Springer Tracts on Transportation and Traffic

Margarita Martínez-Díaz *Editor*

The Evolution of Travel Time Information Systems

The Role of Comprehensive Traffic Models and Improvements Towards Cooperative Driving Environments

Springer

Springer Tracts on Transportation and Traffic

Volume 19

About this Series

The book series "Springer Tracts on Transportation and Traffic" (STTT) publishes current and historical insights and new developments in the fields of Transportation and Traffic research. The intent is to cover all the technical contents, applications, and multidisciplinary aspects of Transportation and Traffic, as well as the methodologies behind them. The objective of the book series is to publish monographs, handbooks, selected contributions from specialized conferences and workshops, and textbooks, rapidly and informally but with a high quality. The STTT book series is intended to cover both the state-of-the-art and recent developments, hence leading to deeper insight and understanding in Transportation and Traffic Engineering. The series provides valuable references for researchers, engineering practitioners, graduate students and communicates new findings to a large interdisciplinary audience.

Indexed by SCOPUS, WTI Frankfurt eG, zbMATH, SCImago.

More information about this series at https://link.springer.com/bookseries/11059

Margarita Martínez-Díaz
Editor

The Evolution of Travel Time Information Systems

The Role of Comprehensive Traffic Models and Improvements Towards Cooperative Driving Environments

Springer

Editor
Margarita Martínez-Díaz
Barcelona Innovative Transportation
Research Group (BIT)
Department of Civil and Environmental
Engineering, Civil Engineering School
Polytechnic University of Catalonia
Barcelona, Spain

ISSN 2194-8119 ISSN 2194-8127 (electronic)
Springer Tracts on Transportation and Traffic
ISBN 978-3-030-89671-3 ISBN 978-3-030-89672-0 (eBook)
https://doi.org/10.1007/978-3-030-89672-0

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Time waits for nobody

Freddie Mercury (1986). Time.

In: Time (Pr. David Clark). London, UK: EMI

To my loving parents

Preface

This book was conceived with some clear objectives in mind. The first one is the proposal of solutions for the immediate improvement of current highway travel time information systems, making them able to provide drivers entering a stretch with an accurate prediction of the time they will need to fulfill their particular trips. This objective involves two challenging tasks, which must be performed in real time: i) the precise estimation of current travel times and ii) the short-term prediction of the highway traffic state evolution.

Two additional goals join this first objective. On the one hand, the key proposed methodology tries to make the most of all available data sources. That is, it takes advantage of the potential of the latest technologies that are being introduced both in vehicles and in the infrastructure, but without disregarding the valuable information that traditional surveillance, still much more common, provides. Therefore, no expenses specifically devoted to travel time estimation are needed, as the methodology adapts to the evolution of traffic surveillance. On the other hand, the method is derived with a vocation of continuity in the sense that, although already applicable, it will also be valid and even more necessary in future cooperative environments. Moreover, the goodness of its results will increase as the role of vehicles as data sources increases.

However, it cannot be overlooked that, up to now, many roads worldwide exclusively depend on loop data. Consequently, travel times on them are still obtained by using spot speed methods. Even when this situation changes, it seems worth taking advantage of these widespread detectors and these well-known procedures. Therefore, the second objective of this book is to enhance their outputs for those cases in which they are already acceptable, i.e., when no traffic transitions exist or, better, for free flowing situations. In this regard, an algorithm aimed at deriving space mean speeds, those truly related to travel times, instead of time mean speeds, from inductive loop detectors is introduced. Based on statistics, neither other detectors nor modifications in the loops are necessary.

Thirdly, this book aims at providing an overview of some key challenges that must be overcome to achieve efficient traffic management in more complex scenarios.

Firstly, in urban areas and, secondly, in future cooperative automated driving environments. In both cases, traffic management, to be effective, will have to be able to react to continuous variations in traffic conditions with a quickness and an accuracy that has not been achieved yet, at least in practice. This overview will help readers understand how management systems have evolved and where we are today. It therefore aspires to serve as a starting point for further research that will take advantage of current technological developments to make further progress towards the efficient traffic management that we all desire.

Special Acknowledgements

First of all, the editor of this book would like to gratefully acknowledge Springer Nature and, particularly, Mr. Holger Shaepe (Senior Editorial Assistant), Mr. Roger P. Roess (Series Editor) and Dr. Thomas Ditzinger (Editorial Director) for this opportunity and for their patient, as well as Ms. Viradasarani Natarajan (Project Coordinator) and Ms. Femina Joshi (Project Manager) for their support throughout the development process.

I would also like to express my sincere and deep gratitude to Prof. Emeritus Jaume Barceló, who generously agreed to collaborate in the writing of this book and whose contributions are, unsurprisingly, invaluable. It has been a great honor to collaborate with him and also a great pleasure: I have learned a lot, but I have also laughed a lot during these months of joint work.

Also, Prof. Barceló and I would like to thank the kind collaboration of Dr. Heribert Kirschfink, Mr. Marco Boero, Dr. Josefa Hernández, Mr. Josep M. Aymamí, Dr. Emmanuel Bert, Professor Guido Gentile and Mr. Lorenzo Meschini, who provided us with very useful information and figures for the writing of some of the chapters.

However, this book would not have been possible without the knowledge and support I have been receiving during the last years from my Ph. D. thesis director and now friend and colleague Prof. Francesc Soriguera. In fact, this book can be considered the continuation of the one he himself published in this Series. His research was an inspiration to me, as for him was that of Prof. Francesc Robusté, to whom I am also deeply grateful for his guidance over these years and for his kind welcome since I first arrived at the Polytechnical University of Catalonia.

Last but not least, I would like to thank wholeheartedly all my dear ones, especially my parents, for their love and patience not only during the writing of this book, but throughout my career. In fact, during my whole life. I love you.

Barcelona, Spain Margarita Martínez-Díaz

Overall Approach and Outline

Overall Approach

Since the beginning of time, human welfare has been linked to mobility opportunities. The first civilizations moved looking for the best places to settle or fleeing from enemies. To certain extent, so do we nowadays too. Mobility represents freedom. It allows us choosing where to live, where to work, where to have fun.

With the advent of Internet and with new technologies becoming increasingly appealing and available, the end of personal mobility was envisaged. However, this has not been the case. Internet enables us to work at home, to buy online, or to hold a meeting with people who are thousands of kilometers away. Nevertheless, we keep on moving. What's more, the mobility rate increases in line with the gross domestic product (GDP). And so does the number of private cars for the moment, although sustainability issues will probably curb the trend in a near future. The reason is that the potential demand for transport systems follows an increasing trend parallel to the development of modern societies. We do not usually move for the simple pleasure of enjoying the trip itself, but to satisfy a need or to obtain a benefit. We move looking for accessibility to those activities inherent to the nature of our societies. These needs, the patterns, the means, etc., change with time. However, mobility requirements/desires remain.

In this context, we must keep in mind that mobility involves costs. Some of them exclusively concern transportation users (e.g., travel time, vehicle amortization, energy needs). However, externalities like pollution and other ecological impacts, noise, safety problems, land occupation, expenses devoted to the construction of the infrastructures, etc., affect the whole population. These costs increase with congestion, which leads to the inefficiency of the system and could even result in a global gridlock. In fact, congestion is already a major problem in most countries worldwide. As an example, Bogota commuters spent more than 191 hours in traffic jams in 2019, and those of Rome more than 166 (Inrix, 2020). There is a need for urgent solutions aimed at ensuring an efficient, safe, inclusive, and environmentally friendly mobility. In other words, at making mobility sustainable. The COVID-19 pandemic made

us more aware of the consequences of misconceived mobility: during the mobility restrictions, we could observe how pollution and noise in urban areas decreased, how some animals returned to the cities, how there was hardly any space in the cities for pedestrians, etc. Undoubtedly, these reflections fed the conscience of many users and administrations, and it was thought that, after the pandemic, things would not go back to the way they were before and that everyone would do their part: users would use more soft means of transport, administrations would promote public transport and reorganize cities, companies would facilitate teleworking, etc. And so it has happened, but perhaps not to the extent dreamed of. In fact, congestion problems are returning to the cities and make clear the need not only for the aforementioned change of awareness but also for the continued implementation of management strategies.

In fact, and focusing on road transportation, traditional measures consisting in incrementing road physical capacity by increasing the number of lanes of an existing infrastructure or by constructing new ones, usually result inefficient nowadays, above all in infrastructural developed countries. Moreover, they are often unrealizable due to the lack of free space. In fact, the current road network worldwide is mostly complete and well developed. That is why today's solutions should thus lie, mainly, in traffic management. More in particular, in the implementation of active *ad hoc* management strategies based on real-time data and adapted to face current traffic conditions. These strategies must aim to the equilibrium between supply and demand by improving the first one and regulating the second one. This idea is not new and, supported by the advent of new technologies, particular strategies have already been designed and/or implemented. Examples of those aimed at optimizing the available capacity are the dynamic management of speed limits, the dynamic lane assignment or incident management. For their part, freeway access management or high-occupancy vehicle (HOV) lanes are cases of demand management strategies.

At this point, the question is why congestion continues to be a severe challenge if we have the technology and the knowledge to avoid or at least to relieve it. The answer is twofold. On the one hand, neither users nor some administrations were so far aware of one key fact: the goal of traffic management must be the optimality of the whole system (Wardrop, 1952), above particular interests. On the other hand, and probably linked to this first reason, the lack of enough efforts (economic, but also planning and operation-related) to implement this kind of strategies has been noticeable thus far. Apart from remarkable exceptions or pilot tests, smart roads, smart transportation systems, smart cities, etc., are not a generalized reality yet even when they have been extensively tackled in research. In fact, the presence of static traffic management strategies based on average historical data still prevails. Although helpful to some extent, they are insufficient to deal with current traffic problems. Therefore, many changes are needed so that the aforementioned appealing scenarios come true.

Fortunately, there is good news too. Firstly, the incredible evolution of technology in the last years allows having more data than ever before. There are not only new gadgets with outstanding capabilities, but their prices are increasingly competitive. Therefore, their penetration rate grows over time. Think for example of the widespread use of smart phones in the last years. Secondly, a change of mind can be glimpsed. Traffic-related undesirable issues as well as previous (and current)

bad experiences are making stakeholders conscious of the need for a real paradigm shift. Additionally, the implementation of varied research results in this regard in several areas is already demonstrating the benefits of dynamic traffic management systems. Finally, the appealing of new technologies also plays a role, as many agents are keen on applying them as soon as they reach the market. The most illustrative example of this attraction is that of "autonomous" (a tricky word that is going to be clarified in this book) vehicles, which are in the spotlight of automakers, technological companies, users, administrations, researchers, etc. The opportunities that technology offers to traffic control and regulation are indubitable. Nevertheless, as it is going to be discussed in the next chapters, the success of any traffic management strategy will not lie in the involved technology but in how this technology is applied. That is, in whether its use is aligned with traffic flow principles.

Given this background, highway travel time information systems are (corridor) demand-side traffic management strategies, and can also be inputs for improving the supply of the road network. However, they are considered special and standalone systems due to the significance of their output. In fact, accurate travel times (both their predictions and their reliability) constitute probably the most important information that both drivers and traffic management centers can handle. For the first ones, it is an easily understandable variable (Turner et al., 1998; van Hinsbergen et al., 2007) that allows them, for example, to change their route, their departure time, or even their mode of transport to avoid congestion. With regard to traffic agencies, travel time reliability is the best indicator of the level of service of a road/stretch. In fact, this concept has already been introduced in, for example, the last editions of the American Highway Capacity Manual (2016) or the German Handbuch für die Bemessung von Straßenverkehrsanlagen (2015). Additionally, real-time travel times are the most important variable to quantify congestion and a critical input for active traffic management strategies.

Travel times have and are being measured in very different ways and with different equipment, mainly according to the surveillance and computational capabilities available. In any case, it must be taken into account that each procedure leads to different results. Not only regarding accuracy but also with respect to the real meaning of the travel times that are being obtained. On the one hand, travel times along a link of a highway can be obtained directly by identifying individual vehicles at the beginning and at the end of the link, or by tracking vehicles during the whole trip. Automatic Vehicle Identification (AVI) devices or tracking technologies such as cell phone geolocation or GPS are, respectively, needed to this end. The penetration rate of the necessary technology was a big issue in the past. Nowadays, its availability is much larger, especially in those highways with big traffic volumes. However, it cannot be overlooked that measured travel times are obtained once vehicles have traversed the target stretch. Therefore, they could be considered "obsolete" information for the next drivers entering it. On the other hand, travel times can also be obtained indirectly. That is, other traffic variables such as speeds, flows, etc., are measured and travel times are afterwards calculated from them. The most common source of data for this purpose are inductive loop detectors, which are present on most roads. In fact, travel time estimation relied almost exclusively on them for a long time. Usually,

punctual measurements of speed are first averaged for predetermined time intervals. The obtained means are then extrapolated to the links between the measuring points (i.e., the points where the loops are located), and the result of these extrapolations is used to estimate the travel times in the links. Finally, link travel times are added up to obtain corridor travel times, i.e., travel times along several consecutive links. These methodologies have two baseline mistakes: i) the procedures used for the spatial generalization of the punctual mean speeds do not consider traffic dynamics and queue evolution, but are just mathematical interpolations and ii) even if the former approach were enhanced, space means and not time mean speeds should be used to calculate travel times, according to traffic flow theory. Research has demonstrated that travel time estimations via spot speed methodologies are only satisfactory when traffic is free flowing or the density of detectors high. As it is precisely in congestion when travel time information is more valuable, and taking into account that very high surveillance density is not available everywhere, the need for new more appropriate procedures is indisputable. In any case, either as a standalone system when no other data source exists, or as a complement to other procedures suitable for congestion, spot speed methods should rely on space mean speeds.

Researchers have tried to overcome the disadvantages of individual (i.e., one source-based) procedures to estimate travel times by means of data fusion. Furthermore, technological progress has provided both new surveillance and more powerful computational capabilities from which highway travel time information systems can benefit. Nevertheless, the penetration rate of these novelties is uneven among networks and even among the links of each single network. Therefore, efforts must be focused not only on very ambitious schemes but also on improving the simplest ones. Very interesting existing initiatives on data fusion for travel time estimation are reviewed in this book. However, until now there was lack of a methodology that, at the same time: i) did not provide past or instantaneous travel times, but (short-term) travel time predictions, ii) were generalizable in current basic scenarios without the need for large investments, and iii) were applicable in the (near) future cooperative driving environments as well, in which accuracy requirements will be much higher, but many more input data will be available. This book introduces such a data fusion methodology, which couples new technologies with basic surveillance and the classic (and key) principles of traffic flow theory.

Finally, it is necessary to draw attention to urban environments. Despite their importance, few cities have travel time information systems similar to those used in highways and freeways. Instead, most agencies deliver coarse estimates of travel time obtained via unreliable and too simplistic approaches. The lack of appropriate surveillance and the complexity of traffic in urban sites are usually behind these procedures (Mori et al., 2015). It must be noted that urban environments, like corridors, add an additional complexity to the traffic system as far as they exhibit a significant difference with respect to, for example, freeways, due to the existence of multiple paths between the origins and destinations of the trips. Understanding behavioral aspects on how drivers choose these paths, namely in terms of travel times, becomes a significant component of traffic management systems. Although this book specially refers to highways or freeways, i.e., to uninterrupted traffic, the proposed

methodologies are also applicable to urban environments. Defined at the link level, the presented algorithms could be suitably modified to appropriately deal with urban or arterial roads and to account for the different boundary conditions they imply. Proper modifications would be necessary depending on the available surveillance and on the characteristics of the site. Also in this respect, the last two chapters of this book extend the point of view of the first ones, since they analyze travel times in their role as inputs for dynamic traffic management systems in their most global concept. These systems play a very important role in road networks, but they are crucial in urban areas, where management is often more complex. Hence, they have been evolving, and continue to do so, with the aim of achieving increasingly accurate outputs and, therefore, being able to define the most appropriate management strategies in real time. As will be seen in detail, one of the critical steps for the correct operation of these systems is to obtain, also in real time, precise dynamic origin-to-destination matrices, which allow the correct definition of the demand of the different links of the network. Again, numerous methodologies have been proposed to obtain these matrices (and, again, the accurate prediction of travel times on the different links/routes is fundamental), but difficulties still arise. Increasingly advanced approaches to obtain these matrices are being tested, most of them trying to make the most of "new" data sources, especially mobile phone data. In fact, data-driven methodologies are attracting the interest or researchers and are called to complement other solutions based on traffic models.

Outline

This book is divided into four parts, each one grouped in different chapters. The relationships between the chapters and their main content are schematically outlined in Fig. 1.

After the preface and this introductory section, Chapters 1 and 2 of the first part of the book constitute a detailed report of the concepts, theories, and tools that are used in the following Chapters 3 and 4. More in particular, Chapter 1 clarifies the different equipment and procedures that are being used for traffic monitoring and reconstruction, whereas Chapter 2 introduces the key variable of this book: the travel time. Its significance for traffic management, the different definitions that exist under its general concept, how these definitions are linked to the way (and equipment) in which they are obtained, etc., are analyzed.

Part II, which is divided into three chapters, contains the main contributions of this book, and constitutes its more novel part. Both Chapter 3 and 4 focus on highway travel time information systems. Chapter 3 introduces a methodology to improve the simplest current procedure for travel time estimation, which uses punctual measurements of speed registered on double-loop detectors. Spot speed methods are commonly used with one important initial error: the use of time mean speeds, while average travel times result from the integration of space mean speeds. The proposed methodology is able to derive space means from the time means provided by loop detectors without the need for extra data sources. Next, Chapter 4 addresses

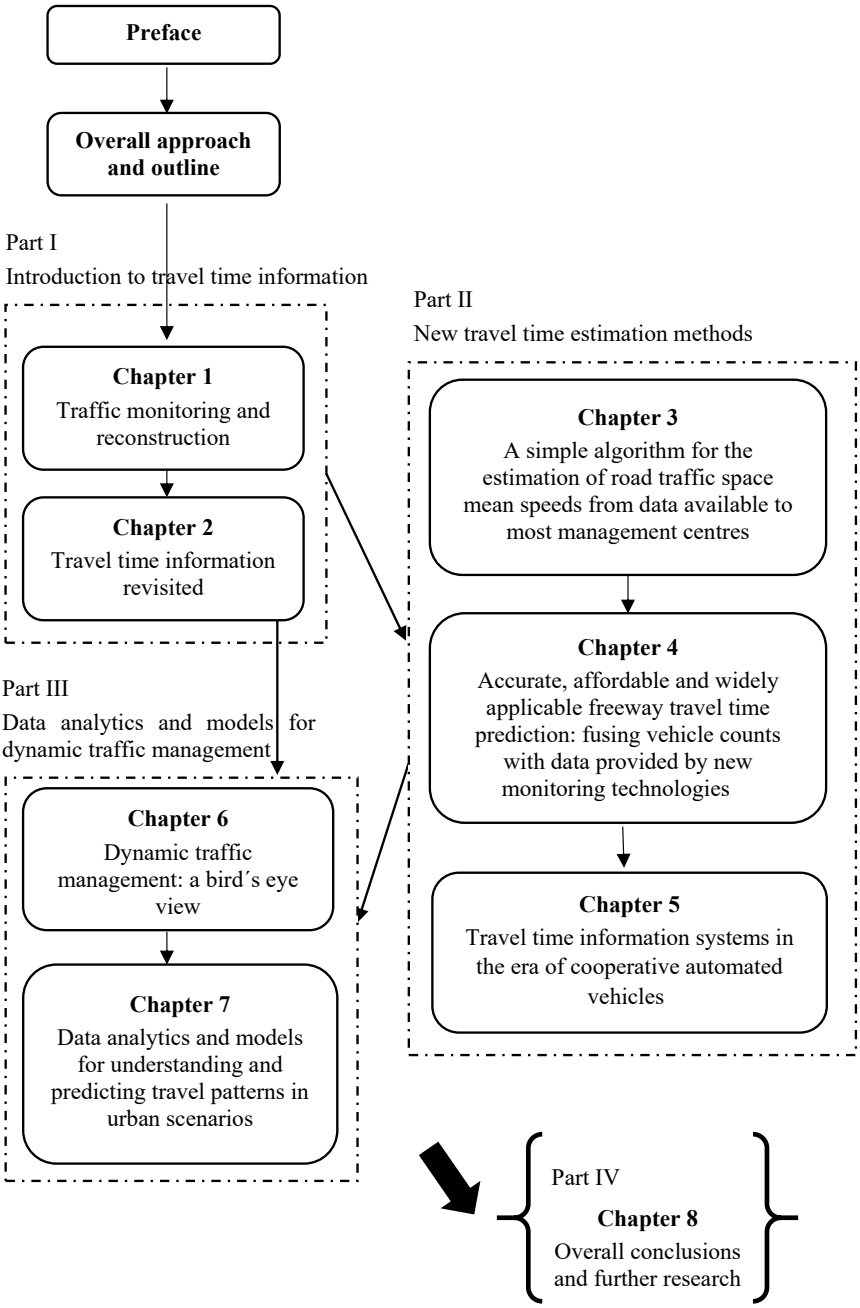

Fig. 1 Book outline

the most challenging and promising part of this book. It introduces a data fusion algorithm that is able to predict travel times in real time. To make the predictions, this algorithm combines data obtained from loop detectors with a comparatively small amount of direct measurements of travel time. Particularly, it fuses eulerian and lagrangian data. Especially the latter are becoming increasingly available in line with vehicle automation. Therefore, the suitability of the methodology both today and in future mixed (i.e., with traditional and intelligent vehicles sharing roads) and cooperative environments is guaranteed. Both Chapter 3 and 4 include *ad hoc* experimental studies with real and simulated data that demonstrate the good performing of the proposed methodologies. Next, Chapter 5 serves as a link between the preceding chapters and those that follow, as it describes what future cooperative automated driving environments will be like and analyzes what the role of travel times will be in these new scenarios and what requirements they must meet in order to remain key information.

The third part of this book broadens the outlook and addresses the evolution dynamic traffic management systems, including those accounting for urban areas and for cooperative driving environments. Firstly, Chapter 6 explains the different approaches with which traffic management systems have been developed so far and addresses those that are gaining interest at present in parallel with the emergence of new data sources. For its part, Chapter 7 expounds on data analytics and models that are being used to understand and predict travel patterns in urban scenarios, which are indispensable to achieve an efficient traffic management. The dynamic estimation of origin-to-destination matrices constitutes a key part of this chapter.

Finally, the most important conclusions drawn from both the research and comprehensive literature review performed are highlighted in Chapter 8 of Part IV. This chapter also outlines possible lines of future research.

References

Handbuch für die Bemessung von Straßenverkehrsanlagen 3. Version (2015) Köln: Forschungsgesellschaft für Straßen- und Verkehrswesen e. V.

Highway Capacity Manual 6th Edition (2016) Washington, D.C.: Transportation Research Board.

Inrix (2020) *Inrix Global Traffic Scorecard of 2019*

Mori U, Mendiburu A, Álvarez M, Lozano JA (2015) A review of travel time estimation and forecasting for advanced traveller information systems. Transp A: Transp Sci 11(2):119–157

Turner SM, Eisele WL, Benz RJ, Holdener, DJ (1998) Travel time data collection handbook. Research Report FHWA-PL-98-035. Washington, D.C.: Federal Highway Administration, Office of Highway Information Management

van Hinsbergen CPIJ, van Lint JWC, Sanders FM (2007) Short term traffic prediction models. Proceedings of the 14th World Congress on Intelligent Transport Systems (ITS), 2007, Beijing, 1–18

Wardrop J (1952) Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers, 1(2), 325–378

Contents

Part I Introduction to Travel Time Information

1 Traffic Monitoring and Reconstruction 3
Margarita Martínez-Díaz

2 Travel Time Information Revisited 31
Margarita Martínez-Díaz

Part II New Travel Time Estimation Methods

3 A Simple Algorithm for the Estimation of Road Traffic Space Mean Speeds from Data Available to Most Management Centers . . . 67
Margarita Martínez-Díaz

4 Accurate, Affordable and Widely Applicable Freeway Travel Time Prediction: Fusing Vehicle Counts with Data Provided by New Monitoring Technologies 101
Margarita Martínez-Díaz

5 Travel Time Information Systems in the Era of Cooperative Automated Vehicles 139
Margarita Martínez-Díaz

Part III Data Analytics and Models for Dynamic Traffic Management

6 Dynamic Traffic Management: A Bird's Eye View 165
Jaume Barceló and Margarita Martínez-Díaz

7 Data Analytics and Models for Understanding and Predicting Travel Patterns in Urban Scenarios 201
Jaume Barceló, Xavier Ros-Roca, and Lidia Montero

Part IV Overall Conclusions and Further Research

8 Overall Conclusions and Further Research 281
Margarita Martínez-Díaz

Glossary 287

Part I
Introduction to Travel Time Information

Chapter 1
Traffic Monitoring and Reconstruction

Margarita Martínez-Díaz

Abstract In traffic engineering, as in so many other disciplines, any good analysis requires data. Regardless of whether the most powerful software is available, it will not produce good results if it does not receive the necessary inputs. It is generally accepted that the more data available, the better results can be achieved. Omitting data-driven techniques, this is true only if the data is adequate and, of course, more or less accurate. In this sense, the equipment that collects the data also plays a fundamental role, since it will determine what data can be collected and in what amount. This chapter provides a simple but very useful classification of the most commonly used sensors and explains the data they can collect. It also gives a brief and simplified introduction to the reconstruction of traffic conditions from these data using the most common techniques. Both aspects will be discussed in more detail throughout this book.

1.1 Introduction

The need for traffic data collection grew in parallel to the widespread adoption of the automobile, and the subsequent development of the road network. At the beginning, the goal was the measurement of traffic volumes aimed at planning. Manual (and visual) counts were the most usual monitoring procedure in those years. The mass production of the Ford Model T on a moving assembly line, initiated by the Ford Motor Company in 1913, constituted a great leap forward for the automobile industry and for society. Mass production allowed companies to lower their selling prices, making thus vehicles affordable for a broader sector of the population. Particularly, Ford had already produced over 15,000,000 Model T automobiles by 1927 (Banham 2002). With more and more vehicles on the roads, the need for (i) the

M. Martínez-Díaz (✉)
UPC-BarcelonaTech, Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, Barcelona, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_1

control of the level of service on the existing roads and (ii) a more rigorous network planning process, arose. Counts, although necessary, became insufficient to meet the requirements of these purposes. Among others, average speeds turned out to be indispensable too (Highway Research Board 1950).

Additionally, manual data collection led either to inaccuracies or to huge staff costs. The support of technology was seen as essential. The first known deployment of a vehicle detection device was that of a semi-actuated signal. Installed in 1928 at an intersection in Baltimore, drivers were required to honk in order to activate the detector, which consisted of a microphone mounted in a small box on a nearby utility pole. The right-of-way was then assigned according to the information collected by this sensor. Although useful, it was too rudimentary so as to be sustained for a long time. On the contrary, a treadle-type detector proved at the same time became a common means for vehicle detection at actuated signals over some years. It was a pressure-sensitive pavement detector with two metal plates that acted as electrical contacts and were forced together by the weight of the passing vehicles (Institute of Transportation Engineers 1991). The next monitoring gadget introduced was an electro-pneumatic detector. Although it also became a popular method for vehicle monitoring for some years, the fact that it was only able to detect vehicle passage did not compensate its high installation costs.

Despite being the best available option, and accepting that weight is the most easily detectable and quantifiable property of vehicles, the treadle-detector also suffered from frequent inconveniences. First, there were mechanical problems with the contact-plate sensor. Second, rock falls or snowplows, for example, usually lifted the plate from the roadway. Additionally, the whole detector had to be reinstalled after any kind of pavement repair. From then on, efforts were focused on the development of detectors that measured more subtle properties, like the following ones (FHWA 2006):

- Sound
- Opacity (optical and infrared sensors and video image processors)
- Geomagnetism (magnetic sensors, magnetometers)
- Reflection of transmitted energy (infrared laser radar, ultrasonic sensors and microwave radar sensors)
- Electromagnetic induction (inductive loop detectors)
- Vibration (triboelectric, seismic and inertia-switch sensors)

All these properties and the related sensors are currently used. Moreover, inductive loops, introduced as a vehicle detection system in the early 1960s, are still the most widespread source of traffic data despite the fact that other more modern devices are gaining ground (e.g., mobile phones). Most of the aforementioned sensors have evolved with time and, in fact, this evolution continues. Think for example of new loops capable of (re)identifying vehicles. However, the complexity of current traffic scenarios, which is expected to increase in line with vehicle automation, as well as the subsequent intricacy of actual traffic management strategies, do not only demand sensing devices. In fact, the impressive development of computing and communication technologies in the last years, which continues, results decisive. They

have already brought about the possibility of collecting, processing and delivering data in real time. That is, they made possible the birth of the so-called *Intelligent Transportation Systems* (ITS).

Although this concept gained impact since the end of the 1990s, the idea of taking advantage of the new communication and computational capabilities for road transportation is older. According to the American Transportation Research Board (Weiland and Purser 2000), it was already in the early 1980s when a small group of Japanese transportation professionals came up with it. They later called their idea the *Japanese Intelligent Vehicle System* (IVS) *Program*. Siemens was also doing some pioneering work on route guidance systems in Berlin in the 1980s. In those years, the Europeans referred to these initiatives as *Road Transport Informatics* and later, with information and communication playing an increasingly important role, as *Advanced Transport Telematics* (ATTL). The United States addressed these topics in the late 1980s, at first referring to them as *Intelligent Vehicle Highway Systems* (IVHS). Afterwards, the designation ITS was chosen, giving recognition to the wider application of technology not only to private vehicles and highways, but also to public transport and general roads. The majority of transport organizations and stakeholders (Andersen and Sutcliffe 2000) have finally adopted this name.

As said, ITS do not only consist of surveillance. They could be defined as comprehensive and complex systems that combine sensors, high-level technology, communications, controllers and advanced mathematical and/or computer science approaches, with the goal of managing traffic in a sustainable way (Sussman 2005). All of them can be integrated in different ways. However, the basic architecture of ITS is shown in Fig. 1.1. First, the physical layer is that composed of all elements of the transportation system, including the surveillance. At first, sensors were exclusively deployed in the infrastructure. Currently, those carried by users (smartphones, tablets, etc.) and that located on-board enormously enrich databases. Particularly the last ones are becoming increasingly important in parallel to vehicle automation. Second, the communications layer makes the interchange of data or information between all

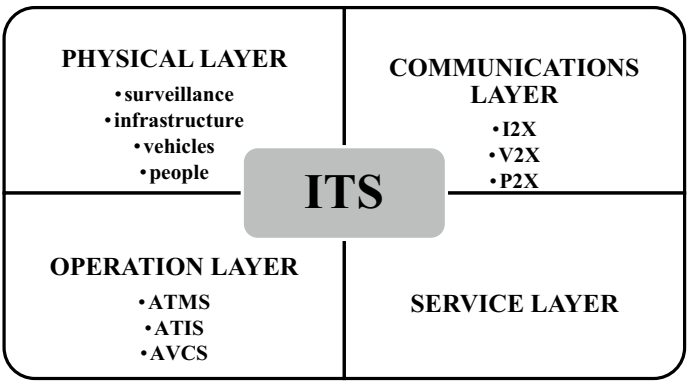

Fig. 1.1 Basic architecture of current intelligent transportation systems (ITS)

involved agents possible. I2X refers to *Infrastructure-to-All communications*, V2X to *Vehicle-to-All communications* and P2X to *Person-to-All* communications (where "person" stands for drivers, passengers, pedestrians, bikers, etc.).

For its part, the operations layer is that of traffic management systems. That is, at this level the collected data are processed and the subsequent information is distributed. Although the general term *Advanced Traffic Management Systems* (ATMS) is often used, they are also grouped in different ways. For example, the following classification is usual:

- *Advanced Transport Management Systems* (ATMS)
- *Advanced Traveler Information Systems* (ATIS)
- *Advanced Vehicle Control Systems* (AVCS)

With this classification, ATMS include strategies aimed at optimizing the available capacity or at managing the demand, excluding those that involve providing users with information. These ones, which deliver for example travel times or incident warnings, constitute ATIS. Finally, AVCS directly communicate management strategies or information to vehicles, and not to drivers. AVCS have meaningfully progressed in the last years and their significance will grow when autonomous or at least highly automated vehicles hit the road (Shladover 1990). When referring to these three groups, the adjective "advanced" is usually substituted by "active", i.e., all these systems collect and process data in real time and, consequently, suggest or order any kind of behavior (or simply inform users), also in real time. This difference with regard to traditional traffic management or information systems (often called "passive"), which work based on statistics or average past data, is their most important feature and the key for their effectiveness. Decisions are taken on the basis of the current traffic state. The shorter the time interval of data aggregation, the faster and more effective the response to any inconvenience may be. But this response will only be appropriate if the amount of data collected is enough and if the information it contains is correctly extracted, also in a short period of time. Finally, the service layer refers to the area where all services are deployed and run, and/or to the public or private responsible operator (Lin et al. 2017).

The preceding paragraphs remark that ITS is much more than just sensors. However, the present chapter is especially devoted to traffic monitoring, i.e., to the sensing system. The reflections of Palen (1997) help to explain why they deserve such attention: "*An Intelligent Transportation System, by definition, involves the use of intelligence to enhance the operation of the transportation system. Intelligence, by definition, requires information. Information, by definition, is data formulated in a formation. Data are generated by surveillance. Therefore, surveillance forms the basis for the formation of information for an ITS. You can't have a usable ITS without surveillance*". These considerations are valid nowadays and will also be applicable to future cooperative driving environments. Traffic data collection (and standardization) is the first and usually more important step in any kind of traffic management strategy, as it is for research and development in the field (Barceló and Kuwahara 2010).

1.2 Eulerian Sensing Versus Lagrangian Sensing

As explained, the amount of different parameters needed to reconstruct traffic states and, thus, to derive proper management strategies, grows in parallel to the complexity of driving scenarios. Therefore, the number and variety of required sensors also increase, and so do the expectations about their capabilities and accuracy. As it is expounded in Chap. 5, the role of the sensing system in future cooperative scenarios, when vehicles will monitor the environment themselves, will even be more crucial than today. If this monitoring task is correctly performed, the first step to lower the accident rate would be taken, as the human factor and its associated errors would disappear. On the contrary, erroneous and/or insufficient data delivered in a cooperative environment could have terrible consequences.

Starting with current scenarios, the most important parameters that sensors collect are counts, speeds, occupancy, size (length and/or weight), location and emissions. Particular sensors are aimed at measuring one or some of them. The way in which these sensors perform must also be considered as, consequently, the obtained measurement can have different and important nuances (see the cases of speeds and travel times, respectively addressed in Chaps. 3 and 4). Many different classifications of these measurements (and consequently of their sources) are possible, but none fits the central topic of this book as good as the one that divides them into eulerian or lagrangian measurements. As on other occasions, mechanics (fluid flow theory) is in the origin of these concepts (Lamb 1895), which have later been applied to traffic. Eulerian data are provided by static traffic sensors, which measure variables through an immovable control volume, i.e., a fixed coordinate system is used. On the contrary, lagrangian sensors collect data along the trajectory of a particle (a vehicle), i.e., the coordinate system they use moves with this particle. Table 1.1. includes the most widespread sensors of each type and the variables they measure, which are next addressed. Experience demonstrates that a strategic combination of several from both types of sensors and the implementation of ad hoc data fusion procedures yield the best results. Future cooperative automated environments are called to rely on these comprehensive schemes.

1.2.1 Eulerian Sensors in Traffic Monitoring

Inductive loop detectors, toll tickets, traditional cameras and Automatic Vehicle Identification (AVI) technologies are examples of eulerian sensors. Currently, toll tickets are seldom used alone for research purposes, as the number of vehicles equipped with Electronic Toll Collection (ETC) systems increases over time. Nevertheless, they can be profitable in a data fusion scheme (see Sect. 2.2.3 in Chap. 2). For their part, traditional cameras are part of the basic control equipment of traffic management centers. However, their current role is often limited to incident detection.

Table 1.1 Examples of (**a**) eulerian and (**b**) lagrangian sensors, and the most usual data they collect

Type	Name		Key measurements	Notes
Eulerian	Induction loop detectors	Simple	Counts (flow), occupancy	
		Trap (double loops)	Counts (flow) , occupancy, spot speeds, vehicle length	Spot measurements are usually averaged over a determined time interval
	(Re)identification of toll tickets		Counts (flow), speeds, travel times	In closed turnpikes
	Video cameras		Counts (flow), speeds, density, lane changes	Mostly human analysis and poor accuracy Seldom for reidentification
	Automated Vehicle Identification Technologies (AVI)	Bluetooth or WIFI-signal detectors	Counts (flow), speeds, travel times	Either on-board Bluetooth/WIFI or that of passengers' smartphones
		ALPR (Automatic License Plate Video Recognit.)	Counts (flow), speeds, travel times, vehicle classification, lane changes	Also known as ANPR or OCR
		(Toll) Tag (re)identifi-cation	Counts (flow), speeds, travel times	With varied systems, either with closed or open toll configurations
		(Re)identi-fication by means of special loops	Counts (flow), speeds, travel times	Reidentification by comparing vehicle's electromagnetic signature or vehicle's length at two different loops

(continued)

Table 1.1 (continued)

Type	Name	Key measurements	Notes
Lagrangian	GPS tracking	Trajectories, from which all kind of data can be derived	Either on-board GPS or that of smartphones that travel in the vehicle
	Cell phone signal geo-localization	Trajectories, from which all kind of data can be derived	Raw measurements have poor accuracy
	Radio Frequency Identification (RFID)	Trajectories, from which all kind of data can be derived	Chips can also be controlled at fixed points, i.e., they can be used as eulerian
	Probe vehicles and dynamic floating vehicles	Counts, speeds, braking forces, lane changes, travel times, weather conditions, trajectories	Ad hoc probe vehicles can carry different sensors and thus collect varied data, depending on the objective sought
	Unmanned aerial vehicles (drones)	Trajectories, from which all kind of data can be derived	Promising. Unknown reliability as of today

More advanced vision techniques included in the AVI group are used to automatically collect other kinds of data. In fact, together with inductive loop detectors, AVI technologies are the most worthy eulerian sensors nowadays.

1.2.1.1 Inductive Loop Detectors

As explained, traditional inductive loop detectors are the main source of traffic data thus far. They are not expected to disappear in future more technological driving environments and the data they provide will continue to be useful within more complex schemes, as it will be demonstrated in Chap. 4. Therefore, it is worthwhile to take a closer look at their operating mode.

Single inductive loops consist of an electrically conducting wire loop installed under the pavement of a particular lane, and an electronic unit that transmits energy to it. The pass of a metallic object (e.g., a vehicle) over the loop induces eddy currents in the wire, and inductance thus decreases. This decline activates the electronic unit, which sends a pulse to the traffic signal controller. In this way, the vehicle is counted.

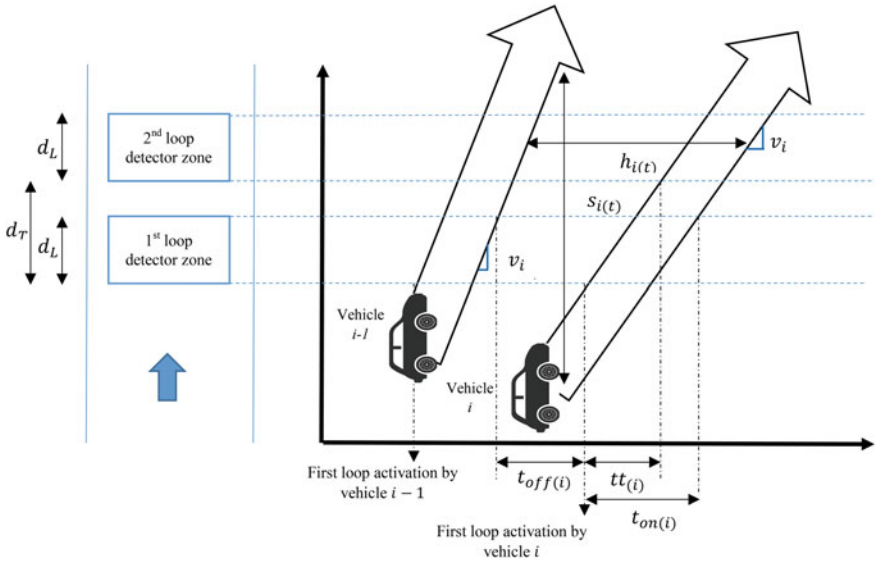

Fig. 1.2 Double-loop detector configuration scheme and some measurements available from it (adapted from Soriguera 2016)

Additionally, as the sensor remains activated until the vehicle leaves the detection zone, occupancy can also be calculated. Finally, it is possible to derive vehicle spot speeds, usually by considering an average constant vehicle length. However, this assumption leads to inaccuracies. Researchers tried to overcome this problem by the modification the loop detector controller (Coifman 2001; Coifman et al. 2003; Hellinga 2002; etc.). In spite of the promising results, such adjustments would be tedious and expensive in practice. On the contrary, vehicle spot speeds can be easily obtained when loops are placed in pairs. Considering this fact, the deployment of double-loop detectors (also called dual-loop detectors or speed traps) has been generalized. Figure 1.2. represents one of these traps and some important parameters that can be measured with them (Soriguera 2016). Important features are shown as, for example, the difference between the detection zone of each loop,d_L, and the distance between equivalent points of each loop of the trap, d_T. This difference reaches 1.5 m with the usual configurations, where $d_T = 3.5$ m and $d_L = 2$ m, and is sometimes overlooked when obtaining subsequent data, which leads to inaccurate results. The four basic measurements that double loops provide are also indicated:

- The instant when a vehicle activates the trap, i.e., when it enters the detection zone
- The time between the activations of each loop of the pair due to the passage of a vehicle i ($tt_{(i)}$)
- The time the first loop has remained off since the preceding vehicle $i-1$ left its detection zone ($t_{off(i)}$)

- The time the first loop remains on because of the passage of vehicle i ($t_{on(i)}$)

The most basic microscopic variables can be obtained from the former measurements (Eqs. 1.1 to 1.4):

$$v_i = \frac{d_T}{tt_i} \tag{1.1}$$

$$h_i = t_{off(i)} + t_{on(i)} \tag{1.2}$$

$$s_i = v_{i-1} * h_i \tag{1.3}$$

$$l_i = v_i * t_{on(i)} - d_L \tag{1.4}$$

where v_i stands for the punctual speed of vehicle i, h_i for its headway regarding vehicle $i-1$, s_i for its spacing also with regard to the preceding vehicle, and l_i for its length. Then, the averaged macroscopic characteristics of the traffic stream, compliant with Edie's (1965) generalized definitions for any region A with length*time dimensions, can also be obtained (Eqs. 1.5 to 1.7, where the formulae just after the first equal sign corresponds to Edie's definitions). When working with loops, this area is d_L*Δt. Δt is the time interval of aggregation for the averages, during which n activations of the loop took place.

$$q = \frac{\sum_{i=1}^{n} x_i}{A} = \frac{n * d_L}{\Delta t * d_L} = \frac{n}{\Delta t} \tag{1.5}$$

$$k = \frac{\sum_{i=1}^{n} t_i}{A} = \frac{\sum_{i=1}^{n} \frac{d_L}{v_i}}{\Delta t * d_L} = \frac{\sum_{i=1}^{n} \frac{tt_i}{d_T}}{\Delta t} = \frac{\sum_{i=1}^{n} tt_i}{\Delta t * d_T} \tag{1.6}$$

$$\overline{v_s} = \frac{1}{\frac{1}{n} * \sum_{i=1}^{n} \frac{1}{v_i}} = \frac{1}{\frac{1}{n} * \sum_{i=1}^{n} \frac{tt_i}{d_T}} = \frac{n * d_T}{\sum_{i=1}^{n} tt_i} \tag{1.7}$$

where x_i is the distance traveled by vehicle i in A and t_i the time this vehicle spent in the region. It can be observed that, with these definitions, the so-called fundamental equation of traffic flow (Eq. 1.8) holds:

$$q = \overline{v_s} * k \tag{1.8}$$

Despite this potential, the controllers of double loops do not usually estimate these macroscopic variables. In addition, individual data are not stored. Traditionally, only the following variables are calculated and sent to the traffic management center each Δt:

- n, the traffic count during Δt

- The time mean speed $\overline{v_t}$ (Eq. 1.9)

$$\overline{v_t} = \frac{\sum_{i=1}^{n} v_i}{n} = \frac{\sum_{i=1}^{n} \frac{d_T}{tt_i}}{n} = \frac{d_T * \sum_{i=1}^{n} \frac{1}{tt_i}}{n} \tag{1.9}$$

- The time the first loop remained on during Δt (Eq. 1.10), i.e., the occupancy (occ)

$$occ = \frac{\sum_{i=1}^{n} t_{on(i)}}{\Delta t} \tag{1.10}$$

- The average across time of vehicles' lengths $\overline{l_t}$(Eq. 1.11)

$$\begin{aligned} \overline{l_t} &= \frac{\sum_{i=1}^{n} l_i}{n} = \frac{\sum_{i=1}^{n} (v_i * t_{on(i)} - d_L)}{n} = \frac{\sum_{i=1}^{n} \frac{d_T}{tt_i} * t_{on(i)}}{n} - d_L \\ &= \frac{d_T * \sum_{i=1}^{n} \frac{t_{on(i)}}{tt_i}}{n} - d_L \end{aligned} \tag{1.11}$$

Depending on the standards, more data can be available. A deeper analysis of these aspects and some related issues is performed in Chap. 3, which also tries to mend a baseline mistake of current procedures. Before, Chap. 2 addresses how instantaneous travel times are being estimated from data provided by loop detectors. Additionally, Chap. 4 includes loop data in a fusion algorithm aimed at obtaining travel time predictions.

