databases and Geographic Information Systems, in particular for routing applications, as well as to develop automated or assisted driving capabilities.

Bibliography

[1] 511 SF Bay. http://511.org.

[2] Cabspotting. http://www.cabspotting.org.

[3] CCTV Information. http://www.cctv-information.co.uk/i/An_Introduction_to_ANPR.

[4] Intellione. http://www.uinformatics.com/.

[5] NAVTEQ Inc. http://www.traffic.com/.

[6] NAVTEQ Inc. http://www.navteq.com.

[7] Next Generation Simulation. http://ngsim-community.org/.

[8] Sensys Networks. http://www.sensysnetworks.com.

[9] Telenav Inc. http://www.telenav.com.

[10] Tom-Tom Inc. http://www.tomtom.com.

[11] Waze Inc. http://www.waze.com.

[12] Travel in London: key trends and developments. Transport for London, 2009.

[13] Roadmap to a Single European Transport Area – Towards a competitive and resource efficient transport system. *SEC*, vol. 358, 2011.

[14] M. Abbas, D. Bullock and L. Head. Real-time offset transitioning algorithm for coordinating traffic signals. *Transportation Research Record*, vol. 1748:26–39, 2001.

[15] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, vol. 19, no. 6:716–723, 1974.

[16] R. E. Allsop. Delay at fixed Time Traffic Signals-I: Theoretical analysis. *Transportation Science*, vol. 6:280–285, 1972.

[17] A. Allström, J. Archer, A. Bayen, S. Blandin, D. Goundlegard, H. Koutsopoulos, J. Lundgren, M. Rahmani and O. Tossavainen. Mobile Millennium Stockholm. In *2nd International Conference on Models and Technologies for Intelligent Transportation Systems*. Leuven, Belgium, 2011.

[18] S. Ardekani and R. Herman. Urban network-wide traffic variables and their relations. *Transportation Science*, vol. 21, no. 1:1, 1987.

[19] J. Aubin, A. Bayen and P. Saint-Pierre. Dirichlet Problems for some Hamilton–Jacobi equations with inequality constraints. *SIAM Journal on Control and Optimization*, vol. 47, no. 5:2348–2380, 2008.

[20] J. Aubin, A. Bayen and P. Saint-Pierre. *Viability Theory: New Directions, 2nd edition*. Springer, 2011.

[21] C. Bails. Optimization of the synchronization of traffic lights. *Report, Ecole Polytechnique, Applied Mathematics Departement, Palaiseau, France,* `http://www.eecs.berkeley.edu/~aude/papers/Bails_Optimization_signals.pdf`, 2011.

[22] X. Ban, R. Herring, P. Hao and A. Bayen. Delay Pattern Estimation for Signalized Intersections Using Sampled Travel Times. In *88th Transportation Research Board Annual Meeting*, vol. 2130, pp. 109–119. Washington, D.C., 2009.

[23] X. Ban, Y. Li, A. Skabardonis and J. Margulici. Performance evaluation of travel time methods for real time traffic applications. In *11th World Congress on Transport Research*. 2007.

[24] H. Bar Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel. *Transportation Research Part C*, vol. 15, no. 6:380–391, 2007.

[25] A. Bayen, J. Butler, A. Patire. and et al. Mobile Millennium Final Report. Tech. rep., University of California, Berkeley, CCIT Research Report UCB-ITS-CWP-2011-6, 2011.

[26] A. Bennett. *Inverse methods in physical oceanography*. Cambridge university press, 1992.

[27] D. S. Berry and D. M. Belmont. Distribution of vehicle speeds and travel times. *Proc. 2nd Berkeley Symposium Mathematical Statistics and Probability*, pp. 589–602, 1951.

[28] P. Bickel, C. Chen, J. Kwon, J. Rice, E. V. Zwet and P. Varaiya. Measuring Traffic. *Statistical Science*, vol. 22, no. 4:581–597, 2007.

[29] A. Biem, E. Bouillet, H. Feng, A. Ranganathan, A. Riabov, O. Verscheure, H. Koutsopoulos and C. Moran. IBM infosphere streams for scalable, real-time, intelligent transportation services. In *International conference on Management of data*, pp. 1093–1104. ACM, 2010.

[30] S. Blandin, D. Work, P. Goatin, B. Piccoli and A. Bayen. A General Phase Transition Model for Vehicular Traffic. *SIAM journal on Applied Mathematics*, vol. 71, no. 1:107–127, 2011.

[31] P. Boggs and J. Tolle. Sequential quadratic programming. *Acta numerica*, vol. 4, no. 1:51, 1995.

[32] F. Boillot, J. M. Blosseville, J. B. Lesort, V. Motyka, M. Papageorgiou and S. Sellam. Optimal signal control of urban traffic networks. In *6th International Conference on Road Traffic Monitoring and Control*, pp. 75–79. 1992.

[33] A. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, vol. 18. Oxford University Press, USA, 1997.

[34] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[35] L. Breheret. Real-time information on road traffic based on Floating Car Data. In *Forum Transport à Haut Niveau de Service (THNS)*. Urba2000, Site éco-mobilité France-Chine, 2008.

[36] R. Bruntrup, S. Edelkamp, S. Jabbar and B. Scholz. Incremental map generation with GPS traces. In *IEEE Intelligent Transportation Systems Conference*, pp. 574 – 579. 2005.

[37] K. Burnham and D. Anderson. *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Verlag, 2002.

[38] C. Chen, K. Petty, A. Skabardonis and P. Varaiya. Freeway performance measurement system: mining loop detector data. In *80th Transportation Research Board Annual Meeting*, vol. 1748, pp. 96–102. Washington, D.C., 2001.

[39] L. Chen, W. Jin, J. Hu and Y. Zhang. An Urban Intersection Model Based on Multicommodity Kinematic Wave Theories. In *11th Intelligent Transportation Systems Conference*, pp. 269–274. IEEE, 2008.