1.2.1.2 Automatic Vehicle Identification Systems

The fact that loop detectors are eulerian is obvious, as they are located at fixed points of the road. Moreover, they provide punctual (also called "spot") measurements. The case of Automatic Vehicle Identification Technologies could at first be confusing. AVI detectors identify individual vehicles at control points by means of the number of their license plate (ANPR or ALPR) (Fig. 1.3.), their Bluetooth signature, their electromagnetic footprint, an on-board electronic toll collection system, etc. As it will be elaborated in Chap. 2, they are increasingly used to collect direct measurements of travel time. To this end, the same vehicle must be identified at two different points, which is called reidentification, pairing or matching. This fact could lead to think that AVI technologies are lagrangian, as the travel time has a space–time component that depends on vehicles' trajectories. However, the key is that they use fixed points as references (the points where the detectors are located, i.e., the control points) and

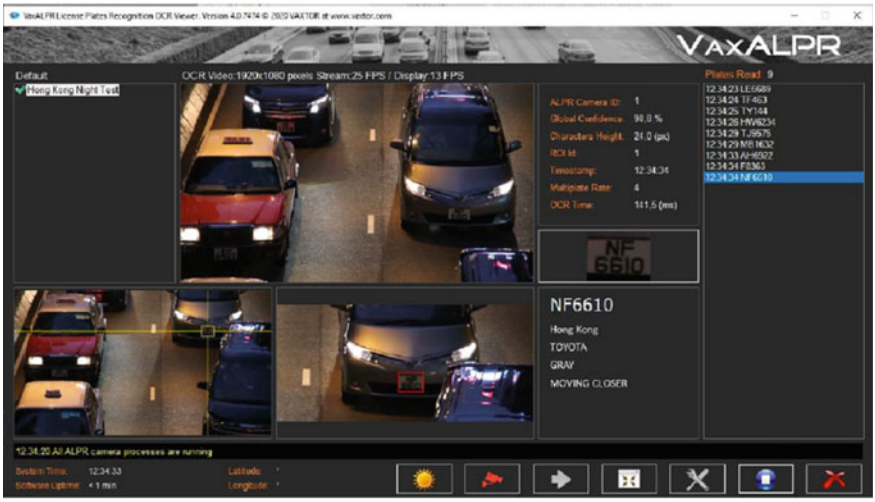

Fig. 1.3 Example of license plate recognition system (VaxALPR 2021)

do not literally track vehicles. Therefore, all AVI technologies belong to the eulerian group.

Although they can provide counts at a fixed point (like loops do), the interest of AVI techniques precisely lies in their ability to distinguish vehicles and, thus, to provide their individual travel times and average speeds between control points. Despite not being able to render the most valuable information, i.e., predictions, they can be used to feed highway travel time information systems. Additionally, individual vehicle identifications are very useful, for example, for the construction of origin–destination matrices, a key input for simulation models that is usually not easy to obtain (Barceló et al. 2010, 2013). This aspect is deeply addressed in Chap. 7. Some issues regarding accuracy rates, sample sizes, detectors' location and the fact that travel times can only be measured over delimited sections, among others, still represent challenges to overcome. Chapters 2 and 4 expound on these topics.

1.2.2 Lagrangian Sensors in Traffic Monitoring

In comparison to the eulerian, lagrangian sensors are being applied to traffic control and management for a short time now. However, they arouse an increasing interest, and the reason is twofold. On the one hand, the majority of them involve meaningful savings in installation and maintenance costs when compared to eulerian sensors, since they mostly work upon already established infrastructures and private devices. On the other hand, lagrangian sensors track vehicles. They can therefore provide individual trajectories and, thus, all important information (e.g., travel times, speeds,

flows, densities, the localization of bottlenecks, the formation or dissipation of shockwaves). However, several issues remain. Some of them affect only particular technologies. On the contrary, others like user concern about privacy are common to all of them. Next sections address the particular functioning and applications of on-board Global Positioning System (GPS) devices, mobile phones or Radio Frequency Identification (RFID) transponders, which are nowadays the most used detectors of this group. Some considerations about probe and/or floating vehicles and unmanned aerial vehicles are also included.

1.2.2.1 Global Positioning System

GPS receivers are part of the equipment of modern vehicles. Additionally, the use of extern GPS navigation devices that can be placed inside traditional vehicles has become common since the beginning of the century. Different private companies as Inrix, Tomtom, Garmin, Mitac, etc. offer varied products aimed at guiding drivers on their routes. These companies have realized the value that traffic data have, and they collect many types of data for other commercial and/or research purposes. In addition, the fact that smartphones include a GPS receiver, definitively boosted both GPS navigation support and the collection of GPS data for traffic management.

GPS is a satellite-based radio-navigation system owned by the government of the United States (US), and operated by the US Air Force. Other countries like China, Russia, Japan and, also, the European Union, have their own satellite-based positioning system. Nevertheless, none has proved as accurate as the American one so far. This is precisely the aim of China with its system Beidou, designed with 35 satellites. At the moment, it is the American GPS that has been adopted practically all over the world. The GPS navigation is based on two different components: GPS receivers and dozens of satellites and ground stations. GPS receivers are the above mentioned on-board devices, mobile phones, tablets, etc. They all have a GPS chipset with a powerful processor. This processor makes all calculations and is also responsible for the user interface, etc. Regarding the satellites, the first one was launched by the US Government in 1978, and the last one for the moment, the 31th, in 2018 (Space Segment 2018). At any given time, there are at least 24 active satellites orbiting over 12,000 miles above the Earth, and the rest are occasionally activated to improve accuracy by the provision of redundant measurements. The positions of the satellites are not coincidental, but they are thought to properly cover the Earth's surface. All satellites contain an extremely accurate atomic clock. Among other pieces of data, a GPS module receives (by radio frequency) a timestamp from each of the visible satellites, along with data on where each one is located in the sky. From this information the processor is able to calculate its distance to each satellite in view. If the receiver's antenna can see at least four satellites, the processor can accurately calculate its position and time. This procedure is called a lock or a fix (Doberstein 2011).

For traffic management, the updating time interval of the system (i.e., how often the receiver calculates and reports its position) is very important. Once per second

is the standard today, but chipsets that provide 10 updates per second are also available. Many variables have influence on the final accuracy. Signal noise, bad satellite positions regarding that of the receiver, bad weather, insufficient light, and obstructions (tunnels, buildings, mountains, etc.) provoke errors in the perceived location. Different GPS assistants (e.g., Inertial Measurement Units, IMUs) have been developed to support GPS in these situations. As in other cases, data fusion results beneficial.

1.2.2.2 Mobile Phones

First of all, a clarification is needed: the term mobile phone refers to either cell phones or to smartphones. On the contrary, these last terms are not interchangeable. Any smartphone is a cell phone with advanced features, but not all cell phones are smart. Basic cell phones connect to a wireless communications network through radio waves or satellite transmissions. Most of them provide voice communications, Short Message Service (SMS) and Multimedia Message Service (MMS). The newest may also provide Internet services such as web browsing, instant messaging capabilities and email. Smartphones are able to do many more things, depending on their own capabilities and on their operating system. They integrate cell phone functions with others typical of handheld computers or Personal Digital Assistants (PDAs). Smartphones additionally allow users, for example, to store information, make photos, install games and programs, etc.

A significant amount of contemporary research studies relies on cell phone or, above all, on smartphone data for traffic monitoring and management. This current focus on smartphones is linked to their penetration rate in society, which increases significantly throughout the world, but within the senior population (Berenguer et al. 2017). Smartphones could also be seen as part of an AVI scheme (i.e., eulerian) as they are able, for example, to transmit Bluetooth signals that can be (re)identified at fixed points of the infrastructure. However, their potential as tracking devices is more valuable. Additionally, their use for traffic monitoring avoids extra installation and maintenance costs, since it implies working upon already established infrastructures (for example, the cell network) and (usually) private devices.

Focusing on their tracking capabilities, Table 1.2. groups some relevant smartphone-based studies depending on the particular methodology used. Broadly speaking, two main different procedures based on dedicated mobile phones can be found in the literature: GPS-based techniques and network-based or cellular signal-based techniques. Smartphones and GPS-enabled cell phones are useful for both techniques. However, the most ancient cell phones could only be used for geo-localization. The basis of both approaches is simple: when one mobile phone is located at two points over time, both “its” speed and “its” travel time between these points can be calculated.

GPS-based techniques follow the steps addressed in the former section. They have already been used to identify traffic conditions, to analyze traffic patterns or drivers’ behavior, etc. (e.g., Yoon et al. 2007; Herrera and Bayen 2010; Herrera

Table 1.2 Examples and comparison of mobile phone-based traffic studies

Method		Fundamentals	Research examples	Pros and cons
GPS-based techniques		The mobile phones' built-in GPS sensor is used to accurately localize them	Yoon et al. (2007) Herrera and Bayen (2010) Tao et al. (2012) Herrera et al. (2010) Paek et al. (2010) Ge and Fukuda (2016) Sanaullah et al. (2016) Woodard et al. (2017)	• PROS: The most accurate localization No modification nor disturb in phone networks • CONS: GPS are power hungry Unreliable performance in urban environments
Network-based techniques	Passive COMM	Cell handovers and dwell times (i.e., cellular signal) are monitored to estimate each phone's location	Hansapalangkul et al. (2007) Janecek et al. (2012)	• PROS: Light modifications in phone networks No network overload • CONS: Less accuracy in phones' localization
	Active COMM	The same techniques as with passive communications. The number of calls and anonymous call data have also been used for different purposes	Ygnace et al. (2001) Cáceres et al. (2012)	• PROS: Bigger dataset: localization accuracy better than with passive comm • CONS: Need for investments in the network to deal with greater communication loads
Other techniques		Those based on other built-in sensors like the accelerometer, or on the Bluetooth or WIFI signals	Hansapalangkul et al. (2007) Lv et al. (2015)	• PROS: Varied. For example energy efficiency • CONS: Varied. Above all, unreliability linked to their condition of isolated studies

et al. 2010; Tao et al. 2012; Ge and Fukuda 2016; Sanaullah et al. 2016; Woodard et al. 2017). Other studies have developed cellular signal-based procedures able, for example, to estimate travel times, to detect congestion (e.g., Hansapalangkul et al. 2007; Cáceres et al. 2012; Janecek et al. 2012) and even for incident monitoring. For this last purpose, Ygnace et al. (2001) examined the number of calls made per unit time on two test sites (respectively urban and rural) in the south of France. They found out that the volume of calls was related to the severity of the incidents. This

is an example of in-use (i.e., there exists voice or data transmission) mobile phones acting as traffic probes. However, other methodologies only require that phones are switched on. Indeed, when speaking about signal-based techniques, it is important to note that two types of communications are possible between the components of the mobile phone system, i.e., between handsets and base stations. By the way, the later usually cover hexagonally shaped areas or "cells", which is the origin of the term "cell phone" (Rose 2006). These communications are:

- The so-called "active" communication, which occurs when a phone is in use. Then, it is continuously associated with its closest base station. That is, when one cell phone moves outside the cell boundary associated with one base station, it is handed over to the next closest one.
- The "passive" communication. When a phone is not in use but switched on, it periodically reports its precise location to the network in the event of an incoming call or an emergency.

Trilateration using signal strength or transmission delay from multiple base stations, the monitoring of cell handover zones, etc., are used to localize phones in both approaches. In the context of traffic monitoring, passive communication is said to be more advantageous than active communication. The reason is that relatively minor hardware additions (cabling and computing) to the mobile phone network are needed. Only those that allow the traffic probe system to (i) receive the raw data stream already collected by the mobile phone system and (ii) process it. Active systems are able to provide more data. When travelling, cell handovers are more numerous than passive location reports. However, these systems require the mobile phones to be polled (i.e., their status must be actively sampled by a client program) to establish their positions, which implies the overloading of the network and adds costs to the operating system.

Overall, methodologies that use mobile phones as GPS-receivers are said to localize them more accurately. Network-based procedures highly depend on the location of the phone base stations, on the road network geometry, on the data processing procedure (e.g., map-matching algorithms), etc. For their part, the main drawbacks of GPS-based procedures are the power-hungry nature of GPS and the common instabilities in urban environments (Paek et al. 2010).

Two important aspects must be highlighted. First, most research performed thus far relied on dedicated smartphones, i.e., devices specifically used for research purposes. Therefore, their number and characteristics (and sometimes the routes "they" followed) were known a priori. Second, the described methodologies are not the only ones that have been tested, but the most common and, thus, those whose results are accepted as more reliable. For example, Lv et al. (2015) developed a completely different methodology to detect road congestion by means of smartphones. On the one hand, these were undedicated, which is a better approach to reality when trying to standardize their use as traffic probes. On the other hand, the system depended on two mobile phone sensors: the accelerometer and the cellular signal. First, an accelerometer-based vehicular movement detection module identified the periods when phone users traveled by vehicle. Second, the cellular signal was

used by a map-matching module to determine the traveled road segments. Finally, another module was able to qualitatively infer the degree of congestion by means of the vehicular dataset. Experimental results based on real-world data demonstrated both the goodness of the system and its energetic efficiency in contrast to other techniques. However, the fact that only qualitative results can be obtained seems insufficient to perform accurate traffic monitoring and, subsequently, to be the only basis of any dynamic traffic management strategy. Techniques that combine several of the mentioned approaches, i.e., GPS and cellular signal, are also on the table (Hansapalangkul et al. 2007).

1.2.2.3 RFID Technologies

Another example of lagrangian sensing is the use of RFID technologies. They allow gathering data about a particular object (e.g., a vehicle) by means of inductive coupling or electromagnetic waves. Three main parts must be present in a RFID-based system: (i) a transponder or tag, which consists of a microchip attached to an antenna, (ii) the reader or transceiver, which receives the information from the tag within a limited range, and (iii) the computing framework, which processes the information. The implemented software varies depending on the final purpose of the system. Nevertheless, anti-collision algorithms that allow reading several tags simultaneously, as well as encryption modules, should always be included. There are several types of tags depending for example on their power source or the characteristics of their memory. Receivers can also have different designs and controllers, and they can be stationary or mobile. As said, two main types of wireless systems are used for the communication between tags and readers (Ilie-Zudor et al. 2006):

- Induction: it only works when the distance between the tag and the reader is relatively short, so that electromagnetic or inductive coupling is possible. Low frequency (LF) and high frequency (HF) bands can be used.
- Propagation: for longer distances, via the propagation of electromagnetic waves. It operates in the Ultra High Frequency (UHF) and microwaves' frequency bands.

RFID technologies are being applied to very different areas like transportation, manufacturing, supply change management, health care, human identification, clothing, etc. Within them all, three main goals can be distinguished, which are (i) object detection and/or identification, (ii) object localization and (iii) data transfer from or to the RFID tag. Regarding transportation, uses and objectives go from the most simple (e.g., smart car key, public transport ticket, toll collection) to the most complex. This last group encompasses the use of RFID as AVI technologies (thus eulerian) but, above all, for vehicle tracking (i.e., lagrangian sensing). However, little research and, subsequently, few applications of RFID-based tracking aimed at monitoring traffic have been developed, either to feed travel time information systems or to design other management strategies, Vehicle control via RFID is mostly circumscribed to logistics, either alone or combined with GPS (Prasanna and Hemalatha 2012). In these applications, the interest resides in goods and not in

traffic management. One use of RFID that combines traffic management and logistics is that of emergency operations management. Effective emergency management requires integrating and analyzing data collected from sources such as emergency inventories, emergency vehicles, distribution centers, and shelters. In this context, RFID can provide a total visibility of vehicle and commodity movement in the disaster supply chain (Ozguven and Ozbay 2015). For example, Ozguven and Ozbay (2013) proposed a real-time online feedback control scheme that was able to analytically control emergency vehicle flows, and to compute the trajectories of emergency inventory levels. The aforementioned uses benefit from the three main advantages of RFID-based approaches, i.e., their suitability for (i) tracking, (ii) identification, sensing, and authentication and (iii) automatic data collection and transfer. However, some challenges must still be overcome to make RFID more competitive for traffic monitoring against other technologies:

- Standardization: it is the most important issue. There are very dissimilar technologies, communication protocols, signal modulation types, data transmission rates, data encoding and frames, collision handling algorithms, etc. This heterogeneity impedes interoperability even within a particular country. There is a need for an agreement among public agencies and private companies in this regard.
- Communication weaknesses: if the communication infrastructure is completely down or if there is not enough radio frequency signal strength, RFID systems are not able to transmit data between the reader and the tag, i.e., they are not able to feed data into the monitoring system. When working in real time, this fact could have severe consequences. If used as a standalone system, both accuracy and reliability should be enhanced by means of a reasonable level of redundancy regarding tags, readers, antennas, and operating modes (Vaidya and Das 2008; Bolic et al. 2010).
- Expenses: actually not a major problem. The implementation of a RFID-based tracking system implies extra costs when compared, for example, to a smartphone-based system. Nevertheless, that also occurs with other sensing equipment and these costs are reasonable and justifiable by the system's advantages.

1.2.2.4 Probe Vehicles and Dynamic Floating Vehicles

Ad hoc equipped probe vehicles and dynamic floating vehicles also act as lagrangian sensors. Before delving into their contribution, it must be remarked that the adjective "probe" refers here to vehicles that are introduced in the traffic stream only with research purposes. On the contrary, "floating" alludes to vehicles that are used for other objectives, but from which valuable data can additionally be obtained. There is confusion between these terms in the literature, with some authors agreeing with this explanation (e.g., Young 2007), others considering both terms as synonyms (e.g., Sunderrajan et al. 2016) and others, especially in the past years, using them just in the opposite way (e.g., Turner et al. 1998).

Taking into account the aforementioned standpoint, ad hoc equipped probe vehicles are mostly used in specific research projects (e.g., current tests of intelligent

vehicles) and/or as a complement to other sensing systems (Nanthawichit et al. 2003; Bachmann 2011; Treiber et al. 2011, etc.). In most cases, particular and powerful sensors not usually available are implemented in traditional vehicles to this end. Therefore, they are able to provide very specific and valuable data. However, their equipment involves extra and, often, significant costs. Additionally, they are in any case a very small sample of the traffic stream, which implies that the information they provide is not fully representative (Rose 2006; Oberauer et al. 2011, etc.). Besides, these data are biased to some extent, as drivers are aware of taking part of a study and their behavior is not always completely natural. This aspect has been assessed on other occasions, for example, when drivers answer driving-related surveys (Lajunen and Summala 2003). For their part, the use of existing floating vehicles does not generally imply large expenses. Monitoring usually relies on already addressed technologies, and the equipment to install (if not already on-board) generally consists of a GPS receiver, a dedicated smartphone or a RFID chip, all of them affordable nowadays. Thus, these technologies are not found in the traffic stream casually, but their presence is known. The involved vehicles usually belong to a particular collective (e.g., taxis, buses, trucks of any company). Therefore, the amount of gathered data is much greater than in the case of ad hoc probe vehicles. Additionally, drivers do not usually behave conditioned, because their driving task has a goal beyond research. However, the fact that vehicles often belong to particular categories must be taken into account, as it implies some bias (Wang et al. 2010; Yuan et al. 2011). Depending on the objective, data could be insufficient to represent the whole traffic stream. In this context, the progressive introduction of more and more intelligent vehicles will have a decisive effect on traffic monitoring. Autonomous (automated) vehicles are expected to have the advantages of both ad hoc probe vehicles and dynamic floating vehicles. That is, they will have powerful on-board sensors and move unconditioned within the traffic stream. Additionally, autonomous vehicles are called to reach the whole vehicle spectrum (private cars, trucks, motorbikes, public services, etc.). However, their high cost as well as the implementation of new mobility paradigms will limit their penetration rate. Chapter 5 addresses all these topics in detail.

1.2.2.5 Unmanned Aerial Vehicles

Finally, the use of unmanned aerial vehicles (UAVs) for traffic monitoring seems promising. The combination of UAV flights over road segments with video image processing techniques has already been used to determine particular traffic flow parameters, complete trajectories, flow patterns, drivers' behavior, etc. (Azevedo et al. 2014; Barmpounakis et al. 2016; Salvo et al. 2017; Kaufmann et al. 2018, etc.). Increasingly used as part of traffic monitoring schemes (sometimes as eulerian sensors, although part of their potential is lost), they probably will not be able to form a standalone system. Currently, some external factors such as an unfavorable climate (e.g., wind, rain), the presence of electromagnetic fields or obstructions (e.g., buildings, urban canyons), etc., physically limit their flights. Additionally, the modest

autonomy of current batteries, their limited payload, and legal issues (e.g., zones where they are not allowed to fly) also restrict their usefulness.

1.3 Traffic Reconstruction

As said, data are essential for traffic reconstruction. The goal of this reconstruction can be very different: to analyze current traffic conditions in order to make some improvements (e.g., avoid recurrent congestion), to predict future traffic performance (e.g., after implementing some management strategy or simply according to changes in traffic streams as time goes by), to plan a new infrastructure, etc. In any case, data assimilation is a key step. Leaving aside trends based on the use of artificial intelligence or data-driven approaches, which are addressed in Chap. 6, data assimilation can be described as the combination of a traffic model with real data to estimate traffic states, and it can be performed in many different ways. Ultimately, the procedure chosen should depend on (i) the particular objective of the reconstruction and (ii) the available parameters. Although an exhaustive description of all possible approaches is out of the scope of this chapter (see Chaps. 6 and 7 for more information), some fundamentals are summarized below.

First of all, many classifications of traffic models are possible. For example, a clear dichotomy exists between the models that make use of traffic flow physics and those that do not, and base their estimations on statistics and current and/or historical data. However, broadly speaking, the distinction between macroscopic and microscopic analyses is the clearest. In both cases and depending on the complexity of the study and on the available resources, the use of analytical or numerical methods with the support of simulation software is common. Heuristics is also applied to particular tasks (e.g., routing, traffic assignment). Table 1.3. outlines some key features and examples of macro and microscopic analyses, some of them integrated into commercial software.

Table 1.3 More distinctive features and examples of macroscopic and microscopic traffic models

	Key variables			Models (examples)	Simulation (examples)
Macro	q	K	$\overline{v}$	Continuous theory (LWR or KWT)	Cell Transmission model, Visum, Aimsun
Micro	h	S	v_i	Car-following models	Vissim, Aimsun
Relation macro–micro	$q = \frac{1}{\overline{h}}$	$k = \frac{1}{\overline{s}}$	$\overline{v} = \frac{\sum_1^n v_i}{n}$	Greenberg model	3rd generation General Motors Car-Following model
Aggregation region	(x,T)	(L, t)	$\overline{v_t}$ (x,T) $\overline{v_s}$ (L,t)		

Macroscopic models are used to obtain global characteristics of the traffic stream, i.e., to describe it by estimating average values of its most defining parameters like the flow q (veh/h), the density k (veh/km) or the average speed $\overline{v}$. They are based on the *Continuous Theory of Traffic Flow*, also known as *Kinematic Wave Theory* (KWT) or the Lighthill, Whitham (1955) and Richards (1956) Theory (LWR). These researchers independently proposed a first-order partial differential equation based on traffic flow physics to describe traffic evolution over time and space (Eq. 1.12). In fact, continuous theories were first developed for fluids. Their application to traffic involves, thus, focusing on the behavior of a stream rather than on individual cars.

$$\frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = 0 \tag{1.12}$$

This equation, which relates changes in flow over space to changes in density over time, is actually a conservation equation: in a closed system (without inflows or outflows) vehicles neither disappear nor are created. It is considered the most important equation of traffic flow theory after the fundamental equation (Eq. 1.8, which is deeply analyzed in Chap. 3). For example, the application of the conservation equation plays a key role in the methodology presented in Chap. 4. In addition to the assumption of conservation (i.e., closed sections), the LWR model supposes homogeneous sections (i.e., no changes in geometry) and the existence of an equation of state. This last assumption means that we must be able to define the traffic state by means of a single input (Eq. 1.13).

$$q = Q(k) \tag{1.13}$$

In other words, a fundamental diagram, i.e., a diagram that relates average values of flow and density for a particular section, is required.

Besides, in order to simplify the application of the theory, instantaneous changes of vehicle speeds are accepted. That is, accelerations and decelerations are neglected. Despite being one of the most important references for traffic engineering, the former assumptions and other features imply limitations for the initial LWR model. For example, as it focuses on traffic streams, vehicle heterogeneity is overlooked. This fact can be especially problematic in free flow, when the differences between vehicles (e.g., light as opposed to heavy vehicles) are more noticeable. With light traffic, vehicles drive with very different speeds, have dissimilar lane-changing behavior, etc. If these differences are meaningful, the LWR model is not accurate. It also results unadvisable for derived studies like the estimation of traffic emissions, highly linked to speeds. Varied extensions of the LWR model have tried to overcome its limitations. For example, Daganzo (1999a, b) developed his theory of "slugs and rabbits", which allows distinguishing between different types of vehicles within a traffic a stream. Second order models accounting for vehicle acceleration and deceleration and, thus, able to represent traffic instabilities (i.e., *stop and go*), were also derived (e.g., Zhang 1998; Aw and Rascle 2000). The most used ones can be numerically discretized by means of schemes like that designed by Godunov

(1959). In case a triangular fundamental diagram is used, Daganzo (1994, 1995) proposed another discretization of the first order LWR model. Particularly, his finite differences model divides highways into cells of (related) dimensions $\Delta x * \Delta t$ (where x stands for length and t for time) to compute the state of the system. Known as the Cell Transmission Model (CTM), its usefulness and simplicity have encouraged its use all over the world and out of the scope of road traffic (e.g., Wei et al. 2013). Moreover, several extensions of the CTM that better fit particular phenomena have also been developed. Many other macroscopic models have been designed. Overall, all of them are suitable, for example, to analyze queuing. In this regard, the work of Newell (1993a, b, c) is worth mentioning. He derived a simplification of the LWR theory that works with (N, t) coordinates (i.e., cumulative number of vehicles vs time) instead of (x, t) coordinates, especially to this end.

For their part, microscopic models consider each particular vehicle in the traffic stream or, more precisely, each unit vehicle-driver. They need the trajectory of every single vehicle, which is thus defined by its own equation. Interactions between near vehicles are also taken into account. In this context, averaged parameters are not useful and individual values of headways h (s^{-1}) or spacings s (m^{-1}) are required. They are the distance between the same end of two consecutive vehicles, respectively in time or space. Generally speaking, a microscopic model is composed of a car-following theory and a lane-changing submodel. The car-following theory is aimed at predicting how vehicles follow another one that has been chosen as the "leader" or the reference. These theories usually provide a relation $s(v)$, i.e., they define the spacing as a function of the speed. In fact, spacings are usually accepted as the most important variables in microscopic approaches, as they have clear implications for both traffic efficiency and safety. Thinking of them individually, small spacings would imply a better use of the available capacity, whereas large spacings would diminish the accident risk. Therefore, a trade-off must be reached. Many car-following models have been designed over time and all of them have their strengths and weaknesses. Well-known examples are those theories of Pipes (1953, 1967), Forbes et al. (1958), Forbes (1963), General Motors (Chandler et al. 1958; Herman et al. 1959; Herman and Potts 1959; Gazis et al. 1959, 1961) and the *Optimal Velocity Theory* (Bando et al. 1995). In a few words, Pipes' (1953) model involves the idea that the minimum safe spacing increases linearly with speed. Forbes (1958) includes the concept of *reaction time*. In his model, the minimum time headway is equal to the reaction time (minimum time gap, i.e., minimum required distance in time between the rear of the leader and the front of the follower) and the time the lead vehicle requires to traverse a distance equivalent to its length. Regarding General Motors, actually a series of models was developed at the research laboratories of the company. They all have the same basic form (Eq. 1.14):

$$response = function\,(stimuli, sensitivity) \tag{1.14}$$

where the response is the acceleration or deceleration of the following vehicle and the stimuli are the relative speeds between the lead and the following vehicles. The

varied modifications of the sensitivity term led to a generalized form of car-following model, whose importance is indisputable. On the one hand, simulation models based on it often fit field data. On the other hand, it was possible to establish a mathematical relationship between the General Motors' model and the macroscopic Greenberg's logarithmic diagram (Greenberg 1959) for speed-density (Chakroborty and Kikuchi 1999). This last point allows analyzing traffic jointly from a multiple point of view. Finally, the idea under the Optimal Velocity Theory is that each driver tries to achieve an optimal speed depending on the distance to the preceding vehicle and the speed difference between both vehicles. Many extensions of the former car-following models as well as others that somehow differ from them are also available. Not that all of them were thought for this unit vehicle-driver. In very simple words, they tried to imitate drivers' behavior. Therefore, these models apply to current vehicle automation levels. However, when vehicles become autonomous (or highly automated), new models should be developed in which the role of the driver should increasingly negligible. An example of disruptive research in this regard in that in Troullinos et al. (2021). With regard to lane-changing models, a great variety has also been developed (e.g., rule-based models, discrete-choice-based models, artificial intelligence models, incentive-based models, etc. See for example Rahman et al. (2013) for further information).

Both approaches, macroscopic and microscopic, have advantages and disadvantages. As said, both the outputs (and accuracy) sought and the available data should support the choice between them. It is important to remark that, in line with the advances of programming and simulation software, there is an increasing tendency to indiscriminately use microscopic models. When the necessary data are accessible, these models allow performing more detailed analyses, as both the behavior of each involved agent (vehicles, pedestrians, etc.) and their interactions are examined. However, this potential power makes them much more complex and the amount of required parameters is much larger too. The accurate calibration of the selected simulation software is also essential. If only a partial amount of the necessary data is available and the rest is substituted by raw estimations or default values, results can be completely flawed. In this context, the use of macroscopic models is much more advisable. They are simpler and can even be solved by hand. And, above all, the fact that they rely on fewer parameters makes them usually more robust.

Trying to find a middle ground, mesoscopic models are being increasingly used. They are intermediate procedures that properly combine parts of macroscopic and microscopic analyses. Many possibilities exist. For example, each lane of a highway could be macroscopically analyzed, whereas a microscopic model would study the relationships between the traffic streams on each lane. Varied simulation software that combines both approaches can be found in the market.

In any case, the chosen model must be combined with real data to estimate or predict traffic states. As said, this process of data assimilation can also be performed by means of different techniques. Most of them are based on a Bayesian statistics analysis that treats the forecast from the model as the prior distribution, and then calculates a posterior distribution based on the available observations. Kalman filters

are often used for this later calculation (Xia et al. 2017). In fact, Kalman filtering techniques have been shown useful for data assimilation both with eulerian (e.g., Gazis and Knapp 1971; Szeto and Gazis 1972; Sun et al. 2004) and lagrangian measurements (e.g., Chu et al. 2005; Nanthawichit et al. 2003; Herrera and Bayen 2010). These techniques, also known as linear quadratic estimations, have different versions. Generally speaking, a Kalman Filter is a recursive (i.e., new measurements are processed as they arrive) algorithm that infers parameters of interest from indirect, inaccurate, and uncertain observations. More in particular, it uses measurements observed over time, containing statistical noise and other inaccuracies, and produces optimal approximations of other unknown variables by estimating a joint probability distribution over them for each updating interval.

References

Andersen J, Sutcliffe S (2000) Intelligent Transport Systems (ITS)—an overview. Proceedings of the International Federation of Automatic Control (IFAC), Technology Transfer in Developing Countries, Pretoria, South Africa 33(18):99–106

Aw A, Rascle M (2000) Resurrection of "second order" models of traffic flow. SIAM J Appl Math 60(3):916–938

Azevedo CL, Cardoso JL, Ben-Akiva M, Costeira JP, Marques M (2014) Automatic vehicle trajectory extraction by aerial remote sensing. Proc Soc Behav Sci 111:849–858

Bachmann C (2011) Multi-sensor data fusion for traffic speed and travel time estimation. PhD dissertation. University of Toronto

Bando M, Hasebe K, Nakyaama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. Phys Rev E 51:1035–1042

Banham R (2002) The ford century: ford motor company and the innovations that shaped the world. New York: Artisan Publishers. ISBN-13: 9781579652012

Barceló J, Kuwahara M (2010) Traffic data collection and its standardization. Springer, Berlin

Barceló J, Montero L, Marquès L (2010) Travel time forecasting and dynamic OD estimation in freeways based on Bluetooth traffic monitoring. Proceedings of the 89th Annual Meeting of the Transportation Research Board, 10–14 January 2010, Washington DC

Barceló J, Montero L, Bullejos M, Serch O, Carmona C (2013) A kalman filter approach for exploiting bluetooth traffic data when estimating time-dependent OD matrices. J Intell Transp Syst 17(2):123–141

Barmpounakis EN, Vlahogianni EI, Golias JC (2016) Extracting kinematic characteristics from unmanned aerial vehicles. Proceedings of the 95th Annual Meeting of the Transportation Research Board, January 2016. Washington, USA

Berenguer A, Goncalves J, Hosio S, Ferreira D, Anagnostopoulos T, Kostakos V (2017) Are smartphones Ubiquitous?: an in-depth survey of smartphone adoption by seniors. IEEE Cons Elect Magaz 6(1):104–110

Bolic M, Simplot-Ryl D, Stojmenović I (2010) RFID systems: research trends and challenges. Wiley, New Jersey

Cáceres N, Romero LM, Benitez FG, del Castillo JM (2012) Traffic flow estimation models using cellular phone data. IEEE Trans Intell Transp Syst 13(3):1430–1441

Chakroborty P, Kikuchi S (1999) Evaluation of the general motors based car-following models and a proposed fuzzy inference model. Trans Res Part c: Emerg Technol 7(4):209–235

Chandler RE, Herman R, Montrol EW (1958) Traffic dynamics: studies in car-following. Oper Res 6(2):165–184

Chu L, Oh S, Recker W (2005) Adaptive Kalman filter based freeway travel time estimation. Transportation Research Board 84th Annual Meeting, 2005, Washington, DC

Coifman B (2001) Improved velocity estimation using single loop detectors. Transp Res Part a: Policy Pract 35(10):863–880

Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. ASCE J Transp Eng 129(5):475–483

Daganzo CF (1994) The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. Transp Res Part b: Methodol 28(4):269–287

Daganzo CF (1995) The cell transmission model, part II: network traffic. Transp Res Part b: Methodol 29(2):79–93

Daganzo CF (1999a). A Behavioral Theory of Multi-Lane Traffic Flow Part I: Long Homogeneous Freeway Sections. Report. California: ITS Berkeley Institute for Transportation Studies, UC Berkeley

Daganzo CF (1999b) A behavioral theory of multi-lane traffic flow part II: merges and the Onset of Congestion. Report. California: ITS Berkeley Institute for Transportation Studies, UC Berkeley

Doberstein D (2011) Fundamentals of GPS receivers: a hardware approach. Springer, New York

Edie LC (1965) Discussion of traffic stream measurements and definitions. Proc. 2nd International Symposium on the Theory of Traffic Flow, OECD, Paris, pp 139–154

Federal Highway Administration (2006) Traffic detector handbook, vol. 1. Report FHWA-HRT-06–108. Virginia: US Department of Transportation

Forbes TW, Zagorski HJ, Holshouser EL, Deterline WA (1958) Measurement of driver reactions to tunnel conditions. Highway Res Board Proce 37:60–66

Forbes TW (1963) Human factor considerations in traffic flow theory. Highway Res Record 15:60–66

Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady state flow. Operat Res 7(4):499–505

Gazis DC, Herman R, Potts RB (1959) Car-following theory of steady state flow. Oper Res 7(4):499–505

Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. Oper Res 9(4):545–567

Gazis D, Knapp C (1971) On-line estimation of traffic densities from time-series of flow and speed data. Transp Sci 5(3):282–301

Ge Q, Fukuda D (2016) Updating origin-destination matrices with aggregated data of GPS traces. Transp Res Part c: Emerg Technol 69:291–312

Godunov S (1959) A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. Matematicheskii Sbornik 47(3):271–306

Greenberg H (1959) An analysis of traffic flow. Oper Res 7(1):79–83

Hansapalangkul T, Keeratiwintakorn P, Pattaraatikom W (2007) Detection and estimation of road congestion using cellular phones. Proceedings of the 7th International Conference on ITS Telecommunications, 2007, Sophia Antipolis, pp 1–4

Hellinga B (2002) Improving freeway speed estimates from single loop detectors. J Transp Eng 128(1):58–67

Herman R, Potts RB (1959) Single lane traffic flow theory and experiment. Proceedings of the Symposium on the Theory of Traffic Flow, 1959, Research Labs, General Motors, pp 147–157. New York: Elsevier

Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car-following. Oper Res 7(1):86–106

Herrera JC, Bayen AM (2010) Incorporation of Lagrangian measurements in freeway traffic state estimation. Transp Res Part b: Methodol 44:460–481

Herrera JC, Work D, Ban X, Herring R, Jacobson Q, Bayen A (2010) Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp Res Part c: Emerg Technol 18:568–583

Highway Research Board (1950) Highway capacity manual: practical applications of research. Washington DC: National Research Council

Ilie-Zudor E, Kemény Zs, Egri P, Monostori L (2006) The RFID technology and its current applications. *Proceedings of the 8th International Conference on The Modern Information Technology in the Innovation Processes of the Industrial Enterprises*, 2006, Laboratory of Engineering and Management Intelligence, Computer and Automation Research Institute (SZTAKI), Budapest, Hungary, pp 29–36

Janecek A, Valerio D, Hummel KA, Ricciato F, Hlavacs H (2012) Cellular data meet vehicular traffic theory: Location area updates and cell transitions for travel time estimation. Proceedings of the 2012 ACM Conference on Ubiquitous Computing, 5–8 September 2012, Pittsburgh, Pennsylvania, pp 361–370

Kaufmann S, Kerner BS, Rehborn H, Koller M, Klenov SL (2018) Aerial observations of moving synchronized flow patterns in over-saturated city traffic. Transp Res Part c: Emerg Technol 86:393–406

Lajunen T, Summala H (2003) Can we trust self-reports of driving? Effects of impression management on driver behaviour questionnaire responses. Transport Res f: Traffic Psychol Behav 6(2):97–107

Lamb H (1895) Hydrodynamics. Cambridge University Press, UK

Lighthill M, Whitham G (1955) On kinematic waves II. A theory of traffic flow on long crowded roads. Proc Royal Soc A 229(1178):317–345

Lin Y, Wang P, Ma M (2017) Intelligent Transportation System (ITS): Concept, Challenge and Opportunity. 2017 IEEE 3rd International Conference on Big Data Security on Cloud, IEEE International Conference on High Performance and Smart Computing and IEEE International Conference on Intelligent Data and Security, 2017, Beijing, pp 167–172

Lv M, Chen L, Wu X, Chen G (2015) A road congestion detection system using undedicated mobile phones. IEEE Trans Intell Transp Syst 16(6):3060–3072

Nanthawichit C, Nakatsuji T, Suzuki H (2003) Application of probe-vehicle data for real-time traffic-state estimation and short-term travel-time prediction on a freeway. Transp Res Rec 1855:49–59

Newell GF (1993a) A simplified theory of kinematic waves in highway traffic. Part I: General Theory. Transp Res Part B: Methodol 27(4):281–287

Newell GF (1993b) A simplified theory of kinematic waves in highway traffic. Part II: Queuing at freeway bottlenecks. Transp Res Part B: Methodol 27(4):289–303

Newell GF (1993c) A simplified theory of kinematic waves in highway traffic. Part III: Multi-destination flows. Transp Res Part B: Methodol 27(4):305–313

Oberauer C, Stottan T, Wagner R (2011) Requirements of processing extended floating car data in a large scale environment. In Advanced Microsystems for Automotive Applications, pp 335–342. Berlin: Springer

Ozguven EE, Ozbay K (2013) A secure and efficient inventory management system for disasters. Transp Res Part c: Emerg Technol 29:171–196

Ozguven EE, Ozbay K (2015) An RFID-based inventory management framework for emergency relief operations. Transp Res Part c: Emerg Technol 57:166–187

Paek J, Kim J, Govindan R (2010) Energy-efficient rate-adaptive GPS-based positioning for smartphones. Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, 15–18 June 2010; San Francisco, CA, pp 299–314

Palen J (1997) The need for surveillance in intelligent transportation systems. Intellimotion 6(1):1–3. University of California PATH, Berkeley, CA

Pipes LA (1953) An operational analysis of traffic dynamics. J Appl Phys 24:274–281

Pipes LA (1967) Car-following models and the fundamental diagram of road traffic. Transp Res b: Methodol 1:21–29

Prasanna R, Hemalatha M (2012) RFID GPS and GSM based logistics vehicle load balancing and tracking mechanism. Proceedings of the International Conference on Communication Technology and System Design, 8–10 December 2012, Beijing, China, pp 726–729

Rahman M, Chowdhury M, Xie Y, He Y (2013) Review of microscopic lane-changing models and future research opportunities. IEEE Trans Intell Transp Syst 14(4):1942–1956
Richards P (1956) Shock waves on the highway. Oper Res 4(1):42–51
Rose G (2006) Mobile phones as traffic probes: practices, prospects and issues. Trans Rev 26(3):275–291
Salvo G, Caruso L, Scordo A, Guido G, Vitale A (2017) Traffic data acquirement by unmanned aerial vehicle. Euro J Remote Sens 50(1):343–351
Sanaullah I, Quddus M, Enoch M (2016) Developing travel time estimation methods using sparse GPS data. J Intell Transp Syst 20(6):532–544
Shladover SE (1990) Advanced Vehicle Control Systems (AVCS) (1990). Vehicle Electronics in the 90s: Proceedings of the International Congress on Transportation Electronics, 1990, Warrendale, US, pp 103–112
Soriguera F (2016) Highway travel time estimation with data fusion. Springer Tracts on Transportation and Traffic, 11, 212 pages. Berlin: Springer-Verlag Berlin Heidelberg
Space Segment (2018) Official US government information about the Global Positioning System (GPS) and related topics. https://www.gps.gov/systems/gps/space/. Accessed 4th April 2021
Sun X, Muñoz L, Horowitz R (2004) Mixture Kalman filter based highway congestion mode and vehicle density estimator and its application. Proceedings of the 2004 American Control Conference, 2004, Boston, MA, 2098–2103
Sunderrajan A, Viswanathan V, Cai W, Knoll A (2016) Traffic state estimation using floating car data. Proc Comp Sci 80:2008–2018
Sussman J (2005) Perspectives on intelligent transportation systems (ITS). Springer Science and Business Media, Berlin
Szeto M, Gazis D (1972) Application of Kalman filtering to the surveillance and control of traffic systems. Transp Sci 6(4):419–439
Tao S, Manolopoulos V, Rodriguez S, Rusu A (2012) Real-time urban traffic state estimation with A-GPS mobile phones as probes. J Trans Technol 2(1):22–31
Traffic Detector Handbook (1991) Washington DC: Institute of Transportation Engineers
Treiber M, Helbing D (2002) Reconstructing the spatio-temporal traffic dynamics from stationary detector data. Cooper@tive Tr@nsport@tion Dyn@mics, 1, 3.1–3.24
Treiber M, Kesting A, Wilson RE (2011) Reconstructing the traffic state by fusion of heterogeneous data. Comput Aided Civ Infrastructure Eng 26(6):408–419
Troullinos D, Papamichail I, Chalkiadakis G, Papageorgiou M (2021) Collaborative Multiagent Decision Making for Lane- Free Autonomous Driving. In Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021, IFAAMAS, p 9
Turner SM, Eisele WL, Benz RJ, Holdener DJ (1998) Travel time data collection handbook. Research Report FHWA-PL-98–035. Washington, DC: Federal Highway Administration, Office of Highway Information Management
Vaidya NH, Das SR (2008) RFID-based networks: exploiting diversity and redundancy. Mobile Comput Commun Rev 12:2–14
Vaxtor (2021) VaxALPR PC Brochure. Av. in https://www.vaxtor.com/products/vaxalpr/on-pc/. Accessed in June 2021
Wang T, Fang T, Han J, Wu J (2010) Traffic monitoring using floating car data in Hefei. Proceedings of the International Symposium in Intelligence Information Processing and Trusted Computing (IPTC), 28–29 October 2010, Huanggang, China, 122–124
Wei P, Cao Y, Sun D (2013) Total unimodularity and decomposition method for large-scale air traffic cell transmission model. Transp Res Part b: Methodol 53:1–16
Weiland RJ, Purser LB (2000) Intelligent Transportation Systems. In: Transportation in the New Millennium: State of the Art and Future Directions, Perspectives from Transportation Research Board Standing Committees. Washington, DC: Transportation Research Board
Woodard D, Nogin G, Koch P, Racz D, Goldszmidt M, Horvitz E (2017) Predicting travel time reliability using mobile phone GPS data. Transp Res Part C Emerg Technol 75:30–44

Xia C, Cochrane C, DeGuire J, Fan G, Holmes E, McGuirl M, Murphy P, Palmer J, Carter P, Slivinsiki L, Sandstede B (2017) Assimilating eulerian and lagrangian data in traffic-flow models. Physica D 346:59–72

Ygnace JL, Benguigui C, Delannoy V (2001) Travel Time/Speed Estimates on the French Rhone Corridor Network Using Cellular Phones as Probes. Lyon: INRETS. Final Report of the SERTI V Program

Yoon J, Noble B, Liu M (2007) Surface street traffic estimation. Proceedings of the 5th International Conference on Mobile Systems, Applications, and Services (MobiSys 2007), June 11–13, 2007, San Juan, Puerto Rico, 22–31

Young S (2007) Real-time traffic operations data using vehicle probe technology. Proceedings of the 2007 Mid-Continent Transportation Research Symposium, August 2007, Ames, Iowa

Yuan J, Zheng Y, Xie X, Sun G (2011) Driving with knowledge from the physical world. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 21–24 August 2011, San Diego, CA, USA, 316–324

Zhang HM (1998) A theory of non-equilibrium traffic flow. Trans Res Part b: Methodol 32(7):485–498

Chapter 2
Travel Time Information Revisited

Margarita Martínez-Díaz

Abstract As with speed and many other traffic engineering parameters, there is no single definition of travel time: there are individual travel times, average travel times referenced to the moment when vehicles start their journey, or to the moment they arrive to destination, past travel times, predicted travel times, etc. Knowing the differences among these definitions is essential, for example, to determine which one to estimate and disseminate in each case according to the objective sought. Also, because in order to calculate each of them, it is necessary to have certain data (i.e. certain equipment) and to apply particular estimation methodologies. This chapter explains the most important definitions behind the general concept of travel time and the most common procedures to obtain each of them. As a preview: a good travel time information system is able to provide accurate real-time predicted travel times. The likelihood of such estimates being robust increases if data fusion methodologies are applied.