[40] A. Chow, G. Gomes, A. Kurzhanskiy and P. Varaiya. Aurora Arterial Modeler a Macroscopic Tool for Urban Traffic Signal Control. In *Control in Transportation Systems*, pp. 361–366. 2009.

[41] C. Claudel and A. Bayen. Guaranteed bounds for traffic flow parameters estimation using mixed lagrangian-eulerian sensing. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pp. 636–645. IEEE, 2008.

[42] C. Claudel and A. Bayen. Lax-Hopf based incorporation of internal boundary conditions into Hamilton-Jacobi equation. Part I: theory. *IEEE Transactions on Automatic Control*, vol. 55, no. 5:1142–1157, 2010. Doi:10.1109/TAC.2010.2041976.

[43] C. Claudel and A. Bayen. Lax-Hopf based incorporation of internal boundary conditions into Hamilton-Jacobi equation. Part II: Computational methods. *IEEE Transactions on Automatic Control*, vol. 55, no. 5:1158–1174, 2010. Doi:10.1109/TAC.2010.2045439.

[44] C. Claudel, A. Hofleitner, N. Mignerey and A. Bayen. Guaranteed bounds on highway travel times using probe and fixed data. In *88th Transportation Research Board Annual Meeting*, 09-3616. Washington D.C., 2009.

[45] C. Claudel, M. Nahoum and A. Bayen. Minimal error certificates for detection of faulty sensors using convex optimization. In *47th Allerton Conference on Communication, Control, and Computing*. Allerton, IL, 2009.

[46] B. Coifman. Estimating travel times and vehicles trajectories on freeways using dual loop detectors. *Transportation Research Part A*, vol. 36, no. 4:351–364, 2002.

[47] D. Cox and D. Oakes. *Analysis of survival data*. Chapman & Hall, London, U.K., 1984.

[48] M. Crandall, L. Evans and P. Lions. Some properties of viscosity solutions of Hamilton-Jacobi equations. *American Mathematical Society*, vol. 282, no. 2, 1984.

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

[50] C. Daganzo. The cell transmission model, part II: network traffic. *Transportation Research Part B*, vol. 29, no. 2:79–93, 1995.

[51] C. Daganzo. *Fundamentals of transportation and traffic operations*. Elselevier, 1997.

[52] C. Daganzo. A variational formulation of kinematic waves: basic theory and complex boundary conditions. *Transportation Research Part B*, vol. 39, no. 2:187–196, 2005.

[53] C. Daganzo. On the variational theory of traffic flow: well-posedness, duality and applications. *Networks and Heterogeneous Media*, vol. 1, no. 4:601, 2006.

[54] C. Daganzo and N. Geroliminis. An analytical approximation for the macroscopic fundamental diagram of urban traffic. *Transportation Research Part B*, vol. 42, no. 9:771–781, 2008.

[55] J. N. Darroch. On the Traffic-light Queue. *The Annals of Mathematical Statistics*, vol. 35, no. 1:pp. 380–388, 1964.

[56] R. De Charette and F. Nashashibi. Traffic light recognition using image processing compared to learning processes. In *2009 IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 333–338. 2009.

[57] C. de Fabritiis, R. Ragona and G. Valenti. Traffic Estimation And Prediction Based On Real Time Floating Car Data. In *IEEE 11th International Conference on Intelligent Transportation Systems. ITSC 2008.*, pp. 197–203. 2008.

[58] A. De Palma and F. Marchal. Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale mesoscopic transportation systems. *Networks and Spatial Economics*, vol. 2, no. 4:347–369, 2002.

[59] S. A. et al. Mobile Century-using GPS mobile phones as traffic sensors: a field experiment. In *15th World congress on ITS, New York, NY*. 2008.

[60] L. C. Evans. *Partial Differential Equations*. Graduate Studies in Mathematics, V. 19. American Mathematical Society, Providence, RI, 1998.

[61] G. Evensen. *Data Assimilation: The Ensemble Kalman Filter*. Springer-Verlag, Berlin Heidelberg, 2007.

[62] Y. Feng, G. Davis and J. Hourdos. Arterial Travel Time Characterization and Real-time Traffic Condition Identification Using GPS-equipped Probe Vehicles. In *90th Transportation Research Board Annual Meeting*, 11-2018. 2011.

[63] R. Follmer. National Transportation Surveys: Activities in Germany. In *Transportation Research Board 91st Annual Meeting*, 12-6949. 2012.

[64] H. Frankowska. Lower semicontinuous solutions to Hamilton-Jacobi-Bellman equations. In *Proceedings of the 30th IEEE Conference on Decision and Control*, pp. 265–270. IEEE, 1991.

[65] C. Furtlehner, J. Lasgouttes and A. de la Fortelle. A belief propagation approach to traffic prediction using probe vehicles. In *10th Intelligent Transportation Systems Conference*, pp. 1022–1027. IEEE, 2007.

[66] N. Gartner and P. Wagner. Analysis of traffic flow characteristics on signalized arterials. *Transportation Research Record*, vol. 1883:94–100, 2004.

[67] N. H. Gartner, J. D. C. Little and H. Gabbay. Optimization of traffic signal settings by mixed-integer linear programming part II: the tetwork synchronization problem. *Transportation Science*, vol. 9, no. 4:344–363, 1975.

[68] N. H. Gartner and C. Stamatiadis. Arterial-based control of traffic flow in urban grid networks. *Mathematical and Computer Modelling*, vol. 35, no. 5-6:657 – 671, 2002.

[69] H. Gault and I. Taylor. The use of output from vehicle detectors to access delay in computer-controlled area traffic control systems. Tech. Rep. Research Report No. 31, Transportation Operation Research Group, University of Newcastle upon Tyne, United Kingdom, 1977.

[70] N. Geroliminis and C. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B*, vol. 42, no. 9:759 – 770, 2008.

[71] N. Geroliminis and A. Skabardonis. Prediction of arrival profiles and queue lengths along signalized arterials by using a Markov decision process. *Transportation Research Record*, vol. 1934, no. 1:116–124, 2006.

[72] N. Geroliminis and A. Skabardonis. Queue spillovers in city street networks with signal-controlled Intersections. In *89th Transportation Research Board Annual Meeting*. 2010.

[73] M. Gonzalez, C. Hidalgo and A. Barabasi. Understanding individual human mobility patterns. *Nature*, vol. 453, no. 7196:779–782, 2008.

[74] A. Halati, H. Lieu and S. Walker. CORSIM-corridor traffic simulation model. In *Traffic Congestion and Traffic Safety in the 21st Century: Challenges, Innovations, and Opportunities*. 1997.

[75] A. Haoui, R. Kavaler and P. Varaiya. Wireless magnetic sensors for traffic surveillance. *Transportation Research Part C*, vol. 16, no. 3:294 – 306, 2008.

[76] D. Heidemann. Queue length and delay distributions at traffic signals. *Transportation Research Part B*, vol. 28, no. 5:377–389, 1994.

[77] D. Helbing. Derivation of a fundamental diagram for urban traffic flow. *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 70, no. 2:229–241, 2009.

[78] B. Hellinga and L. Fu. Reducing bias in probe-based arterial link travel time estimates. *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4:257–273, 2002.

[79] B. Hellinga, P. Izadpanah, H. Takada and L. Fu. Decomposing travel times measured by probe-based traffic monitoring systems to individual road segments. *Transportation Research Part C*, vol. 16, no. 6:768 – 782, 2008.

[80] J. Herrera and A. Bayen. Traffic flow reconstruction using mobile sensors and loop detector data. In *87th Transportation Research Board Annual Meeting*. Washington, D.C., 2008.

[81] J. Herrera, D. Work, R. Herring, X. Ban, Q. Jacobson and A. Bayen. Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C*, vol. 18, no. 4:568–583, 2010.

[82] R. Herring, A. Hofleitner, P. Abbeel and A. Bayen. Estimating arterial traffic conditions using sparse probe data. In *13th Intelligent Transportation Systems Conference*, pp. 929–936. IEEE, Madeira, Portugal, 2010.

[83] R. Herring, A. Hofleitner, S. Amin, T. A. Nasr, A. Abdel Khalek, P. Abbeel and A. Bayen. Using Mobile Phones to Forecast Arterial Traffic through Statistical Learning. In *89th Transportation Research Board Annual Meeting*, 10-2493. Washington D.C., 2010.

[84] A. Hofleitner and A. Bayen. Optimal decomposition of travel times measured by probe vehicles using a statistical traffic flow model. In *14th Intelligent Transportation Systems Conference*, pp. 815 –821. IEEE, 2011.

[85] A. Hofleitner, C. Claudel and A. Bayen. Probabilistic formulation of estimation problems for a class of Hamilton-Jacobi equations. In *51st Conference on Decision and Control*. IEEE, 2012.

[86] A. Hofleitner, E. Come, L. Oukhellou, J.-P. Lebacque and A. Bayen. Automatic inference of map attributes from mobile data. In *15th IEEE Intelligent Transportation System Conference (ITSC 20112)*. IEEE, 2012.

[87] A. Hofleitner, R. Herring and A. Bayen. A hydrodynamic theory based statistical model of arterial traffic. Tech. Rep. UCB-ITS-CWP-2011-2, UC Berkeley, 2011.

[88] A. Hofleitner, R. Herring and A. Bayen. Arterial travel time forecast with streaming data: a hybrid flow model - machine learning approach. *Transportation Res. B*, 2012.

[89] A. Hofleitner, R. Herring and A. Bayen. Learning the dynamics of arterial traffic from probe data using a Dynamic Bayesian Network. *IEEE Transactions on Intelligent Transportation Systems*, 2012.

[90] A. Hofleitner, R. Herring and A. Bayen. Probability distributions of travel times on arterial networks: a traffic flow and horizontal queuing theory approach. In *91st Transportation Research Board Annual Meeting*, 12-0798. Washington, D.C., 2012.

[91] B. Hoh, M. Gruteser, R. Herring, J. Ban, D. Work, J. Herrera and A. Bayen. Virtual trip lines for distributed privacy-preserving traffic monitoring. In *The Sixth Annual International conference on Mobile Systems, Applications and Services (MobiSys 2008)*. Breckenridge, U.S.A., 2008.

[92] B. Hoh, M. Gruteser, H. Xiong and A. Alrabady. Enhancing security and privacy in traffic-monitoring systems. *IEEE Pervasive Computing*, vol. 5, no. 4:38–46, 2006.

[93] R. Horst and H. Tuy. Global optimization: deterministic approaches. *Journal of the Operational Research Society*, vol. 45, no. 5:595–596, 1994.

[94] E. Horvitz, J. Apacible, R. Sarin and L. Liao. Prediction, expectation, and surprise: Methods, designs, and study of a deployed traffic forecasting service. In *21st Conference on Uncertainty in Artificial Intelligence*. 2005.

[95] B. Hull, V. Bychkovsky, Y. Zhang, K. Chen, M. Goraczko, A. Miu, E. Shih, H. Balakrishnan and S. Madden. CarTel: a distributed mobile sensor computing system. In *SenSys '06: 4th international conference on Embedded networked sensor systems*, pp. 125–138. ACM, New York, NY, USA, 2006.

[96] T. Hunter, T. Moldovan, M. Zaharia, S. Merzgui, J. Ma, M. J. Franklin, P. Abbeel and A. M. Bayen. Scaling the Mobile Millennium system in the cloud. In *2nd ACM Symposium on Cloud Computing*, vol. 28, pp. 1–8. ACM, 2011.

[97] J. Ibanez-Guzman, O. Le Marchand and C. Chen. Metric evaluation of automotive-type GPS receivers. 2008.