2.1 Travel Time Information and Traffic Management

Traffic management is indispensable to ensure a sustainable and efficient mobility. ATIS play a key role in this regard, as they assist users in making pre-trip and en-route decisions (Mori et al. 2015). Although the information they provide is varied (e.g. warnings, advices), the dissemination of travel times stands out as particularly beneficial, due the intuitiveness and clarity behind the concept: the travel time is the time required to traverse a route between any two points of interest, taking into account the stops, queuing delays and intersection delays (Zhu et al. 2009). For this purpose, it is especially important that the information delivered is accurate and estimated in real time, on the basis of current traffic states. The knowledge of travel

M. Martínez-Díaz (✉)
Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Catalonia, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_2

times in real time is not only important for drivers, but also for management centres that, based on them, can take dynamic decisions aimed at improving traffic efficiency.

Even off-line travel time information is profitable. In fact, engineers have used travel time and delay studies since the late 1920s to evaluate transportation facilities and to plan improvements (Turner et al. 1998). Today, the concept of travel time reliability has taken over that of the level of service (LOS) as the most indicative parameter to define the quality of a given road or network in terms of efficiency. Travel time reliability can be defined as "the absence of variability in the travel time between a determined origin and destination, independently of the rest of the conditions" (Elefteriadou and Cui 2007). Depending on the researcher, on the administration, etc., slightly variations of this general definition can be found in the literature (e.g. "the consistency of the travel time for a given route", "the consistency or dependences of travel times, measured day by day and/or throughout different times of the day ", etc.). It is important to note that (i) travel time reliability has no direct relationship with travel time variability and (ii) travel time reliability depends on the driver's expected travel time, that is, on the driver's information.

In practice, travel time reliability is described by an index (i.e. a number) that represents to what extent travel times depend on factors other than the path length (Fig. 2.1). This dependence must be measured over different periods of the same day, in different days, months or even years. In fact, engineering, safety-related and economic studies have shown that the better route between two points is not the one that allows the shorter travel time under "normal" conditions, but the one in which the travel time remains within a stable and acceptable range, regardless of the boundary conditions.

This knowledge allows users to optimize their trips by rescheduling them, by selecting the best route or even a different transportation mode, etc. Thus, it helps to avoid situations of stress during the driving, which are unequivocally linked to higher accident rates. This would otherwise not be possible and, additionally, would

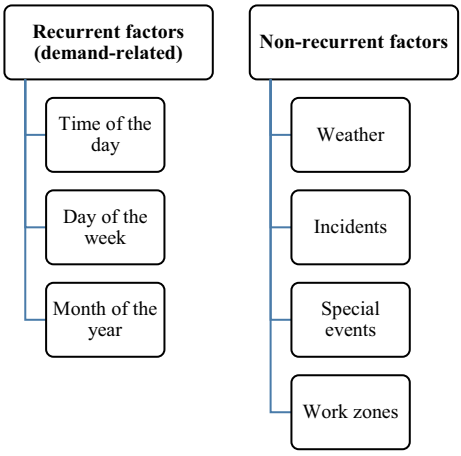

Fig. 2.1 External factors that affect travel times between two points of a determined route

lead drivers to have a bad opinion of the system. Even if it generally performs well, by nature, they would mostly remember the longer travel times (Fig. 2.2).

Although other methods exist (e.g. the use of histograms), the construction of probability density functions or cumulative density functions (Fig. 2.3) of travel times over long periods, from which particular indexes can be derived, is the most common and suitable way to assess reliability. Studies carried out so far indicate that travel time probability density functions obtained from empirical data usually fit the lognormal, bimodal, Burr, normal, gamma or Weibull distributions. In any case, once constructed, two different types of indicators can be extracted from these functions (Ryus et al. 2013).

- The variability of travel times along a particular route in the form of percentiles (P).
- More direct measurements of reliability, such as the number of trips that meet or fail to meet a certain operating standard (e.g. a minimum expected speed).

Figure 2.4 and Table 2.1. include some of these measures. Some other are also used in practice depending on the administration. In any case, the use of one or another should be linked to the purpose of the reliability analysis. For example, the PTI provides the most interesting information (delays with high demand) for a person who commutes to work at rush hour. However, the BTI is the best option for a user who is going to make an occasional leisure trip and can choose the departure time. According to its value, she/he will know the extra time she/he needs to reach her/his destination at the desired moment.

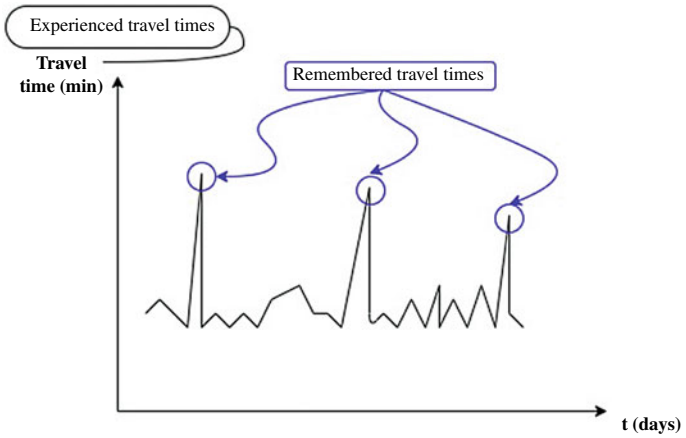

Fig. 2.2 Traveller travel time feeling

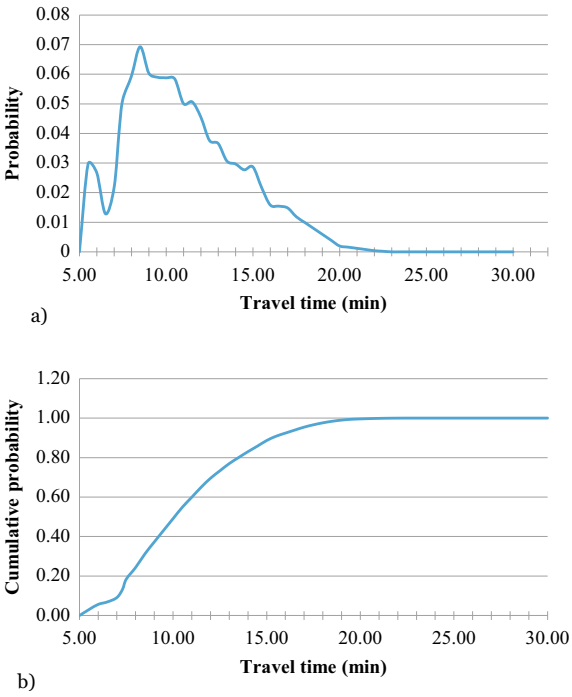

Fig. 2.3 Probability functions of the travel time between two points on a particular route: **a** Probability density function; **b** Cumulative density function

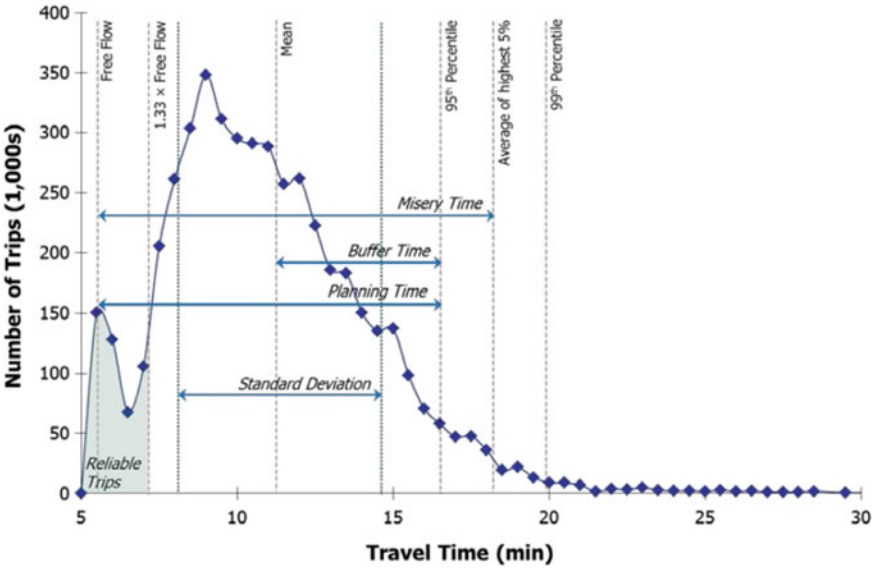

Fig. 2.4 Example of travel time probability density function with some measures used to quantify reliability (from Ryus et al. 2013. Exhibit 36–5, p. 36–18. Copyright, National Academy of Sciences. Reproduced with permission of the Transportation Research Board)

Table 2.1 Most used travel time reliability measures

Measure	Unit	Formulation	Description
Percentiles P50, P80, P95	Minutes	–	Travel time that is only exceeded respectively by 50, 20 or 5% of trips
Standard deviation (SD)	Minutes	$\sqrt{\frac{1}{N} * \sum_{i=1}^{N} (t_i - \bar{t})^2}$	Variation in travel time compared to the average. A 5-min SD indicates that 5 more minutes than in average conditions will probably be necessary to reach a destination
Planning time index (PTI)	–	$\frac{P_{95}t}{t_f}$	Extra time required to arrive "on time" 95% of the trips. Calculable for facilities, road segments or itineraries
Buffer time index (BTI)	–	$\frac{P_{95}t - \bar{t}}{\bar{t}^*}$ (* mean or median)	Extra time required to arrive "on time" 95% of the trips compared to the average. A BT of 1.5 indicates that 95% of the trips will require 50% more time to reach the destination than in average conditions
Misery index (MI)	–	$\frac{\text{Average of the 5\% longest } t_i}{t_f}$	How longer it takes to reach the destination in the 5% of trips made under the worse conditions. A MI of 4 indicates that the worst trips last 4 times longer than they would without congestion

Legend: t_i: value i of travel time; t_f: free flow travel time

2.1.1 Value of Travel Time Information

Travel time reliability has a direct relationship with the value of travel time information. Even if drivers cannot modify their trips and are trapped in a queue, if they were well informed, that is, if they knew in advance that they were going to be delayed, their stress is reduced or can even be non-existent. The value of the information will be higher if other costs apart from stress can also be reduced, for example, if they can use it to change their routes or their departure times in order to avoid congestion and arrive on time to their destinations.

Therefore, the value of travel time information can be split in two different components: the objective and the subjective component. The objective part of this value is directly linked to economy. To give an illustrative example, the delay suffered by a truck that transports material to a factory is more detrimental in economic terms

than that of a family that travels by car to spend a day on the beach. From this point of view, an accurate travel time information would be more valuable for the truck. However, the subjective valuation of travel time information could not agree with this. To continue with the former example, perhaps the members of the family have few opportunities to make such a trip, and would even pay some money to guarantee their arriving soon to the beach. That is, travel time information would be very valuable for them too. The subjective concept of "value of travel time savings" (VTTS) can be therefore defined as the willingness-to-pay to reduce the travel time (Jara-Díaz 2000). The VTTS usually depends on trip purpose and trip length and differs between modes of transportation. For example, higher values are usually estimated for commuting trips than for leisure or shopping trips (Shires and Jong 2009; Abrantes and Wardman 2011). Regarding the means of transport, there is no consensus. In fact, the VTTS for commuting by car is in some studies lower but in others higher than by public transportation. Focusing on cars, passengers tend to have a lower VTTS compared to drivers (Mackie et al. 2003; Shires and Jong 2009; Abrantes and Wardman 2011). Traffic conditions play also a role: several empirical studies found that the VTTS of business travelers and commuters is higher in congestion than for free-flowing traffic (Abrantes and Wardman 2011; Hensher 2011; Rizzi et al. 2012).

Another factor that influences the value of travel time information is the number of drivers that receive it. The more the drivers that receive this information, the lower its value (Wardrop 1952). For example, it could occur that all or many drivers making the same particular route would change their paths also in the same way, according to a given information, trying to avoid some long travel times previously announced. Therefore, congestion would be translated to the new route, at least partially. How travel time information is managed is also a very important topic for highway travel time information agencies.

In short, the accurate estimation of travel times, both in real time and off-line, is indispensable to ensure an efficient performance of the road traffic system, with the subsequent benefits for the economy, the environment, safety, etc., that this fact implies. However, the real-time estimation of travel times or, more precisely, their real-time prediction, is especially valuable and useful. In this regard, "new" technologies are revolutionizing transport, particularly road transport, and offer an increasing range of possibilities. Autonomous vehicles or cooperative driving are clear examples of this transformation. Additionally, the diversity and quantity of the information handled at present would be unthinkable a few years ago, and this trend is expected to continue. In this context, very accurate travel time predictions will be not only possible, but also essential.

2.2 Travel Time Definitions and Estimation Methods

Next sections describe different ways to get travel time estimations aimed, usually, at feeding highway travel time information systems. However, there is a need for a clarification regarding some terms before delving into the topic. Many researchers

refer to "travel time estimates" when they (i) indirectly (including by data fusion) obtain past or instantaneous travel times or (ii) make travel time predictions. In this context, directly obtained past or instantaneous travel times are referred to as "travel time measurements". Most of them also accept interchangeably "measurements" and "estimates" in this last case, taking into account that measuring always implies a certain degree of error. Nevertheless, there exists a minor group of authors that use "travel time estimations" for past or present values, regardless of the way in which they were obtained, whereas they call future travel times "travel time predictions" (van Lint 2004; Mori et al. 2015). It is important to remark that this chapter follows the first of these terminologies. The reason is twofold: (i) it is more usual and (ii) there exist models able to provide past, instantaneous or future travel times, depending on the nature of the input data. If the first terminology is chosen, they can be simply referred to as "travel time estimation models", avoiding more complex jargons.

2.2.1 Direct Travel Time Measurements

Direct travel time measurements are the result of the most straightforward ways to obtain travel times: the reidentification of vehicles at two different control points, or their tracking along their individual trajectories. Both methods are addressed in detail and independently in the next sections. As a preview and considering them jointly, Fig. 2.5. introduces the main advantages and disadvantages of the direct measuring of travel times.

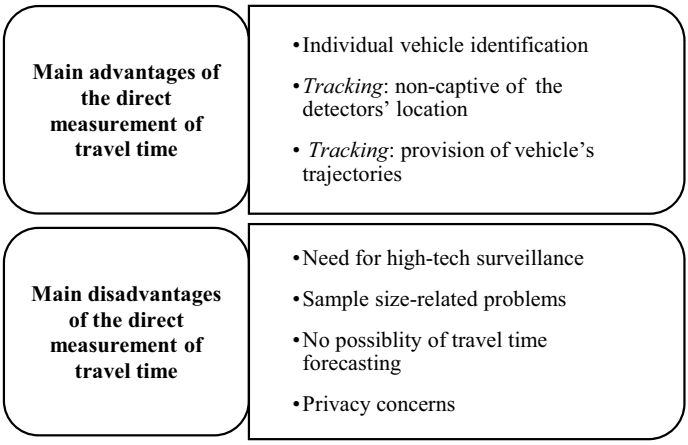

Fig. 2.5 Main advantages and disadvantages of the methods used for the direct measurement of travel times

According to it, two distinctive characteristics must be highlighted:

- Each measurement is linked to a particular vehicle. This type of individualized information is very valuable. However, this fact also implies the need for a meaningful amount of measurements to obtain representative averages of the monitored stretch. More taking into account that not all vehicles that travel along it are identified and/or tracked due to different reasons (lack of technology, failures, bad weather, etc.).
- The obtained measures have not only a temporal implication (i.e. they depend on the moment of measurement), but also a spatial implication too. That is, they are the result of the boundary conditions in the target stretch. However, the correctness of this asseveration is pointed out below for the case of reidentification at control points.

2.2.1.1 Vehicle Reidentification at Control Points

The first way to obtain the individual travel time of a vehicle over a stretch consists in registering the times at which it enters and exists this stretch, and making the proper subtraction. These kinds of measurements were first made manually, with two people respectively positioned at each end of the target stretch. Nevertheless, the data obtained from such operations contained significant errors, unless traffic flow was minimal. Technological progress allowed the automatization of these measurements. Nowadays, there exist different types of detectors that are placed on the roads with the goal of reidentifying vehicles, thus mimicking the manual procedure. Known as *Automatic Vehicle Identification* (AVI) technologies, they recognize individual vehicles on the basis of different features that distinguish them univocally. In any case, clock synchronization at control points is a key issue to obtain accurate measurements.

Although slightly addressed in Chap. 1 as examples of eulerian sensors, it is important to remark the main features regarding travel time measuring by means of AVI technologies. First, the accuracy of AVI systems is much higher than that of manual procedures. However, none of them is perfect and the number of pairings in a stretch does not normally equal the number of vehicles that effectively crossed it. In this regard, the characteristics of the road (geometry, number of lanes, average illumination level, weather, average flow, etc.) should be taken into account to choose the more suitable technology. Nevertheless, this choice is usually based on the available budget. Although their price decreases over time, AVI systems are still quite expensive. The use of insufficient or inappropriate technologies boosts unsuccessful pairing. In fact, sample size is another inconvenience of these methods (Quiroga and Bullock 1998; Cheu et al. 2002; Shuo Li et al. 2002; Jiang et al. 2006, etc.). If it is too small, the obtained travel times are probably not representative of the reality in the stretch. Besides, the detection of outliers (e.g. a driver stops to make a call) is unfeasible. This was a major problem in the past, as most AVI systems require not only infrastructural equipment, but also the presence of a particular on-board technology (e.g. Bluetooth). Nowadays, most of these gadgets are widespread within the

traffic stream, which facilitates the achievement of bigger samples. However, there is still a need for investment, especially in the infrastructure. This statement could be verified during the writing of this book. When trying to obtain data of Bluetooth signal identifications, it was found out that only a couple of freeways in Spain have Bluetooth detectors. On the contrary, it is well known that most vehicles or their drivers' smartphones deliver a Bluetooth signal. That is, valuable data for traffic management are not being collected. The situation of AVI readers, usually placed on gantries, to obtain representative samples must also be carefully analyzed (Sherali et al. 2006; Li and Ouyang 2011, 2012; Xing et al. 2013). Think for example of a short stretch (e.g. 2 km) with detectors at both ends but a junction in the middle. The number of pairings between reidentifications could be low due to significant inflows and outflows. No measurements (apart from counts) of those vehicles using the junction would be available. Nevertheless, if quite an acceptable configuration is available, data fusion can be a solution (see Sect. 2.3 and Chap. 4). In any case, reidentification failures must be avoided as much as possible to make the most of the penetration rate of these technologies both on-board and in the road network.

Additionally, AVI systems have an unavoidable disadvantage: travel time measurements are captive of the control points. That is, they can only be measured between the fixed points on the road where the AVI detectors are located. Again, the significance of the number of detectors and their location arises. In fact, this constraint has more implications if detectors are very far apart, as there is no possibility of knowing how a measured travel time is distributed along the path. Think for example of a stretch with excessive travel times: discerning if a general or a punctual problem is behind these long measurements would be unfeasible. Therefore, detectors placed in the infrastructure at very large distances would be insufficient to manage traffic in a proper way if no other data source is available. Respectively distances of 8 or 2 km between AVI detectors have been proven to be the maximum admissible in interurban and metropolitan (i.e. with more junctions) freeways in order to obtain profitable measurements (Turner et al. 1998). Finally, a non-negligible issue related to interval detectors is data treatment regarding privacy, as particular vehicles are identified and, thus, the location of their passengers (Turner et al. 1998; Yim 2003; Hoh et al. 2012). Traffic management centres usually have protocols to ensure population's rights in this regard. Besides, the use of encryption methodologies is increasingly frequent.

As mentioned in Chap. 1, examples of AVI systems are the reidentification of on-board transponders of different types by means of roadside beacons (Longfoot 1991; Nishiuchi et al. 2006), automatic license plate identification (ALPR or ANPR) by means of especial video cameras and license plate matching techniques, also known as optical character recognition (OCR) (Buisson 2006; van Hinsbergen et al. 2009), or Bluetooth- (Barceló et al. 2010; Bhaskar and Chung 2013) and WIFI- based (Abbottjard et al. 2013) detection systems. Other techniques have tried to take advantage of the notable density of loop detectors in road networks by using them as identification devices. One of these technologies tried to recognize individual vehicles from their lengths (Coifman and Cassidy 2002; Coifman and Ergueta 2003; Coifman and Krishnamurthya 2007). Although loop traps provide them, only singular vehicles can

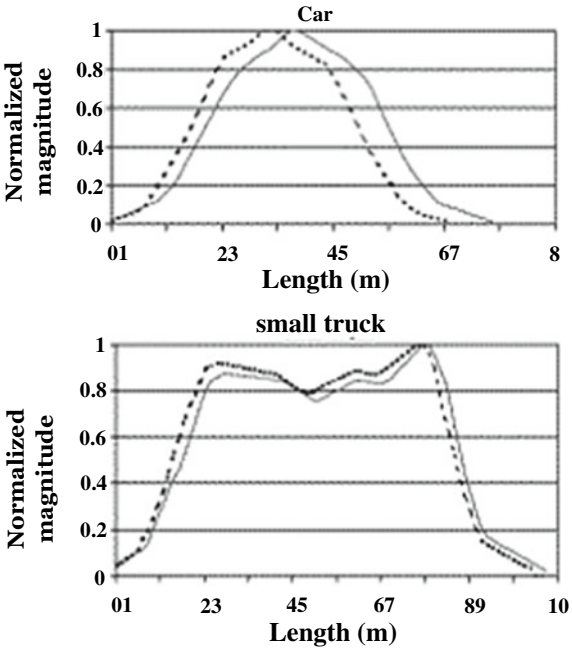

Fig. 2.6 Example of normalized signatures of a car and a small truck at two-loop detectors (inspired in Abdulhai and Tabib 2003)

be distinguished, especially in free-flowing situations. Under these conditions, high speeds, lane changing, merges and diverges increase the bias of the registrations. Other methodology tried to identify platoon structures. This resulted unfeasible in congested situations, as such a structure is lost (Petty et al. 1998; Lucas et al. 2004). Distinguishing vehicle's electromagnetic signature (inductance, Fig. 2.6) has been shown to be more successful (Abdulhai and Tabib 2003; Kwon 2006; Ndoye et al. 2011). Additionally, such methods have important advantages in comparison to other AVI technologies: (i) privacy is not violated, as the identification of the vehicles is anonymous in the sense that it is not linked to a license plate, an owner, etc., and (ii) no on-board technology is necessary and thus, failures apart, all vehicles contribute to the sample size. However, traditional loops are not capable of performing this task (Ndoye et al. 2011). Therefore, the acquisition of other non-intrusive technologies is gaining more success (Vanajakshi 2004; Soriguera et al. 2010).

However, even in the most ideal scenario (i.e. proper technology with high penetration rate, accurate measurements, etc.), AVI interval detectors have a major disadvantage: they have no predictive capabilities. In fact, they provide somehow "obsolete" travel times. Note that each travel time measurement is obtained once a vehicle has crossed the target stretch. Additionally, all direct measurements obtained in this stretch during a predetermined time interval ΔT are usually averaged, being this average the output of the system for this ΔT. A non-insignificant amount of highway travel time information systems rely on this kind of outputs. However, these so-called

Arrival-Based Travel Times (ATT) or *Measured Travel Times* (MTT), trajectory-based at the corridor level, are not the information that a driver at the entrance of the stretch should receive. Unless traffic is stationary, ATT can be very different from the travel times that these next drivers at the entrance of the stretch will experience. Thus, ATT represents a past measurement and involves a delay in the real-time dissemination of the information (Soriguera and Robusté 2011a). The effects of this delay can be considerably in the event of long travel times, resulting either from congested states or long stretches.

2.2.1.2 Vehicle Tracking

Another possibility to obtain direct measurements of travel time consists in continuously tracking vehicles while they circulate along the road network. The position of each vehicle is recorded every few seconds and, thus, individual trajectories are obtained. Therefore, tracking provides not only travel times, but also further valuable information. Indeed, all interesting traffic parameters, as they can be derived from these trajectories.

The most common technologies for tracking and their particular features were described in detail in Chap. 1, namely GPS-based tracking, RFID-based tracking, geo-location of cell phones, tracking by means of probe or floating vehicles and UAV-based tracking. In any case, lagrangian sensors that move with vehicles. No monitoring equipment is generally placed in the infrastructure, as no in situ control points are necessary. Central server stations, either in the traffic management centres or not, directly receive and process the information. At first, radio channels (e.g. General Packet Radio Service—GPRS—system) mostly performed the data location transmission from the vehicle to the control centres. Thus, interruptions in the transmission or data losses were quite frequent. Nowadays, transmissions performed via wireless networks are much more reliable. The absence of control points on the road implies another important advantage of vehicle tracking in comparison to AVI systems: it is possible to measure travel times between any two points. For their part, the main disadvantages of tracking equal some of those of AVI technologies, that is, privacy concern and the possibility of obtaining too small or biased (e.g. only of trucks) sample sizes. As said, the increasing penetration rate of the necessary devices, especially that of smartphones, helps to improve the situation over time (Yim 2003; Yim and Crayford 2006; Bar-Gera 2007; Herrera et al. 2010, Unde and Borkar 2014, etc.).

Highway travel time information systems that rely on the aforementioned methods also make averages of the individual travel times of all vehicles that traverse the target stretch in a fixed ΔT. As the last available information is used for their calculation, these averages can be considered "more recent" than those obtained from interval detectors. Called *Instantaneous Travel Times* (ITT), they involve less delay in the dissemination of real-time information. In fact, they are the best directly obtainable estimation of future traffic states and, thus, a better approach to the *Predicted Travel*

Times (PTT), that is, the travel times that vehicles asking for the information will actually experience (Soriguera and Robusté 2011a).

2.2.2 Indirect Travel Time Estimation

In the absence of technologies that allow the direct measurement of travel times, their indirect estimation from other characteristic variables of traffic flow is a very common procedure. In fact, even novel methodologies that perform such indirect estimations relying on new technologies are being developed. For example, Herrera et al. (2010) used information derived from GPS-enabled smartphones not to measure travel times directly, but to obtain them indirectly from previously constructed speed fields. However, apart from such valuable contributions, two main groups of methodologies based on basic surveillance are more commonly used for the indirect estimation of travel times: (i) those that rely on punctual values of traffic flow parameters for the estimation (basically spot speeds, densities or flows), or (ii) those that use accumulations over time of the anonymous counts gathered at the measuring points. Although not exclusively, inputs can be obtained from inductive loop detectors in both cases, which facilitates the implementation of these methods. The obtained estimates have two main characteristics:

- They are not linked to particular vehicles. In fact, not individual but average travel times, whether per lane or per section, are estimated. This fact avoids the need for identification, but it implies the loss of important information.
- When calculated from punctual measurements of traffic flow variables (the most usual procedure in practice), the obtained travel times do not truly have a spatial implication. Indeed, punctual inputs must be extrapolated to some extent and in different ways depending on the chosen procedure. Therefore, the disregard of events that occur between the measuring points during the trips involves some inaccuracy. This problem can be satisfactorily overcome using those methods based on the construction of cumulative count curves, as it is further explained in Chap. 4.

Figure 2.7 advances the respective advantages and disadvantages of each approach. Both these and their basics are described in the following sections more in detail.

2.2.2.1 Spot Speed-Based Methods

Traffic management centres have traditionally used these methods for the estimation of travel times, trying to take advantage of their loop detector equipment. This surveillance was in the past the only one for most traffic agencies, a situation that continues in a significant number of road networks. However, the simplicity of spot speed methods as well as human inertia cause that many highway travel time information

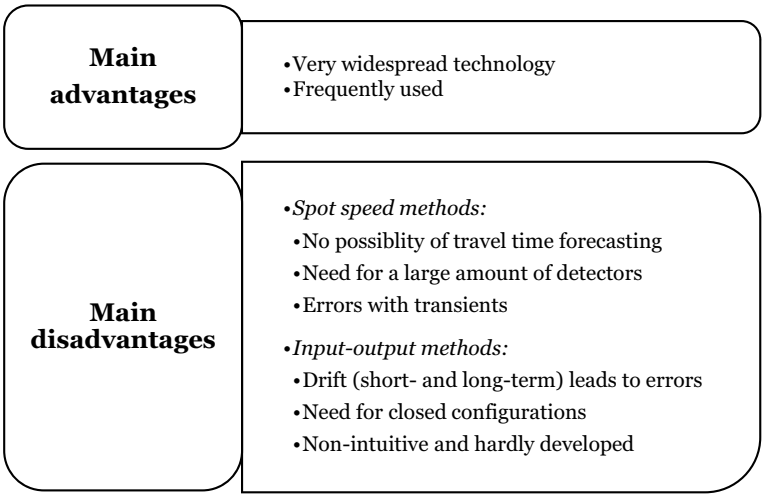

Fig. 2.7 Main advantages and disadvantages of the methods used for the indirect estimation of travel times

systems are still based on them, even in cases in which more advanced equipment is available.

Although a couple of variants exist, all methods follow the same basic steps. First, the punctual speeds gathered by the loops are averaged for predetermined time intervals of aggregation Δt, whose duration differs among the administrations. Second, these punctual means are generalized for the links delimited by each pair of loops. Third, indirect travel time measurements in these links are obtained. Three main factors involve inaccuracy from the very beginning:

- The punctual speed measurements at the loops are not always accurate enough. This is especially important in the case of single loops (see Chap. 1). When double loops are available, spot speeds are generally satisfactory.
- Different procedures are being used for the spatial generalization of the punctual speed means. However, all of them introduce bias in the subsequent travel time estimates, especially those consisting in mere mathematical interpolations between detectors. In fact, any interpolation involves assigning to the link some average traffic conditions that do not correspond to reality, as they arise from an artificial combination of the conditions at particular points. Only extrapolations based on traffic dynamics and queue evolution could result satisfactory.
- In any case, the average speed accepted for the link is given a spatial nature. Then, average travel times for a link, $\overline{tt}$, are simply obtained by dividing the length of this link by this generalized speed. However, the variable that would allow such calculation is a generalized space mean speed. Space means, $\overline{v_s}$, can be calculated as the harmonic mean of individual spot speeds (Eq. 2.1, where L stands for the length of the link, n for the number of vehicles that passed over the detector in the considered time interval and v_i for the individual spot speeds). However, loop

detectors provide time mean speeds, $\overline{v_t}$, i.e. arithmetic means of individual spot speeds for determined time intervals of aggregation Δt. Unless traffic is stationary, time means speeds are higher than space mean speeds. Therefore, travel times are underestimated. Chap. 3 of this book further elaborates on this issue, and it is precisely devoted to solve this inconvenience with no extra expenses.

$$\overline{tt} = \frac{\sum_{i=1}^{n} tt_i}{n} = \frac{\sum_{i=1}^{n} \frac{L}{v_i}}{n} = L * \frac{1}{n} \sum_{i=1}^{n} \frac{1}{v_i} = \frac{L}{\overline{v_s}} \tag{2.1}$$

These drawbacks can lead to completely flawed travel time estimations, especially in unstable congested situations (stop and go instabilities, formation or dissipation of shock waves, etc.). The smaller the Δt, the higher the possibility that detectors only measure speeds over one of the traffic instabilities. In such cases, the punctual speeds that are extrapolated to the whole link are not representative of the traffic conditions along its entire length. The subsequent errors in travel time estimations can reach 30% (Rakha and Zhang 2005). These errors are smaller with longer Δt, as these unstable non-stationary traffic states are generally smoothed. However, longer Δt are undesirable for real-time highway traffic information systems, for which information's immediacy is essential. Situations of severe congestion can also lead to wrong results, as loops only measure speeds of moving vehicles. That is, they do not account for the periods in which they are stopped. Average speeds would be overestimated and, consequently, travel time estimates underestimated. All aforementioned considerations correspond to the trends usually observed in such situations. However, other could occur, in line with the uncertainty linked to unstable traffic phenomena. This fact makes difficult the implementation of proper corrections. Regarding loop density, better results are obtained for the freeways where an extensive net of loop detectors is uniformly displayed (at least a double loop every 500 m), especially taking into account the possibility of sensor failure (Gentili and Mirchandani 2018). However, longer links between loops result more favourable in the event of traffic instabilities, as they are usually smoothed. Different speed interpolation models (Table 2.2) have tried to solve the former issues, but none of them resulted satisfactory enough.

Travel time estimations from all these methods are acceptable when free-flowing conditions prevail, that is, when there is spatial stationarity in the link. However, those that completely overlook traffic dynamics provide highly underestimated travel times in fully congested situations. During congestion onsets and offsets, i.e. in partially congested situations, each method leads to different results, but they all tend to be inaccurate and unreliable in any case. The reason is that most of them neglect where transitions really occur within the links. The proposals of Coifman (2002) and Treiber and Helbing (2002) partially face this issue and, in fact, their results are better with fully congested sections. However, further improvement is needed, as they continue to fail with transients. Lacking a better approach based on traffic dynamics and queue evolution, an intelligent smoothing of the noisy loop detector data implemented together with midpoint, linear or truncated quadratic methods helps

Table 2.2 Most common methods used for the interpolation of spot speeds for travel time estimation, and examples of application developed by particular authors

Method	Reference example	Comments
Constant interpolation between detectors (e.g. conservative approach: choosing the lowest spot speed from those calculated at the end detectors)	Cortés et al. (2002)	– Traffic dynamics and queue evolution are overlooked – Speeds are supposed to change instantaneously – Speeds' discontinuities are assumed to take place at the detectors – A weighted average speed was used in this case
Piecewise constant interpolation between detectors (e.g. midpoint algorithm, thirds method)	Kwon (2004)	– Traffic dynamics and queue evolution are overlooked – Speeds are supposed to change instantaneously – The speed discontinuity is assumed to take place within the section – Performed with the thirds method in this case
Piecewise linear interpolation of speeds between detectors	van Lint and van der Zijpp (2003)	– Traffic dynamics and queue evolution are overlooked – Speed changes are distributed in the traffic transition along the section. However, this distribution does not match drivers' real response to changes in the traffic state (much more concentrated in time and space) – Anticipation of slower or faster speed regimes due to the assumption of constant acceleration between detectors – Better than the piecewise constant interpolation, at least with simulated data. Nevertheless, inaccuracies remained in fully or partially congested stretches

(continued)

to reduce fluctuations linked to short Δt, which, although desirable for real-time information systems, are a major source of errors for spot speed methods (Soriguera and Robusté 2011b).

Finally, it cannot be overlooked that the final information that highway travel time information systems provide are not travel times in each link between loop

Table 2.2 (continued)

Method	Reference example	Comments
Quadratic interpolation of speeds between detectors (e.g. basic or with a truncated definition of the speed evolution)	Sun et al. (2008)	– Traffic dynamics and queue evolution are still overlooked – Attempt to mimic drivers' real response to changes in the traffic state, allowing smooth accelerations and decelerations between detectors – Changes in speed and acceleration have no relation to traffic evolution, but are linked to a mathematical function
Punctual speed generalized over the section according to classical continuum traffic flow theory	Coifman (2002)	– Linear approximation of the flow-density relationship. Thus, the speed at which transitions between two congested states propagates is known (the slope of the right branch of the diagram) – Satisfactory results in congestion. Unsatisfactory results in partially congested situations
Punctual speed generalized over the section according to classical continuum traffic flow theory + smoothing filter	Treiber and Helbing (2002)	– A non-linear spatio-temporal lowpass filter is applied to the input detector data. This filter exploits the fact that, in congested traffic, perturbations travel upstream at a near-constant speed, while in free flow, information propagates downstream. All higher fluctuations are smoothed – Speeds are finally obtained as smooth functions of space and time – Good results in congestion in short sections and for short Δt. Unsatisfactory results in partially congested situations – Conservation is overlooked

detectors, but corridor (i.e. several links) travel times. For this purpose, the travel times (estimated at the same instant) of all links that form the target corridor are simply added up in real time. Note that each link travel time is obtained from the average speed measured over the last few minutes. This implies that (i) the most recent information is used and (ii) the obtained corridor travel time is not trajectory-related. It is like a picture of the current travel times along its path. Furthermore, it is possible that no vehicle trajectory fits this travel time. This temporal alignment-based concept of travel time is again known as ITT (*Instantaneous Travel Time*), as in the case of tracking. However, in that case trajectories and their associated travel times were real. Anyway, ITT may be considered again the best approximations to the desired real-time "future" information, in the absence of forecasting methods. This asseveration implies that traffic conditions are supposed to remain constant between consecutive time intervals. To support it, both short Δt and short corridor lengths are advisable (Soriguera and Robusté 2011a).

2.2.2.2 Input–Output Methods

Input–Output Methods are explained in detail in Chap. 4. Therefore, this section only introduces their most basic fundamentals, so that they can be broadly compared to the other travel time estimation methods revisited. Figure 2.8 shows a typical cumulative count input–output diagram of a road section limited by two inductive loop detectors. For both of them and starting at any time $t = 0$, corresponding to the pass of a reference vehicle over the entrance detector, it is possible to count and accumulate over time all vehicles passing their respective locations. If this process is graphically represented, cumulative count curves are generated. More in particular, they are drawn little by little after the end of predetermined time intervals Δt, respectively

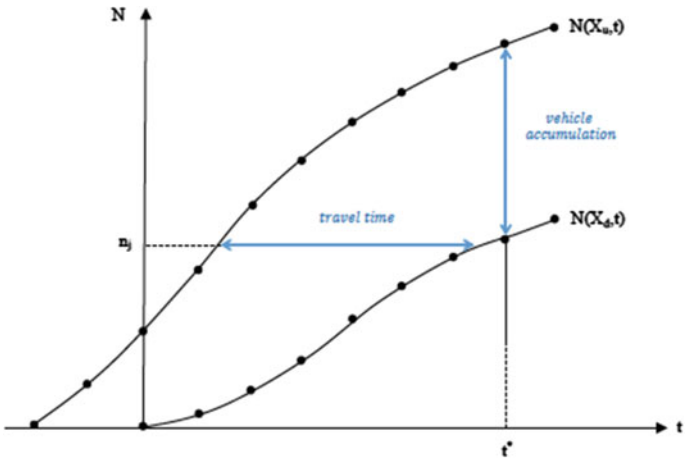

Fig. 2.8 Basic graphical interpretation of input–output diagrams

adding the number of vehicles that passed the detectors during these periods. They are, therefore, stepwise functions, which can be smoothed in practice taking into account the large amounts of vehicles that is usually registered. In the example of Fig. 2.8, $N(X_u, t)$ is the "arrivals" count curve measured at the upstream detector X_u of the section and accumulated over time. For its part, $N(X_d, t)$ is the "departures" cumulative count curve measured at the downstream detector X_d. The travel time in the section of any vehicle n_j can be directly obtained from the diagram, as it is the horizontal distance between the curves at the height of vehicle n_j. For its part, the vertical distance between the curves at any time t^* is the vehicle accumulation in the section at this time.

Two important assumptions, whose consequences are also discussed in Chap. 4, are behind these diagrams:

- There is no passing in the section, i.e. the system is FIFO (First in-First out). This assumption allows obtaining travel times from the curves. In this regard, it is important to remark that input–output diagrams do not actually consider individual vehicles, but positions within the traffic stream or, as it is usually said, they consider "labels".
- Vehicle conservation holds in the section, i.e. vehicles do not enter or leave the section at any halfway point. This assumption allows obtaining vehicle accumulation from the curves.

Input–output methods have significant advantages in comparison to other procedures:

- First, technological requirements are minimum. Not only because they rely on inductive loop detectors, but also because these can even be single loops. Moreover, they can be located at relatively large distances. Therefore, no expensive or rare surveillance is necessary. The only condition in this regard is that all junctions must be monitored, in order to ensure the assumption of conservation (see Chap. 4). Even if this were not the case, the extra expenses needed for the deployment of some additional loops would be affordable for most administrations.
- Second, but more important, input–output methods can provide not only trajectory-based instantaneous travel times (ITT), but they also have predictive capabilities linked to the spatial nature of one of their outputs: the vehicle accumulation. The value of this parameter at any time t^* indicates the number of vehicles that should be served (i.e. that should be able to traverse the section) in the next time interval $t^* + \Delta t$ under free-flowing conditions. Again, Chap. 4 further elaborates on this concept.

The usefulness of such methods for different purposes has already been demonstrated by outstanding researchers of varied disciplines and, particularly, of that of road traffic (e.g. Daganzo, unkn. date; Newell 1993a, 1993b, 1993c; Lawson et al. 1997; Newell 1999; Cassidy et al. 2002). However, they are not used in practice for travel time prediction. The fact that they are not intuitive is one possible reason. Also, that detector drift, a common problem of loop detectors, could lead

to significantly biased results. The methodology introduced in Chap. 4 of this book allows benefiting from the advantages of input–output methods for travel time prediction, while solving the consequences of drift by means of data fusion.

2.2.3 Data Fusion for Travel Time Estimation

Well-developed data fusion procedures are gaining acceptance in most research fields, supported by technological progress and advances in computing capacity. All of them are aimed at combining data from different sources, sometimes using different algorithms/models, so that the strengths of some of them compensate for the weaknesses of others. Thus, the goal is to obtain better outputs than that available from a single data source.

Regarding travel times, a significant amount of studies have developed data fusion methodologies to obtain travel time distributions, i.e. travel time reliability, on particular highways. In fact, travel time information systems are also usually fed with this kind of data. If not, as seen, either with direct measurements of travel time or with travel time estimates obtained from a single data source. Less research can be found that used data fusion for travel time estimation in real time. Table 2.3. summarizes some of the most important pieces of research (according to their impact on the field) on the topic performed in the last years, overlooking for the moment those based on Big Data methodologies. Although most procedures can perform satisfactorily in uncomplicated environments (free flow, very recurrent traffic state patterns, etc.), they do not adapt well to varying or congested situations, especially if they are linked to sudden events. In addition, this kind of schemes would not be accurate enough in future scenarios like cooperative driving or seamless intermodal on-demand systems. Therefore, there is a need for the development of more accurate, reliable, robust and standardizable (e.g. economically and technically feasible and non-site-dependent) data fusion procedures for travel time prediction. The methodology presented in Chap. 4 has this goal. More significant pieces of research from previous years are also available, which coincide in time with the boosting of ITS (from the end of the 1990s until the early twenty-first century) and with the recognition of the travel time as one of the most (if not the most) important parameters for traffic management. Despite their outstanding contribution, significant changes in the boundary conditions (e.g. available surveillance, technological level, computational capabilities, etc.) make their re-evaluation and probably their update advisable.