[98] Z. Jia, C. Chen, B. Coifman and P. Varaiya. The PeMS algorithms for accurate, real-time estimates of g-factors and speeds from single-loop detectors. In *Intelligent Transportation Systems Conference*, pp. 536–541. IEEE, 2001.

[99] E. L. Kaplan and P. Meier. Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, vol. 53, no. 282:pp. 457–481, 1958.

[100] R. M. Kimber and E. M. Hollis. *Traffic queues and delays at road junctions*. Traffic Systems Division, Traffic Engineering Department, Transport and Road Research Laboratory, 1979.

[101] A. Krause, E. Horvitz, A. Kansal and F. Zhao. Toward Community Sensing. In *7th international conference on Information processing in sensor networks*, pp. 481–492. IEEE Computer Society, 2008.

[102] J. Kwon and K. Murphy. Modeling Freeway Traffic with Coupled HMMs. Tech. rep., University of California, Berkeley, 2000.

[103] K. Kwong, R. Kavaler, R. Rajagopal and P. Varaiya. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transportation Research Part C*, vol. 17, no. 6:586–606, 2009.

[104] S. Lämmer, R. Donner and D. Helbing. Anticipative control of switched queueing systems. *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 63:341–347, 2008.

[105] J. Lebacque. First-order macroscopic traffic flow models: Intersection modeling, network modeling. In *16th International Symposium on Transportation and Traffic Theory*. 2005.

[106] J. S. V. Leeuwaarden. Delay analysis for the fixed-cycle traffic-light queue. *Transportation Science*, vol. 40, no. 2:189–199, 2006.

[107] J. Lewis, S. Lakshmivarahan and S. Dhall. *Dynamic data assimilation: a least squares approach*, vol. 104. Cambridge Univ Pr, 2006.

[108] M. Lighthill and G. Whitham. On Kinematic Waves. II. A theory of traffic flow on long crowded roads. In *Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 229, pp. 317–345. 1955.

[109] W. Lin, A. Kulkarni and P. Mirchandani. Short-term arterial travel time prediction for advanced traveler information systems. In *Intelligent Transportation Systems*, vol. 8, pp. 143–154. Taylor & Francis, 2004.

[110] H. Liu, A. Danczyk, R. Brewer and R. Starr. Evaluation of Cell Phone Traffic Data in Minnesota. *Transportation Research Record*, vol. 2086:1–7, 2008.

[111] H. Liu and W. Ma. A Virtual Probe Approach for Time-Dependent Arterial Travel Time Estimation. In *87th Transportation Research Board Annual Meeting*. 2008.

[112] H. Lo, E. Chang and Y. Chan. Dynamic network traffic control. *Transportation Research Part A*, vol. 35, no. 8:721–744, 2001.

[113] Y. Lou, C. Zhang, Y. Zheng, X. Xie, W. Wang and Y. Huang. Map-matching for low-sampling-rate GPS trajectories. In *17th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 352–361. ACM, 2009.

[114] F. J. Massey. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, vol. 46, no. 253:68–78, 1951.

[115] P.-E. Mazaré, A. Dehwah, C. Claudel and A. Bayen. Analytical and grid-free solutions to the Lighthill Whitham Richards traffic flow model. *Transportation Research Part B*, vol. 45, no. 10:1727 – 1748, 2011.

[116] D. R. McNeil. A Solution to the Fixed-Cycle Traffic Light Problem for Compound Poisson Arrivals. *Journal of Applied Probability*, vol. 5, no. 3:624–635, 1968.

[117] A. J. Miller. Settings for Fixed-Cycle Traffic Signals. *Operations Research*, vol. 14, no. 4:373–386, 1963.

[118] L. Mimbela, L. Klein, P. Kent, J. Hamrick, K. Luces and S. Herrera. *Summary of Vehicle Detection and Surveillance Technologies used in Intelligent Transportation Systems*. Federal Highway Administration's (FHWA) Intelligent Transportation Systems Program Office, 2007.

[119] X. Min, J. Hu, Q. Chen, T. Zhang and Y. Zhang. Short-term traffic flow forecasting of urban network based on dynamic STARIMA model. In *12th Intelligent Transportation Systems Conference*. IEEE, 2009.

[120] L. Muñoz, X. Sun, R. Horowitz and L. Alvarez. Traffic density estimation with the cell transmission model. In *American Control Conference, 2003. Proceedings of the 2003*, vol. 5, pp. 3750–3755. IEEE, 2003.

[121] L. Muñoz, X. Sun, R. Horowitz and L. Alvarez. Piecewise-linearized cell transmission model and parameter calibration methodology. *Transportation Research Record*, vol. 1965:183–191, 2006.

[122] L. Muñoz, X. Sun, D. Sun, G. Gomes and R. Horowitz. Methodological calibration of the cell transmission model. In *American Control Conference*, vol. 1, pp. 798–803. IEEE, 2004.

[123] G. Newell. A simplified theory of kinematic waves in highway traffic. *Transportation Research Part B*, vol. 27, no. 4:281–313, 1993.

[124] R. Noland and J. Polak. Travel time variability: a review of theoretical and empirical issues. *Transport Reviews*, vol. 22, no. 1:39–54, 2002.

[125] C. Oh and S. Ritchie. Anonymous vehicle tracking for real-time traffic surveillance and performance on signalized arterials. In *82nd Transportation Research Board Annual Meeting*. 2003.

[126] K. Ohno. Computational Algorithm for a Fixed Cycle Traffic Signal and New Approximate Expressions for Average Delay. *Transportation Science*, vol. 12, no. 1:29–47, 1978.

[127] B. Park, C. J. Messer and T. Urbanik. Traffic signal optimization program for oversaturated conditions: genetic algorithm approach. *Transportation Research Record*, vol. 1683:133–142, 1999.

[128] T. Park and S. Lee. A Bayesian Approach for Estimating Link Travel Time on Urban Arterial Road Network. In *Computational Science and Its Applications*, pp. 1017–1025. 2004.

[129] F. Porikli and X. L. Traffic congestion estimation using HMM models without vehicle tracking. In *Intelligent Vehicles Symposium*, pp. 188 – 193. IEEE, 2004.

[130] A.-S. Puthon, F. Nashashibi and B. Bradai. A complete system to determine the speed limit by fusing a GIS and a camera. In *14th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pp. 1686 – 1691. 2011.

[131] M. Rahmani, H. Koutsopoulos and A. Ranganathan. Requirements and potential of GPS-based floating car data for traffic management: Stockholm case study. In *13th Intelligent Transportation Systems Conference*, pp. 730–735. IEEE, 2010.

[132] M. Ramezani and N. Geroliminis. Arterial route travel time distribution estimation with a markov chain procedure. In *11th Swiss Transport Research Conference*. 2011.

[133] W. J. Rankine. On the Thermodynamic Theory of Waves of Finite Longitudinal Disturbance. *Philosophical Transactions of the Royal Society of London*, vol. 160:277–288, 1870.

[134] P. Richards. Shock Waves on the Highway. *Operations Research*, vol. 4, no. 1:42–51, 1956.

[135] D. I. Robertson and R. D. Bretherton. Optimizing networks of traffic signals in real time-the SCOOT method. *IEEE Transactions on Vehicular Technology*, vol. 40, no. 1:11 –15, 1991.

[136] S. Rogers, P. Langley and C. Wilson. Mining GPS data to augment road models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '99, pp. 104–113. ACM, 1999.

[137] G. Rose. Mobile phones as traffic probes: practices, prospects and issues. *Transport Reviews*, vol. 26, no. 3:275–291, 2006.

[138] N. M. Rouphail. Delay models for mixed platoon and secondary flows. *Journal of Transportation Engineering*, vol. 114, no. 2:131–132, 1988.

[139] S. Samaranayake, S. Blandin and A. Bayen. A tractable class of algorithms for reliable routing in stochastic networks. *Transportation Research Part C*, vol. 20, no. 1:199 – 217, 2012.

[140] R. Schäfer, K. Thiessenhusen and P. Wagner. A traffic information system by means of real-time floating-car data. In *ITS world congress*, vol. 2. 2002.

[141] D. Schrank and T. Lomax. Urban Mobility Study. *Texas Transportation Institute*, 2007.

[142] D. Schrank, T. Lomax and Texas Transportation Institute. *2009 Urban mobility report*. Citeseer, 2009.

[143] G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, vol. 6, no. 2:461–464, 1978.

[144] A. Sen, P. Thakuriah, X. Zhu and A. Karr. Frequency of probe reports and variance of travel time estimates. *Journal of Transportation Engineering*, vol. 123:290, 1997.

[145] V. Sisiopiku and N. Rouphail. Travel time estimation from loop detector data for advanced traveler information system applications. Tech. rep., Illinois University Transportation Research Consortium, 1994.

[146] A. Skabardonis and R. Dowling. Improved Speed-Flow Relationship for Planning Applications. *Transportation Research Record*, vol. 1572:18–23, 1997.

[147] A. Skabardonis and N. Geroliminis. Real-time estimation of travel times on signalized arterials. In *16th International Symposium on Transportation and Traffic Theory*. University of Maryland, College Park, MD, 2005.

[148] A. Skabardonis and N. Geroliminis. Real-time Monitoring and Control on Signalized Arterials. *Journal of Intelligent Transportation Systems*, vol. 12, no. 2:64–74, 2008.

[149] B. L. Smith and S. T. Laboratory. *Cellphone probes as an ATMS tool*. Center for ITS Implementation Research, University of Virginia, VA, 2003.

[150] J. M. Staniewicz and H. S. Levinson. Signal delay with platoon arrivals. *Transportation Research Record*, vol. 1005:28–37, 1985.

[151] X. Sun, L. Munoz and R. Horowitz. Highway traffic state estimation using improved mixture Kalman filters for effective ramp metering control. In *42nd Conference on Decision and Control*, pp. 9–12. IEEE, Maui, HI, 2003.

[152] X. Sun, L. Munoz and R. Horowitz. Mixture Kalman Filter Based Highway Congestion Mode and Vehicle Density Estimator and its Application. In *American Control Conference*, pp. 2098–2103. IEEE, Boston, MA, 2004.

[153] N. Tanaka. Energy Technology Perspectives 2008 – Scenarios and Strategies to 2050. *International Energy Agency (IEA), Paris*, 2008.

[154] A. Thiagarajan, L. Sivalingam, K. LaCurts, S. Toledo, J. Eriksson, S. Madden and H. Balakrishnan. VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones. In *7th Conference on Embedded Networked Sensor Systems*. ACM, Berkeley, CA, 2009.

[155] Trans Res Board. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, D.C., 2000.

[156] M. Van Den Broek, J. Van Leeuwaarden, I. Adan and O. J. Boxma. Bounds and approximations for the fixed-cycle traffic-light queue. *Transportation Science*, vol. 40, no. 4:484496, 2006.

[157] H. Van Zuylen, F. Zheng and Y. Chen. Using Probe Vehicle Data for Traffic State Estimation in Signalized Urban Networks. *Traffic Data Collection and its Standardization*, pp. 109–127, 2010.

[158] F. Viti and H. van Zuylen. Modeling queues at signalized intersections. *Transportation Research Record*, vol. 1883:68–77, 2004.

[159] F. Viti and H. J. Van Zuylen. The Dynamics and the Uncertainty of Queues at Fixed and Actuated Controls: A Probabilistic Approach. *Journal of Intelligent Transportation Systems*, vol. 13, no. 1, 2009.

[160] Y. Wang and M. Papageorgiou. Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. *Transportation Research Part B*, vol. 39, no. 2:141–167, 2005.

[161] J. Wasson, J. Sturdevant and D. Bullock. Real-Time Travel Time Estimates Using MAC Address Matching. *Institute of Transportation Engineers Journal*, vol. 78, no. 6:20–23, 2008.