Anyway, the review of all previous works allows classifying them according to different features. Table 2.4. addresses this division depending on five aspects: the level/levels at which the fusion is performed, the mathematical nature of the methodology, the general approach to the problem, the role of the contextual information in the fusion and the behavior of the fusion operator. Describing in detail all existing models is out of the scope of this Chapter, but this differentiation as well as the brief explanations provided below provide a good overview. Each particular data fusion methodology could be included in one category of each classification. Additionally,

Table 2.3 Significant data fusion procedures for travel estimation developed in the last years

Reference	Data	Procedure	Results and challenges
Soriguera and Robusté (2011a) Transportmetrica	Spot speeds and counts from inductive loop detectors and toll tickets	Simple and feasible. Short-term prediction. Two levels of fusion by fuzzy logic and a probabilistic approach based on Bayes' theory	Applied with reasonable and accurate results, but rounding implies negative effects. Need for a learning process that improves the probabilistic fusion
Lim and Lee (2011) IET Intelligent Transport Systems	Spot speeds from inductive loop detectors and AVI direct measurements of travel time	Based on a traffic flow model and a k-nearest neighbourhood (k-NN) model. Attempt to correct the time lag of AVI measurements	Short-term predictions better than direct measures. Too many assumptions behind the traffic flow model. Need for more surveillance/calculations per iteration
Yildirimoglum and Geroliminis (2013) Transp. Research Part B	Spot speeds from inductive loop detectors and historical data	Short-term prediction. Bottleneck identification, development of stochastic congestion maps for clustered data and online congestion search algorithms	Promising travel time predictions under varying traffic conditions, but it is a very complex method, difficult to implement. Need for good historical data
Zhang and Ge (2013) Computer-Aided Civil and Infrastructure Engineering	Speed, volume, and occupancy from microwave detectors. Travel time measures from toll tags	Freeway corridor travel time. Combination of a Takagi–Sugeno–Kang fuzzy logic system and a neural network	Good short-term predictions, but the method is complex and requires quite a large amount of data
Chen and Rakha (2016) Transp. Research Part C	Varied. From on-board GPS, loops, mobile phones, etc	Agent-based modelling approach that performs multi-step travel time predictions	Accurate and efficient predictions in real time (better than ITT), but it requires too many data (here bought to a private company). Unfeasible for generalization
Tak et al. (2016) Computer-Aided Civil and Infrastructure Engineering	Varied. From toll collection systems, loops and DSRC (dedicated short-range communications)	Long-term prediction horizon (6 h). Hierarchical pattern matching (Multilevel k-nearest neighbour, Mk-NN) method	Relatively accurate and robust prediction of long-term travel times with short computation time. Short-term predictions are more accurate, especially if no historical data are considered. However, data sources are site-restrictive

(continued)

Table 2.3 (continued)

Reference	Data	Procedure	Results and challenges
Wang et al. (2016) Transportmetrica A: Transport Science	Varied. From a particular Intelligent Monitoring and Recording System and the Multifunction Automatic Detecting and Recording System	A regression model is built and integrated into an existing travel time estimation model. The goal is to differentiate link travel times depending on the particular traffic stream directions	Useful in urban networks, where different directions are present, but not on highways. It depends on the accuracy of the implemented travel time model. Linked to very particular data sources and, thus, not standardizable
Pirc et al. (2016) IET Intelligent Transport Systems	Loops spot speeds, AVI travel time direct measurements, qualitative level of service (LOS) data	Multiple linear regression modelling the relationship between explanatory variables (MTT, travel time estimated from spot speed measurements, LOS) and a response variable (departure-based travel time) by fitting a linear equation to observed data	Implemented in a real case. 9% better accuracy than usual travel time prediction algorithms. It depends on qualitative data. Doubts about the possibility of its standardization
Shao et al. (2018) Transportmetrica A: Transport Science	AVI or GPS-based travel time measurements and historical data	Generalized least squares problem with non-linear constraints. Solution algorithm based on the penalty function method	Network-wide travel times. Potential to enhance their accuracy and quality, which are worse than that of other link travel time estimation methods. Only numerical examples. The number and location of sensors impact the accuracy and quality of the estimates, which could be problematic for a real implementation

with the exception of the categories linked to the general approach to the problem, the rest would be valid for any other fusion models not aimed at travel time estimation. Descriptions are thus generalized for the sake of simplification.

It must first be taken into account that the fusion can operate on the inputs of the travel time estimation model or on the outputs of one source-based estimations. In this regard, three types of models can be distinguished, which are represented in Fig. 2.9. for the example case of three data sources:

Table 2.4 Classification of data fusion travel time estimation procedures

Fusion level/s	Single-estimation models	A single travel time estimation model fuses heterogeneous data
	Multiple-estimation models	Fusion of the outputs of several travel time estimation models (one per data source)
	Hybrid models	Fusion of the outputs of both single- and multiple-estimations models
Mathematical nature	Probabilistic models	Based on empirical probability functions with their associated conditional probabilities
	Evidential logic models	Based on approximate probability functions. Probabilities are given a confidence level
	Fuzzy logic models	Statement logic. The truth-values of variables may range any real number between 0 and 1
	Artificial intelligence (AI) schemes	Machines mimic human cognitive functions such as learning and problem solving
General approach to the problem	Naïve models	No model. Based on very restrictive assumptions that are not always true
	Traffic flow-based models	A traffic model developed on the basis of theoretical principles predicts traffic evolution
	Data-based models	Data themselves determine the function that relates the inputs to the outputs
	Hybrid models	Combination of the former approaches performed at the same or at different steps
Role of contextual information	Context-independent models	External information plays no role
	Context-dependent models	External information influences the fusion
Operator behaviour	Indulgent fusion operator models	The credibility of the inputs is increased
	Severe fusion operator models	Criteria is simultaneously satisfied
	Cautious fusion operator models	They behave like a compromise

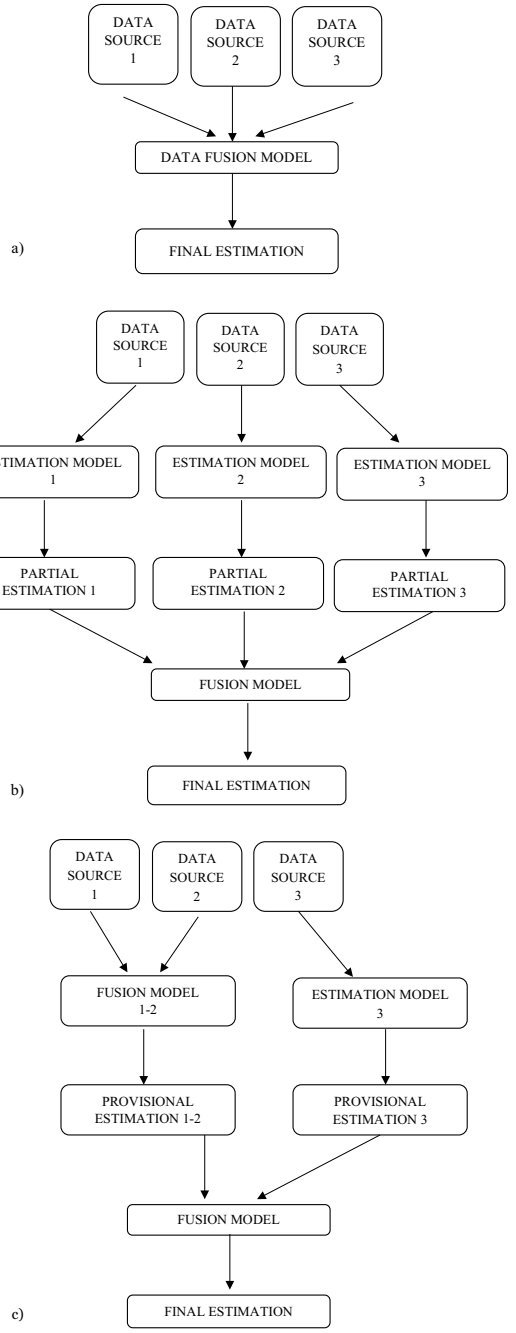

Fig. 2.9 Classification of data fusion models depending on the levels at which the fusion is performed: **a** single-estimation models; **b** multiple-estimation models; **c** hybrid models

- *Single-estimation models* are those that fuse data from different sources. Therefore, a sole model processes all data to obtain a unique output. Previous or integrated algorithms aimed at detecting and removing outliers from the inputs are advisable. Neural networks, state-space models and several traffic theory-based models usually perform within this category.
- *Multiple-estimation models* are those that obtain a final estimation by fusing previous "partial" estimations. That is, they fuse the estimations provided by simpler models, whose inputs were respectively supplied by a single data source. The removal of outliers can be in this case performed within the raw data, once obtained the partial estimations or, at best, in both cases. Models based on the Bayesian theory, on the Dempster-Shafer theory, some fuzzy logic schemes, weighed linear combinations or network equilibrium models are normally applied in this way.
- *Hybrid models* combine the approaches of single and multiple-estimation models. That is, provisional outputs obtained from an amalgam of models of the aforementioned categories are fused to obtain the final estimation. Developing them by the modification of an existing multiple estimation model is a usual procedure.

Second, it is possible to classify data fusion models (also one source-based models) depending on their mathematical baseline. In this sense, fusion operators can be based on:

- *Probabilistic logic models*, which are those based on the classic probability theory. Therefore, they are characterized by their robust mathematical foundations. Additionally, they require empiric data both to construct probability density functions and to define laws for the conditional probabilities. They become complex when dealing with very complicated cases. Excessive computation power and time as well as the possibility of obtaining counter-intuitive results advise against their use in these contexts.
- *Evidential logic-based models*, which try to overcome the difficulties of the classic probability theory to deal with ignorance. A complete empirical probability model is not required. Approximations to the probability functions, which are given a particular level of credibility, are thus acceptable. Additionally, evidential logic tries to take advantage of sets of hypotheses and not of each hypothesis separately. This facilitates the reallocation of the probability of belief in the hypotheses when evidences change.
- *Fuzzy logic-based models*, which range the veracity of an assertion between 0 and 1. This assertion is a consequence of the reasoning from the available level of knowledge, usually inaccurate or partial. Thus, and in contrast to classical logic, for which conclusions are either true or false, fuzzy logic is multivalent. Additionally, fuzzy logic models use these degrees of truth as a mathematical model of vagueness. They also differ in this sense from probabilistic or evidential approaches, in which the probability functions, either empirical or approximate, model ignorance.
- *Artificial intelligence-based schemes*, which are very different from mathematical models. Moreover, AI is a key part of computer science influenced

by mathematics, but also by philosophy, psychology, etc. Primary AI is based on four pillars (Nilsson 1980). First, the search of a required "state" among all states produced by all possible actions. Second, the use of genetic algorithms, in which the candidate solutions to an optimization problem are iteratively modified to achieve better and better solutions. Third, artificial neural networks, which are computing systems that "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules. And finally, there is a need for reasoning through a formal logic and for abstract thinking. AI can have very different approaches (e.g. machine learning, cybernetics, soft computing, computational intelligence, embodied intelligence), in which the integration of probabilistic logic models, evidential logic models, fuzzy logic models, etc., can be suitable.

Advances in computing capacity have made the combinations of the former approaches feasible. However, as in other cases, the most complex methodologies are seldom the most advisable. Simple procedures, even deterministic, based on a good knowledge of all boundary conditions, thus avoiding the need for assumptions or coarse estimations, are not only more feasible, but they may also be more accurate.

Although related to the mathematical logic applied, for which, as said, categories are not exclusive of travel time estimation models, other differentiation in this case particularized for them, is not only suitable, but meaningful (Fig. 2.10). Depending on the general approach they have to the problem, i.e. on how they estimate travel times, models can be (van Hinsbergen et al. 2007; Mori et al. 2015):

- *Naïve models*, which are ad hoc very simple travel time prediction methods, used even without fusion, i.e. with data coming from a single source. They are based on the acceptance of very restrictive assumptions, whose suitability is not confirmed. Despite its usually low accuracy, they are widely applied in practice (but barely in the research field, where they are mainly restricted to comparison) because of their minimal computational demand, their speed and their easy implementation. For example, some methods assume that traffic remains constant for a period and accept ITT as predictions. Others accept averages estimated in different ways from historical data. The combination of instantaneous and historical data (not guided by a model) or simple clustering techniques also belong to this group.
- *Traffic flow-based models*, which use the principles of traffic theory to recreate present traffic conditions as well as their evolution over time. Once this is obtained, travel times are derived. The success of these models depends on the accuracy when forecasting traffic evolution, which is not a simple task. Simulation tools are usually applied for this purpose but, as it was explained in Chap. 1, they must be properly chosen and calibrated with empirical data. These methods could be very profitable for traffic centres, as the knowledge of traffic conditions could support related decision-making tasks. However, the fact that they are not intuitive (a good background on traffic theory is necessary to manage them properly), their demanding computational requirements and difficulties in calibration hinder their generalization.

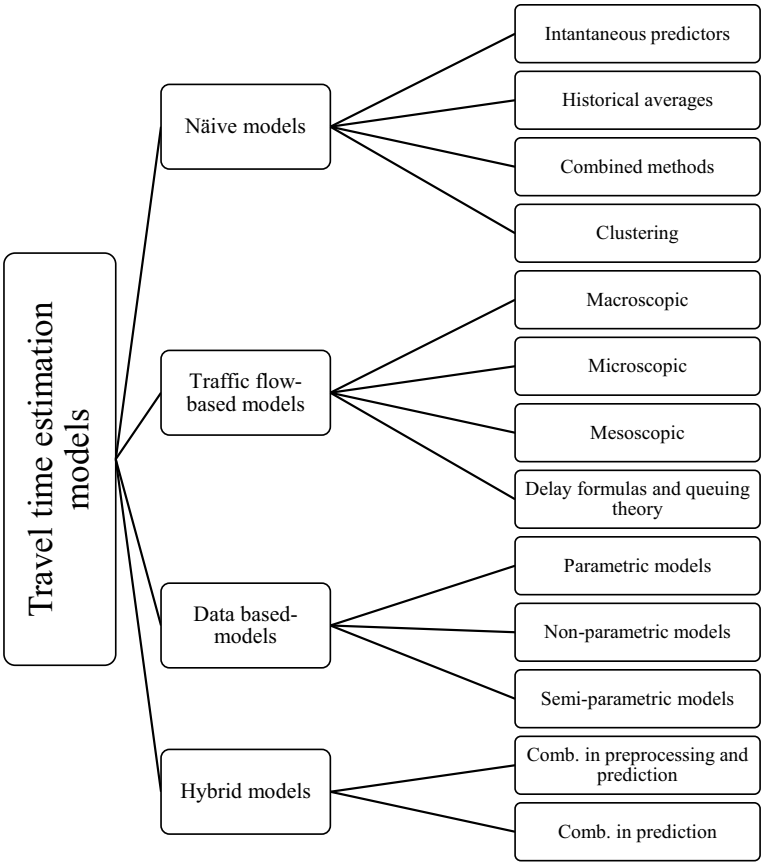

Fig. 2.10 Classification of (data fusion) models depending on their general approach to the estimation of travel times

- *Data-based models,* which are those that, by means of statistical and machine-learning techniques, derive from the available data a function that relates the explicative variables with the target variable. Traffic principles are therefore completely neglected. Although in fashion, these models (i) require huge amounts of data to derive a proper function and (ii) are site-dependent. If the structure of the function is predetermined and the data are used only to "adjust" it, the models are called *parametric*. In *non-parametric* models, the function is freely derived from the data and has no predefined structure. An intermediate approach is found in *semi-parametric* models, which neglect some of the assumptions that would determine the structure of the function in a parametric model to make it more flexible.
- *Hybrid models*, which combine some of the aforementioned approaches in different ways, trying to take advantage of their strengths and compensate for their weaknesses. These combinations can take place in different steps of the

fusion (e.g. one model fuses/processes the data and another one predicts travel times) or in the same/s.

Fourth, data fusion models can be divided into groups depending on the role of the contextual information in their estimations:

- A *context-independent* fusion operator only takes into account the values to fuse, neglecting any other external information.
- A *context-dependent* fusion operator acts according to the contextual information of the fusion and, thus, it needs this information to perform properly.

Particularizing for the case of travel time prediction models, contextual information plays an important role. It is usually related to the study site (number of lanes, number and location of junctions, accuracy of the available detectors, etc.), especially for those models based on traffic flow principles. Data-based models, on the contrary, pay more attention to weather or incident information, the calendar, etc. The consideration of both types of contextual information, if possible, leads to better results (Mori et al. 2015).

Finally, fusion operators can be distinguished according to their behaviour. In this regard, Bloch (1996) used the qualitative adjectives "severe", "cautious" and "indulgent" to make the distinction. Although valid for any number of inputs to fuse, she gave a simple example with only two for the sake of simplicity. She referred to the credibility associated to each of these inputs as x and y, which could take values into the interval $[0, 1]$. $F(x, y)$, with possible values within the same interval, was the credibility associated to the result of the fusion. In this context:

- A fusion operator is *indulgent* if it acts with a disjunctive behavior (Eq. 2.2):

$$F(x, y) \geq \max(x, y) \tag{2.2}$$

  These operators express redundancy between criteria. They increase the certainty we have about an information.
- A fusion operator is *severe* if it has a conjunctive behavior (Eq. 2.3):

$$F(x, y) \leq \min(x, y) \tag{2.3}$$

  This kind of operators represents a consensus between inputs or, in other words, focuses on their common or redundant aspects. It searches for a simultaneous satisfaction of criteria.
- A fusion operator is *cautious* if its behavior is intermediate (Eq. 2.4):

$$\min(x, y) \leq F(x, y) \leq \max(x, y) \tag{2.4}$$

  Such an operator provides a global measure, intermediate between the inputs to fuse. By the introduction of weights, it is able to include interactions and dependencies between them, thus avoiding combination bias.

It is important to note that some operators become severe, indulgent or cautious depending on the nature of the inputs to fuse (e.g. their sign). Namely, AI depends on the values of the inputs, but is independent of the contextual information. For example, some fuzzy logic techniques depend on both aspects.

2.3 Dissemination of Travel Time Information

The dissemination of the information is a fundamental task of any travel time information system. This section is particularly aimed at reviewing the basics of this dissemination from a practical point of view. Considerations about how information should be delivered in order to reach the system optimum (Wardrop 1952) are not included here. In this context, Fig. 2.11 summarizes the three most important variables to be considered so that travel time information is successfully communicated. In fact, interdependences among these variables exist. Note that most of the next reflections are also applicable to the dissemination of other types of traffic-related information.

First of all, strategies designed to deliver information to drivers before they begin their trips will mostly be different from those used to fulfill their information demands while travelling. Indeed, the required information is not the same either, at least normally. Pre-trip information allows planning. Travel time reliability information is commonly demanded in this context, for different purposes like (i) trip planning for habitual trips such as commutes, or (ii) trip planning before following an unfamiliar route, to be aware of the typical travel times on it. Drivers can decide on its basis their mode of travel, their departure time or even, if possible, if they make the journey or not. Travel time reliability information can also be useful as on-trip information,

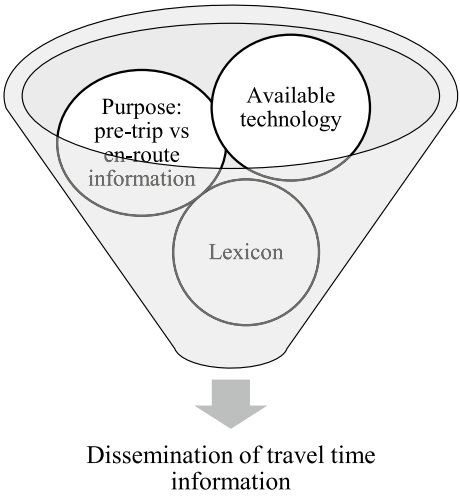

Fig. 2.11 Features to take into account before implementing plans for the dissemination of travel time information

allowing users to change their trip while in progress, prior to a route or mode choice point. However, instantaneous travel time estimations (note that predictions are not generally available, as it is discussed in Chap. 4) are more demanded when a trip has already begun. Anyway, on-trip information enables drivers to modify the initial planning according to current traffic conditions, provided that this possibility exists. Both pre-trip and en-route, the main goal of the dissemination of travel time information is to avoid late arrivals, which can have subsequent impacts if the person/freight is awaited at the destination at a certain time. It can also be useful to avoid stress and, therefore, accidents. A good travel time information system should thus allow users to receive the particular information they ask for in the form that results more useful for them.

In line with the former statement, the lexicon used for any kind of travel time information delivery must be clear and concise, and adapted to the target driver. These features are especially important for en-route information, as (i) the time a driver has to process the information is usually short and (ii) he will possibly take decisions in real time on its basis. Ambiguous or too complicated messages could, in the best case, result useless. In the worst case, they could lead to incidents. In this regard, many studies tried to define the most suitable length of any kind of warning or information delivered in real time to drivers, so that it results useful and does not represent a distraction from the driving task. For the case of Dynamic Message Signs (DMSs) displayed on the road, the optimal distance between each pair has also been analyzed and, in fact, is regulated by standards. Research on lexicon for auditory messages is also in process. Nowadays, auditory messages are already used by some apps or navigations systems, but they are, generally, not subjected to any rules. However, this type of communication is expected to be increasingly important (at least before self-driving vehicles hit the road) and, therefore, the development of some guidelines seems advisable.

Finally, a good travel time information system must have appropriate media and technology interfaces to cover all possible demands, namely (i) different types of travel time information, (ii) a changing (often high) amount of requirements and (iii) queries performed through all possible communication channels. The available media platforms evolve in line with technological progress, but they are sometimes constrained by limited budgets. Some of these platforms are only useful to disseminate pre-trip information. This is the case of press or TV information, which were the traditional channels of communication in the past. Although their role is nowadays less important, they still result useful to warn in advance against traffic disturbances linked to a special event (e.g. a concert, a football match, construction sites). Information through websites, although theoretically feasible for en-route information dissemination via smartphones, is normally used before starting a route too. Since the first websites for traffic information appeared in the mid-1990s, outstanding improvements have been made. Representative examples are the public website of the Department of Transportation of California (Caltrans Performance Measurement System, http://pems.dot.ca.gov/) or that of the private company Inrix (http://inrix.com/web-portals/), where the most important information is only fee-based available. However, the administrators of this kind of websites have usually developed

a smartphone app with the same or even more targeted information too. For their part, DMSs are only used to deliver information to those people already travelling. Therefore, they are considered as especially directed to the interested users, as they are linked to the routes these users follow. However, several studies state that they are not good platforms to deliver travel time reliability information, as drivers could confuse it with real-time information (Kuhn et al. 2017).

The remaining travel time dissemination technologies are useful before and during a trip. Information points, traffic call centres, radio broadcasts, car navigators and smartphones belong to this group. The last ones are those more successful nowadays. Information points have the problem of their low accessibility. Additionally, they oblige travellers to stop and, if necessary, to wait to be served. Their main advantage is that information can be very complete and specific. Another on-demand service able to provide extensive and particular information are call centres. However, they are usually paid services and, additionally, some waiting time is often required. Therefore, they are mainly used in the event of an accident, and not to obtain travel time information. Radio broadcasts have similarities with press or TV reports: information times are discrete and information is not user-specific. Furthermore, signal losses while driving make them unreliable to deliver important information, at least as a single source. For their part, car-navigators provide continuous and immediately accessible information. Besides, they can be manipulated by drivers to deliver the specific information they want with their favourite interface. However, smartphones overtake them today as the most preferred source of information. First, most travellers have a smartphone for their daily communications and routines, while car-navigators must be bought on purpose to support the driving task, at least for the majority (medium-class) of vehicle models. Second, many modes of information are available via smartphones: short message services (SMS), social media platforms (e.g. Twitter), specific traffic apps, etc., especially interesting for the dissemination of real-time information. Apps are undoubtedly on the top. They are normally user-friendly and include user input and output screens and data entry mechanisms, such as drop-down text boxes and scrolling menus, specifically designed for the touchscreen or keyboard supported by the operating system (Kuhn et al. 2017). Although a fee is required to use the best ones (e.g. Inrix app, Fig. 2.12), very satisfactory options can be found free of charge. As said, being aware of their usefulness, several traffic agencies have developed their own public mobile applications.

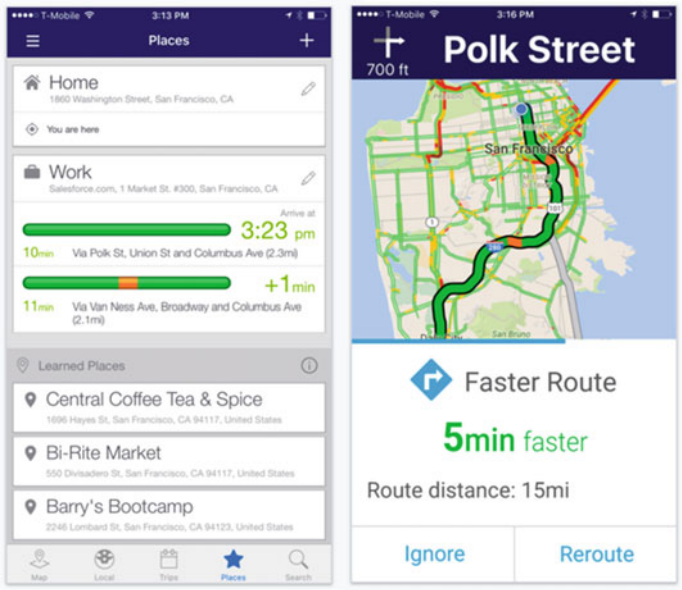

Fig. 2.12 Example of use of the Inrix app

References

Abbott-jard M, Shah H, Bhaskar A (2013) Empirical evaluation of bluetooth and wifi scanning for road transport. In: Proceedings of the 36th Australasian transport research forum, October, 1–14, Brisbane

Abdulhai B, Tabib SM (2003) Spatio-temporal inductance-pattern recognition for vehicle reidentification. Transp Res Part C: Emerg Technol 11(3–4):223–239

Abrantes PAL, Wardman MR (2011) Meta-analysis of UK values of travel time: an update. Transp Res Part a: Policy Pract 45(1):1–17

Barceló J, Montero L, Marquès L (2010) Travel time forecasting and dynamic OD estimation in freeways based on bluetooth traffic monitoring. In: Proceedings of the 89th annual meeting of the transportation research board, 10–14 January 2010, Washington D.C.

Bar-Gera H (2007) Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel. Transp Res Part C: Emerg Technol 15(6):380–391

Bhaskar A, Chung E (2013) Fundamental understanding on the use of bluetooth scanner as a complementary transport data. Transp Res Part C: Emerg Technol 37:42–72

Bloch I (1996) Information combination operators for data fusion: comparative review with classification. IEEE Trans Syst, Man Cybern-Part A 26(1):52–67

Buisson C (2006) Simple traffic model for a simple problem: sizing travel time measurement devices. Transp Res Record: J Transp Res Board 1965:210–218

Cassidy MJ, Anani SB, Haigwood JM (2002) Study of freeway traffic near an off-ramp. Transp Res Part A: Policy Pract 36(6):563–572

Chen H, Rakha HA (2016) Multi-step prediction of experienced travel times using agent-based modelling. Transp Res Part C: Emerg Technol 71:108–121

Cheu RL, Xie C, Lee D-H (2002) Probe vehicle population and sample size for arterial speed estimation. Comput-Aided Civ Infrastruct Eng 17(1):53–60

Coifman B (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transp Res Part a: Policy Pract 36(4):351–364
Coifman B, Cassidy M (2002) Vehicle reidentification and travel time measurement on congested freeways. Transp Res Part a: Policy Pract 36(10):899–917
Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. ASCE J Transp Eng 129(5):475–483
Coifman B, Krishnamurthya S (2007) Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. Transp Res Part C: Emerg Technol 15(3):135–153
Cortés CE, Lavanya R, Oh J, Jayakrishnan R (2002) General-purpose methodology for estimating link travel time with multiple-point detection of traffic. Transp Res Record: J Transp Res Board 1802:181–189
Daganzo CF (Unknown) Queuing of two conflicting traffic streams. Unpublished notes from CE150. UC Berkeley
Elefteriadou L, Cui X (2007) A framework for defining and estimating travel time reliability. In: Transportation research board 86th annual meeting, January 21–25 2007, Washington DC, paper #07–1675
Gentili M, Mirchandani PB (2018) Review of optimal sensor location models for travel time estimation. Transp Res Part C: Emerg Technol 90:74–96
Hensher DA (2011) Valuation of travel time savings. In: Palma AD, Lindsey R, Quinet E, Vickerman R (ed) A handbook of transport economics. Edward Elgar Publishing Limited, Cheltenham, UK
Herrera JC, Work D, Ban X, Herring R, Jacobson Q, Bayen A (2010) Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp Res Part C: Emerg Technol 18:568–583
Hoh B, Iwuchukwu T, Jacobson Q, Work D, Bayen AM, Herring R, Herrera JC, Gruteser M, Annavaram M, Ban J (2012) Enhancing privacy and accuracy in probe vehicle-based traffic monitoring via virtual trip lines. IEEE Trans Mob Comput 11(5):849–864
Jara-Díaz SR (2000) Allocation and valuation of travel-time savings. In: Hensher DA, Button KJ (ed) Handbook of transport modelling. Elsevier Science Ltd., Simi Valley, CA
Jiang G, Gang L, Cai Z (2006) Impact of probe vehicles sample size on link travel time estimation. Proceedings of the 2006 IEEE Intelligent Transportation Systems Conference, September 17–20 2006, Toronto, 505–509
Kuhn B, Pate A, Higgins L, Krile R, Vickich M (2017) Disseminating traveler information on travel time reliability. Report FHWA-HOP-16–067. Federal Highway Administration, Washington DC
Kwon E (2004) Development of operational strategies for travel time estimation and emergency evacuation on freeway network. Final Report MN/RC—2004–49. Prepared for the Minnesota Department of Transportation
Kwon TM (2006) Blind deconvolution processing of loop inductance signals for vehicle reidentification. In: Proceedings of the 85th annual meeting of the transportation research board, January 22–26 2006, Washington D.C.
Lawson TW, Lovell DJ, Daganzo CF (1997) Using the input–output diagram to determine the spatial and temporal extents of a queue upstream of a bottleneck. Transp Res Rec 1572:140–147
Li X, Ouyang Y (2011) Reliable sensor deployment for network traffic surveillance. Transp Res Part b: Methodol 45(1):218–231
Li X, Ouyang Y (2012) Reliable traffic sensor deployment under probabilistic disruptions and generalized surveillance effectiveness measures. Operations Res 60(5):1183–1198
Lim S, Lee C (2011) Data fusion algorithm improves travel time predictions. IET Intel Transport Syst 5(4):302–309
Longfoot J (1991) An automatic network travel time system-ANTSS. In: Proceedings of the 1991 vehicle navigation and information systems conference, 2, 20–23 October 1991, Dearborn, Michigan, 1053–1061
Lucas DE, Mirchandani PB, Verma N (2004) Online travel time estimation without vehicle identification. Transp Res Record: J Transp Res Board 1867:193–201

Mackie P, Wardman M, Fowkes A, Whelan G, Nellthorp J, Bates J (2003) Values of travel time savings in the UK. Full Report

Mori U, Mendiburu A, Álvarez M, Lozano JA (2015) A review of travel time estimation and forecasting for advanced traveller information systems. Transportmetrica a: Transport Science 11(2):119–157

Ndoye M, Totten VF, Krogmeier JV, Bullock DM (2011) Sensing and signal processing for vehicle reidentification and travel time estimation. IEEE Trans Intell Transp Syst 12(1):119–131

Newell GF (1993a) A simplified theory of kinematic waves in highway traffic. Part I: General Theory. Transp Res Part B: Methodol 27(4):281–287

Newell GF (1993b) A simplified theory of kinematic waves in highway traffic. Part II: queuing at freeway bottlenecks. Transp Res Part B: Methodol 27(4):289–303

Newell GF (1993c) A simplified theory of kinematic waves in highway traffic. Part III: multi-destination flows. Transp Res Part B: Methodol 27(4):305–313

Newell GF (1999) Delays caused by a queue at a freeway exit ramp. Transp Res Part b: Methodol 33(5):337–350

Nilsson N (1980) Principles of artificial intelligence. Tioga Publishing, Palo Alto

Nishiuchi H, Nakamura K, Bajwa S, Chung E, Kuwahara M (2006) Evaluation of travel time and OD variation on the Tokyo Metropolitan Expressway using ETC data. In: Research into practice: 22nd ARRB conference proceedings information, 29.10–2.11 2006, Australian Road Research Board, Canberra, Canada

Petty KF, Bickel P, Ostland M, Rice J, Schoenberg F, Jiang J, Rotov Y (1998) Accurate estimation of travel time from single loop detectors. Transp Res Part a: Policy and Pract 32(1):1–17

Pirc J, Turk G, Žura M (2016) Highway travel time estimation using multiple data sources. IET Intel Transport Syst 10(10):649–657

Quiroga CA, Bullock D (1998) Determination of sample sizes for travel time studies. ITE J 68(8):92–98

Rakha H, Zhang W (2005) Estimating traffic stream space-mean speed and reliability from dual and single loop detectors. Transp Res Record: J De Transp Res Board 1925:38–47

Rizzi LI, Limonado JP, Steimetz SSC (2012) The impact of traffic images on travel time valuation in stated-preference choice experiments. Transportmetrica 8(6):427–442

Ryus P, Bonneson J, Dowling R, Zegeer J, Vandehey M, Kittelson W, Rouphail N, Schroeder B, Hajbabaie A, Aghdashi B, Chase T, Sajjadi S, Margiotta R (2013) Proposed chapters for incorporating travel time reliability into the highway capacity manual, SHRP 2 reliability project L08. Transportation Research Board of the National Academies, Washington, D.C.

Shao H, Lam WHK, Sumalee A, Chen A (2018) Network-wide on-line travel time estimation with inconsistent data from multiple sensor systems under network uncertainty. Transportmetrica a: Transport Sci 14(1–2):110–129

Sherali HD, Desai J, Rakha H (2006) A discrete optimization approach for locating automatic vehicle identification readers for the provision of roadway travel times. Transp Res Part b: Methodol 40(10):857–871

Shires JD, Jong GC (2009) An international meta-analysis of values of travel time. Savings Eval Program Plan 32(4):315–325

Shuo Li PE, Zhu K, van Gelder BHW, Nagle J, Tuttle C (2002) Reconsideration of sample size requirements for field traffic data collection using GPS devices. Transp Res Record: J Transp Res Board 1804(1):17–22

Soriguera F, Robusté F (2011a) Highway travel time accurate measurement and short-term prediction using multiple data sources. Transportmetrica 7(1):85–109

Soriguera F, Robusté F (2011b) Requiem for freeway travel time estimation methods based on blind speed interpolations between point measurements. IEEE Trans Intell Transp Syst 12(1):291–297

Soriguera F, Rosas D, Robusté F (2010) Travel time measurement in closed toll highways. Transp Res Part b: Methodol 44(10):1242–1267

Sun L, Yang J, Mahmassani H (2008) Travel time estimation based on piecewise truncated quadratic speed trajectory. Transp Res Part a: Policy Pract 42(1):173–186

Tak S, Kim S, Oh S, Yeo H (2016) Development of a data-driven framework for real-time travel time prediction. Comput-Aided Civ Infrastruct Eng 31:777–793

Treiber M, Helbing D (2002) Reconstructing the spatio-temporal traffic dynamics from stationary detector data. Cooper@tive Tr@nsport@tion Dyn@mics, 1, 3.1–3.24

Turner SM, Eisele WL, Benz RJ, Holdener DJ (1998) Travel time data collection handbook. Research Report FHWA-PL-98–035. Federal Highway Administration, Office of Highway Information Management, Washington, D.C.

Unde MD, Borkar B (2014) Remote vehicle tracking and driver health monitoring system using GSM modem and google maps. (IJCSIT) Int J Comput Sci Inf Technol 5(3):2828–2832

van Hinsbergen CPIJ, van Lint JWC, Sanders FM (2007) Short term traffic prediction models. In: Proceedings of the 14th world congress on intelligent transport systems (ITS), 2007, Beijing, 1–18

van Hinsbergen CPIJ, van Lint JWC, van Zuylen HJ (2009) Bayesian committee of neural networks to predict travel times with confidence intervals. Transp Res Part C: Emerg Technol 17(5):498–509

van Lint J, van der Zijpp N (2003) Improving a travel-time estimation algorithm by using dual loop detectors. Transp Res Record: J Transp Res Board 1855:41–48

van Lint JWC (2004) Reliable travel time prediction for freeways. PhD Dissertation. Delft University of Technology, The Netherlands

Vanajakshi LD (2004) Estimation and prediction of travel time from loop detector data for intelligent transportation systems applications. PhD dissertation. Texas A&M University

Wang D, Fu F, Luo X, Jin S, Ma D (2016) Travel time estimation method for urban road based on traffic stream directions. Transportmetrica a: Transp Sci 12(6):479–503

Wardrop J (1952) Some theoretical aspects of road traffic research. Proc Inst Civ Eng 1(2):325–378

Xing T, Zhou X, Taylor J (2013) Designing heterogeneous sensor networks for estimating and predicting path travel time dynamics: an information-theoretic modeling approach. Transp Res Part b: Methodol 57:66–90

Yildirimoglu M, Geroliminis N (2013) Experienced travel time prediction for congested freeways. Transp Res Part b: Methodol 53:45–63

Yim Y (2003) The state of cellular probes. Technical Report. July. Berkeley: Institute of Transportation Studies, University of California

Yim Y, Cayford R (2006) Field operational test using anonymous cell phone tracking for generating traffic information. In: Proceedings of the 85th annual meeting of the transportation research board, January 22–26 2006, Washington D.C.

Zhang Y, Ge H (2013) Freeway travel time prediction using Takagi–Sugeno–Kang fuzzy neural network. Comput-Aided Civ Infrastruct Eng 28(8):594–603

Zhu T, Kong X, Lv W (2009) Large-scale travel time prediction for urban arterial roads based on kalman filter. In: Proceedings of the IEEE international conference on computational intelligence and software engineering, December 11–13 2009, Wuhan, China, pp 1–5

Part II
New Travel Time Estimation Methods

Chapter 3
A Simple Algorithm for the Estimation of Road Traffic Space Mean Speeds from Data Available to Most Management Centers

Margarita Martínez-Díaz

Abstract In traffic engineering, a lot of valuable information is obtained after appropriate processing of data collected by certain sensors. However good the data may be, the information extracted can be completely wrong if the processing is inadequate. One of the most common simplifications in the field, which, for example, affects some travel time estimation methodologies, is the use of temporal average speeds as equivalent to spatial averages. This chapter explains the causes of this bad practice, which is linked to the most traditional (and most extended) road equipment and procedures. To correct this trend, a highly applicable solution in the form of an algorithm is proposed. Although the results of the algorithm are not fully robust, they are favorable in a wide variety of cases, with the added bonus that no additional investment would be required.

3.1 Introduction

As society progresses, new requirements and needs may appear. With regard to road transport, researchers, administrations, and private companies are aware that controlling the evolution of traffic results in an increase of productivity and safety allows exploiting synergies among different means of transport and contributes to a more sustainable growth (SHRP 2 2013). Many different initiatives such as real-time calculation of travel times, active traffic management or automated driving emerge as examples of key achievements.

Although these lines of research are very different, they have two commonalities, namely the need for (i) appropriate data and (ii) well-founded calculations. The

M. Martínez-Díaz (✉)
UPC-BarcelonaTech, Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, Barcelona, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_3

development of new technologies and computer software offers the possibility of collecting varied data and combining them to obtain accurate results (Yuan et al. 2014). Mobile phones, GPS (Global Positioning System), Bluetooth, Optical Character Recognition (OCR) cameras, and many other devices are sources of traffic data usable for calculations. As seen, GPS-enabled cell phones, RFID technologies, etc. have opened a new way of collecting traffic data, as they are able to register individual vehicle trajectories (Hiribarren and Herrera 2014). Furthermore, vehicles themselves will also act as high-level "sensors" in future cooperative scenarios. With these lagrangian sensors moving within the traffic stream, increasing amounts of data will be available. Therefore, it will be possible to design much more precise methodologies, either for real-time travel time estimation or for any other purpose aimed at the dynamic management of traffic.

However, currently, neither totally accurate data nor the most complex programs are usually available, at least in a sufficient amount, in less trafficked areas. This is the case for example on secondary roads, in rural areas or for small traffic management centers. In fact, the majority of these centers in developed countries depend on equipment such as loop detectors and common cameras (unable to identify vehicles). That is, loops are the main sources of data. Traffic researchers have demonstrated the advisability of deploying double loops (in pairs in each section of each lane) rather than single loops to obtain more data and thus better results in later calculations (Chen et al. 2003). Fortunately, at present this trend is usually fulfilled. Moreover, the fact that there is a single data source on any road is expected to gradually disappear. Anyway, until today's scenarios evolve, some modifications can be performed in the procedures currently implemented in traffic centers so that they better manage traffic. In the case of this chapter, travel time estimation by means of spot speed methods will be improved only with the application of traffic flow theory, and maintaining loops as the unique data source. First, a remainder of the basics of these detectors is included next.

All inductive loop detectors are similar. They consist of a wire loop installed under the pavement of a lane, which is able to detect the presence of a vehicle (in essence a metallic object) because of the change that it causes in the electromagnetic properties of the loop. The main differences among loops are related to the software that manages and stores these data, which can be programed in several ways. As explained in Chap. 1, the data usually available in previously determined time intervals of aggregation, Δt, with the double-loop configuration are follows:

- Number of vehicles that pass over the detectors.
- Lengths of these vehicles: the software that manages the information usually classifies them into groups and keeps only the number of vehicles in each group. For example, in Spain the usual classification is as groups of vehicles shorter than 6 m, between 6 and 10 m and larger than 10 m.
- Spot speed measurements: again, although at first individual spot speeds are detected, the software calculates and registers only their mean, i.e., the time mean speed, $\overline{v}_t$, the average speed of all vehicles passing over a particular spot.

- Number of vehicles that pass over the detectors with a speed lower than a particular reference speed. It is common to have two different references. Only the number of vehicles that meet this requirement is stored. It must be highlighted that the chance of obtaining these data directly from the software of loop detectors is not a standard in the USA, but it is quite common in Europe. As an example, all Spanish freeway traffic centers manage them.

The duration of the time intervals of aggregation ranges from 20–30 s in the USA up to 15 min in some European countries. Intervals between 3–5 min have proven to be the most suitable (Soriguera and Robusté 2013): both shorter and longer durations have some advantages but also some disadvantages, as it will be discussed in Sect. 3.5.

Variation of traffic speeds at various places over time turns out to be one of the basic inputs for subsequent studies, for example, the indirect estimation of travel times. However, the problem is that most studies are based on the fundamental equation of traffic flow (Eq. 3.1, introduced as Eq. 1.8 in Chap. 1); it provides the relationship between flow, q, and density, k, by means of a specific type of speed, the so-called space mean speed, $\overline{v_s}$, which is really a harmonic mean calculated under particular conditions (Wardrop 1952). Further explanation about this point is included in Sect. 3.2.

$$q = \overline{v_s} * k \tag{3.1}$$

The use of data provided by loop detectors involves various difficulties when determining the evolution of speeds:

- Individual speeds are measured at fixed points of a road and must be extrapolated to some extent to achieve the spatial implication needed. This spatial generalization is extremely complicated, particularly in case of congestion.
- As mentioned, the software delivers time mean speeds. The use of these time means as substitutes of the space means required for calculations can cause a considerable loss of accuracy in the final results.
- Although loops are simpler, more economical and more common than other devices used to collect traffic data, their utility depends on their density on the road (Bachmann et al. 2013). Some research has resulted in the development of simple search algorithms that efficiently select sensor locations in order to obtain suitable data when the number of available sensors is limited (Viti et al. 2014). Nevertheless, difficulties remain on those roads already constructed.

The goal of the algorithm introduced in this chapter is to calculate spot space mean speeds exclusively from the data provided by double-loop detectors, avoiding extra expenses for the administrations. Specifically, it is focused on the calculation of the variance of the speeds with respect to the time mean, which allows using the relationship between time mean speeds and space mean speeds in the event of stationarity defined by Rakha and Zhang (2005). As explained in Chap. 2, further improvements must be implemented to obtain more accuracy in the final objectives, in this case, in the travel time estimations. Once working with space mean speeds,

a procedure for the generalization of these speeds over the links between detectors based on traffic dynamics and queue evolution would be the next challenge to face. Anyway, improvements in this first basic input have, as it is next demonstrated, satisfactory consequences.