[162] F. V. Webster. *Traffic signal settings*. Paper No. 39, Road Research Laboratory, England, published by HMSO, 1958.

[163] D. Work and A. Bayen. Impacts of the mobile internet on transportation cyber-physical systems: traffic monitoring using smartphones. In *National Workshop for Research on High-Confidence Transportation Cyber-Physical Systems: Automotive, Aviation, & Rail*, pp. 18–20. 2008.

[164] D. Work, O. Tossavainen, S. Blandin, A. Bayen, T. Iwuchukwu and K. Tracton. An ensemble Kalman filtering approach to highway traffic estimation using GPS enabled mobile devices. In *47th Conference on Decision and Control*, pp. 5062–5068. IEEE, Cancun, Mexico, 2008.

[165] D. B. Work, S. Blandin, O.-P. Tossavainen, B. Piccoli and A. M. Bayen. A Traffic Model for Velocity Data Assimilation. *Applied Mathematics Research eXpress*, vol. 2010, no. 1:1–35, 2010.

[166] X. Xie, R. Cheu and D. Lee. Calibration-free arterial link speed estimation model using loop data. *Journal of Transportation Engineering*, vol. 127, no. 6:507–514, 2001.

[167] H. Xiong and G. Davis. Travel time estimation on arterials. In *87th Transportation Research Board Annual Meeting*. 2008.

[168] T. Yamamoto, K. Liu and T. Morikawa. Variability of travel time estimates using probe vehicle data. In *Proceedings of the Fourth International Conference on Traffic and Transportation Studies (ICTTS)*, pp. 278–287. 2006.

[169] J. Ygnace. Travel time estimation on the San Francisco Bay Area network using cellular phones as probes. *California PATH Program, Institute of Transportation Studies, University of California at Berkeley, UCB-ITS-PWP-2000-18*, 2000.

[170] Y. Yim and R. Cayford. Investigation of vehicles as probes using global positioning system and cellular phone tracking: field operational test. *California PATH Program, Institute of Transportation Studies, University of California at Berkeley, UCB-ITS-PWP-2001-9*, 2001.

[171] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun and Y. Huang. T-drive: driving directions based on taxi trajectories. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 99–108. ACM, 2010.

[172] H. Zhang. A link journey speed model for arterial traffic. *Transportation Research Record*, vol. 1676:109–115, 1998.

[173] H. Zhang and T. Kim. A car-following theory for multiphase vehicular traffic flow. *Transportation Research Part B*, vol. 39, no. 5:385 – 399, 2005.

[174] F. Zheng and H. van Zuylen. Reconstruction of delay distribution at signalized intersections based on traffic measurements. In *13th Intelligent Transportation Systems Conference*, pp. 1819 –1824. IEEE, 2010.

[175] A. Zhigljavsky and A. Zilinskas. *Stochastic global optimization*. Springer, 2007.

Appendix A

Derivation of the probability distribution of total delay between arbitrary locations in the congested regime

We derive the probability distribution of travel times for vehicles traveling from a location x_1 to a location x_2 on the link. As in the previous notations, x represents the distance to the intersection.

We call n_s the maximum number of stops in the remaining queue experienced by the vehicles between the locations x_1 and x_2, and omit the indices x_1 and x_2 for notational simplicity. In the duration of a light cycle, the distance traveled by vehicles stopped in the queue is $l_{\max}$. Thus, the maximum number of stops in the remaining queue, between x_1 and x_2,

$$n_s = \left\lceil \frac{\min(x_1, l_r) - \min(x_2, l_r)}{l_{\max}} \right\rceil .$$

The delay experienced when reaching the triangular queue is readily derived from the expression of the delay in the undersaturated regime. The delay experienced when reaching the remaining queue is the duration of the red time R. The expression of the delay at location x is then

$$\delta^c(x) = \begin{cases} R & \text{if } x \leq l_r \\ R\,\frac{l_r + l_{\max} - x}{l_{\max}} & \text{if } x \in [l_r, l_r + l_{\max}] \\ 0 & \text{if } x \geq l_r + l_{\max} \end{cases}$$

A.1 Case 1: x_1 is upstream of the total queue and x_2 is in the remaining queue

Condition 1: $x_1 \geq l_r + l_{\max} \quad x_2 \leq l_r$

Since x_1 is upstream of the total queue and x_2 is in the remaining queue, all the vehicles stop once in the triangular queue between x_1 and x_2. We define the critical location x_c as the location in the triangular queue such that

- Vehicles reaching the triangular queue upstream of x_c stop n_s times in the remaining queue. They represent a fraction $\frac{l_r + l_{\max} - x_c}{l_{\max}}$ of the vehicles entering the link in a cycle.
- Vehicles reaching the triangular queue downstream of x_c stop $n_s - 1$ times in the remaining queue. They represent a fraction $\frac{x_c - l_r}{l_{\max}} = 1 - \frac{l_r + l_{\max} - x_c}{l_{\max}}$ of the vehicles entering the link in a cycle.

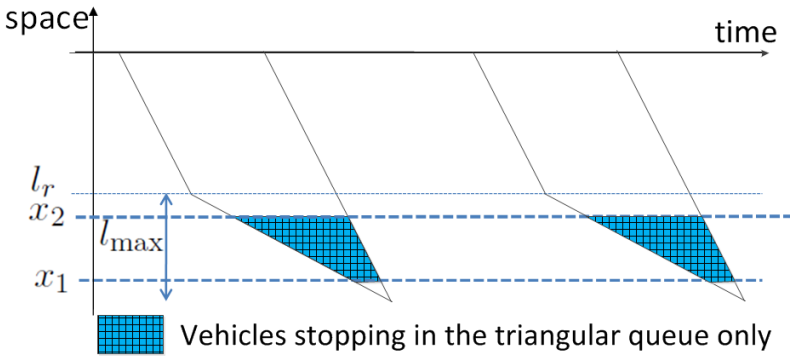


Figure A.2: Case 2: Some vehicles stop in the triangular queue. The others do not experience delay.

The location x_c is given by $x_c = x_2 + n_s l_{\max}$.

The values of the minimum and maximum delays are given by $\delta_{\min} = (n_s - 1)R + \delta^c(x_c)$ and $\delta_{\max} = n_s R + \delta^c(x_c)$. The delay experienced by the vehicles is uniformly distributed on $[\delta_{\min}, \delta_{\max}]$.
We note that $n_s \geq 1$ since $x_2 \leq l_r$.

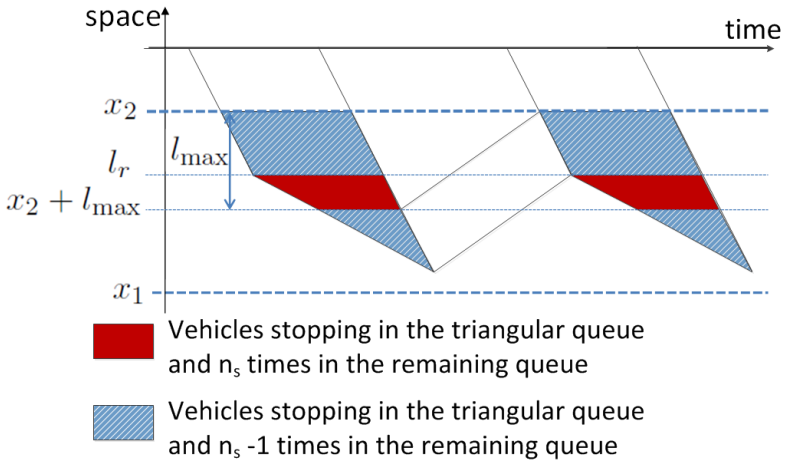


Figure A.1: Case 1: All the vehicles stop in the triangular queue. A fraction stops n_s times in the remaining queue, the other ones stop $n_s - 1$ times.

A.2 Case 2: x_1 and x_2 are upstream of the remaining queue

Condition 2: $x_1 \geq l_r \quad x_2 \geq l_r$

Given that x_2 is upstream of the remaining queue, this case is similar to the under-saturated regime. A fraction of the vehicles does not experience delay between x_1 and x_2. The vehicles reaching the queue between x_1 and x_2 experience a delay in the triangular queue. This delay is a random variable, uniformly distributed on $[\delta^c(x_1), \delta^c(x_2)]$.

The fraction of vehicles experiencing delay is $\eta^c_{x_1,x_2} = \frac{\min(l_{\max}+l_r,x_1)-\min(l_{\max}+l_r,x_2)}{l_{\max}}$

A.3 Case 3: x_1 is in the remaining queue, and thus so is x_2

Condition 3: $x_1 \leq l_r$ (which implies $x_2 \leq l_r$)

The path starts downstream of the triangular queue. Some vehicles stop n_s times and experience a delay $n_s R$ and the other vehicles stop $n_s - 1$ times and experience a delay $(n_s - 1)R$.

We define the critical location x_c as the location in the remaining queue such that

- Vehicles joining the queue between x_1 and x_c stop n_s times between x_1 and x_2. Their stopping time is $n_s R$ and they represent a fraction $(x_1 - x_c)/l_{\max}$ of the vehicles entering the link in a cycle.
- Vehicles joining the queue between x_c and $x_c - l_{\max}$ stop $n_s - 1$ times between x_1 and x_2. Their stopping time is $(n_s - 1)R$ and they represent a fraction $1 - (x_1 - x_c)/l_{\max}$ of the vehicles entering the link in a cycle.

The critical location x_c is given by $x_c = x_2 + (n_s - 1)l_{\max}$.

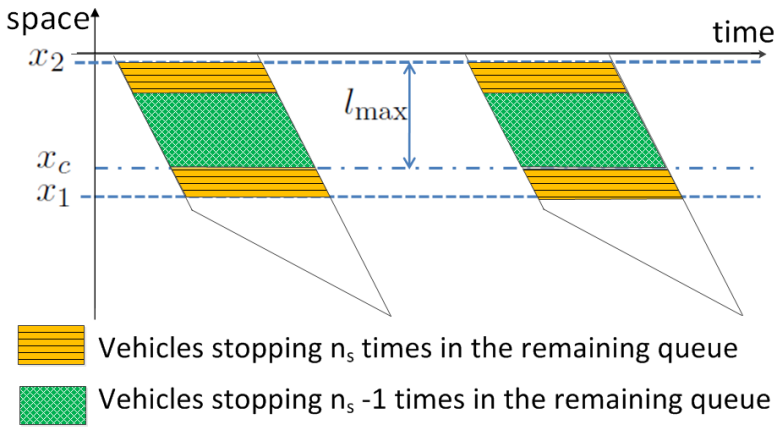


Figure A.3: Case 3: A fraction of the vehicles stop n_s times in the remaining queue. The rest stop $n_s - 1$ times in the remaining queue.

A.4 Case 4: x_1 is in the triangular queue, x_2 is in the remaining queue

We distinguish two different cases to derive the probability distribution of travel times. We define the critical location x_c as $x_c = x_2 + n_s l_{\max}$ and derive probability distributions of travel times for the two subcases 4a ($x_c \leq x_1$, Figure A.4 (top)) and 4b ($x_c \geq x_1$, Figure A.4 (bottom)).

Case 4a. $x_c \leq x_1$. The delay patterns are the following:

- One stop in the triangular queue and n_s stops in the remaining queue. The queue is first reached between x_1 and x_c. The delay is a random variable with uniform distribution with support $[\delta^c(x_1) + n_s R,\ \delta^c(x_c) + n_s R]$. The vehicles following this pattern represent a fraction $\frac{x_1 - x_c}{l_{\max}}$ of the vehicles entering the link in a cycle.