The remaining sections of this chapter are as follows. Section 3.2 gives the background of different traffic speed definitions and summarizes their relationships according to various researchers. Section 3.3 develops the proposed algorithm, whose implementation is demonstrated in Sects. 3.4 and 3.5 with artificial and real data, and also compared with other methodologies. After the discussion of the results, attempts to find new relationships between mean speeds are performed in Sect. 3.6. Finally, Sect. 3.7 includes the conclusions and a proposal for new lines of research.

3.2 Background

Since 1952, when Wardrop (1952) stated his two principles concerning the idea of traffic equilibrium previously developed by Knight (1935), the differences between the time mean speed and the space mean speed have been widely demonstrated. The space mean speed, $\overline{v_s}$, is the average speed of all vehicles in a particular stretch of a road at a specific instant (Homburger et al. 1996). The time mean speed, $\overline{v_t}$, is the average of the speeds of all vehicles that pass over a section of a road during a certain time interval. It is easy to deduce that the time mean speed is greater than the space mean speed (Daganzo 1997) because vehicles that are faster contribute more to the time-mean than the slow ones. On the contrary, vehicles of all speeds contribute equally to the space-mean. Space averages equal time averages only in case of space–time homogeneous traffic (Breiman 1969).

As it has been explained before, loops on a road detect and average spot speeds in stipulated time intervals, thus providing time mean speeds. However, if the individual spot speeds were stored, $\overline{v_s}$ could be calculated by giving them certain spatial nature and by considering stationary traffic in the section (Edie 1965) as Eq. 3.2 shows:

$$\overline{v_s} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} tt_i} = \frac{n * dx}{\sum_{i=1}^{n} \frac{dx}{v_i}} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} \frac{1}{v_i}}, \tag{3.2}$$

where,

x_i = distance covered by vehicle *i*,

tt_i = time used by vehicle i to cover the distance x_i,

v_i = spot speed of vehicle *i*,

n = number of vehicles that pass over the detector during the time interval,

dx = differential length taken up by the detector.

Therefore, in these conditions the space mean speed could be calculated as the harmonic mean of the individual spot speeds. It must be highlighted, however, that in the origin of this formulation neither a time mean nor a space mean was established,

but a generalized definition of the average speed. The fact of labelling this generalized definition of the average speed as space mean speed $\overline{v_s}$ is an abuse of notation. Actually, $\overline{v_s}$ does not share the spatial implications of the original space mean speed definition unless traffic is stationary. Some limitations have been imposed for that reason, considering that this identification is only performed when the average speed is computed over a narrow rectangular strip in the $x-t$ plane with a spatial width dx and a time length T, which corresponds to the measurement region of a loop detector on a highway. Taking this definition into account, the space mean speed appears for example in the mathematical formulation of the average travel time $\overline{tt}$ of n vehicles that cover a specific distance of a road L at a constant speed v_i (Eq. 3.3, already introduced in Chap. 2 as Eq. 2.1):

$$\overline{tt} = \frac{\sum_{i=1}^{n} tt_i}{n} = \frac{\sum_{i=1}^{n} \frac{L}{v_i}}{n} = L * \frac{1}{n} \sum_{i=1}^{n} \frac{1}{v_i} = \frac{L}{\overline{v_s}}. \tag{3.3}$$

In consequence, travel times would be underestimated if $\overline{v_t}$ were used instead of $\overline{v_s}$ (Soriguera and Robusté 2011). This substitution could lead to other inaccuracies such as wrong estimates of jam densities or shock wave speeds (Knoop et al. 2009). The data aggregation process is in fact an influential source of noise and errors present in conventional measures of the traffic state (Coifman 2014). Many authors have stated the importance of correctly using time-based or space-based data, no matter their source. For example, the inverse of the harmonic mean of instantaneous speeds from probe vehicles is an unbiased and consistent estimator of the mean segment travel time when sampling by space, whereas it is biased upward when sampling by time (Jenelius et al. 2015).

Clearly, upgrades in the loop software would allow these devices to store individual data or even to directly calculate space mean speeds. However, the large number of loops deployed worldwide and human inertia have so far precluded those modifications. Therefore, many researchers have tried to calculate space mean speeds from the time mean speeds provided by the loops, especially in case of stationarity, which is the common hypothesis of all the following methodologies.

The first of these relationships, shown in Eq. 3.4, is due to Wardrop (1952):

$$\overline{v_t} = \overline{v_s} + \frac{\sigma_s^2}{\overline{v_s}}, \tag{3.4}$$

where σ_s^2 is the variance of the speed with regard to the space mean for the specific time interval of aggregation chosen. The accuracy of the formula has been experimentally verified, but most traffic management centers cannot use it because individual speeds are needed in order to calculate the variance with regard to the space mean. This formula was actually devised to calculate time means from space means, what is not usually necessary in real life.

Another formula postulated to relate both means is that of Garber (2002) shown in Eq. 3.5:

$$\overline{v_t} = 0.966 * \overline{v_s} + 3.541. \tag{3.5}$$

The main problem of this relationship is that it was established based only on experimental data; thus, it cannot be extrapolated to many situations in which the boundary conditions differ from the original ones. It must be continuously calibrated and, ultimately, it is not worth using.

Equation 3.6 has been used in several traffic studies. It was first derived by Khisty (2003), but they were Rakha and Zhang (2005) who proved it analytically:

$$\overline{v_s} = \overline{v_t} - \frac{\sigma_t^2}{\overline{v_t}}. \tag{3.6}$$

In this equation σ_t^2 is the variance of the speed with regard to the time mean for the specific time interval of aggregation. However, the impossibility of calculating the variance arises again. Nevertheless, and taking into account the utility of the formula, Soriguera and Robusté (2011) were able to estimate this variance by imposing the common hypothesis of stationary traffic in each time interval of aggregation and additionally assuming normality of the speed distribution. Then, the variance with regard to the time mean speed is given by Eq. 3.7:

$$\sigma_t = \frac{v^a - \overline{v_t}}{F^{-1}\left[\frac{n_v{}^a}{n}\right]}, \tag{3.7}$$

where.

σ_t = standard deviation of the speed with regard to the time mean,

v^a = value of the speed chosen by traffic management centres,

F^{-1} = inverse of the cumulative standard normal distribution,

n_v^a= number of vehicles that pass over the detectors with a speed lower than v^a in each time interval of aggregation,

n = number of vehicles that pass over the detectors in each time interval of aggregation.

Although this methodology performs well in specific conditions, Soriguera and Robusté (2011) warned that it is inappropriate to use it indiscriminately, especially in cases of shock wave onsets or offsets or with "stop and go" situations. As Cassidy (1998) declared, stationarity ensures some otherwise senseless relationships. However, the relationship established by Rakha and Zhang (2005) has been proven useful under certain conditions even with non-spot data such as those from GPS (Poomrittigul et al. 2008).

Another fact that must be taken into account to establish relationships between speeds is that they more or less fit common statistical distributions. The normal, log-normal, gamma and bimodal distributions appear in the majority of the traffic studies. The normal distribution is undoubtedly the most used because of its simplicity, and it performs well when traffic conditions are homogeneous. Consequently, it is also the common assumption of multivariate normal distributions for link travel times

(Jenelius et al. 2013). However, the log-normal and gamma distributions are usually more suitable because they have additional advantages (Haight 1962):

- They avoid the appearance of negative speeds.
- They keep their shape if either time speeds or space speeds are fitted.

In the case of the log-normal distribution, another important advantage is the fact that the distribution of travel times based on speeds that fit this distribution maintains the same shape (El Faouzi et al. 2007). If the log-normal speed distribution has a mean μ and a standard deviation σ, the distribution of travel times will follow Eq. 3.8:

$$f_t(t) = \frac{1}{\sqrt{2 * \pi} * \sigma * t} * e^{\left[-\frac{(Lnt + \mu)^2}{2 * \sigma^2} \right]}. \tag{3.8}$$

In the cases where traffic is too heterogeneous (for example, because there are many different vehicle types that may behave differently or because phases of free flow follow congestion periods), unimodal distributions should be avoided (Dey et al. 2006). Bimodal or even multimodal distributions might be used. Each of their components would often be a normal or log-normal distribution (May 1990).

Many other complex distributions have been used in research, but their complexity prevents them from being put into practice (Zou and Zhang 2012). Even for log-normal distributions, some improvements can be expected if the distributions are truncated because only a range of speeds makes sense. In addition, the variances of these truncated distributions are always smaller than those of the original ones (Wang 2012).

3.3 Simple Algorithm for the Estimation of Space Mean Speeds from the Data Provided by Double-Loop Detectors

Having analyzed previous investigations and taking into account the data available, the author decides to use the equation of Rakha and Zhang (2005) to solve the problem of not having an explicit value of the variance. The motivation is that the validity of this formula has been widely demonstrated in experimental studies. However, a particular analysis has been performed in order to compare it with other possible relationships. Section 3.6 contains the results of this comparison, which effectively verifies the goodness of this formula against the others.

To be able to estimate the variance, two important hypotheses are assumed. In each time interval of aggregation T:

- Traffic is stationary.
- The speed distribution is log-normal.

The validity of these hypotheses will be discussed in Sect. 3.5.4. The first one has also been taken for granted in the other methodologies discussed in the chapter. With regard to the second, the author exploits the advantages of the log-normal distribution mentioned in Sect. 3.2. Assuming that the distribution of individual speeds v_i in each time interval of aggregation T is log-normal, the distribution of the logarithms of these speeds $x = Lnv$ is a normal distribution $N(\mu_x, \sigma_x)$. Therefore, the probability density function of the speeds, their mean and their variance are given by Eqs. 3.9 to 3.11, respectively,

$$f_v(v) = \frac{1}{\sqrt{2*\pi}*\sigma_x * v} * e^{\left[-\frac{(Lnv-\mu_x)^2}{2*{\sigma_x}^2}\right]} \quad with\ v > 0, \tag{3.9}$$

$$\mu_v = \overline{v_t} = e^{\mu_x + \frac{\sigma_x^2}{2}}, \tag{3.10}$$

$$\sigma_v^2 = \sigma_t^2 = \left(e^{\sigma_x^2} - 1\right) * e^{2*\mu_x + \sigma_x^2}, \tag{3.11}$$

where

v = individual speed,

μ_x = arithmetic mean of the logarithms of the speeds,

σ_x^2= variance of the logarithms of the speeds with regard to the mean.

Note that the goal of the algorithm is to estimate σ_v^2, which corresponds to the variance with regard to the time mean speed, termed σ_t^2 by Rakha and Zhang (2005). Therefore, μ_x and σ_x are needed. μ_v is supplied by the loops (the time mean speed, termed $\overline{v_t}$ by Rakha and Zhang (2005)).

Let n_v^a be the number of vehicles that pass over the detectors in a section with a speed lower than v^a in one time interval of aggregation T. The probability that a vehicle passes over the detector with such a speed is shown in Eq. 3.12:

$$P\left[V \le v^a\right] \approx \frac{n_v^a}{n} \approx P\left[e^X \le e^{x^a}\right] \approx P\left[Lne^X \le Lne^{x^a}\right] \approx P\left[X \le x^a\right] = F\left[Z\left(x^a\right)\right]$$
$$= F\left[Z\left(Lnv^a\right)\right] = F\left[\frac{Lnv^a - \mu_x}{\sigma_x}\right], \tag{3.12}$$

where

v^a = speed chosen as a reference,

n = number of vehicles that pass over the detectors in each time interval of aggregation,

x^a = logarithm of the speed v^a,

F = cumulative standard normal distribution,

(Z)= standardized value.

Rearranging Eqs. 3.10 and 3.12 yields a system with two equations (Eqs. 3.13 and 3.14) and two unknowns:

$$2\mu_x + \sigma_x^2 = Ln\overline{v_t}^2, \tag{3.13}$$

$$\mu_x + F^{-1}\left[\frac{n_v^a}{n}\right] * \sigma_x = Lnv^a, \tag{3.14}$$

where

F^{-1}= inverse of the cumulative standard normal distribution.

Finally, Eq. 3.15 is obtained

$$\sigma_x^2 - 2 * F^{-1}\left[\frac{n_v^a}{n}\right] * \sigma_x + Ln\left(\frac{v^a}{\overline{v_t}}\right)^2 = 0. \tag{3.15}$$

Solving Eq. 3.15, two possible values of σ_x arise. For two reference values of speed (v^{a1} and v^{a2}), four values are provided. In practice, some of these are nullified during the calculations because there are some mathematical limitations for the algorithm. In each time interval of aggregation T:

- n cannot be too small or the initial substitution of the theoretical probability by the accumulated frequency (Eq. 3.12) is problematic and the confidence interval of the estimations is too small.
- It is necessary that $n_v^a \neq 0$ and $n_v^a \neq n$. This keeps the inverse of the cumulated standard distribution from tending to infinite.
- $\left(F^{-1}\left[\frac{n_v^a}{n}\right]\right)^2$ must be greater than $Ln\left(\frac{v^a}{\overline{v_t}}\right)^2$ to avoid square roots of negative numbers when solving Eq. 3.15.
- It is necessary that $\frac{v^a}{\overline{v_t}} \neq 0$ to avoid natural logarithms of zero.

In those cases when more than one value of σ_x results, an action protocol must be established that helps to choose the most suitable. One possibility is to keep the value with the smallest confidence interval for a specific level of confidence. Once a value of σ_x is found and introduced into Eq. 3.13, the corresponding μ_x can be calculated. By using both values in Eq. 3.11, σ_t^2 is finally obtained and can be introduced into Eq. 3.6 to estimate $\overline{v_s}$. The flow chart in Fig. 3.1 summarizes the main steps of the algorithm.

As noted, in practice it is not easy to choose the best estimate of σ_t^2 from more than one possible value. There are no simple methods to calculate confidence intervals for the variance of log-normal distributions. Bayesian procedures seem to be the most suitable (Harvey et al., 2012), although quite difficult to implement.

A naïve solution could be to consider the confidence intervals of a parameter calculated in a previous step of the method, for example σ_x. If the best σ_x is chosen, the best σ_t^2 and thus a more accurate $\overline{v_s}$ will be obtained. Because the variable x is normally distributed, the solution for the confidence interval limits of σ_x proposed

by Soriguera and Robusté (2011) and developed in Eqs. 3.16 and 3.17 can be used:

$$\varepsilon_{\sigma_{x(1)}} = -\frac{(v^a - \mu_x) * \varepsilon_{z(1)}}{Z(Z + \varepsilon_{z(1)})}, \tag{3.16a}$$

$$\varepsilon_{\sigma_{x(2)}} = -\frac{(v^a - \mu_x) * \varepsilon_{z(2)}}{Z\left(Z + \varepsilon_{z(2)}\right)}, \tag{3.16b}$$

where

$$\varepsilon_{Z(1)} = F^{-1}\left(p + \varepsilon_p\right) - F^{-1}(p), \tag{3.17a}$$

$$\varepsilon_{Z(2)} = F^{-1}\left(p - \varepsilon_p\right) - F^{-1}(p). \tag{3.17b}$$

The variable p is the probability of a vehicle with a speed smaller than v^a passing over the detector in the time interval of aggregation. The circulation of vehicles over the detectors can be observed as a Bernouilli process; the possibilities are their driving slower than a reference speed or not, these trials being independent. Thus, the estimator of p, $\widehat{p}$, matches Eq. 3.18:

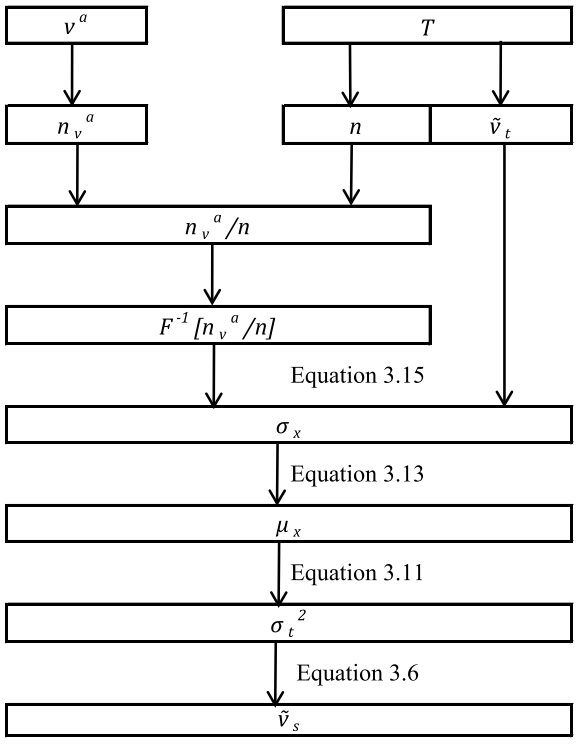

Fig. 3.1 Steps of the algorithm to obtain space mean speeds from loop detector data

$$\widehat{p} = \frac{n_v^a}{n}. \tag{3.18}$$

The proposed methodology relies heavily on the availability of n_v^a. If n_v^a is not reported to the traffic management center in the normal functioning of the system, the method cannot be applied. Obviously, carrying out modifications in the controllers in order to achieve these data lacks any sense, as it would be simpler, in this case, to introduce other modifications in order to directly obtain $\overline{v_s}$. Nevertheless, in those countries where n_v^a is available (a substantial number), the fact of using the estimated $\overline{v_s}$ instead of working with $\overline{v_t}$ (the current procedure) for later calculations would imply a higher level of accuracy without the need of any re-coding.

3.4 Implementation of the Algorithm with Artificial Data

To first verify the proper functioning of the algorithm, it was tested successfully with data generated with Matlab and readjusted to fulfil the main hypotheses of the method, i.e., the stationarity of the traffic and the log-normality of the speed distribution in each time interval of aggregation T as well as the mathematical requirements detailed in Sect. 3.3. For this last reason, the reference values were set at 101 km/h and 110 km/h (90 and 98% of the total time mean speed), ensuring enough vehicles participating in the calculations. The steps followed and the results are shown in Table 3.1, whereas Fig. 3.2 shows them in comparison with time means and real space mean speeds.

The estimated space mean speeds are much closer to the real space mean speeds than the time mean speeds that the loops provide. The error introduced by the latter is 2.17%, compared to 0.65% for the estimations of the algorithm. The validity of the algorithm has been therefore demonstrated in these ideal conditions.

The mean relative error was calculated taking into account absolute values of the differences. In addition, regarding the estimated space means, only values with differences smaller than the maximum difference incurred by the loops were admitted. This procedure was followed also in Sect. 3.5 with real data.

3.5 Implementation of the Algorithm with Real Data

The validity of the algorithm has been demonstrated in an ideal situation where all the initial conditions that were assumed when defining the method were met. However, it is also necessary to test it with different combinations of real data for which one or more of these conditions probably will not apply.

Table 3.1 Estimation of the space mean speeds and comparison of the results obtained with the data provided by the loops and with the real values

Time period T	Number of vehicles n	Classification of vehicles according to speeds			n_{va}/n		Time mean speed (Km/h)	Estimated space mean speed (Km/h)	Real space mean speed	Difference time mean - real space mean	Error (%)	Difference estimated space mean - real space mean	Error (%)
		V < 101	**101 < = V < 110**	**V > = 110**	**V^a= 101**	**V^a= 110**							
07:30:00	28	0	4	24.00	0.0000	0.1429	123.56	122.22	122.47	1.09	0.89	-0.25	0.20
07:45:00	25	3	9	13.00	0.1200	0.4800	111.10	110.41	110.30	0.79	0.72	0.11	0.10
08:00:00	33	11	8	14.00	0.3333	0.5758	106.51	105.23	105.62	0.89	0.84	-0.39	0.37
08:15:00	32	1	11	20.00	0.0313	0.3750	114.48	113.98	113.58	0.89	0.79	0.39	0.35
08:30:00	36	1	5	30.00	0.0278	0.1667	119.43	118.55	118.41	1.02	0.86	0.14	0.12
08:45:00	45	2	4	39.00	0.0444	0.1333	124.11	122.77	122.67	1.45	1.18	0.10	0.08
09:00:00	36	9	10	17.00	0.2500	0.5278	110.18	108.62	108.55	1.63	1.50	0.07	0.06
09:15:00	51	14	18	19.00	0.2745	0.6275	107.80	106.71	106.87	0.93	0.87	-0.16	0.15
09:30:00	43	7	14	22.00	0.1628	0.4884	111.38	110.37	110.05	1.32	1.20	0.32	0.29
09:45:00	39	11	9	19.00	0.2821	0.5128	108.55	107.13	106.93	1.62	1.52	0.21	0.19
10:00:00	31	8	6	17.00	0.2581	0.4516	112.33	109.87	108.34	3.99	3.68	1.53	1.41
10:15:00	22	10	2	10.00	0.4545	0.5455	104.23	101.48	100.36	3.87	3.86	1.11	1.11
10:30:00	32	18	3	11.00	0.5625	0.6563	103.71	99.80	95.59	8.12	8.49	4.21	4.40
10:45:00	29	4	11	14.00	0.1379	0.5172	110.39	109.70	109.61	0.78	0.71	0.09	0.08
11:00:00	16	5	1	10.00	0.3125	0.3750	114.11	112.97	111.29	2.82	2.54	1.69	1.52
11:15:00	24	2	6	16.00	0.0833	0.3333	114.12	113.28	112.84	1.28	1.14	0.44	0.39
11:30:00	26	8	3	15.00	0.3077	0.4231	111.88	108.45	108.36	3.52	3.25	0.10	0.09
11:45:00	29	7	6	16.00	0.2414	0.4483	116.54	109.99	110.89	5.65	5.09	-0.90	0.81
TOTAL VEHICLES	**577.0**	**121.0**	**130.0**							**MEAN ERROR**	**2.05**		**0.59**

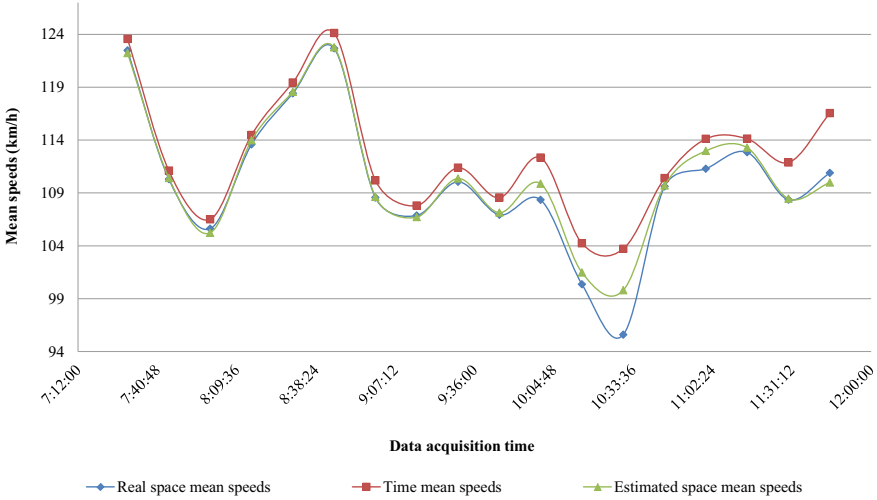

Fig. 3.2 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm from data that completely fulfil the initial conditions of the method

3.5.1 The Data

The data used for this study were collected during two days, on March 31th, 2014 and April 1st, 2014 in a section with double loops (P.K. 86 + 211, with two lanes in the direction toward A Coruña) of the AP-9 freeway, which runs north and south along the west coast of Galicia in Spain. The data were provided per lane and for aggregation time intervals T of 15 min. It must be noted that the fact that the data is a few years old has no special implication. In fact, the traffic control center in charge of this freeway still generates this type of information on a daily basis.

During the normal management of this freeway, the common data available were and are as follows:

- Number of vehicles that pass over the loops (n).
- Number of vehicles with lengths L shorter than 6 m, between 6 and 10 m or longer than 10 m.
- Time mean speeds $\overline{v_t}$: in an initial stage these speeds are averaged every 5 min, but then they are smoothed for time intervals of 15 min.
- Number of vehicles (n_V^a) that pass over the loops with speeds lower than 50 km/h (V^{a1}) and 100 km/h (V^{a2}), respectively.

Specifically for investigation purposes however, on this occasion the individual speeds and lengths were also provided, thus allowing an analysis of the algorithm with a wide range of different boundary conditions, as well as the comparison of the estimated space mean speeds with the real ones. The algorithm was executed with data obtained on different days, in different lanes (the left, for the fastest vehicles, and the right, for medium–low speed vehicles) and for all vehicles or only those whose lengths L were within a specified range. In addition, different time intervals

Table 3.2 Cases analyzed to test the algorithm

Case	Day	Lane	T(')	L	N	V^{al}	V^{a2}
I	31 March	Right	15	all	4,662	50	100
II	01 April	Right	15	all	2,841	50	100
III	01 April	Right	15	all	2,841	98	107
IV	01 April	Right	5	all	2,841	50	100
V	01 April	Right	5	L < 10 m	2,489	50	100
				L > = 10 m	352	50	100
VI	31 March	Left	15	all	769	50	100
VII	31 March	Left	15	all	769	110	120
VIII	01 April	Left	15	all	596	50	100
IX	01 April	Left	5	all	595	50	100
X	01 April	Left	5	all	595	50	115

Note that neither the stationarity of the traffic flow nor the log-normality of the speeds is guaranteed. This issue is discussed in Sect. 3.5.4

of aggregation (T, in minutes) and reference speeds (V^{a1} and V^{a2}) were used. N is the number of vehicles detected during the entire data acquisition period. Table 3.2 shows the cases that have been analyzed:

3.5.2 The Results

Table 3.3 shows the difference between using the time mean speeds provided by the loop detectors or the space mean speeds estimated with the algorithm as substitutes for real space mean speeds. This difference is shown as in Sect. 3.4, i.e., by determining the mean relative error in each case.

In 8 out of the 11 cases analyzed (and taking into account that case V has been subdivided) the algorithm implies an improvement, but there are 2 cases where the results have been worse and another in which no reasonable value has been obtained. This behavior was analyzed and understood; it is discussed in Sect. 3.5.4.

Note that in most cases it is not possible to determine the validity of the algorithm by focusing only on one of the boundary conditions; attention to the combination of all of them is required. Nevertheless, once all the conditions for the calculation have been established, its performance can be improved by changing only one of them. As an example, between cases VI (Fig. 3.3) and VII (Fig. 3.4) only the reference speeds are different. However, the algorithm only shows a good performance in the latter case. The reason underlying this fact is that, in case VI, the sample includes fewer vehicles because most of them were driving at speeds higher than 50 km/h. Another example is based on cases IV (Fig. 3.5) and V (Fig. 3.6). Segregating the sample according to the vehicle length improves the performance for light vehicles

Table 3.3 Comparison between the errors derived from the use of time means and those of the algorithm

Case	I	II	III	IV	Va	Vb	VI	VII	VIII	IX	X
Weighted error derived from the use of time means (%)	1.35	1.19	1.21	2.04	1.68	0.27	0.56	0.47	0.59	1.48	0.93
Weighted error of the algorithm (%)	0.79	0.87	0.99	0.59	0.46	–	0.86	0.44	0.78	0.58	0.50

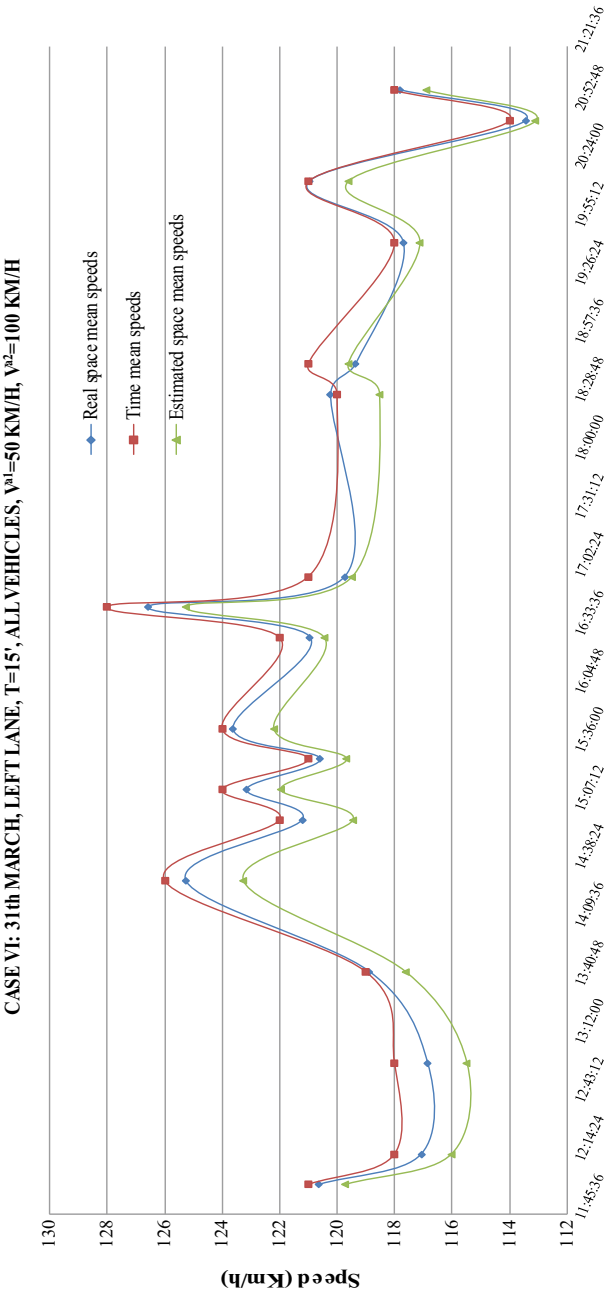

Fig. 3.3 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm in case VI

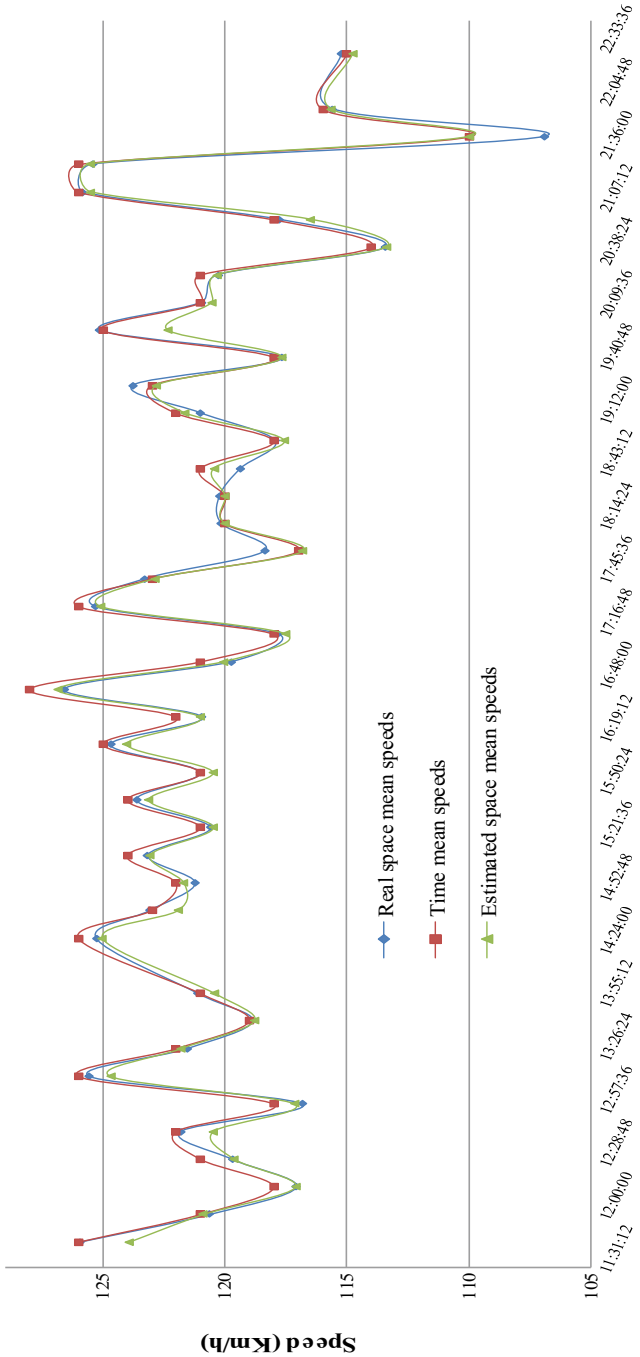

Fig. 3.4 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm in case VII

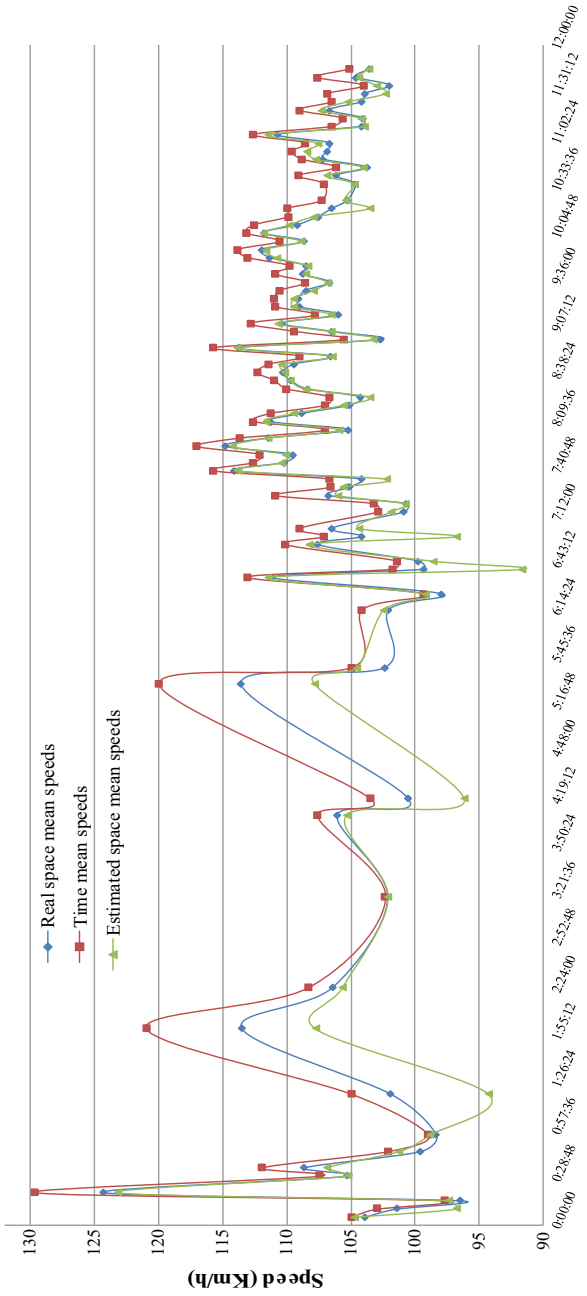

Fig. 3.5 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm in case IV

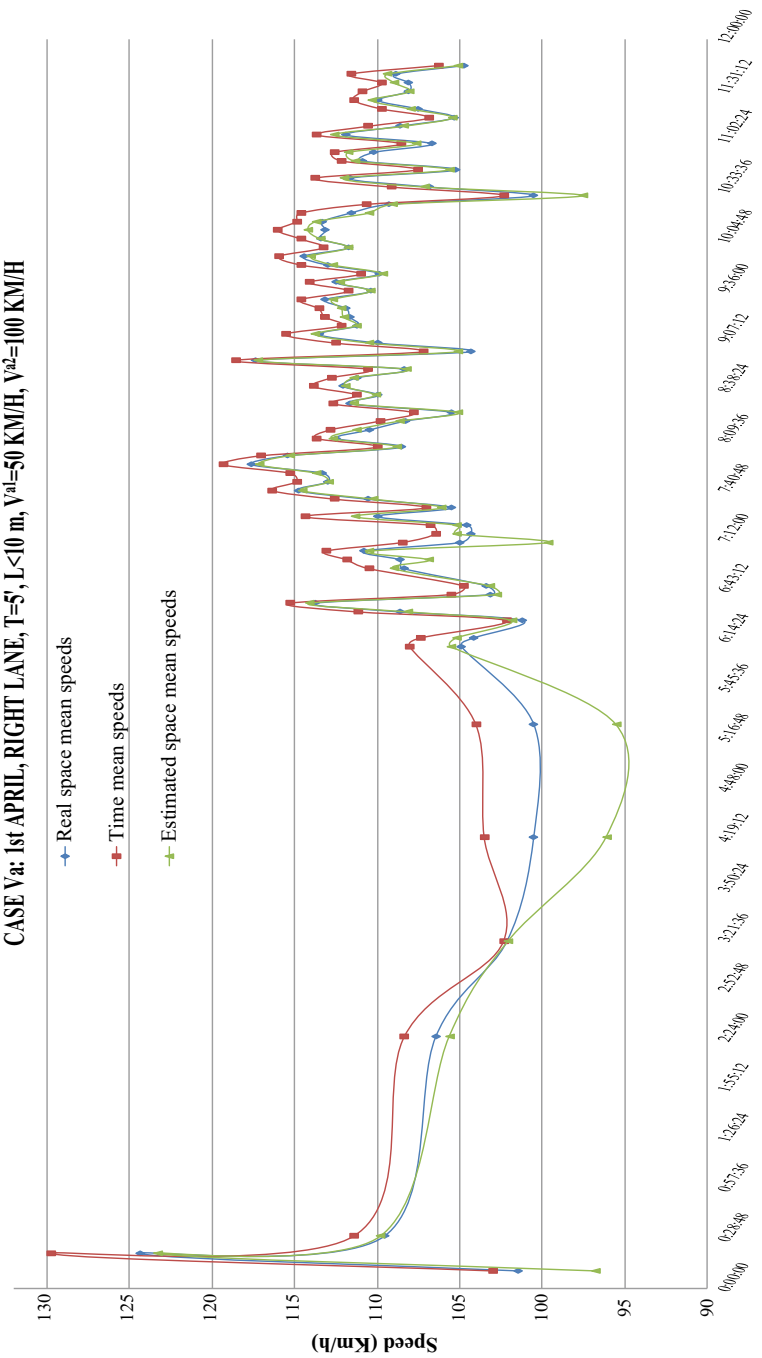

Fig. 3.6 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm in case Va

because the hypothesis of log-normality is better achieved. As for heavy vehicles, the algorithm in this specific example does not even run due to the small sample size of these vehicles. The influence of the length of the time interval of aggregation can be observed for example between cases II and IV (Figs. 3.7 and 3.5). The results of case IV, where $T = 5$ minutes, are much better.

3.5.3 Comparison Between the Proposed Algorithm and Other Methods

Because the proposed algorithm is somewhat more complicated than that introduced by Soriguera and Robusté (2011), a comparative analysis was performed to verify that it is worth using. In case I for example, the proposed algorithm demonstrated good behavior, diminishing the error incurred by the use of time mean speeds by 0.58%. Figure 3.8 and Table 3.4 Comparison of the errors introduced by different methodologies in case I. compare these results with that obtained with the methodology of Soriguera and Robusté (2011), which, as mentioned before, assume normality and stationarity in each time interval of aggregation T.

In spite of being conscious of the dependence of the formula of Garber on the boundary conditions, Table 3.4 also includes the results that would be obtained from its application, only for comparison purposes. The equation of Wardrop, as it has been previously stated, is clearly useful only to calculate $\overline{v_t}$ from $\overline{v_s}$, what is not necessary in practical uses.

3.5.4 Discussion

Given the accuracy of the estimates achieved in each case, some conclusions must be drawn. It seems that the algorithm is worth using in numerous situations because results are usually more accurate than the currently accepted time mean speeds. However, while it clearly performs better in some of these cases, it does not do so well in others. The analysis was carried out taking into account the following boundary conditions:

- Sample size.
- Log-normality of the speed distribution.
- Speeds chosen as references.
- Length of the time interval of aggregation.
- Prevailing type of vehicles.
- General traffic conditions.
- Place, day and moment of data acquisition.

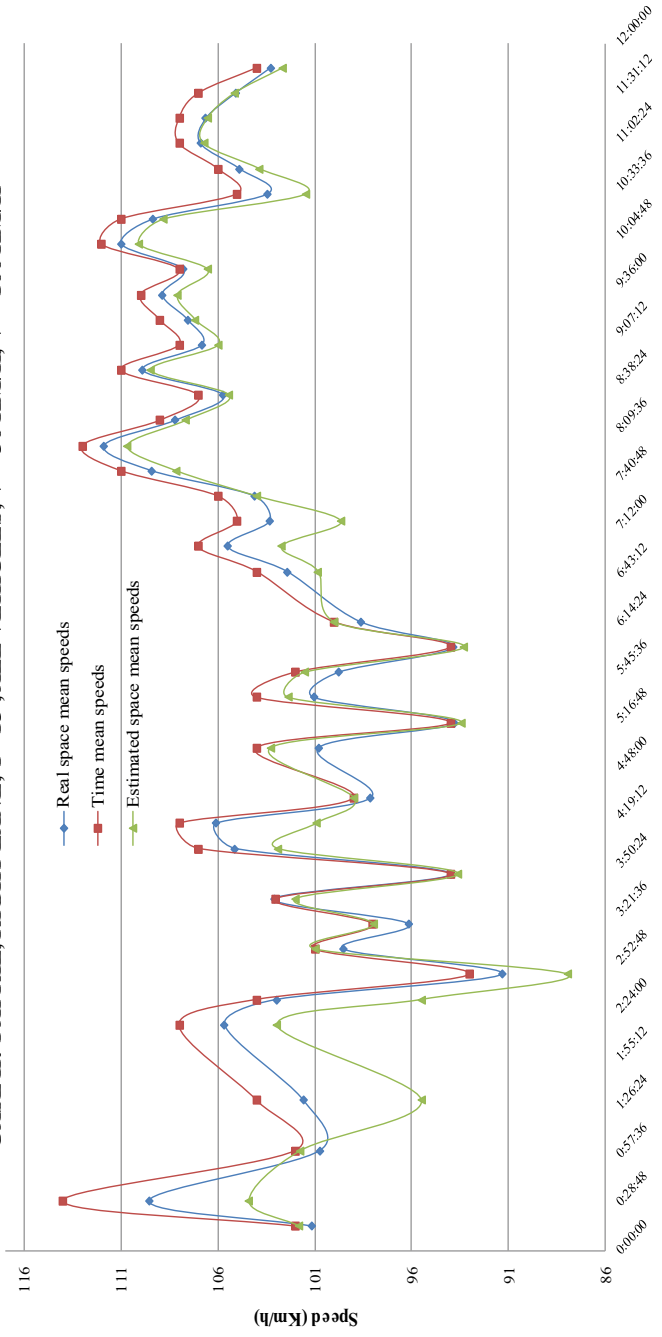

Fig. 3.7 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the algorithm in case II

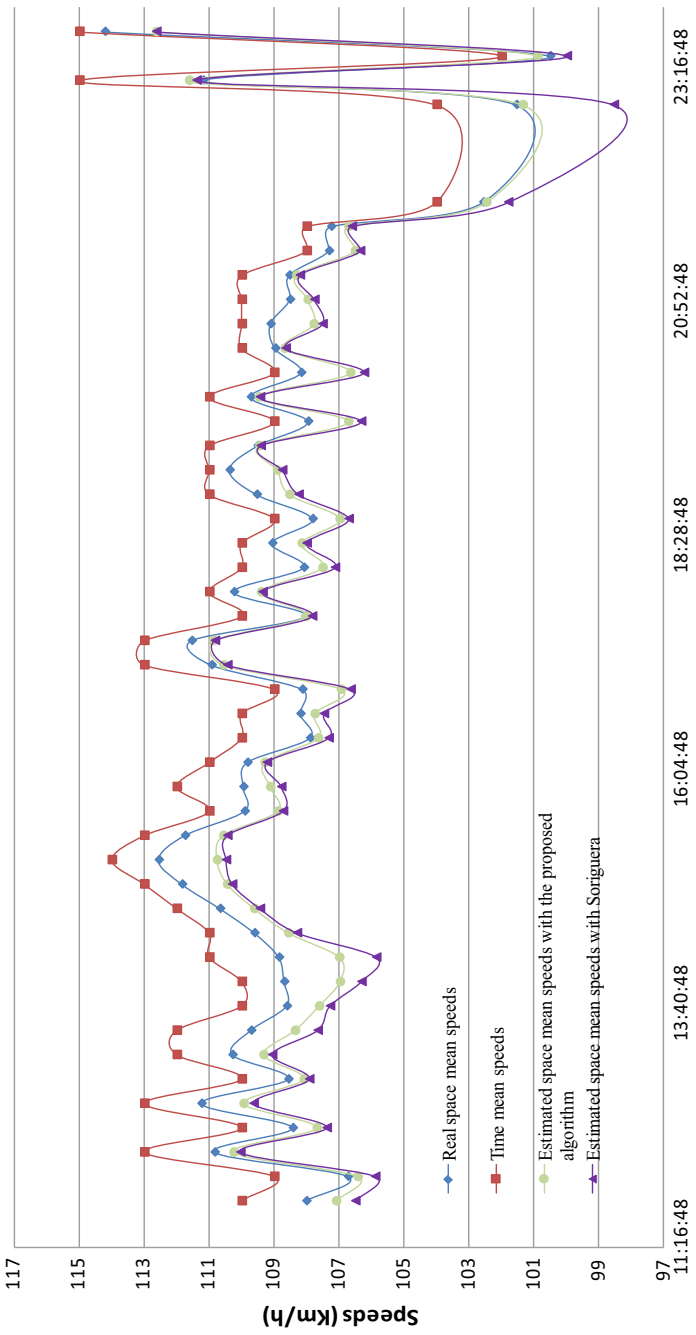

Fig. 3.8 Comparison of the real space mean speeds, the time mean speeds and the space mean speeds estimated with the proposed algorithm and the algorithm of Soriguera and Robusté (2011) in case I

Table 3.4 Comparison of the errors introduced by different methodologies in case I

Methodology	Vehicles suitable for calculations		Weighted mean error (%)
	Number	% of total vehicles	
Use of time mean speeds directly delivered by loop detectors	4,662	100	1.35
Use of the equation (Eq. 3.5) proposed by Garber (2002)	4,662	100	1.56
Use of the algorithm (Eq. 3.7 proposed by Soriguera (2011)	4,547	97.53	1.05
Use of the algorithm (Eq. 3.6 plus Eq. 3.15 and precedent) proposed in this paper	4,628	99.27	0.79

Regarding the sample size, the larger the sample, the better the algorithm performs. The main reasons are that the probability of having a log-normal distribution of speeds in each time interval of aggregation increases and because fewer mathematical inconsistencies appear during the calculations.

The log-normality of the speed distribution in each time interval of aggregation is one of the main hypotheses of the method and, therefore, it must be met. This can be more or less difficult depending on the conditions established for the calculations. For example, with low traffic densities, the behaviors of fast (e.g. cars) and slow (e.g. trucks, buses, vans) vehicles can be very different (Dey et al. 2006). If the estimation is made with samples from all lanes, bimodal or even multimodal distributions will probably appear. Therefore, the analysis must be made by lane (Soriguera and Robusté 2011). However, with high-medium densities, log-normality could appear even in the whole section because the faster vehicles will not be able to reach their usual speeds. As previously mentioned, log-normality is more suitable with large samples. To illustrate the importance of fulfilling this hypothesis, two time intervals of $T = 5$ minutes of case Va were chosen (time intervals between 7.40 and 7.45 a.m. and between 11.10 and 11.15 a.m.). The errors of estimation in these intervals were among the smallest (0.04% and 0.03%, respectively). The logarithms of the speeds were tested with the Kolgomorov-Smirnov (KS) Test. Table 3.5 shows the results, where the p-value in both cases was greater than 0.05, indicating normality of the logarithms and thus log-normality of the speeds. Figures 3.9 and 3.10 also roughly represent this trend.

The election of the speeds chosen as a reference must be made in a logical way with the only purpose of having a sufficient number of vehicles in the sample. In the specific case of the AP-9 freeway, the values used were 50 and 100 km/h. As it is

Table 3.5 KS test results for two time intervals with accurate estimates

Test KS	7:45	11:15
Z Kolmogorov-smimov	0.481	0.764
P value (bilateral)	0.975	0.604

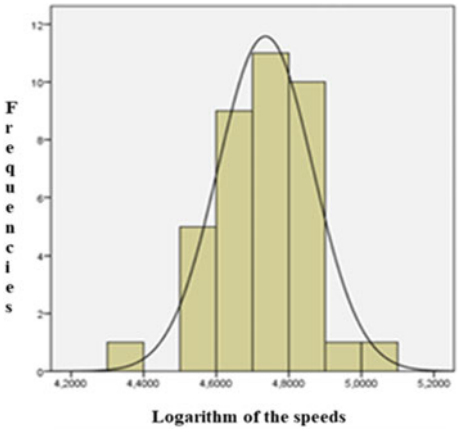

Fig. 3.9 Log-normal trend for time interval between 7.45 and 7.50 a.m

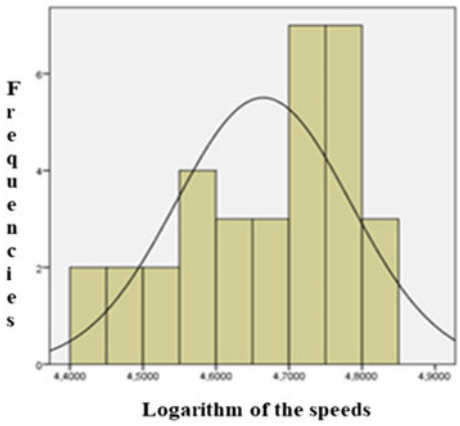

Fig. 3.10 Log-normal trend for time interval between 11.15 and 11.20 a.m

obviously uncommon for a vehicle to drive slower than 50 km/h on a freeway, some data will still be missed. Since the individual speeds were available, other values have been chosen for some of the analyses, what has led to better results. In this research, values of 90 and 98% of the average speed were chosen. In practice, these values could be based on (recent) historical data.

As for the lengths of the time intervals of aggregation, both long and short intervals show advantages and disadvantages. Short durations are more likely to comply with the other main hypothesis of the method, i.e., the stationarity of the traffic flow, and yield more accuracy in subsequent calculations in real time (for example, in travel time calculations). On the contrary, longer periods involve a greater sample size and a lower need for calculation capacity because a smaller number of iterations will be run each day.

Again, the prevailing type of vehicle is related to the convenience of making the estimations per lane or in a whole section to help to ensure the appearance of log-normal distributions. If possible, it is always advisable to work per lane and even to divide the vehicles into groups by their usual speeds, although this last step adds some extra effort. In case of working per lane, later estimates for the section can be obtained with equations such as Eq. 3.19, where the superscript i labels the lanes of the section (Soriguera and Robusté 2011):

$$\overline{v_s^{section}} = \frac{1}{\left[\frac{1}{\sum_i n^i}\right] * \sum_i \left(n^i / \overline{v}_s^i\right)}. \tag{3.19}$$

A preliminary analysis of the behavior of each type of vehicle should be done to avoid useless work. In this study, dividing the vehicles into the three sizes established by the Galician traffic management center generally provided the same results as classifying them into only two sizes (presumably the fast and slow ones), or even worse ones in some time intervals of aggregation lacking of vehicles of specific groups in the sample.

Note that the hypothesis of stationarity for the traffic flow has conditioned most of the steps followed when deriving the algorithm and, thus, is essential to achieve a good performance. This stationarity is assumed for each time interval of aggregation, and it is quite likely to occur. Nevertheless, there will also be frequent occasions in which transients (shock waves, stop and go behavior, etc.) will be present, and, thus, in which the algorithm as it is will not provide accurate estimates and would need some complex changes. To detect these situations, some simple measures can be taken. One parameter that can help to detect the presence of transients is the coefficient of variation (CV) (Eq. 3.20):

$$CV_v = \frac{\sigma_v}{\overline{v}}, \tag{3.20}$$

where

CV_v = speed coefficient of variation,

σ_v= speed standard deviation,

$\overline{v}$= mean speed.

Theoretically, if stationary traffic is assumed, this parameter tends to increase as the mean speed does; although it is in the denominator, the more the mean increases, the more the deviation does. Besides, the coefficient of variation indicates the importance of distinguishing time mean speeds from space mean speeds based on the relationships established by Wardrop (1952) or Rakha and Zhang (2005), as Eq. 3.21 shows:

$$\overline{v_t} - \overline{v_s} = \frac{\sigma_t^2}{\overline{v_t}} = \frac{\sigma_s^2}{\overline{v_s}} = CV * \sigma = CV^2 * \overline{v}. \tag{3.21}$$

The formula indicates that greater differences will occur with high CV_S and high mean speeds. However, empirically, it is common that the greatest differences appear with high CV_S and low mean speeds, a supposedly incompatible pairing. This fact indicates that the traffic is not stationary (May 1990; Rakha and Zhang 2005; Soriguera and Robusté 2011). Figure 3.11 shows the relationship between the mean speed and the CV in case VI, in which the algorithm did not perform well. In this case the CV diminishes with the mean, indicating the presence of transients and thus explaining the poor functioning of the method. In case IX (Fig. 3.12), the trend agrees with the assumption (stationarity) and the algorithm provides good results.

Although similar trends are usually obtained by directly comparing average speeds with the difference between time and space means (Figs. 3.13 and 3.14), the fact of not taking into account the variance of the speeds could result in an exaggerated impression of the magnitude of the relationship. The use of CV is strongly advised.

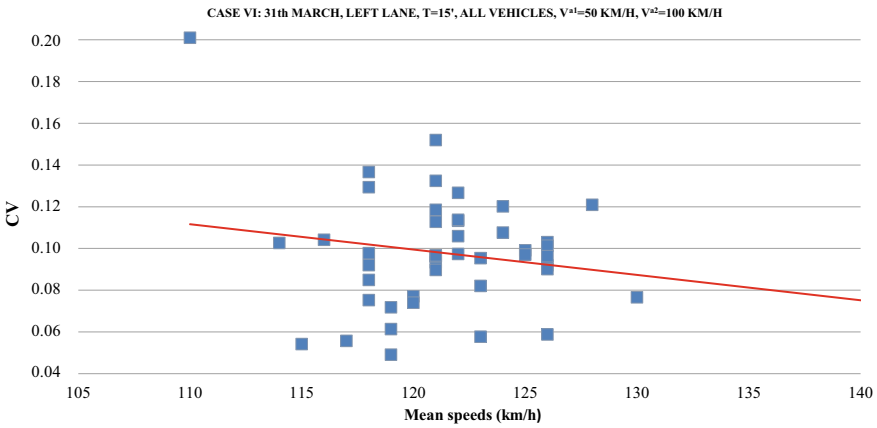

Fig. 3.11 Mean speeds versus the coefficient of variation in case VI

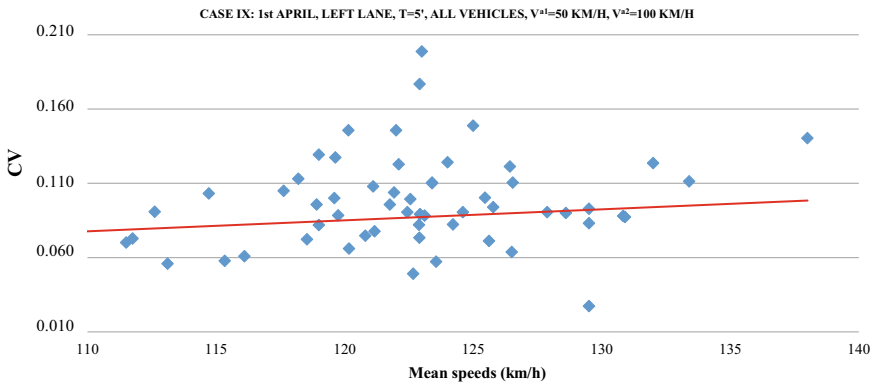

Fig. 3.12 Mean speeds versus the coefficient of variation in case IX

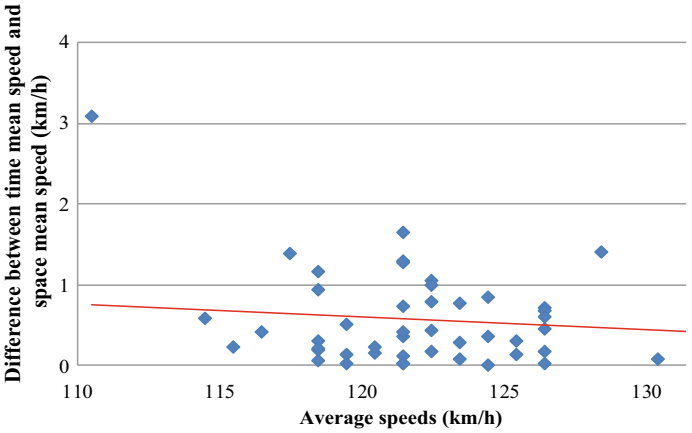

Fig. 3.13 Average speeds versus the difference between time mean speed and space mean speed in case VI

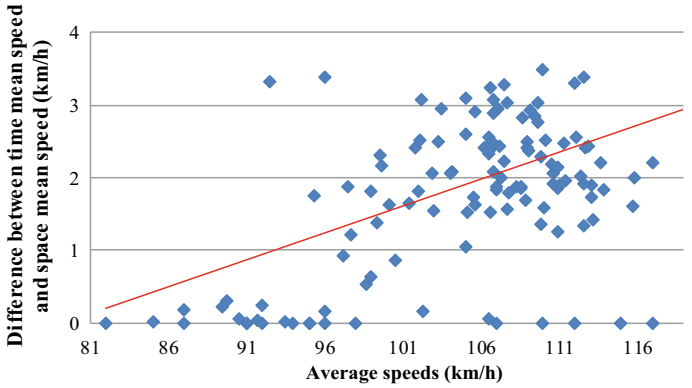

Fig. 3.14 Average speeds versus the difference between time mean speed and space mean speed in case IX

Finally, the place, day and moment when the data are collected is related to some of the issues previously mentioned. For example, the number and type of vehicles that drive on a freeway toward a capital on a workday morning in March will be very different from that on an August Sunday on a secondary road surrounding a small town. Therefore, speeds and traffic conditions will also be very different.

3.6 In Search of Other Relationships Between Mean Speeds

As explained, the algorithm proposed in Sect. 3.3 draws from the premise that the formula derived by Rakha and Zhang (2005) is the one that best defines a relationship between time mean speeds and space mean speeds, under different boundary conditions. Several researchers (e.g. Soriguera and Robusté 2011) reached the same conclusion, and this chapter has demonstrated the goodness of this formula, for example, compared to that of Garber's. However, the author wanted to check whether it would be possible to find a formula that would yield better results for the real case study analyzed in Sect. 3.5, even assuming a priori the impossibility of extrapolation. As the reference speeds played no role in this analysis, a different and more concise nomenclature has been defined (Table 3.6).

As explained in Sect. 3.5.1, individual spot speed data were in this case available. This allowed the calculation of the exact time mean speeds (arithmetic means) and space mean speeds (harmonic means). Then, space mean speeds were estimated from time means by using Garber's and Rakha and Zhang's relations. The mean absolute and mean relative errors in relation to the real space mean for each case were also calculated.

In addition, an attempt was made to find another kind of correlation between both means. More in particular, the possibility of a linear, quadratic, cubic, logarithmic, inverse, exponential or power-type relationship was analyzed (Table 3.7).

Both the corrected coefficient of determination, R_c^2, and the p-value were determined for this purpose. As it is already known, R_c^2 is a downward correction of R^2 based on the sample size n and on the number of independent variables k', as shown in Eq. 3.22 below:

$$R_c^2 = R^2 - \left[\frac{k' * (1 - R^2)}{(n - k' - 1)}\right]. \tag{3.22}$$

Table 3.6 Cases analyzed to verify the best relationship between the time mean speeds and the space mean speeds

Case	Day	Lane	T(')	L	N
1	31 March	right	15	all	4,662
2	31 March	left	15	all	769
3	01 April	right	15	all	2,841
4	01 April	left	15	all	596
5	31 March	left	5	all	769
6	01 April	right	5	all	2,841
7	01 April	right	5	L < 10 m	2,489
8	01 April	right	5	L > = 10 m	352

Table 3.7 Tested correlations between space and time mean speeds

Correlation	Outline
Lineal	$v_{ms} = a*v_{mt} + b$
Logarithmic	$v_{ms} = a*\text{Ln}(v_{mt}) + b$
Inverse	$v_{ms} = a*(1/v_{mt}) + b$
Quadratic	$v_{ms} = a*v_{mt}^2 + b*v_{mt} + c$
Cubic	$v_{ms} = a*v_{mt}^3 + b*v_{mt}^2 + c*v_{mt} + d$
Power	$v_{ms} = b*v_{mt}^a$
Exponential	$v_{ms} = b*\exp(a*v_{mt})$

The p-value is related to the contrast of the regression (ANOVA). In this case, the null hypothesis stands for a value of R^2 that equals zero. If the significance (p-value) in the statistical F-test is lower than 5% (for a confidence level of 95%), the null hypothesis can be rejected and, therefore, the existence of a correlation is proved.

In each of the cases studied, the estimated space mean speeds and the errors for the most suitable correlation were calculated. In this way, the best relationship both in general and for each particular case was determined. In order to remove the possible outliers, a slight smoothness was also made.

It should be highlighted that the variance with regard to the time mean for each specific time interval of aggregation introduced in Rakha and Zhang's equation was again calculated from individual spot speeds, which, as said, are not usually available. It is also important to notice that the data used in this study fit different types of distributions depending on the time interval of aggregation, being lognormal and normal distributions the most commonly found, as expected.

Table 3.8 shows the results of the curvilinear estimation. The corrected coefficient of determination indicates that the quadratic correlation is the most suitable in most cases. The coefficients of the quadratic correlation for each analysis are included in Table 3.9. New estimates of space mean speeds were calculated with these values. A level of significance was given to each coefficient, being the value of the null

Table 3.8 Coefficients and their significance for quadratic relationships

Case	Non-standarized coefficients and p-value					
	a	p	b	p	c	P
1	-0.011	0.058	3.487	0.010	-137.497	0.057
2	0.000	0.623	1.083	0.000	-8.046	0.380
3	-0.006	0.008	2.095	0.000	-51.618	0.028
4	0.002	0.000	0.485	0.000	29.919	0.000
5	0.001	0.083	0.861	0.000	8.133	0.124
6	0.003	0.036	0.410	0.126	33.357	0.034
7	0.002	0.000	0.574	0.000	23.899	0.000
8	0.002	0.008	0.689	0.000	13.764	0.009

Table 3.9 Corrected coefficient of determination and significance for each correlation

Case		Corrected R2 and p-value													
		Lineal		Logarithm		Inverse		Cuadratic		Cubic		Power		Exponent	
1	March 31th-right lane-T = 15’	0.971	0.000	0.972	0.000	0.973	0.000	0.973	0.000	0.973	0.000	0.973	0.000	0.970	0.000
2	March 31th-left lane-T = 15’	0.996	0.000	0.990	0.000	0.968	0.000	0.996	0.000	0.996	0.000	0.994	0.000	0.985	0.000
3	April 1st-right lane-T = 15’	0.964	0.000	0.968	0.000	0.966	0.000	0.969	0.000	0.969	0.000	0.969	0.000	0.961	0.000
4	April 1st-left lane-T = 15’	0.992	0.000	0.984	0.000	0.967	0.000	0.994	0.000	0.994	0.000	0.993	0.000	0.994	0.000
5	March 31th-left lane-T = 5’	0.986	0.000	0.965	0.000	0.915	0.000	0.986	0.000	0.986	0.000	0.984	0.000	0.973	0.000
6	April 1st-right lane-T = 5’	0.977	0.000	0.960	0.000	0.930	0.000	0.979	0.000	0.980	0.000	0.979	0.000	0.978	0.000
7	April 1st-right lane-T = 5’ light veh	0.988	0.000	0.970	0.000	0.936	0.000	0.989	0.000	0.989	0.000	0.987	0.000	0.985	0.000
8	April 1st-right lane-T = 5’ heavy veh	0.997	0.000	0.994	0.000	0.987	0.000	0.997	0.000	0.997	0.000	0.997	0.000	0.997	0.000

hypothesis equal to zero. As shown in Table 3.9 most coefficients are significant (p-value < 0.05), that is, they are needed to establish a good correlation. A linear relationship could achieve the same results only in two cases (as coefficient a is non-significant).

Finally, the mean absolute and mean relative errors with respect to the real space mean speeds (i.e., those calculated from individual speeds) encountered with the formula of Rakha and Zhang, that of Garber and with the quadratic correlation were also compared. The results are included in Table 3.10. It can be observed that the relationship of Rakha, despite being the most complex practice because of the need of estimating the variance with regard to the time mean, is worth considering. Both the absolute and relative errors are at the lowest level in all the cases analyzed in this study. Therefore, it has been again demonstrated its appropriateness to be part of the algorithm presented in Sect. 3.3.

Table 3.10 Errors observed with the different estimations of space mean speeds from time mean speeds

Case		Rakha		Garber		Quadratic correlation	
		Mean abs error	Mean relative error (%)	Mean abs error	Mean relative error (%)	Mean abs error	Mean relative error (%)
1	March 31th-right lane-T = 15’	0.09	0.09	2.60	2.40	0.32	0.29
2	March 31th-left lane-T = 15’	0.13	0.11	1.76	1.44	0.45	0.38
3	April 1st-right lane-T = 15’	0.45	0.44	2.47	2.35	0.79	0.76
4	April 1st-left lane-T = 15’	0.10	0.08	1.70	1.38	0.42	0.35
5	March 31th-left lane-T = 5’	0.30	0.25	1.54	1.27	0.61	0.51
6	April 1st-left lane-T = 5’	0.67	0.63	2.16	2.04	0.87	0.82
7	April 1st-left lane-T = 5’ light veh	0.56	0.52	1.93	1.75	0.84	0.77
8	April 1st-left lane-T = 5’ heavy veh	0.06	0.07	0.40	0.46	0.16	0.18

3.7 Conclusions and Further Research

The development of road networks and new technologies has proven to be a useful tool to respond to the increasing demands of society regarding the total control of traffic evolution. Nevertheless, fundamental traffic theory must be correctly incorporated in modern methodologies in order to obtain accurate results. This chapter introduces an algorithm that estimates space mean speeds in a specific time interval of aggregation as a first step, for example, for the calculation of travel times or occupancies. After analyzing the results obtained, three main conclusions can be drawn:

- It is possible to improve the current procedure followed by most traffic management centers, i.e., considering time means equal to space means. It can be done inexpensively by exploiting all the data delivered by loop detectors. Specifically, the proposed algorithm allows an estimation of space mean speed values that are accurate in most cases, or, at least, much closer to the real values than time mean speeds. Consequently, the use of these data also improves the results of subsequent calculations.
- The good performance of the algorithm depends on the fulfilment of its initial hypotheses, i.e., stationarity of the traffic stream and log-normality of speeds in each time interval of aggregation. The boundary conditions for data acquisition and for the calculations can be established to a certain extent in order to achieve these characteristics.
- In case of transients, for example the formation or dissipation of shock waves, most of the steps followed to design the algorithm are not valid (starting from the extrapolation of the spot speeds to a section). Thus, other specific methodologies should be used. Data fusion appears promising in this respect, as well as other completely different approaches that try to explain the propagation of traffic oscillation by means of car-following models (Li et al. 2014).

Further research can be carried out to improve the accuracy of the results or to enlarge the sphere of application of the proposed algorithm. Some lines could be:

- Including a smoothing process to remove erroneous data derived from the tendency of traffic loops to drift.
- Including in the algorithm the steps necessary to calculate the confidence interval for the means in order to be able to choose the most accurate when more than one value is obtained.
- Designing other algorithms adapted to other common speed distributions in addition to that introduced in this chapter and that in Soriguera and Robusté (2011). Thus, after the application of a prior step that may help to find the most suitable distribution for the speeds, the appropriate algorithm could be chosen in each case.

As noted, it is necessary to develop different and more evolved methodologies to estimate space mean speeds in case of transients. Loop data are probably insufficient in these situations. Other researchers have achieved good results with various techniques of data fusion (Soriguera and Robusté 2011; Bachmann et al. 2013; Yuan et al. 2014). However, there is still much work to do, since it is difficult to put most of them into practice because of their complexities and/or high costs. Of course, the same issues arise when thinking of data-driven approaches.

In view of the results, usual spot speed methods enhanced by the proposed algorithm would be satisfactory to estimate travel times in stable traffic conditions. Their combination with more elaborated methodologies that only partially rely on loop data would allow making the most of these widespread detectors on other occasions. For example at present when congestion exists or even in future driving environments.

References

Bachmann C, Abdulhai B, Roorda M, Moshiri B (2013) A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. Transp Res Part C: Emerg Technol 26:33–48

Breiman L (1969) Space-time relationships in one-way traffic flow. Transp Res 3:365–376

Cassidy M (1998) Bivariate relations in nearly stationary highway traffic. Transp Res Part B: Methodol 32(1):49–59

Chen C, Kwon J, Skabardonis AV (2003) Detecting errors and imputing missing data for single loop surveillance systems. Transp Res Record: J Transp Res Board 1855:160–167

Coifman B (2014) Revisiting the empirical fundamental relationship. Transp Res Part B: Methodol 68:173–184

Daganzo CF (1997) Fundamentals of transportation and traffic operations. Pergamon, Oxford

Dey P, Chandra S, Gangopadhaya S (2006) Speed distribution curves under mixed traffic conditions. J Transp Eng 132(6):475–481

Edie LC (1965) Discussion of traffic stream measurements and definitions. Proc. 2nd International Symposium on the Theory of Traffic Flow, OECD, Paris, pp 139–154

El Faouzi N, Maurin M (2007) Reliability of travel time under log-normal distribution: methodological issues and path travel time confidence derivation. Transportation Research Board 86th Annual Meeting (CD- ROM). Washington DC: Transportation Research Record

Garber NH (2002) Traffic and highway engineering. California: Brooks/cole. ISBN 0-534-38743-8

Haight FM (1962) A practical method for improving the accuracy of vehicular vehicle speeds distribution measurements. Highw Res Board Bull 341:92–116

Harvey J, van der Merwe A (2012) Bayesian confidence intervals for means and variances of log-normal and bivariate log-normal distributions. J Stat Plan Inf, 1294–1309

Hiribarren G, Herrera JC (2014) Real time traffic states estimation on arterials based on trajectory data. Transp Res Part B: Methodol 69:19–30

Homburger W, Hall J, Loutzenheiser W, Reilly W (1996) Fundamentals of traffic engineering. California: Institute of Transportation Studies. University of California, Berkeley. ISSN 0192-5911

Jenelius E, Koutsopoulos HN (2013) Travel time estimation for urban road networks using low frequency probe vehicle data. Transp Res Part B: Methodol 53:64–81

Jenelius E, Koutsopoulos HN (2015) Probe vehicle data sampled by time or space: Consistent travel time allocation and estimation. Transp Res Part B: Methodol 71:120–137

Khisty CL (2003) Transportation engineering: an introduction. Prentice Hall, New Jersey

Knight F (1935) The ethics of competition and other essays. New York and London: Harper & Bros and George Allen & Unwin. ISBN-10: 0836910885

Knoop, V., Hoogendoorn, S. and H. Zuylen (2007). *Empirical Differences Between time Mean Speed and Space Mean Speed*. In *Traffic and Granular Flow'07*, 351–356. Springer. ISBN 978-3–540-77074-9

Li X, Cui J, An S, Parsafard M (2014) Stop-and-go traffic analysis: theoretical properties, environmental impacts and oscillation mitigation. Transp Res Part B: Methodol 70:319–339

May A (1990) Traffic flow fundamentals. Prentice Hall, New Jersey

Poomrittigul S, Pan-ngum S, Phiu-Nual K (2008) Mean Travel Speed Estimation using GPS Data without ID number on Inner City Road. ITS Telecommunications 2008:56–61

Rakha H, Zhang W (2005) Estimating traffic stream space-mean speed and reliability from dual and single loop detectors. Transp Res Record: J De Transp Res Board 1925:38–47

SHRP2 (2013) Travel time reliability 2030: innovations and strategies for today and tomorrow. Washington, DC: Transportation Research Board

Soriguera F, Robusté F (2011) Estimation of traffic stream space-mean speed from time aggregations of double loop detector data. Transp Res Part C: Emerg Technol 19(1):115–129

Soriguera F, Robusté F (2013) Freeway travel-time information: design and real-time performance using spot-speed methods. IEEE Trans Intell Transp Syst 14(2):731–742

Viti F, Rinaldi M, Corman F, Tampère CMJ (2014) Assessing partial observability in network sensor location problems. Transp Res Part B: Methodol 70:65–89

Wang Y (2012) A note on speed and travel time estimation based on truncated normal and log-normal distributions. Transp Res Record J Transp Res Board 2315:66–72

Wardrop J (1952) Some theoretical aspects of road traffic research. Proceedings of the Institute of Civil Engineers 1(2):325–378

Yuan Y, van Lint H, Hoogendoorn S, van Wageningen-Kessels F (2014) Network-wide traffic state estimation using loop detector and floating car data. J Intell Transp Syst 18(1):41–50

Zou Y, Zhang Y (2011) Use of Skew-Normal and Skew-t distributions for mixture modeling of freeway speed data. Transportation Research Record: Journal of the Transportation Research Board, pp 65–75

Chapter 4
Accurate, Affordable and Widely Applicable Freeway Travel Time Prediction: Fusing Vehicle Counts with Data Provided by New Monitoring Technologies

Margarita Martínez-Díaz

Abstract Today, technology allows highly accurate direct travel time measurements. These can be attained by identifying vehicles at several locations on the freeway or by directly tracking vehicles' trajectories. The penetration rate of these technologies is higher than ever before and continuously growing so that the traditional problem of data significance (i.e. not having enough measurements during a short updating period) is being attenuated. This fact has encouraged traffic administrations and private companies to deploy real-time information systems based on these data. However, even in an ideal scenario, direct measurements of travel times are representative of near past traffic conditions for vehicles entering the target stretch, while the objective of real-time information systems is to transmit information about traffic conditions in the near future. This chapter aims to fuse the information provided by input–output diagrams obtained from loop detectors with direct measurements of travel times obtained from automatic vehicle identification (AVI) or tracking technologies. This fusion allows exploiting the accuracy of the direct measurements to correct the count drift in loop detectors. Then, corrected input–output curves can be used to obtain reliable short-term predictions of travel time from vehicles' accumulation. The proposed data fusion method has been applied to a test site in the AP7 freeway near Barcelona using real and simulated data. Results show that the method is able to provide predicted travel times that anticipate changes in traffic conditions much faster than the simple dissemination of measured travel times, implying lower average and maximum errors of the real-time information systems. The benefits of using the method grow with the severity of congestion and in low surveillance environments, which represent the scenarios where the travel time information is more precious and more difficult to obtain.

M. Martínez-Díaz (✉)
Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_4

4.1 Introduction and Background

Traffic monitoring is crucial in any kind of traffic analysis. Regardless of the final purpose (e.g. management, planning, etc.), the lack of accurate and adequate data prevents agencies and researchers from obtaining valuable results. Since the earliest traffic volume measurement, which started more than a century ago, the need for traffic data has grown together with the car fleet and the complexity of the networks. This tendency continues, and the envisioned cooperative and automated driving environments will largely rely on advanced traffic data. Fortunately, technological progress has also reached traffic surveillance. Today, surveillance systems are composed of a broad assortment of sensors and communication systems capable of the real-time gathering and processing of huge amounts of traffic data. As detailed in this book, these have supported the rise of (i) Advanced Traffic Management Systems (ATMS), which allow the dynamic management of the traffic streams, (ii) Advanced Traffic Information Systems (ATIS), which provide users with valuable real-time information and (iii) Incident Management Systems (IMS), aimed at the coordination of personnel, facilities, equipment, procedures, and communications in the event of an incident (Hall 1993). These systems are usually integrated in a single freeway management and information system, which receives data from detectors located either in the infrastructure or in the vehicles. Although on-board sensors are called to play an important role in the near future, the increasing need for traffic data has traditionally been faced by the gradual installation of new surveillance in the infrastructure. This "spontaneous" approach, resulting from the lack of planning, budget limitations and the fast technology evolution, turned into the existence of a wide range of surveillance levels and technologies. And this situation complicates the standardization of traffic management and information systems. In spite of this, wireless communications and the widespread introduction of advanced in-vehicle devices, like car navigators or smartphones capable of positioning, speed measurement, etc., considerably contribute to alleviate the infrastructural surveillance deficiencies nowadays (Herrera and Bayen 2010; Herrera et al. 2010; Ge and Fukuda 2016; Woodard et al. 2017).

One of the key elements of ATIS is the travel time information. In fact, the travel time is the best indicator of both, the level of service of a road link (in terms of travel time reliability) and of current traffic conditions (real-time travel time information). It is also the worthiest information for drivers and will continue to be key after the advent of autonomous driving. In such a future ideal scenario, the driver becomes a passenger and is able to use the travel time for other things than driving. The value of the trip time will change, but the information of the expected travel time will still be valuable for passengers in order to plan their activities accordingly. In addition, a travel time increase in the network will continue to be indicative of congestion, and detrimental in terms of costs, safety and sustainability. However, despite the importance of travel time information and related to the aforementioned surveillance heterogeneity, the availability of travel times is limited to small parts of the road networks. Moreover, the received information exhibits variable levels of accuracy,

often being not satisfactory. Currently, the travel times disseminated by ATIS have different origins. Although this aspect has been explained in detail in Chap. 2, next paragraphs revisit all of them with the aim of putting the algorithm proposed in this chapter into context.

On the one hand, direct measurements of travel time can be obtained either from AVI (Automated Vehicle Identification) technologies or from tracking. The case of AVI systems is straightforward: each individual vehicle is identified at two control points, which delimit a section of the road. Its travel time in this section is calculated as the difference between the vehicle's time stamps at both locations. Bluetooth (Barceló et al. 2010) or toll tag identification (Longfoot 1991; Nishiuchi et al. 2006; Soriguera et al. 2010), license plate automatic recognition (Buisson 2006; van Hinsbergen et al. 2009) or the individual identification of vehicles based on several distinctive characteristics such as the length by means of traditional inductive loop detectors (Coifman and Cassidy 2002; Coifman and Ergueta 2003; Coifman and Krishnamurthya 2007) or the inductance signature (Kuhne and Immes 1993; Abdulhai and Tabib 2003; Kwon 2006) are examples of AVI technologies. Additionally, individual vehicle identification is very useful, for example, for the estimation of O/D matrices. The past inconveniences of the low penetration rates and high costs of these technologies have been overcome, at least partially. However, some issues remain. First, travel time measurements directly obtained from AVI systems are captive of the control points, i.e., they are only available along the sections delimited by the sensors. Neither partial nor longer trips can be measured. Second, travel time measurements in a given section are obtained once vehicles have completely covered it. These are called measured travel times (MTT) or, equivalently, arrival-based travel times (ATT), and represent a somehow outdated information for vehicles entering the section, especially if this section is long, or when congestion exists. In case tracking technologies are available, direct measurements of travel times between any two points on a highway can be obtained. In fact, these systems can provide individual vehicles' trajectories every few seconds. Especially GPS technologies, either in vehicle navigation systems or smartphones (Herrera et al. 2010), but also phone signal geopositioning (Yim 2003) or video surveillance via drones (Kaufmann et al. 2018) are used to this end. In fact, these technologies represent the evolution of traditional probe cars and are expected to be the prevailing source of traffic information in the future. Travel time measurements provided by tracking technologies are based on the last information available from vehicles' trajectories. These instantaneous travel times (ITT) still do not provide the desired information to drivers starting their trip, as ITT do not imply a prediction about the conditions that these vehicles will face.

On the other hand, it is possible to indirectly estimate travel times from other traffic variables such as speeds or flows. In fact, this is still the most common approach for travel time estimation, as the necessary inputs can be obtained from inductive loop detectors, which continue to be the most widespread data source on many highways. The most widely used methodology estimates the travel time in a section between two loops from the estimated average speed in the section. This average speed is obtained by extrapolating the punctual speeds measured by the loops at both ends. A high density of double loops (single loops do not provide accurate spot speeds),

ideally a loop every 500 m, is necessary to obtain acceptable results. Still, large errors can appear in congestion or in traffic state transitions (Soriguera and Robusté 2011c). One possibility to improve spot speed methods was developed by Coifman (2002) and Treiber and Helbing (2002). They suggested a procedure to account for traffic dynamics in the spatial interpolation of the speed between loop measurements. This approach should be useful for the travel time estimation in these low surveillance environments, increasing the accuracy of the estimation without the huge investment that more intensive monitoring implies. However, it can only be applied when all the section between detectors is either free flowing or fully congested, as traffic state transitions are overlooked. Because the lengths of these links can be of several kilometers, the fact of not considering traffic transitions implies significant errors. In any case, the travel time along a stretch (i.e. composed of several sections) is calculated by adding up the partial travel times, finally obtaining an ITT measurement still without predictive capabilities.

A potentially useful tool to compute travel times from traditional loop detector measurements is the use of cumulative count input–output curves and the vehicle conservation equation (Nam and Drew 1996; Oh et al. 2003; van Arem et al. 1997). The advantages of input–output methods are twofold. First, they can be applied in low surveillance environments and using data from the already installed loop detectors. By low surveillance it is meant that intensive monitoring (i.e. closely spaced double-loop detectors) is not necessary although some minimum monitoring requirements exist. For example, in case the target section has junctions, loop detectors must be installed in all on/off ramps. These loops complement those needed in the main trunk, which must be present on every section between junctions (one for section would suffice). Additionally, all of them can be single loops, as they are enough to obtain the vehicle count (Fig. 4.1).

The second main advantage of input–output methods is that they provide the vehicle accumulation between detectors. This allows predicting the evolution of travel times in the short term, which is a key feature for highway travel time information systems when providing real-time information. Note that none of the previous travel time estimation methods presents such predictive capabilities, providing all

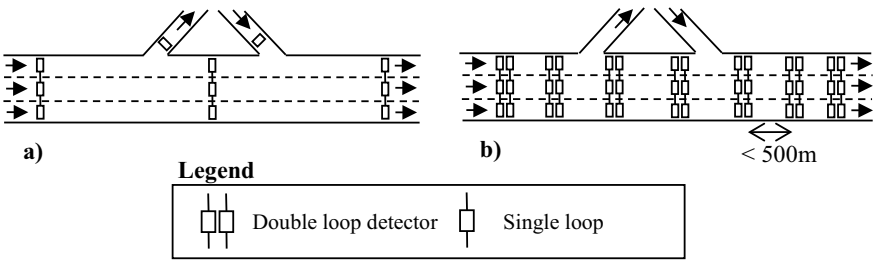

Fig. 4.1 Highway surveillance levels: **a** Minimum requirements for input–output methods; **b** Intensive monitoring for spot-speed methods. Note: In Fig 4.1a ramp detectors could be substituted by another sectional main-trunk detector. In such case, main-trunk detectors should be placed before and after the junction

of them, to some extent, a past measurement. In spite of these advantages, input–output methods are not currently used for freeway travel time estimation. Not being an intuitive method or the need for closed detector configurations can be some of the reasons for this situation. Nevertheless, the most problematic factor is that loop detectors suffer from drift (i.e. small counting errors). Detector drift is not an important problem when data is analysed for a single detector. However, in input–output methods, where the accumulation is computed from the relative difference between the cumulative counts at consecutive detectors, even a small drift accumulated over time can lead to meaningless results. No significant research aimed at understanding and solving the undesirable loop detector drift for travel time estimation has been performed yet. Therefore, there is a need for additional efforts to develop an adequate real-time methodology that (i) is aimed at the prediction of travel times from cumulative count curves obtained from traditional loop detectors even on low surveillance highway stretches, (ii) performs drift correction and (iii) accounts for typical difficulties in input–output methods like the existence of inner junctions in the closed sections or the effects of passing.

The present chapter addresses these objectives by proposing a data fusion method. In this regard, data fusion arises as a promising choice, as it has already been proved useful in traffic state estimation (Ambühl and Menéndez 2016; Deng et al. 2013; Nantes et al. 2016; Sun et al. 2017) and even in travel time prediction under different circumstances (Chen and Rakha 2016; El Faouzi et al. 2007; Soriguera and Robusté 2011b). The method proposed here fuses data from traditional loops with direct travel time measurements. The latter can be obtained either from AVI or tracking technologies, being necessary only minor modifications to the algorithm depending on the case. These adjustments are explained in detail in the next sections. Because in general the spatial–temporal aggregation and coverage of the different data sources is not uniform, the method needs to include the spatial and temporal alignment of the data as a first step. Then, direct measurements are used to correct the drift in the cumulative curves constructed from detector counts. From these corrected input–output curves, predicted travel time estimations are obtained. This allows exploiting the predictive capabilities of input–output methods while keeping the accuracy in the estimation. Because direct measurements are not used as the final information, a low penetration of the used measurement technology, resulting in periods with few or any data, does not imply significant drawbacks. However, it is important to remark that this algorithm has a vocation of continuity and is also applicable in other more equipped environments, either as a central or as a backup methodology. Actually, it will also be applicable in future environments, where AVs will provide direct measurements of travel time, either after their reidentification at control points or, better, because of their condition of lagrangian sensors (trackers).

The structure of the remainder of the chapter is as follows: next, Sect. 4.2 reviews the main concepts in the input–output method to estimate travel times. Section 4.3 introduces the proposed data fusion algorithm. It starts explaining how to predict travel times from vehicle accumulation, continues with the drift correction methodology and ends with the activation conditions of the algorithm. An experimental study with real data provided by AVI detectors is presented in Sect. 4.4 including

the discussion of the results. Similarly, a case study with data provided by tracking technologies, in this case simulated, is performed in Sect. 4.5. The conclusions of this chapter are drawn in Sect. 4.6, where further research is also proposed.

4.2 Travel Time from Input–Output Cumulative Curves

Cumulative count curves are well-known tools used in different disciplines like hydraulics, hydrology or geotechnics. Since Karl Moskowitz (1954) introduced them in traffic engineering for the first time, their simplicity and usefulness have been extensively proved. The works of prof. G. F. Newell represented the definitive popularization of the tool. For instance Newell (1982) developed applications of queueing theory, Newell (1988) analysed the effects of the interruption of traffic streams at signals, Newell (1993a, b, c) proposed a simplified kinematic wave theory of traffic flow, or Newell (1999) analysed delays at freeway off-ramps, in all cases using cumulative count curves.

A cumulative count curve is obtained by counting all vehicles passing over a particular location, x. One of these vehicles is chosen as a reference and is said to have passed at time $t = 0$. From this instant, the counts corresponding to the following vehicles are accumulated over time. This allows defining the function $N(x, t)$ that renders the accumulated number of vehicles that have passed the location x by time t. The graphical representation of this function is the so-called cumulative count curve at x, also known as N-curve. $N(x, t)$ is, by definition, a monotonically increasing stepwise function, from which valuable information can be obtained. For example, the average slope of $N(x, t)$ during a certain time interval Δt (i.e. $[N(x, t + \Delta t) - N(x, t)]/\Delta t$) represents the average flow at x. Because the number of passing vehicles in real applications is large, especially in freeways, N-curves can be smoothed and treated as continuous functions (see Fig. 4.2). This is simply done by interpolating through the crest of every individual step of the discrete function. The continuity of the smoothed version of $N(x, t)$ allows taking derivatives and, for instance, defining the instantaneous flow as the time derivative of $N(x, t)$ at a particular instant (i.e. $\partial N(x, t)/dt$).

The usefulness of cumulative curves for the analysis of queuing systems increases when using input–output diagrams. These diagrams are the result of depicting jointly the N-curves at both ends of a given section between detectors. In Fig. 4.2, x_u and x_d are, respectively, the locations of the upstream and downstream detectors, defining a closed section. Thus, $N(x_u, t)$ corresponds to the vehicles entering the section (i.e. the arrivals curve), whereas $N(x_d, t)$ refers to the exiting vehicles (i.e. the departures curve). Assuming vehicle conservation in the section, the vertical distance between these curves at a time t depicts the accumulation of vehicles between detectors at this instant. Additionally, assuming that vehicles have a FIFO behavior, (i.e. no passing), the horizontal distance between the curves at the height of any vehicle k represents the travel time of the kth vehicle between detectors. Note that this time includes the free-flow travel time and the delay, if any.

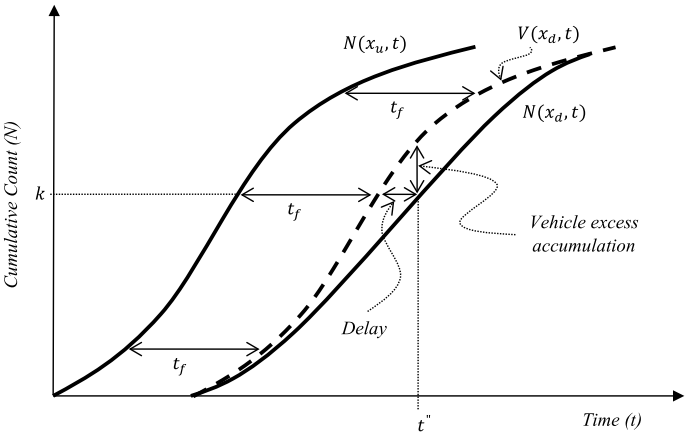

Fig. 4.2 Input–output diagram and determination of delay via the virtual downstream cumulative curve

Because the analysis of delays is usually desired, the introduction of a third curve results useful. The "virtual" downstream cumulative curve, $V(x_d, t)$, is obtained by translating the arrivals curve a distance t_f forward in time, where t_f is the average free flow travel time in the section. $V(x_d, t)$ represents the cumulative number of vehicles that would have exited the section by time t if free flowing conditions would have prevailed. Note that $V(x_d, t) = N(x_u, t - t_f)$. Thus, the horizontal distance between $V(x_d, t)$ and $N(x_d, t)$ at the height of any vehicle k represents the delay experienced by this vehicle in the section. The vertical distance between these same curves at time t is the vehicles' excess accumulation in the section at this instant. Obviously, $V(x_d, t) = N(x_d, t)$ if no delay exists (Daganzo 1987, 1997).