– One stop in the triangular queue and $n_s - 1$ stops in the remaining queue. The queue is first reached between x_c and l_r. The delay is a random variable with uniform distribution with support $[\delta^c(x_c) + (n_s - 1)R,\ \delta^c(l_r) + (n_s - 1)R]$. Noticing that $\delta^c(l_r) = R$, we derive that the support of the delay distribution is $[\delta^c(x_c) + (n_s - 1)R,\ n_sR]$. The vehicles following this pattern represent a fraction $\frac{x_c - l_r}{l_{\max}}$ of the vehicles entering the link in a cycle.
– No stop in the triangular queue and n_s stops in the remaining queue. The queue is first reached between l_r and $x_1 - l_{\max}$. The delay is n_sR. The vehicles following this pattern represent a fraction $\frac{l_r - (x_1 - l_{\max})}{l_{\max}}$ of the vehicles entering the link in a cycle.

We can check that the weights of the different components sum to 1:

$$\frac{x_1 - x_c}{l_{\max}} + \frac{x_c - l_r}{l_{\max}} + \frac{l_r - (x_1 - l_{\max})}{l_{\max}} = 1$$

We remark that, $x_2 \leq l_r$ implies that $n_s \geq 1$. Then using the definition of x_c, $x_c = x_2 + n_s l_{\max}$ and the fact that $x_1 \geq x_c$, we prove that $x_1 - l_{\max} \geq x_2$ and all vehicles reach the queue between x_1 and $x_1 - l_{\max}$.

Case 4b. $x_c \geq x_1$. The delay patterns are the following:

– One stop in the triangular queue and $n_s - 1$ stops in the remaining queue. The queue is first reached between x_1 and l_r. The delay is a random variable with uniform distribution on $[\delta^c(x_1) + (n_s - 1)R,\ \delta^c(l_r) + (n_s - 1)R]$, *i.e.* uniform distribution on $[\delta^c(x_1) + (n_s - 1)R,\ n_sR]$. The vehicles following this pattern represent a fraction $\frac{x_1 - l_r}{l_{\max}}$ of the vehicles entering the link in a cycle.
– No stop in the triangular queue and n_s stops in the remaining queue. The queue is first joined between l_r and $x_c - l_{\max}$. The delay is n_sR. The vehicles following this pattern represent a fraction $\frac{l_r - (x_c - l_{\max})}{l_{\max}}$ of the vehicles entering the link in a cycle.
– No stop in the triangular queue and $n_s - 1$ stops in the remaining queue. The queue is first joined between $x_c - l_r$ and $x_1 - l_{\max}$. The delay is $(n_s - 1)R$. The vehicles following this pattern represent a fraction $\frac{x_c - x_1}{l_{\max}}$ of the vehicles entering the link in a cycle.

We can check that the weights of the different components sum to 1:

$$\frac{l_r - (x_c - l_{\max})}{l_{\max}} + \frac{x_1 - l_r}{l_{\max}} + \frac{x_c - x_1}{l_{\max}} = 1.$$

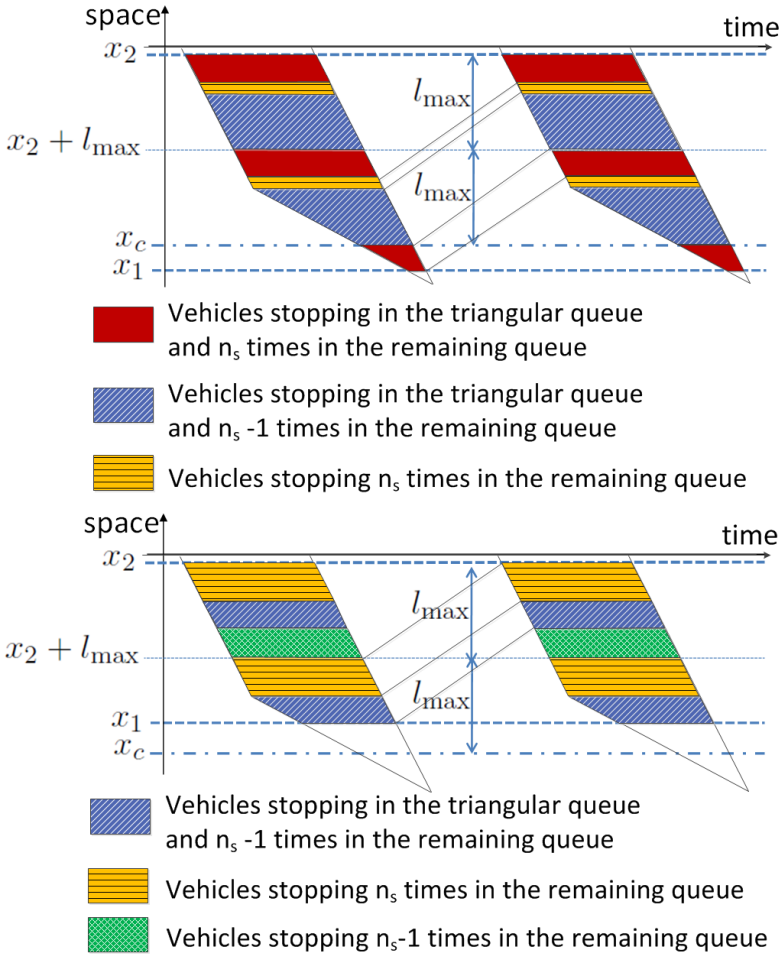


Figure A.4: Case 4: (**Top**) Case 4a: a fraction of the vehicles stop in the triangular queue and n_s times in the remaining queue, a fraction of the vehicles stop in the triangular queue and n_s times in the remaining queue, the rest stop n_s times in the remaining queue. (**Bottom**) Case 4b: a fraction of the vehicles stop in the triangular queue and $n_s - 1$ times in the remaining queue, a fraction of the vehicles stop n_s times in the remaining queue, the rest stop $n_s - 1$ times in the remaining queue.