The reminder of this section is aimed at presenting the application of these concepts to the travel time estimation from the typical loop detector counts and how to deal with the difficulties that frequently arise, namely the effects of inner junctions in the conservation assumption, or the violation of the FIFO assumption due to overtaking.

4.2.1 Travel Time Definitions from Input–Output Diagrams

In practice, input–output diagrams can only be used to obtain aggregated variables of a traffic stream, averaged over time and space. This is not an important drawback, since average values better illustrate traffic conditions and are thus more valuable. There are several reasons for this limitation. First, due to passing, cumulative count curves do not truly represent individual vehicles, but positions or "labels" within the traffic stream (Daganzo 1997) (see Sect. 4.2.2.1). Second, loop detectors do not provide individual measurements, but aggregated counts and average speeds over

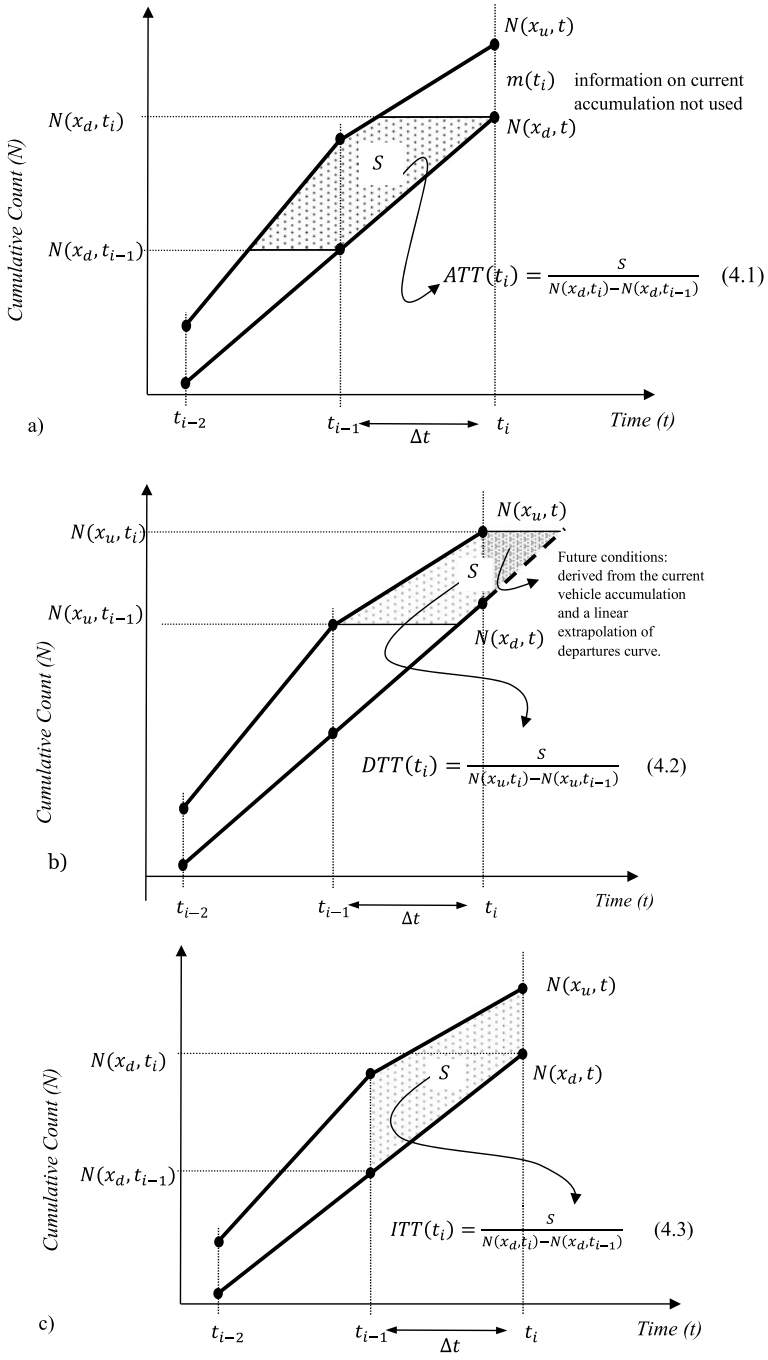

Fig. 4.3 Travel time definitions from input–output diagrams (inspired in Soriguera 2016): **a** Arrival-based travel time (ATT); **b** Departure-based travel time (DTT); **c** Instantaneous travel time (ITT)

predetermined time intervals, Δt. This responds to data treatment standards aimed at avoiding excessive communications and storage needs. Δt highly varies among countries and administrations, commonly lasting between 1–3 min in Europe and typically 30 s in the USA (Soriguera and Robusté 2011a).

In any case, the procedure to obtain the average travel time for a particular Δt consists in (i) computing the area S enclosed between curves $N(x_u, t)$ and $N(x_d, t)$ (i.e. the total aggregated travel time) and (ii) dividing this area by the total number of involved vehicles. However, S can be computed in several ways, thus affecting different groups of vehicles and giving rise to distinct travel time averages (Soriguera 2016). Although these differences are meaningful for any subsequent application, they have traditionally been overlooked (Nam and Drew 1996; Oh et al. 2003; van Arem et al. 1997).

For example, if the considered vehicles are those that exit the control section during Δt, arrival-based travel times (ATT) are obtained (see Fig. 4.3a and Eq. 4.1). These travel time measurements are equivalent to those directly obtained from AVI detectors. On the contrary, if the average involves those vehicles that have departed from x_u during Δt (i.e. that have entered the control section), departure-based travel times (DTT) are obtained (see Fig. 4.3b and Eq. 4.2). In this case, some assumptions regarding future traffic conditions are necessary. Typically, it is assumed that the outflow will remain constant in the short term. Trying to use only the most recent information, another possibility is to consider all the vehicles contained between x_u and x_d from time t_{i-1} to time t_i (where $t_i - t_{i-1} = \Delta t$) (see Fig. 4.3c). In this way, instantaneous travel times (ITT) are obtained. In this case, S represents the vehicles' total travel time in the time–space region $(x_u, x_d) * (t_{i-1}, t_i)$. Note that not all the vehicles considered in S travel the whole distance between x_u and x_d in Δt, and it is even possible that no vehicle covers the entire distance if travel times largely exceed Δt. These situations complicate the estimation of the average travel time. Anyway, the average travel time can be obtained by dividing the total time travelled, S, by the global distance travelled by all the vehicles involved. The result of this division represents an average pace, which needs to be multiplied by Δx (i.e. the distance from x_u to x_d) to obtain the average travel time. Daganzo (2010) proved that the total distance travelled by a group of vehicles in the $(x_u, x_d) * (t_{i-1}, t_i)$ space–time region can be computed as Δx times the number of vehicles exiting the section in the time period (i.e. $N(x_d, t_i) - N(x_d, t_{i-1})$). So, the ITT estimation is obtained as in Eq. 4.3. This definition of ITT is equivalent to the average that would be directly obtained from tracking systems.

Other definitions for the average travel time have been proposed. Nam and Drew (1996) suggest only considering the travel times of those vehicles crossing both x_u and x_d within Δt (see Fig. 4.4a and Eq. 4.4). Evidently, this definition is only useful if individual travel times are significantly shorter than Δt. This is only feasible in case of short distances between loop detectors and free flowing conditions or if the considered time period, Δt, is long. However, Δt needs to be short because long time periods imply less frequent information updates and this is not acceptable for accurate real-time information systems.

Thus, the limited usability of this definition is evident. Alternatively, van Arem et al. (1997) focused again on the most recent information (Eq. 4.5). To this end, the average travel time in the target section is defined as the arithmetic average between the travel times of the last vehicles that have respectively passed over x_u and x_d in Δt. Note that the travel time of the last vehicle to enter the section (i.e. the last crossing x_u in Δt) must be estimated, as this is future information. Both the vehicle accumulation in the section and the outflow at the end of Δt (usually assumed constant in the short-term, like in the case of DTT) are used for this purpose. This average travel time estimation, known as predicted travel time (PTT) (see Eq. 4.6 and Fig. 4.4b), is more suitable than the former for real-time travel time information systems, as for a vehicle entering a section the most valuable information is its expected travel time.

Note that in all the previous definitions, if the curve $V(x_d, t)$ is used instead of $N(x_u, t)$, the enclosed area would not be equal to the total travel time of the involved vehicles in the time interval, but to the total delay they suffered. Equations 4.1–4.6 would then correspond to different definitions of the average delay.

4.2.2 Main Difficulties When Using Input–Output Diagrams for Travel Time Estimation

The estimation of travel times from input–output diagrams is a powerful method to feed real-time travel time information systems. As vehicles' accumulation, which is related to the near-future traffic evolution, is available from these diagrams, the method is able to provide predicted travel times. In addition, the required surveillance is limited, and the only requirement is that the target section needs to be "closed" in the sense that all vehicles that enter or exit the section must be detected, even those using inner junctions. The procedure is independent of the geometry of the freeway, which may influence the bottlenecks' location but not the measurement of the average delay. In this regard, neither the derivation of empirical parameters nor the calibration require a major effort. Despite these advantages, some challenges usually appear in practice. Next sections face them and propose the corresponding solutions.

4.2.2.1 The Effects of Passing

Theoretically, the use of input–output diagrams in queuing systems assumes FIFO behaviour. In practice, some passing exists in traffic streams, that is, non-FIFO situations can take place. In these cases, vehicles change their relative positions. However, if it is considered that cumulative curves do not count particular vehicles but "order labels" (i.e. that vehicles change with each passing manoeuver), important features of the input–output methods such as the assumption of FIFO traffic or the monotonically increasing nature of the cumulative curves are maintained

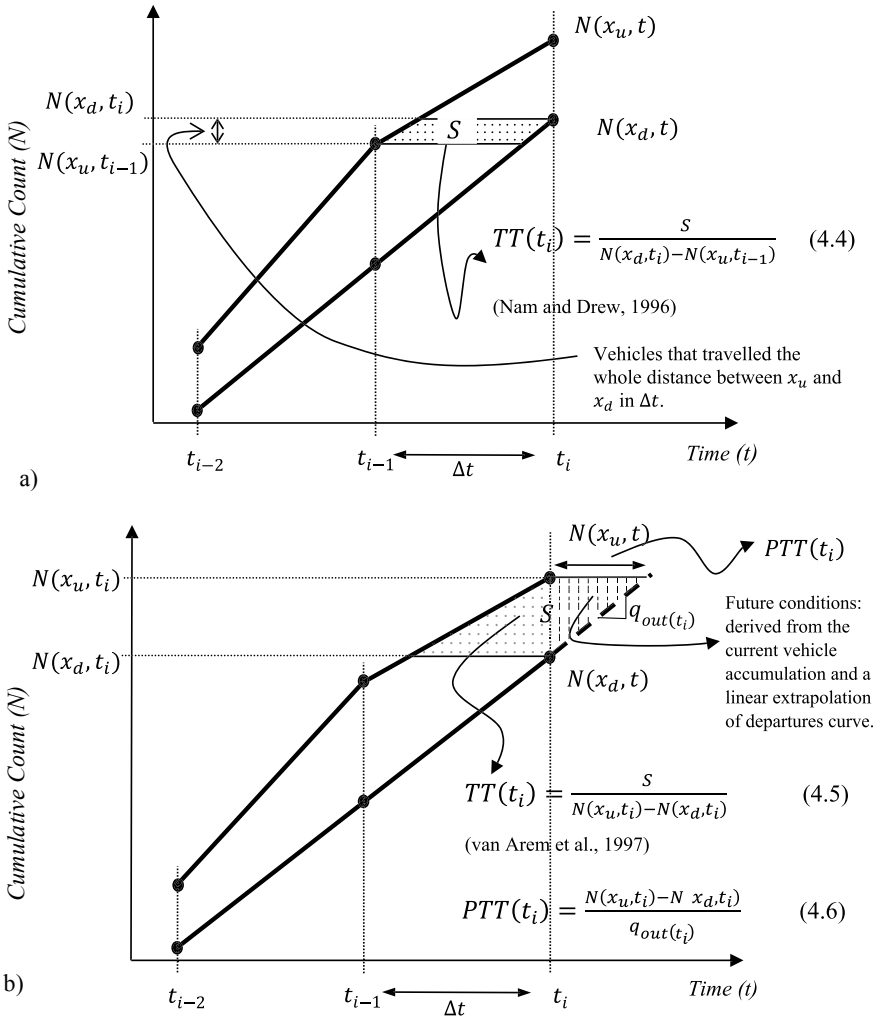

Fig. 4.4 Other travel time definitions from input–output diagrams (inspired in Soriguera 2016): **a** Nam and Drew (1996); **b** van Arem et al. (1997) and predicted travel time (PTT)

(Daganzo 1997). This procedure prevents the use of input–output diagrams to obtain individual travel times, but average travel times will not be significantly affected despite passing. Note that it is possible that some of the "order labels" considered in the calculation of the average travel time for a certain Δt, actually correspond to vehicles that have exited or entered the section in the time period immediately before or after Δt. However, if the percentage of these vehicles is small, the resulting average travel time continues to be a good estimation. In this regard, Muñoz and Daganzo (2002) empirically demonstrated that freeways could be mostly considered

FIFO systems and, thus, the consequences of passing in cumulative count curves, negligible.

4.2.2.2 The Effects of Inner Section On/Off Ramps

As mentioned, vehicle conservation is a requirement of input–output methods. If the target section between detectors contains one or more ramps, entering and exiting flows must be monitored (see Fig. 4.1a). Difficulties arise because these flows can take place at points of the section different from its extremes. Then, two groups of vehicles can be distinguished: (i) those that cover all the distance between detectors and (ii) those using the junction and that only travel a portion of this distance. To take this difference into account, it would be necessary to divide the section into subsections, before and after each junction, and to construct input–output diagrams for each subsection. Even if ramp flows are being monitored, this cannot be achieved without a main-trunk detector at the junction location, which is frequently missing.

Given this partial monitoring layout, one solution could be to model cumulative curves before and after the junction, given the measurements upstream (i.e. $N(x_u, t)$), downstream (i.e. $N(x_d, t)$) and the ramp counts. Newell (1993a, b, c) proposed a method to shift a N-curve to any desired location in between measurements. However, the method requires the flow-density relationship in the section (i.e. the fundamental diagram of traffic) as well as the a priori knowledge of the existence of bottlenecks, their precise location and their capacity. These requirements make the generalized application of Newell's methodology unfeasible in real time.

A simpler approximate procedure consists in directly adding the net input counts at the junction (i.e. entrances minus exits) to the detector measurements, assuming the junction to be located at x_u or at x_d (i.e. one of the detector locations that define the target section). This means that, given a junction located at $x_i \in (x_u, x_d)$ with cumulative net input counts defined by $J(x_i, t)$, the section input–output cumulative curves would be defined by the following Eqs. 4.7 and 4.8:

$$\begin{aligned} Input &= N(x_u, t) + J(x_i, t) \\ Output &= N(x_d, t) \end{aligned} \quad \text{if } x_i \approx x_u, \tag{4.7}$$

$$\begin{aligned} Input &= N(x_u, t) \\ Output &= N(x_d, t) - J(x_i, t) \end{aligned} \quad \text{if } x_i \approx x_d. \tag{4.8}$$

The disadvantage with this approach is that either all vehicles are considered to compute the average travel time (Eq. 4.7), or none (Eq. 4.8), assuming that vehicles using the junction travel the whole section or do not travel in the section, respectively. This will introduce some bias in the average travel time estimation, although this bias will only be significant if vehicles using the junction represent a significant fraction of total flow and their travel are very different with respect to the others (e.g. partial congestion after the junction).

An adequate selection between the alternatives in Eq. 4.7 (i.e. $x_i \approx x_u$) and Eq. 4.8 (i.e. $x_i \approx x_d$) helps to reduce the bias of the estimation. In general, the best option is to assume the junction to be located at the closest detector location. However, if the junction is somehow in the middle and the approximate location of recurrent congestion is known, the previous criterion could be modified. Note that the alternative represented by Eq. 4.7 will involve a higher bias if congestion is concentrated between x_u and x_i. In contrast, the bias will be lower if congestion concentrates between x_i and x_d. Equation 4.8 presents the opposite behaviour. Therefore, the selection should be done accordingly.

4.2.2.3 Detector Count Drift

Errors in loop detector counts have been extensively reported (Nam and Drew 1996; Oh et al. 2003; van Arem et al. 1997). Typically, detectors miss some few vehicles every Δt, without a systematic pattern and with different tendencies to undercount depending on each particular detector. This small measurement drift does not imply important drawbacks when using each detector, isolated, to measure average flows or speeds. However, input–output methods are used to compute the vehicles' accumulation between a pair of detectors. Even a small drift becomes significant when the count difference between detectors is accumulated over time. Note that this difference is usually small, even if a large number of vehicles has been registered. Thus, the error introduced by the detector drift in the estimation of average travel times can be huge. In fact, the accumulated count error could be larger than the vehicle accumulation itself. Precisely, this chapter presents a data fusion scheme to correct the count drift at detectors, enabling the of use input–output methods to estimate accurate average travel time predictions.

4.2.2.4 Initialization of the Input–Output Diagram

Despite the application of drift correction methods, a frequent reset of the input–output diagram is needed to account for the increasing bias in the vehicle accumulation. At each reset, the value of the initial accumulation in the section must be estimated (see Fig. 4.5). However, this is not easily achieved with loop detector data. In case of using the virtual arrivals curve $V(x_d, t)$ instead of $N(x_u, t)$ and, thus, computing delays instead of travel times, the initialization is simplified because the excess accumulation is null in free flowing traffic. Given this situation, it is advisable to only use input–output methods to compute delays in congested conditions. The method should turn on just before the congestion onset, still in free flowing conditions, and with null initial excess accumulation. Still, the measurements on curve $V(x_d, t)$ will be displaced t_f time units with respect to those on curve $N(x_d, t)$ (see Fig. 4.5), and this requires the estimation of the travel time in free flowing conditions, t_f. The construction can be further simplified by considering that traffic flows evolve smoothly between consecutive Δt, and that t_f is generally small when compared

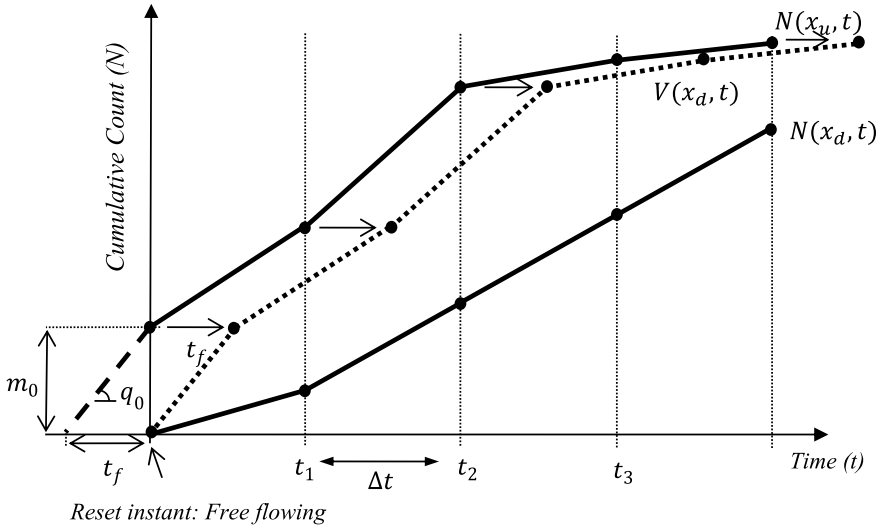

Fig. 4.5 Initialization of input–output diagrams (inspired in Soriguera 2016)

to the precision required for the travel time estimation (e.g. t_f for a 2 km section could be of the order of 1 min). Then, $V(x_d, t)$ can be constructed simultaneously with $N(x_d, t)$ every Δt, neglecting the time lag between the measurements on both curves.

The applicability of the method only in congested conditions is not restrictive in any sense. In free flowing traffic any other methodology (e.g. spot speed methods or direct measurements) would perform well (Soriguera 2016). Furthermore, real-time information is less meaningful in free flowing periods, as the uncertainty faced by the users is much lower (Soriguera 2014).

4.2.2.5 N-Curves Linear Interpolation

The introduction of small errors into the method is accepted when interpolating through the original stepwise N-curve by means of a piecewise linear function. However, the implications of this simplification are negligible, as the aggregation periods, Δt, are usually short and traffic evolution within them can be considered gradual and smooth (Soriguera 2016).

4.3 A Data Fusion Algorithm for the Short-Term Prediction of Freeway Travel Times

This section describes the proposed data fusion method to correct the drift in input–output curves by using direct travel time measurements. This includes the description of the turn-on and turn-off conditions to ensure the periodic reset of the cumulative curves, the temporal and spatial alignment of the different sources of information, the data fusion algorithm, with slight differences in case of having AVI or tracking direct measurements, and the computation of the predicted travel time from the input–output diagrams. The flowchart included in Fig. 4.6 summarizes the main steps of the methodology. A detailed explanation of all of them is provided in the next sections.

4.3.1 Data Inputs for the Algorithm

The proposed travel time short-term prediction method requires two types of real-time inputs. First, data coming from loop detectors, which generally include the time mean speed, v (note that its denomination in Chap. 3, $\overline{v_t}$, has been modified here for the sake of simplification) vehicle count, n, and detector occupancy, ρ, over an aggregation period, Δt. Second, direct travel time measurements obtained from AVI or tracking technologies, which consist in the arithmetic average of the measured travel times on a target stretch during an aggregation period, ΔT. In general, $\Delta T > \Delta t$, because, as explained in former chapters, only a fraction of the vehicles can be automatically identified or tracked, and a representative sample size is needed to compute average travel times. The method assumes that ΔT is an integer multiple of Δt. As frequently $\Delta t = 30$ s or 1 min, and ΔT lasts few minutes, this condition is generally fulfilled. Recall that the directly measured travel times would be ATT (i.e. arrival-based) in case of using AVI, and ITT (i.e. instantaneous) in case of vehicle tracking (see Sect. 4.2.1 and Fig. 4.3). From now on, the description of the proposed method will assume that direct measurements are ATT. If instead, ITT measurements are available, the method is simplified. This last case is described in Sect. 4.3.4.

Figure 4.7 sketches the typical freeway surveillance layout when direct measurements are provided by AVI technologies. AVI devices, like Bluetooth detectors, are usually installed on gantries, and they are fewer in comparison with loops. In this context, direct travel time measurements define the target freeway stretch for which real-time information is to be provided (i.e. between AVI devices). Then, loop detectors divide this stretch into sections, which may range from 500 m to 2 km. It is assumed that the location of AVI devices coincides with that of one loop detector, which is frequently the case.

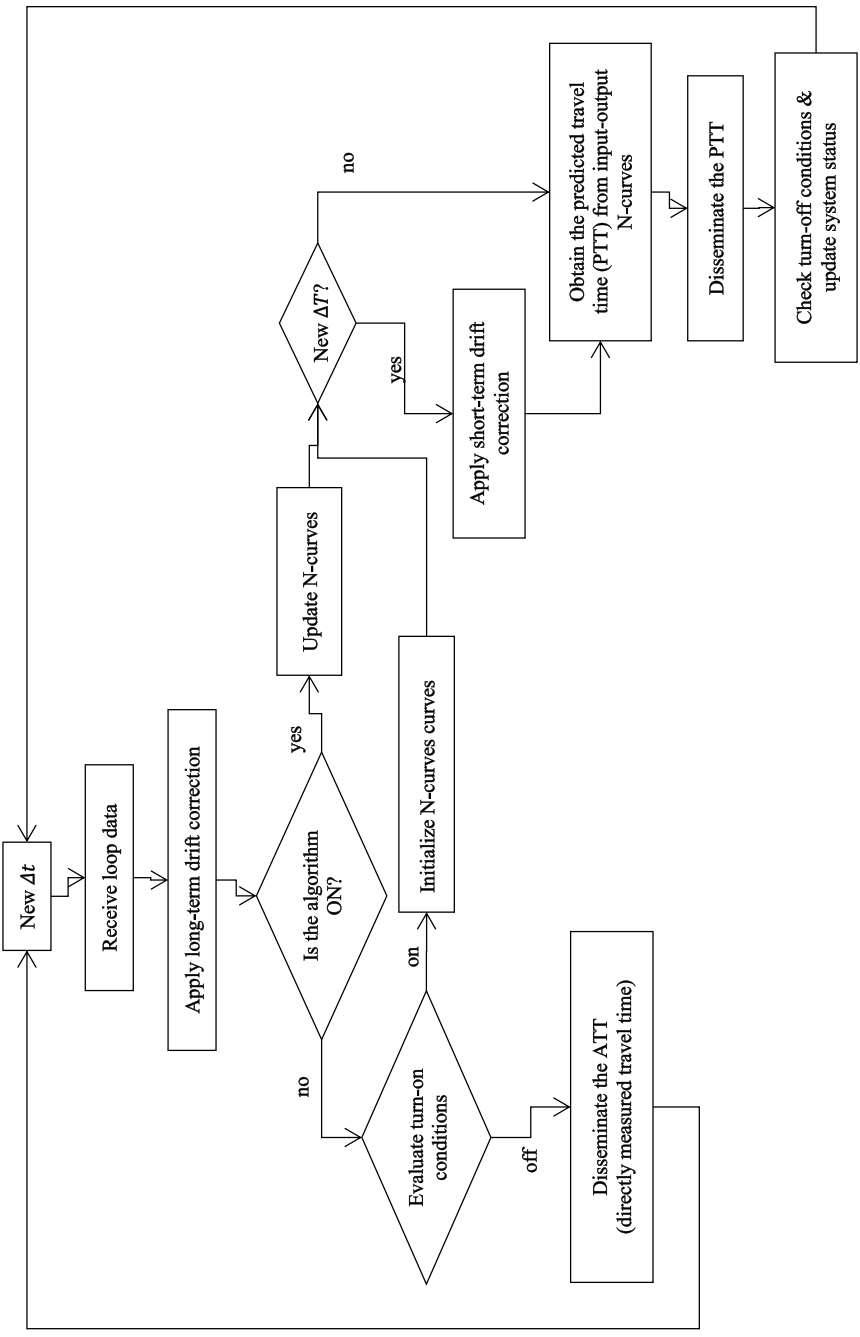

Fig. 4.6 Steps of a data fusion algorithm for the short-term prediction of freeway travel times using input–output cumulative count curves

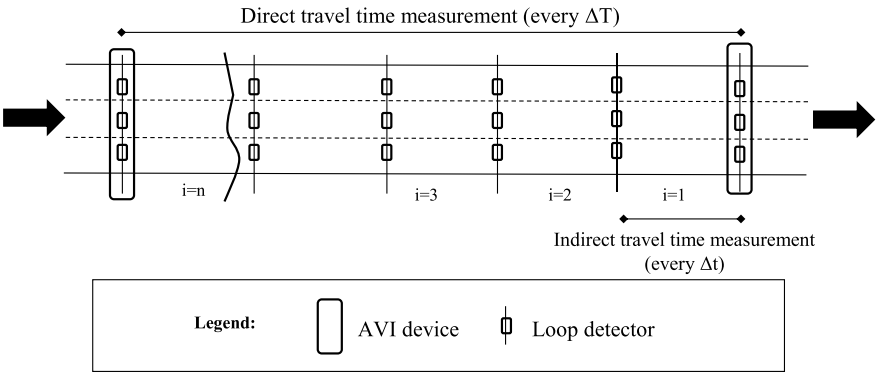

Fig. 4.7 Typical freeway surveillance layout

4.3.2 The Short-Term Travel Time Prediction

From input–output cumulative count curves, the vehicle excess accumulation, $Q_i(t)$, can be obtained (e.g. see Eq. 4.9 and Fig. 4.8). Note that the subscript i refers to a section between loops and t to the instant when a particular Δt time interval ends.

$$Q_i(t) = V_i^*(t) - D_i(t) \tag{4.9}$$

$V_i^*(t)$ defines the virtual arrivals cumulative count curve and $D_i(t)$ the departures curve at the downstream detector of section i. The * superscript in $V_i^*(t)$ indicates that this curve is corrected for the detector drift. The correction procedure is addressed in the next section.

As commented before, vehicle accumulation exhibits predictive capabilities because a vehicle entering i just after t will have $Q_i(t)$ vehicles in front of it until it can be served. Thus, the predicted delay, $pw_i(t)$, is computed according to Eq. 4.10,

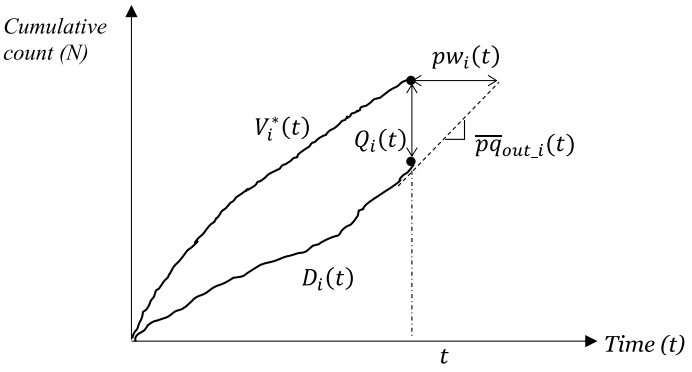

Fig. 4.8 Predicted delay from input–output cumulative count curves

where $\overline{pq}_{out\_i}(t)$ is the predicted average outflow from section i just at t.

$$pw_i(t) = \frac{Q_i(t)}{\overline{pq}_{out\_i}(t)} \tag{4.10}$$

$\overline{pq}_{out\_i}(t)$ is estimated assuming that traffic conditions will not change in the immediate future. This is the best one can do by only using real-time information. So, $\overline{pq}_{out\_i}(t)$ should represent the average outflow of the current traffic state.

Because traffic is a random process, the measurement of the outflow is subject to statistical fluctuations. This means that the robustness of the estimation increases with the sample size or, equivalently, with the time considered. Thus, by computing $\overline{pq}_{out\_i}(t)$ as the average outflow considering several Δt (i.e. the last ones), the statistical significance of the average would increase. However, by doing this, the algorithm would be slower to react to changes in the outflow, which should be a relevant property of the method. Therefore, a trade-off exists between a more robust estimation that smooths out the very last changes in traffic conditions, or a more volatile one that is sensible to them.

Finally, the predicted travel time on section i, $ptt_i(t)$, is obtained by adding the free flow travel time, tt_{f_i} to the predicted delay (Eq. 4.11) and the total predicted travel time on the target stretch is simply the sum of the travel times on all the enclosed sections (i.e. $i = 1, 2, \ldots n$) (Eq. 4.12).

$$ptt_i(t) = tt_{f\_i} + pw_i(t) \tag{4.11}$$

$$ptt(t) = \sum_{i=1}^{n} ptt_i(t) \tag{4.12}$$

The benefits of the proposed approach to deliver real-time travel time information on the target freeway stretch with respect to simply disseminating the direct measurements are multiple. First, and most importantly, a predicted travel time (PTT) is obtained instead of the delayed ATT. The latter would be completely flawed as real-time information in travel time evolving conditions. Note that this is especially important if the target stretch (i.e. the distance between AVI devices) is long. Second, the real-time information can be updated more frequently, as $\Delta t < \Delta T$. Finally, partial travel times within the stretch (i.e. in sections between detectors) can be obtained.

4.3.3 The Data Fusion Method to Correct Detector Drift

Typically, detector drift between a pair of loop detectors that define a closed section is corrected by imposing that, over the long term, inputs must be equal to outputs, provided that the vehicle accumulation has not changed significantly. In this context,

the long-term drift correction factor, β_i, in a section i delimited by its upstream, i_u, and downstream, i_d, detectors, is defined as in Eq. 4.13:

$$\beta_i = \frac{\sum_{t=0}^{long\ term} \widehat{n}_{i_d}(t)}{\sum_{t=0}^{long\ term} \widehat{n}_{i_u}(t)}, \tag{4.13}$$

where $\widehat{n}_{i_u}$ and $\widehat{n}_{i_d}$ are respectively the raw counts at the upstream and downstream detectors, and taking into account that a minimum of 24 h is usually considered by "long term". Then, the corrected upstream count is computed as (Eq. 4.14):

$$n_{i_u}(t) = \beta_i * \widehat{n}_{i_u}(t). \tag{4.14}$$

Note that the correction factor is only applied to arrivals (i.e. to the counts at the upstream detector), while departures (i.e. counts at the downstream detector) are assumed correct. The opposite approach would lead to identical results. This assumption is maintained throughout the chapter so that departure cumulative curves are considered correct, and only (virtual) arrival curves are rectified to account for the drift.

However, this long-term drift correction is not enough for the travel time estimation from input–output cumulative curves. The drift factor typically varies in the short term, especially in congested conditions, so that the long-term average does not suffice to correct the curves. Henceforth, long-term drift correction is considered as the default one and a short-term correction needs to be applied on top of that.

Direct travel time measurements are used to correct the short-term drift in loop detector counts. The concept of the data fusion scheme is simple: the accurate (but delayed) direct measurement is compared to an equivalent estimation obtained from the input–output diagram. From this comparison, the virtual arrivals curve, $V_i(t)$, is modified until both travel time estimations coincide. The corrected virtual arrivals curve, $V_i^*(t)$, is then used to compute the PTT (see Sect. 4.3.2).

The main difficulty of the method consists in obtaining equivalent estimations so that they are comparable. To this end, measurements need to be aligned temporally and spatially. Consider $\widehat{tt}(T)$ to be the direct ATT measurement obtained at T (i.e. at the end of a ΔT time period) for a particular freeway stretch composed of i sections ($i = 1, 2, \ldots, n$). The "$\wedge$" notation will be used to describe directly measured variables. $\widehat{tt}(T)$ can be expressed as the free flow travel time, tt_f, plus the total delay in the stretch, $\widehat{w}(T)$ (see Eq. 4.15):

$$\widehat{tt}(T) = tt_f + \widehat{w}(T). \tag{4.15}$$

An alternative estimation for this total delay, $w(t)$, must be obtained from input–output cumulative curves. To that end, $w(t)$ is simply the addition of the partial delays obtained at the different sections i that compose the target stretch (Eq. 4.16):

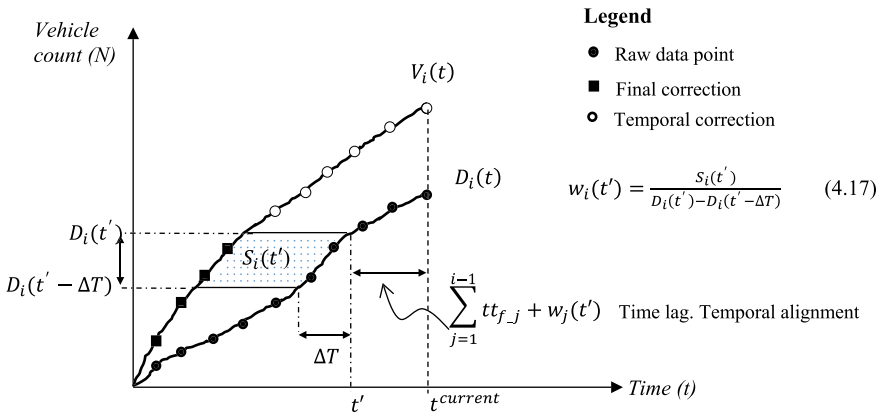

Fig. 4.9 Estimation of arrival-based delays on the section $i (i \in 1 \div n)$ from input–output cumulative curves

$$w(t) = \sum_{i=1}^{n} w_i(t'). \tag{4.16}$$

Because direct measurements are ATT, $w_i(t')$ must be the arrival-based average delays computed from input–output curves (see Fig. 4.3a). In addition, time moving coordinates (denoted with an accent, e.g. t') are needed because ATT is a trajectory-based measurement and, thus, a temporal alignment is necessary for sections different than the most downstream one. Figure 4.9 and Eq. 4.17 describe the computation of $w_i(t')$ from input–output curves. Note that the computation process must start from the most downstream section (i.e. $i = 1$) for which the time alignment is met by default (i.e. $t = t'$) and the time lag is null.

From Eqs. 4.16 and 4.17, $w(t)$ is obtained and can be compared to the equivalent but normally more accurate (even delayed) $\widehat{w}(T)$ in order to correct the drift in the cumulative curves. Note that this comparison can only be updated every ΔT (i.e. when $t' = T'$) and, consequently, this will be the update period for the short-term drift correction factor. The problem with this comparison is that only the total delay difference for the whole stretch is obtained so that the errors cannot be assigned to any particular detector.

In order to solve this problem, it is assumed that the error in the total delay is shared between the n sections that compose the stretch proportionally to the raw partial delays (Eq. 4.18). This assumption allows obtaining an estimation for $\widehat{w}_i(T')$, the direct measurement travel time estimation for section i.

$$\hat{w}_i(T') = \frac{w_i(T')}{\sum_{j=1}^{n} w_j(T')} * \hat{w}(T). \tag{4.18}$$

Then, every ΔT, two equivalent estimations for the sectional delay are available. $w_i\left(T'\right)$ from input–output curves, which includes the drift error, and $\widehat{w}_i(T')$, an accurate estimation from direct measurements. With this information, the virtual arrivals curve, $V_i(t)$, is modified to obtain the corrected $V_i^*(t)$ according to Eq. 4.19:

$$V_i^*(t) = V_i^*(t - \Delta t) + \alpha_i\left(t'\right) * n_{i_u}(t) \ \forall t \in \left(t' - \Delta T, t^{current}\right), \tag{4.19}$$

where $\alpha_i(t')$ is the short-term drift correction factor, which is estimated imposing that $w_i\left(T'\right) = \widehat{w}_i(T)$. A small tolerance, τ, for the final difference must be set to ensure a fast convergence to the solution. Note that the correction is applied to the different counts existing between $t' - \Delta T$ and $t^{current}$. However, two types of corrections are defined. If $t \in (t' - \Delta T, t')$, the correction is said to be final and, thus, will not be modified in future corrections. Otherwise, if $t \in (t', t^{current})$, the correction is temporal and could be modified in next iterations.

4.3.4 *Simpler Process if the Direct Travel Time Measurements Are ITT*

As remarked in Sect. 4.3.1, some of the aforementioned steps slightly vary when fusing ITT instead of ATT. Particularly, the process is simpler because the temporal and spatial alignment is not necessary if direct travel time measurements are ITT (e.g. from the tracking of the trajectories of a fraction of vehicles). Every ΔT (which is generally longer than the loop detector measurement interval, Δt), a directly measured ITT, $\widehat{tt}_i(t)$, is obtained from GPS tracking data. To this end and provided that each GPS measurement provides, at least, the vehicles' position and timestamp, the GPS measurements are filtered to only consider vehicles with two or more data points on section i and time interval $(T - \Delta T, T)$. Next, for each time interval, the travel time of vehicle j in section i, $\widehat{tt}_{ij}(T)$, is obtained as the difference in timestamps between the last and first GPS measurements of this vehicle registered in the database. Similarly, the distance covered by vehicle j, $\widehat{d}_{ij}(T)$, is obtained from the difference in its registered positions. Then, $\widehat{tt}_i(t)$, is obtained with Eq. 4.20:

$$\widehat{tt}_i(T) = \Delta x_i * \frac{\sum_{\forall j} \widehat{tt}_{ij}(T)}{\sum_{\forall j} \widehat{d}_{ij}(T)}. \tag{4.20}$$

Using Eq. 4.15, $\widehat{w}_i(T)$ can be directly measured and compared to an ITT estimation of the delay, $w_i(T)$, obtained from input–output curves as described in Fig. 4.3c and in Eq. 4.3. Then, the short-term drift correction factor (as proposed in Eq. 4.19) can be directly computed.

The vocation of continuity of the method is thus demonstrated. The use of arrival-based direct measurements coming from AVI technologies requires the deployment

of special surveillance on the highway. On the contrary, intelligent vehicles will provide ITT without the need for extra support. Therefore, taking into account the aforementioned abundance of loop detectors in the road networks, the possibility of applying this methodology is also guaranteed in future environments, in this case alongside more advanced ones.

4.3.5 The Algorithm Turn-On and Turn-Off Conditions

In free flowing traffic, there is no need to use the input–output cumulative curves to compute travel times. In free flow, either the direct travel time measurements, or travel times based on the punctual speed measurements at loop detectors, or even a fixed estimation of the free flow travel time in the stretch, would suffice to feed the real-time information system with enough accuracy. It is when congestion and delays appear that the predicted travel time information is meaningful, and when none of the previous methods is appropriate.

Therefore, while free flowing conditions prevail, delays and excess accumulation are assumed to be null, and the data fusion algorithm is turned off. This also allows its required reset. The algorithm should turn on just before the appearance of delays. With this objective, turn-on conditions focus on the quick detection of the congestion onset. Note that these conditions cannot be based on direct travel time measurements because this would delay the detection of the congestion onset.

Three turn-on conditions are defined based on loop detector measurements at i_u and i_d, the upstream and downstream detectors of section i (see Eqs. 4.21–4.23). The fulfilment of any of them suffices to activate the algorithm at the instant $t - \Delta t$, from which cumulative curves are initialized with null excess accumulation, as described in Sect. 4.2.2.4.

$$\lceil v_{i_u}(t) \rceil \leq v_{ref} \tag{4.21}$$

$$\lceil v_{i_d}(t) \rceil \leq v_{ref} \tag{4.22}$$

$$\frac{\partial V_i(t)}{\partial t} > \frac{\partial D_i(t)}{\partial t} \tag{4.23}$$

Equations 4.21 and 4.22 detect congestion at the detector locations by measuring a speed, v_{i_u} or v_{i_d}, lower than a given threshold, v_{ref}. v_{ref} should be calibrated in any particular application of the algorithm, but generally a low percentile of the free flow speed distribution suffices (e.g. 1st quartile of the distribution). The $\lceil \rceil$ brackets in Eqs. 4.21 and 4.22 indicate that an upper bound of the speed estimation is considered, in order to account for the statistical fluctuations in the speed measurement. This upper bound can be computed as the higher limit of a confidence interval of the speed estimation, as in Eq. 4.24:

$$\lceil v_i(t) \rceil = v_i(t) + prob.level * \frac{cv_v(t)}{\sqrt{n_i(t)}}, \tag{4.24}$$

where cv_v is the coefficient of variation of the speed measurements (i.e. the standard deviation over the mean, which can be obtained from a pre-sample) and n_i is the vehicle count in the time period considered (i.e. the sample size). The $prob.level$ defines the confidence of the interval (e.g. 68% for a $prob.level = 1$).

Equation 4.23 detects congestion within the section by measuring that inflows, $\frac{\partial \lfloor V_i(t) \rfloor}{\partial t}$ are higher than outflows, $\frac{\partial \lceil D_i(t) \rceil}{\partial t}$ (i.e. growing accumulation). Note that Eq. 4.22 is equivalent to $\frac{\partial Q_i(t)}{\partial t} > 0$. In order to account for the statistical fluctuations in the flow estimation and to avoid multiple false positives, a lower bound is considered for the inflows, and an upper bound for the outflows. Again, these can be computed from the confidence interval in the flow estimation, as in Eqs. 4.25 and 4.26:

$$\frac{\partial \lfloor V_i(t) \rfloor}{\partial t} = \frac{\partial V_i(t)}{\partial t} - prob.level * \frac{\partial V_i(t)}{\partial t} * \sqrt{\frac{\gamma_q}{n_{i_u}(t)}}, \tag{4.25}$$

$$\frac{\partial \lceil D_i(t) \rceil}{\partial t} = \frac{\partial D_i(t)}{\partial t} + prob.level * \frac{\partial D_i(t)}{\partial t} * \sqrt{\frac{\gamma_q}{n_{i_d}(t)}}, \tag{4.26}$$

where γ_q is the index of dispersion of the flow estimation (i.e. the ratio of the variance with respect to the mean), which can be computed from a pre-sample or simply assumed as $0.2 \div 0.3$, as these are typical values in freeway traffic.

For its part, there is only one turn-off condition. This is based on the achievement of null excess accumulation (see Eq. 4.27), noting that $\sqrt{\gamma_{q*n_i(t)}}$ is the statistical variability of the flow estimation for a 68% confidence level.

$$Q_i(t) = V_i^*(t) - D_i(t) < Min\left[\sqrt{\gamma_q * n_{i_u}(t)}, \sqrt{\gamma_q * n_{i_d}(t)}\right]. \tag{4.27}$$

4.4 Implementation of the Algorithm with Real AVI Data on the AP7 Freeway in Spain

This section is aimed at demonstrating the goodness and applicability of the algorithm with common data managed by traffic management centres.

4.4.1 Layout, Available Data and Considered Parameters

The data used to test the algorithm was measured on April 25th, 2010, in a 3-lane stretch of the AP7 freeway in its southbound direction towards Barcelona, Spain.

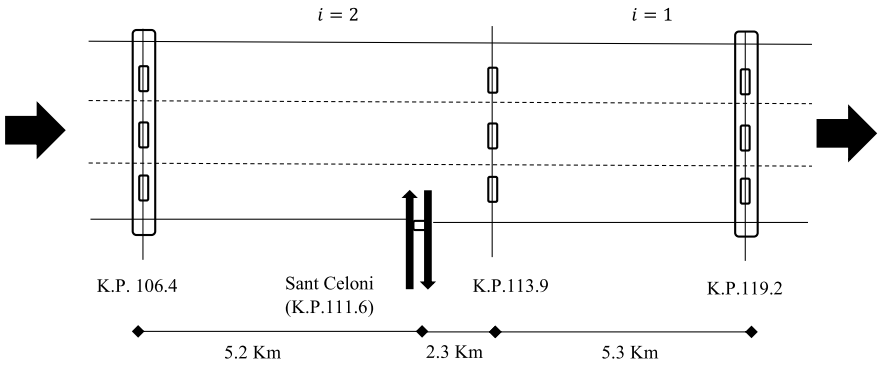

Fig. 4.10 Test site layout

This was a sunny Sunday with some light evening congestion in the AP7, due to the large demand returning the city after the weekend. The layout of the test site is illustrated in Fig. 4.10 and has not varied in the last years. AVI devices (Bluetooth detectors) at both ends limit the target stretch, with a length of 12.8 km. In addition, there are double-loop detectors at the same location, plus one more in between (i.e. at Kilometer Post K.P. 113.9), which divides the stretch into two sections ($i = 1, 2$). Section $i = 2$ (i.e. the upstream one) contains the St. Celoni junction at K.P. 111.6, near its downstream end. Entrances and exits through this junction are monitored. This layout configuration is adequate to test the goodness of the proposed method to obtain short-term travel time predictions.

Vehicle counts, n, and time-mean speeds, v, are available for time aggregations of $\Delta t = 3$ min at loop detector locations. In addition, the net input counts at the junction are also computed for the same Δt. These are transferred to the nearest detector (i.e. at K.P. 113.9). Because this is the downstream detector of section $i = 2$, the junction's net input counts must be subtracted from the detector counts at this location (see Sect. 4.2.2.2).

Average AVI direct travel time measurements are available every $\Delta T = 6$ min. In spite of this, and for the only purpose of this research, individual vehicles' travel time measurements were recorded. These data allowed obtaining the ground truth predicted travel time, $\widehat{gtt}(t)$, defined as the departure-based travel time of those vehicles entering the target stretch between t and $t + \Delta t$. Note that these vehicles are the ones that will receive the real-time information at t. All the calibration parameters of the method are summarized in Table 4.1.

4.4.2 Obtained Results and Discussion

Figure 4.11 shows the performance of the proposed method for the evening congestion episode, between 6 and 10 pm. Table 4.2 summarizes the maximum and mean

Table 4.1 Parameters considered in the application of the method to the AP7 freeway near Barcelona, Spain

Variable description	Notation	Uuits	Value	Comments
Loop detector aggregation period	Δt	[min]	3	Typical aggregation period in Spanish toll highways
AVI travel time updating period	ΔT	[min]	6	From Bluetooth recognition devices. Required to obtain a significant sample size according to the on-board Bluetooth penetration rate
Free flow travel time	tt_{f_i}	[min]	$i = 1 \rightarrow 2.28$ $i = 2 \rightarrow 3.30$	i refers to the section. The free flow travel time is obtained considering the 95% percentile of the observed speeds. The free flow travel time over the entire stretch is $tt_f = 5.58$ min
Long-term drift correction factor	β_i	[–]	$i = 1 \rightarrow 0.9989$ $i = 2 \rightarrow 0.9870$	i refers to the section. Considering cumulative counts over a 24 h period
Speed threshold for the activation of the algorithm	v_{ref}	[km/h]	$i = 1 \rightarrow 102.15$ $i = 2 \rightarrow 100.50$	i refers to the section. Computed as the 60% percentile of the speed distribution. If only free-flow periods are considered, the speed percentile for v_{ref} would be much lower (e.g. 1st quart.)
Coefficient of variation of the speed distribution	σ_v/v	[–]	0.119	Required to compute the confidence interval in the speed estimation to evaluate the algorithm's turn-on conditions
Index of dispersion of the flow estimation	γ_q	[km/h]	0.25	Variance-to-mean ratio (VMR) for the flow estimations. Required to compute the confidence interval in the flow estimation to evaluate the algorithm's turn-on / -off conditions

(continued)

Table 4.1 (continued)

Variable description	Notation	Uuits	Value	Comments
Confidence interval significance	$prob.level$	–	1	Corresponding to a 68% statistical confidence of the interval. Required to compute the confidence intervals in the turn-on/ turn-off conditions
Average predicted outflow moving average intervals	–	–	7	Number of Δt time intervals considered in the moving average to compute the predicted outflow, $\overline{pq}_{out}$
Tolerance in the computation of the short-term drift correction factor	τ	[s]	9	Accepted difference between $w_i(T')$ (from input–output N-curves) and $\widehat{w}_i(T)$ (from AVI direct measurements) in the computation of the short-term drift correction factor, $\alpha_i(t')$

absolute errors.

The results shown in Table 4.2 and Fig. 4.11 demonstrate that the proposed algorithm satisfactorily accomplishes its objectives and provides drivers with a better prediction of their travel times over the freeway stretch, improving the information given by the dissemination of directly measured travel times. Figure 4.11 clearly shows the delay of AVI direct measurements in predicting travel times evolution. This implies large errors when travel times change rapidly, especially at congestion dissolve episodes. In contrast, short-term predicted travel times are able to respond quicker, providing better travel time predictions to drivers, especially by the reduction of maximum errors in rapid evolving conditions.

However, short-term predictions still show some delay. This is due to the adoption of a long averaging period in the computation $\overline{pq}_{out}$ (i.e. 7 * Δt; Table 4.1).

As discussed in Sect. 4.3.2, longer averaging periods imply more delay, in exchange for a more robust estimation. Shorter averaging periods would reduce this delay in the short-term predictions, but would also increase their fluctuations. For this particular application of the method, the 7 * Δt selection turned out to be the optimal in terms of reducing the average and maximum errors.

In fact, it has been found in this case that the fluctuations in the predicted travel times were much larger than expected. Some fluctuations remain even considering a long averaging interval in the computation of $\overline{pq}_{out}$ (e.g. the underestimation around 19:15, or the overestimation around 19:45). Further analysis of these periods unveiled that these fluctuations were related to abnormal net counts at the junction, representing a significant fraction of the main trunk detector counts. Whether the junction

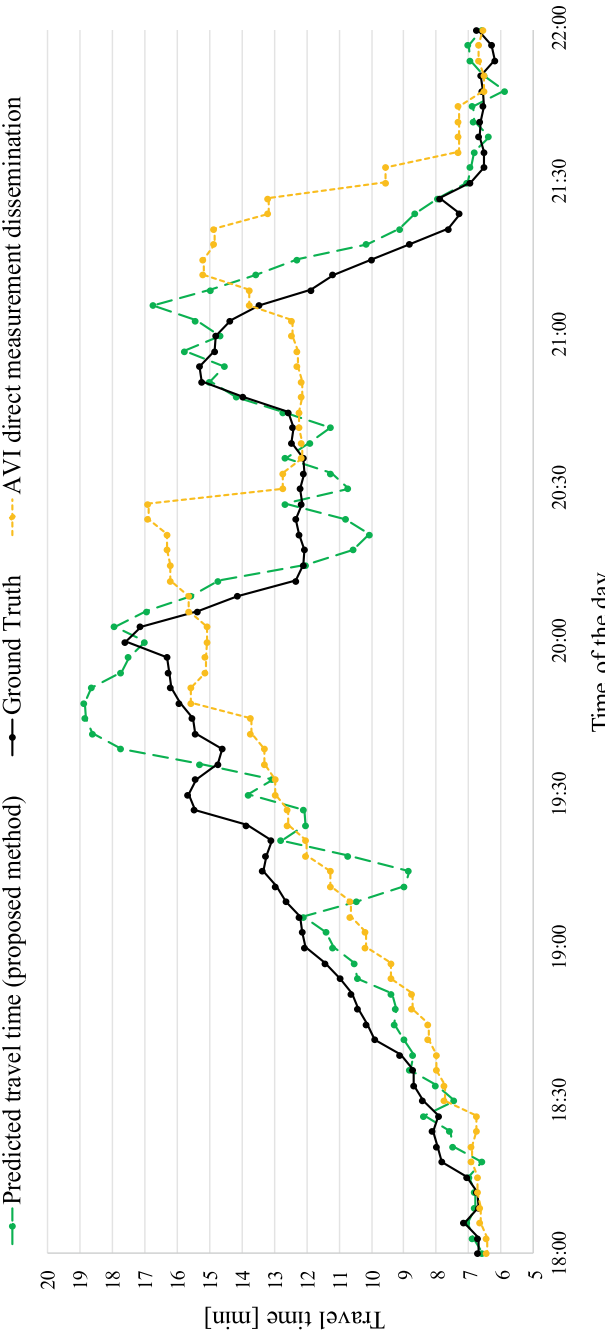

Fig. 4.11 Comparison of the performance as real-time information of predicted travel times versus direct measurements

Table 4.2 Comparison of the errors of predicted travel times versus direct measurements with respect to ground truth predicted travel times

Method	Mean absolute error	Max. error	Units
Predicted travel time (proposed method)	1.16	4.5	[min]
AVI direct measurement dissemination	1.81	−7.23	[min]

counts are accurate or not during these periods is unknown. In any case, these results highlight that the proposed method is sensitive to large changes in the vehicle counts. Therefore, while the method handles adequately small random drift errors in the loop detectors, it will produce bad estimations in case of detector failure (e.g. partial data loss for some period). While this fact needs to be considered, it is not dramatic for the application of the methodology. Note that, still, it manages to produce an estimation of the predicted travel time and that, if the target section is composed of several sections, the relative importance of the error in one of them will be lower, provided that not all the detectors fail at the same time. Instead, drift sooner or later affects all detectors.

Another issue that implies fluctuations is the assumption of the net junction counts taking place at the downstream detector of section $i = 2$. Note that this affects the computation of $\overline{pq}_{out}$, and the predicted travel time is very sensitive to this average outflow (see Eq. 4.10). Even if there is a small distance between the junction and the detector, in congested conditions the time lag between both counts, which are subtracted every Δt (see Eq. 4.8), can be a significant fraction of Δt. In addition, the loop detector outflow in congested conditions can be small so that the junction counts could be a significant fraction of it. All this together might cause fluctuations in the predicted travel time. A possible solution to reduce this type of errors is to avoid considering junction counts in the computation of $\overline{pq}_{out}$ if loop detector counts are small.

Finally, it is worth mentioning that the errors of the method are invariant with respect to the total delay. Therefore, in case of more severe congestion, the relative errors committed would be lower. In contrast, the errors of directly disseminating AVI information would grow proportionally with the delay. It must be additionally highlighted that the benefits of using the method would also be more significant if the AVI detectors were located further apart. In fact, this is the case in most freeways in Spain and around the world.

4.5 Implementation of the Algorithm with Simulated ITT Data

As explained before, in case ITTs from tracking methodologies are fused with counts obtained from loop detectors, the proposed methodology is simplified. The main reason is that no spatial alignment is necessary. Note that, thanks to the trajectories,

it is possible to obtain ITTs between any desired points. Therefore, it is possible to obtain ITT for the sections between loops so that the direct and indirect measures of travel times to be fused are referred to equivalent sections.

4.5.1 Layout, Available Data and Considered Parameters

The case study included in this section has been carried out for the same layout as that of Sect. 4.4 and represented in Fig. 4.10. However, the PTV-Vissim traffic microsimulator generated the data for 3 h of heavy traffic with congestion in this case. Particularly, the data collected from the simulation consisted in the vehicle counts, n, and time-mean speeds, $\overline{v}$, at all detector locations, as well as vehicle counts at the junction, all of them every Δt. For the construction of input–output diagrams, the net vehicle count at the junction was again added to the closer detector (the one at K.P. 113.9). Additionally, the position, timestamp, and speed of a sample of 15% of the vehicles were registered and provided with a frequency ζ. These data are equivalent to that obtained either from GPS or mobile phone tracking.

The fact that only 15% of vehicles were tracked is not casual: the goal was to prove the high degree of applicability of the method, i.e., to prove that its goodness is not linked to a much data demand. This goal was also behind the selection of not demanding ζ: GPS chipsets allow up to 10 Hz signal updates, being 1 Hz the standard (i.e. 1 measurement per second; Martínez-Díaz 2018). However, not all traffic management centres could afford at present the huge storage and computing requirements that these high frequencies would imply. Moreover, different combinations among the typical values of Δt and ΔT were tested in order to analyse their influence on the travel time predictions. Table 4.3 summarizes the different cases analysed regarding the values of ζ, Δt and ΔT. The combination of small values of ζ with short ΔT has been avoided, as it would lead to small samples of vehicles per time interval and, thus, to unreliable ITTs.

Again, a minor number of parameters has been previously calibrated for the case study (Table 4.4).

Table 4.3 Analyzed combinations of values for the GPS frequency and the time intervals of aggregation

GPS frequency ζ (Hz)	Δt (min)	ΔT (min)
1/12	1	1
	1	3
	3	3
1/36	1	3
	3	3
1/60	1	3
	3	3

Table 4.4 Parameters considered in the simulated application of the method to the AP7 freeway near Barcelona, Spain

Variable	Units	Value
Fraction of tracked vehicles	[%]	15
Free-flow speed, v_f	[Km/h]	110
Free-flow travel times, tt_{f_i}	[min]	$i = 1 \rightarrow 2.89$ $i = 2 \rightarrow 4.09$
Turn-on speed threshold, v_{ref}	[Km/h]	80
Duration of the moving average time window to compute $\overline{pq}_{out}$	[min]	5–18 (best calibration for 10)
Variable	Units	Value
Fraction of tracked vehicles	[%]	15

4.5.2 *Obtained Results and Discussion*

The results obtained for all combinations indicated in Table 4.3 are included in Table 4.5. Particularly, the average absolute errors and the maximum errors with respect to the actually experienced travel times that a real-time information system would make providing (i) the short-term predictions estimated with the proposed fusion methodology or (ii) the ITT estimations directly obtained from the GPS sample are shown. It must be again highlighted that ITT measurements and thus, drift correction factors, are obtained per ΔT. However, each Δt a new short-term predicted travel time is estimated. Ground truth travel times at any, available off-line in this case study, were calculated by averaging those travel times really experienced by all vehicles entering the target section per Δt.

A particular example is represented in Fig. 4.12 for identical time intervals of aggregation of 1 min both for the direct and indirect travel time measurements and supposing a sampling frequency for the GPS of 1/12 Hz. For its part, Fig. 4.13 focuses on predicted travel times and represents the different predictions for times interval of aggregation respectively of $\Delta t = 1$ min and $\Delta T = 3$ and all considered GPS frequencies.

From the results in Table 4.5 and the visual example in Fig. 4.12 it is clearly demonstrated that the application of the proposed data fusion methodology is, again, much more advantageous than the current practice of simply disseminating ITT measurements, specially taking into account that it could be set up immediately by most traffic management centres.

A more particular analysis provides very interesting insights. For example, it must be taken into account that the percentage of tracked vehicles (15% in this case), ζ and the duration of ΔT determine the sample of GPS measurements per time interval to compute ITTs. Additionally, ΔT determines the updating frequency of the drift correction factor. With a low percentage of vehicles tracked, a frequent update is more suitable, as it can be derived from in Table 4.5 with better results for short ΔT. However, these short ΔT require a higher frequency of GPS sampling in order

Table 4.5 Comparison of the errors of predicted travel times versus direct measurements with respect to ground truth "predicted" travel times

GPS frequency ζ (s)	Δt (min)	ΔT (min)	Method	Mean absolute error [min]	Mean absolute percentage error [%]	Maximum error [min]	Maximum percentage error [%]
12	1	1	Directly disseminated ITT	22.34	29.53	−40.37	−48.89
			Predicted travel times (data fusion algorithm)	7.48	13.36	−21.86	−27.26
	1	3	Directly disseminated ITT	21.69	28.60	−40.14	−48.62
			Predicted travel times (data fusion algorithm)	12.24	18.56	−28.24	−35.21
	3	3	Directly disseminated ITT	21.66	28.66	−38.50	−46.44
			Predicted travel times (data fusion algorithm)	12.27	19.61	−32.44	−37.76
36	1	3	Directly disseminated ITT	18.12	31.05	48.46	91.83
			Predicted travel times (data fusion algorithm)	12.23	18.50	−28.14	−35.09
	3	3	Directly disseminated ITT	22.19	29.21	−38.35	−46.61
			Predicted travel times (data fusion algorithm)	12.49	18.73	−30.35	−35.32
60	1	3	Directly disseminated ITT	19.06	29.59	48.32	91.57

(continued)

Table 4.5 (continued)

GPS frequency ζ (s)	Δt (min)	ΔT (min)	Method	Mean absolute error [min]	Mean absolute percentage error [%]	Maximum error [min]	Maximum percentage error [%]
			Predicted travel times (data fusion algorithm)	8.77	15.05	−21.88	−28.57
	3	3	Directly disseminated ITT	23.21	30.48	−38.80	−46.79
			Predicted travel times (data fusion algorithm)	12.79	19.08	−29.16	−33.94

to ensure that enough measurements are available for each time interval. Note that a minimum of $\xi = \frac{1}{12}$ Hz is required for $\Delta T = 1$ min. On the contrary, whether Δt nor ζ play an important role for ΔT equal or longer than 3 min, at least with a minimum percentage of vehicles tracked, like in this study. Regarding input–output method, shorter Δt allow a more frequent update of the travel time information but have no influence on the final accuracy.

It is important to remark that this analysis has proven that working with very high GPS frequencies and, thus, with huge storage, communication and/or computation capacities, is not always necessary (Sanaullah et al. 2016). However, this is usually thought linked to interest raised by Big Data. Without detracting from the advances that such methodologies may imply, it is important to be aware that there are improvements available to all traffic management centres that should not be neglected and that could be implemented immediately.

4.6 Conclusions and Further Research

The proposed travel time estimation methodology is based on the use of input–output cumulative curves to determine the vehicles' accumulation in a freeway section. This means that the method requires "closed" freeway sections, in the sense that all the inflows and outflows must be monitored (i.e. upstream and downstream detectors in the main freeway trunk, as well as inner junctions). From vehicles' accumulation, predicted travel times are computed using the principle of the vehicles' conservation.

The derivation of travel times from cumulative count curves is not new. However, input–output methods have not been used in practice so far, mainly due to the flawed results caused by the loop detector drift. In this sense, the main contribution of the present book is the design of a data fusion scheme aimed at correcting the detector

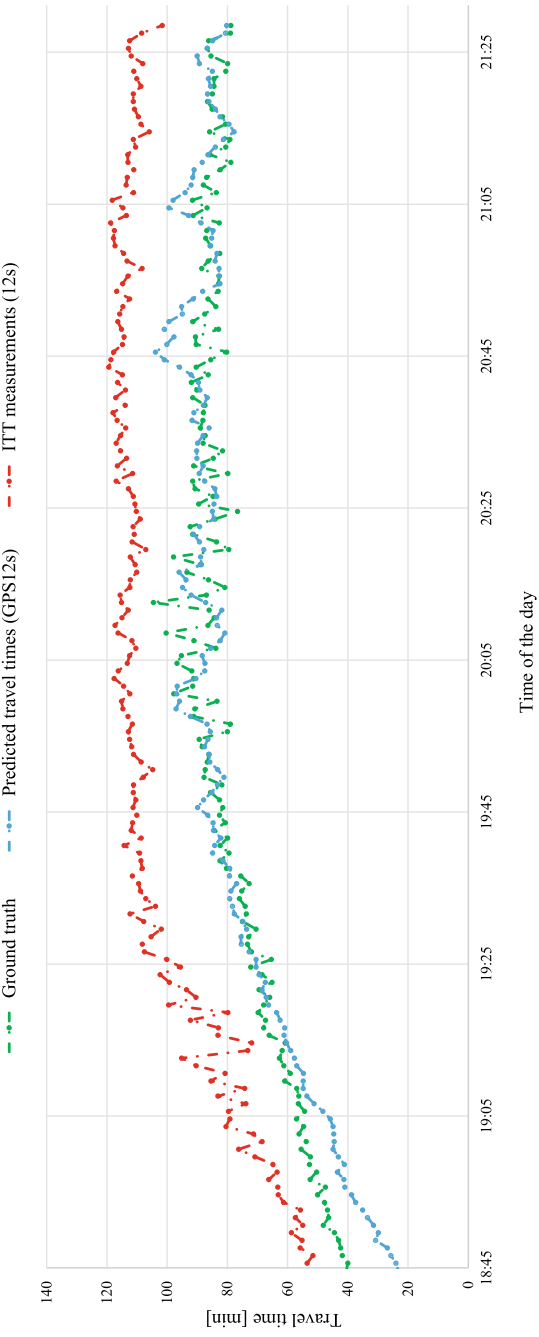

Fig. 4.12 Comparison of the suitability as real-time information of the directly obtained ITT travel times versus the predicted travel times for the case $\zeta = 1/12$ Hz, $\Delta t = 1$ min and $\Delta T = 1$ min

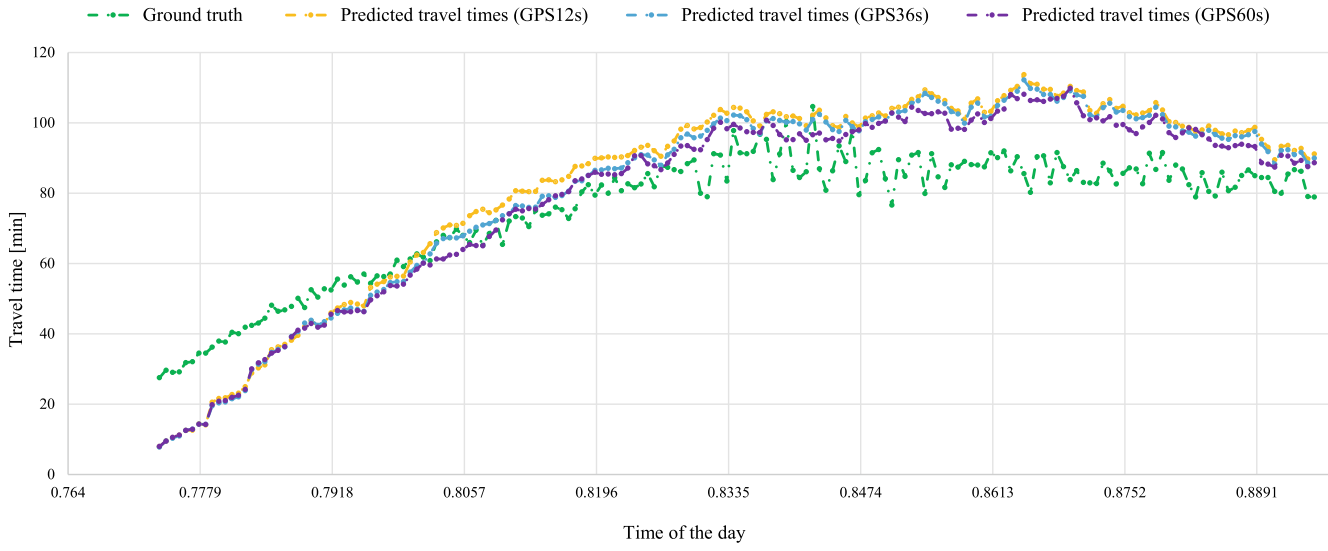

Fig. 4.13 Comparison of the Predicted travel times obtained for the times interval of aggregation $\Delta t = 1$ min and $\Delta T = 3$ min and for GPS frequencies of respectively $\zeta = 1/12$ Hz, $\zeta = 1/36$ Hz and $\zeta = 1/60$ Hz, all of them versus ground truth travel times

drift in cumulative count curves. Direct travel time measurements, supplied either by AVI systems or tracking devices, are used to this end.

The benefits of the proposed methodology with respect to the simple dissemination of directly measured travel times include a quicker detection of travel time changes, a higher updating frequency of the information and, for the case of AVI systems, the possibility of obtaining partial travel times for sections within those delimited by AVI devices. These properties greatly contribute to reduce the errors of the disseminated information, especially the maximum errors that arise in rapidly evolving traffic conditions (i.e. congestion onset/dissolve episodes). The benefits are even larger in situations of severe congestion with large delays and, for the case of ATT, if freeway stretches between AVI devices are long. In free flowing conditions, the proposed method could be turned off, as all types of travel time estimations suffice (e.g. direct measurements, or travel times estimated from spot speed measurements at loop detectors).

The methodology has first been tested with real data from the AP7 freeway towards Barcelona (Spain). Travel times from Bluetooth detectors have been used as direct measurements. Results show that, with respect to directly disseminating the ATT measured with the Bluetooth devices, the predicted travel times obtained with the proposed methodology better match the real travel times that drivers receiving the information will experience. These benefits could be observed even in a scenario with light congestion, despite the methodology being especially suited for medium to severe congestion episodes. With the available real data, the mean and maximum absolute errors achieved with the predicted travel times (i.e. 1.16 min and 4.5 min respectively) represented approximately 10% and 33% of the experienced travel times. In contrast, the simple dissemination of direct measurements would imply that these relative errors would reach 16% and 95%.

The maximum errors of the proposed methodology in this particular application resulted from fluctuations in the predicted travel times, which could be attributed to detector malfunctioning or data loss. This issue highlights that the method deals well with the random detector drift (which invariably affects all detectors) but is sensitive to more severe detector failures. In this respect, the author would like to design a complementary algorithm to identify detector malfunctioning so that the lost data could be replaced. In such situation, the method would be more accurate, not only by reducing the maximum errors, but also because the calibration parameters could then prioritize immediacy in reporting travel time changes with respect to smoothing artificial fluctuations.

Using the same layout, the methodology has also been applied to simulated data. In this case, the drift correction of the cumulative curves was performed using ITT direct measurements estimated in a conservative scenario with only 15% of tracked vehicles and for a situation with heavy congestion. Additionally, different combinations of time intervals of aggregation for the loops and for the GPS measurements, as well as several GPS sample frequencies, were tested. Again, the predicted travel times were much more accurate in all the scenarios when compared with the direct travel time measurements that are usually disseminated. In fact, the MAPE was approximately reduced to half (i.e. from 30 to 15%). Furthermore, the analysis confirmed that

whether a high frequency of GPS sampling nor very long time intervals for the GPS measurements are necessary to achieve good results. This fact reinforces the premise that the proposed method could be used immediately by most traffic management centres without the need for large investments.

In addition to the mentioned side algorithm to account for possible loop data loses, another interesting topic for future research would be the performance of a comprehensive sensitivity analysis of this methodology, aimed at optimizing its transferability. Factors such as the level of congestion from which the method is worthwhile, the optimal arrangement of AVI detectors when fusing ATT, the most suitable section lengths, the optimal percentage of tracked vehicles in the case of ITT fusion or the treatment of on-/off-ramps and their relative flows, among others, should be assessed. Next, further integration of the method with AI to address particular traffic management challenges in the era of autonomous driving would be another ambitious goal to keep in mind. Because of the method's simple fundamentals and non-demanding data requirements, it could be an effective complement and backstop to more smart and complex systems.

References

Abdulhai B, Tabib SM (2003) Spatial-temporal inductance-pattern recognition for vehicle reidentification. Transp Res Part C: Emerg Technol 11(3–4):223–239

Ambühl L, Menéndez M (2016) Data fusion algorithm for macroscopic fundamental diagram estimation. Transp Res Part C: Emerg Technol 71:184–197

Barceló J, Montero L, Marquès L (2010) Travel time forecasting and dynamic OD estimation in freeways based on Bluetooth traffic monitoring. In: Proceedings of the 89th annual meeting of the transportation research board, 10–14 January 2010, Washington D.C.

Buisson C (2006) Simple traffic model for a simple problem: sizing travel time measurement devices. Transportation Research Record: Journal of the Transportation Research Board 1965:210–218

Chen H, Rakha HA (2016) Multi-step prediction of experienced travel times using agent-based modelling. Transp Res Part C: Emerg Technol 71:108–121

Coifman B (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. Transp Res Part A: Policy Pract 36(4):351–364

Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. ASCE J Transp Eng 129(5):475–483

Coifman B, Cassidy M (2002) Vehicle reidentification and travel time measurement on congested freeways. Transp Res Part A: Policy Pract 36(10):899–917

Coifman B, Krishnamurthya S (2007) Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. Transp Res Part C: Emerg Technol 15(3):135–153

Daganzo CF (1987) Increasing model precision can reduce accuracy. Transp Sci 21:100–105

Daganzo CF (1997) Fundamentals of transportation and traffic operations. Pergamon, Oxford

Daganzo CF (2010) Public transportation systems: basic principles of system design, operations planning and real-time control. Institute of Transportation Studies, University of California, Berkeley, California

Deng W, Lei H, Zhou X (2013) Traffic state estimation and uncertainty quantification based on heterogeneous data sources: a three detector approach. Transp Res Part B: Methodol 57:132–157

El Faouzi N, Maurin M (2007) Reliability of travel time under log-normal distribution: methodological issues and path travel time confidence derivation. In: Transportation research board 86th annual meeting (CD- ROM). Transportation Research Record, Washington D.C.

Ge Q, Fukuda D (2016) Updating origin-destination matrices with aggregated data of GPS traces. Transp Res Part C: Emerg Technol 69:291–312

Hall RW (1993) Non-recurrent congestion: how big is the problem? Are traveler information systems the solution? Transp Res Part C: Emerg Technol 1(1):89–103

Herrera JC, Bayen AM (2010) Incorporation of Lagrangian measurements in freeway traffic state estimation. Transp Res Part B: Methodol 44:460–481

Herrera JC, Work D, Ban X, Herring R, Jacobson Q, Bayen A (2010) Evaluation of traffic data obtained via GPS-enabled mobile phones: the mobile century field experiment. Transp Res Part C: Emerg Technol 18:568–583

Kaufmann S, Kerner BS, Rehborn H, Koller M, Klenov SL (2018) Aerial observations of moving synchronized flow patterns in over-saturated city traffic. Transp Res Part C: Emerg Technol 86:393–406

Kuhne R, Immes S (1993) Freeway control systems using section-related traffic variable detection. In: Proceedings of pacific rim transportation technology conference, vol 1. ASCE, pp 56–62

Kwon TM (2006) Blind deconvolution processing of loop inductance signals for vehicle reidentification. In: Proceedings of the 85th annual meeting of the transportation research board, January 22–26 2006, Washington D.C.

Longfoot J (1991) An automatic network travel time system-ANTSS. In: Proceedings of the 1991 vehicle navigation and information systems conference, vol 2, 20–23 October 1991, Dearborn, Michigan, pp 1053–1061

Martínez-Díaz M (2018) Highway travel time information systems: from traditional to cooperative driving environments. PhD Dissertation, University of A Coruña (Spain)

Moskowitz K (1954) Waiting for a gap in a traffic stream. Proc Highway Res Board 33:385–395

Muñoz JC, Daganzo CF (2002) The bottleneck mechanism of a freeway diverge. Transp Res Part A: Policy Pract 36(6):483–505

Nam DH, Drew DR (1996) Traffic dynamics: method for estimating freeway travel times in real time from flow measurements. ASCE J Transp Eng 122(3):185–191

Nantes A, Ngoduy D, Bhaskar A, Miska M, Chung E (2016) Real-time traffic state estimation in urban corridors from heterogeneous data. Transp Res Part C: Emerg Technol 66:99–118

Newell GF (1982) Applications of queuing theory, 2nd edn. Chapman and Hall, London

Newell GF (1988) Theory of highway traffic signals. Institute of Transportation Studies, UC Berkeley

Newell GF (1993a) A simplified theory of kinematic waves in highway traffic. Part I: General theory. Transp Res Part B: Methodol 27(4):281–287

Newell GF (1993b) A simplified theory of kinematic waves in highway traffic. Part II: Queuing at freeway bottlenecks. Transp Res Part B: Methodol 27(4):289–303

Newell GF (1993c) A simplified theory of kinematic waves in highway traffic. Part III: Multi-destination flows. Transp Res Part B: Methodol 27(4):305–313

Newell GF (1999) Delays caused by a queue at a freeway exit ramp. Transp Res Part B: Methodol 33(5):337–350

Nishiuchi H, Nakamura K, Bajwa S, Chung E, Kuwahara M (2006) Evaluation of travel time and OD variation on the Tokyo metropolitan expressway using ETC data. In: Research into practice: 22nd ARRB conference proceedings information, 29 October to 2 November 2006, Australian Road Research Board, Canberra, Canada

Oh JS, Jayakrishnan R, Recker W (2003) Section travel time estimation from point detection data. In: Paper presented in the 82nd transportation research board annual meeting, Washington D.C.

Sanaullah I, Quddus M, Enoch M (2016) Developing travel time estimation methods using sparse GPS data. J Intell Transp Syst 20(6):532–544

Soriguera F (2014) On the value of highway travel time information systems. Transp Res Part A: Policy Pract 70:294–310

Soriguera F (2016) Highway travel time estimation with data fusion. In: Tracts on transportation and traffic, vol 11. Springer-Verlag, Berlin, Heidelberg, 212 pages
Soriguera F, Robusté F (2011a) Estimation of traffic stream space-mean speed from time aggregations of double loop detector data. Transp Res Part C: Emerg Technol 19(1):115–129
Soriguera F, Robusté F (2011b) Highway travel time accurate measurement and short-term prediction using multiple data sources. Transportmetrica 7(1):85–109
Soriguera F, Robusté F (2011c) Requiem for freeway travel time estimation methods based on blind speed interpolations between point measurements. IEEE Trans Intell Transp Syst 12(1):291–297
Soriguera F, Rosas D, Robusté F (2010) Travel time measurement in closed toll highways. Transp Res Part B: Methodol 44(10):1242–1267
Sun Z, Jin W, Ritchie S (2017) Simultaneous estimation of states and parameters in Newell's simplified kinematic wave model with Eulerian and Lagrangian traffic data. Transp Res Part B: Methodol 104:106–122
Treiber M, Helbing D (2002) Reconstructing the spatio-temporal traffic dynamics from stationary detector data. Cooper Transp Dyn 1:3.1–3.24
van Arem B, van der Vlist MJM, Muste MR, Smulders SA (1997) Travel time estimation in the GERIDIEN project. Int J Forecast 13:73–85
van Hinsbergen CPIJ, van Lint JWC, van Zuylen HJ (2009) Bayesian Committee of neural networks to predict travel times with confidence intervals. Transp Res Part C: Emerg Technol 17(5):498–509
Woodard D, Nogin G, Koch P, Racz D, Goldszmidt M, Horvitz E (2017) Predicting travel time reliability using mobile phone GPS data. Transp Res Part C: Emerg Technol 75:30–44
Yim Y (2003) The state of cellular probes. Technical Report, July. Institute of Transportation Studies, University of California, Berkeley

Chapter 5
Travel Time Information Systems in the Era of Cooperative Automated Vehicles

Margarita Martínez-Díaz

Abstract Vehicle automation, together with the development of communications, is already leading to the emergence of new forms of mobility and allows aspiring to new paradigms that are increasingly efficient, safe, sustainable, and inclusive. Obviously, these changes must also reach traffic management for these positive effects to become a reality. In this regard, major challenges such as the processing of huge amounts of data in real time are already the subject of research around the world. In some cases, it will be enough to improve the accuracy of existing management strategies based on these data and on fusion methodologies, AI, etc. However, it will also be necessary to open the mind and explore new ideas, new forms of management to which only cooperative mobility gives meaning. All these topics could give rise to a book, or many. This chapter is intended only to serve as a transition to the following ones and to highlight some of the characteristics of such cooperative scenarios, as well as some expected mobility impacts.

5.1 Introduction

In 1925, Francis P. Houdina developed a remote-controlled driverless car called American Wonder. Huge crowds attended its demonstration on Broadway in New York City. A decade later, autonomous vehicles (AVs) became popular in the American pop culture: films, books, and comics were full of self-driving cars that helped heroes to overpower their enemies or picked children up at school. In fact, the idea of an autonomous conveyance is much older, as it can be deduced from Leonardo da Vinci´s schemes. He devised a kind of self-propelled car that was not designed to transport people but to move autonomously (possibly by means

M. Martínez-Díaz (✉)
Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Catalonia, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_5

of a winding mechanism) during the so-called "Renaissance court festivals". Seen yesterday as a dream, AVs are closer and closer to becoming a reality.

Because of technological and communication developments, the concept of autonomous vehicles (AVs) emerged again and was finally seen as developable. In fact, they were immediately identified as the solution to all traffic-related problems, especially congestion and accidents. Additional benefits were also predicted, such as making transport more inclusive, increasing productivity and comfort while driving and directly and indirectly reducing environmental damage, among others (Fagnant and Kockelman 2015). This vision of the vehicle as a panacea was especially supported by car manufacturers, but also by the public itself, given the undoubted appeal of these vehicles already envisioned by science fiction.

With the passage of time, many of these supposed benefits have been nuanced. The advantages of vehicle automation are undeniable. However, it does not guarantee benefits on its own, but only if well introduced and if accompanied by many other measures (Papageorgiou 2015). The first doubts about such optimistic vision of these vehicles that would not only perform the driving tasks themselves, (which is actually positive and still a goal when talking about highly automated vehicles) but also make their own decisions (i.e., choose their speeds, maneuvers, routes, etc.) for the benefit of their passengers, arose in the light of the results of simulation-based studies. Obviously, such vehicles would have to drive on the basis of very conservative parameters for the sake of safety and, also, of comfort (Hyde et al. 2017). For example, they would only make a lane change or merge in the main trunk of a freeway under ideal conditions. In addition, they should maintain long time gaps with respect to other vehicles ahead, about 2 s (note that humans' time gap is only slightly longer than 1 s). This smooth behavior would result in a capacity underutilization of around 600 vehicles/h/lane on motorways (Diakakis et al. 2015). Other undesirable effects of vehicles taking their own decisions could be the unbalanced distribution of flows across the road network, leading to congestion in some stretches while others remain unused.

Although the term "autonomous vehicles" is still used as an abuse of language, all experts agree that the mobility of the future must be based on vehicles with a high degree of automation that cooperate with each other and with the rest of the traffic agents not in pursuit of individual benefit, but of the collective optimum. In this sense, the adjective "autonomous" loses its relevance. In fact, the term "automated vehicles" (also AV) or, more in particular, "connected automated vehicles" (CAVs) has gained momentum, especially in research. It must be highlighted that the whole content of this book, unless specifically indicated, is referred to these cooperative driving environments. However, the adjective "autonomous" has been maintained in some paragraphs with the only aim of respecting some cited authors' nomenclature.

In any case, and as it is explained in the next section and in Chaps. 6 and 7, vehicle automation will play and in fact already plays a key role in traffic management. Ad hoc traffic management strategies for future cooperative driving environments are already being designed. These are very different and consider very distinct scenarios, for example, with different penetration rates of these highly automated vehicles and different communication capabilities. These new strategies also include travel time

information systems, which are expected to have two major differences from the current ones: (i) the information will be much more precise, thanks, among other things, to the detailed data provided in real time by vehicles themselves and (ii) the predominant dissemination will be directly to the vehicle. Other more complex and generalist traffic management systems will also benefit from these features (see Chaps. 6 and 7).

Although it is not the central objective of this book, Sect. 5.2 briefly summarizes the structure of the cooperative driving environments, in order to give the reader a better understanding of its implications for traffic management. Their major expected impacts are concisely explained in Sect. 5.3. For its part, Sect. 5.4 briefly explains key points of the consequences that CAVs will have for travel time information systems. These sections serve as an introduction to Chaps. 6 and 7, which elaborate on these aspects as the culmination of a comprehensive description of traffic management systems.

5.2 Cooperative Automated Driving Structure and Technological Aspects

As indicated above, the vehicle is only one of the agents involved in cooperative automated driving. In a very simplified way, this driving can be structured based on four additional protagonists: communications, infrastructure, cloud, and other mobile agents with communication capabilities (Fig. 5.1). Next sections describe their main features and roles.

5.2.1 The Vehicles

In parallel to technological progress, new vehicles include increasingly advanced systems that allow them substituting drivers in particular driving tasks. Although several administrations developed their own vehicle automation classifications, that of the Society of Automotive Engineers (SAE), released in 2014 (SAE 2014) for the first time and updated two years later (SAE 2016), has been worldwide adopted. Six automation levels (from 0 to 5) are distinguished depending on the on-board driver assistance systems, i.e., on the distribution of the driving tasks between the vehicle and the driver. Vehicles of levels 0 to 2 are called "traditional", because they are drivers who monitor the environment. At level 0 there are no automated driving functions and only those systems that issue warnings intervene. At level 1, vehicles are able to assume either the lateral or the longitudinal control of the vehicle, whereas they can undertake both tasks at level 2. However, these transfers are only possible under certain boundary conditions: uniform road design, free flow, good weather, etc. From level 3 onwards it is the vehicles that monitor the environment. This is a key

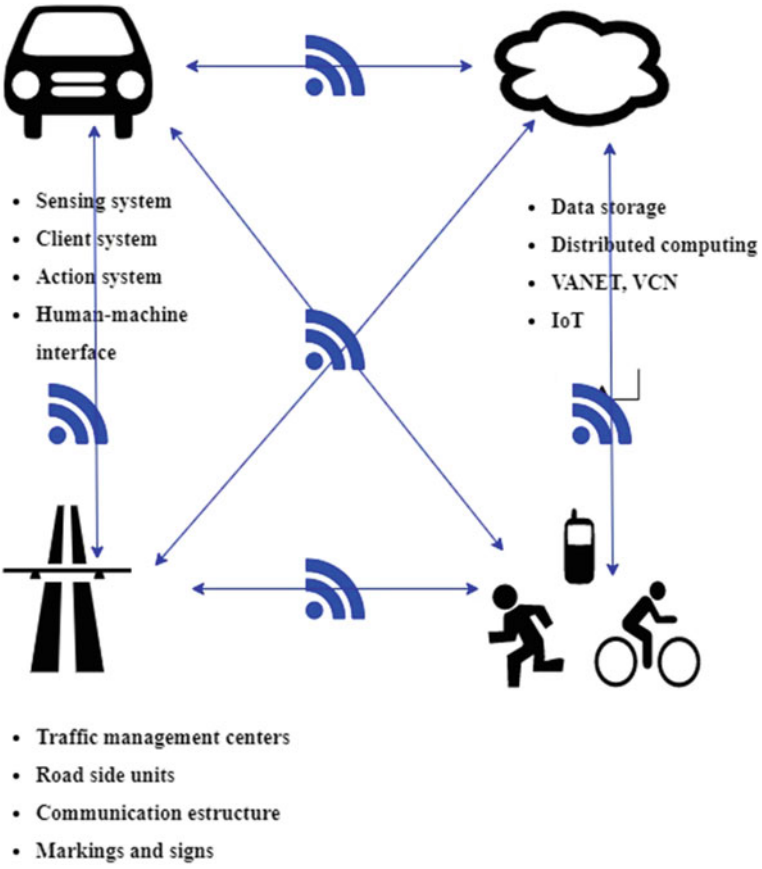

Fig. 5.1 Agents involved in cooperative driving environments

change, as it involves that they must collect all available data from the environment and interpret them. Furthermore, they can take responsibility for the driving task to certain limits. However, both at level 3 and 4 vehicles are expected to request drivers to resume the driving task when unmanageable situations arise. That is, drivers are still in charge of (responsible for) the driving supervision and must be aware of this during the whole trip. Really self-driving vehicles would be those that reach level 5: they are supposed to be able to perform the whole driving task autonomously, on all types of roads, in all speed ranges and under any boundary conditions. If these vehicles become a reality is still under question (Shladover 2016). While writing this chapter, and despite it being common to read/listen about "autonomous" or "self-driving" vehicles, the truth is that the vehicle on the market with the highest level of automation is a Honda with level 3.

In any case, the basic architecture of any highly automated vehicle has already been outlined and it is made up of four parts: (i) the *Sensing System*, (ii) the *Client System*, (iii) the *Action System* and (iv) the *Human–Machine Interface* (HMI).

The *Sensing System* is responsible for the collection of data from other vehicles and from the environment. This task must be performed in real time and in all boundary conditions (i.e., weather, traffic state, speeds, etc.). In fact, the refinement of this system is one of the targets to achieve. To this end, vehicles are equipped with diverse sensors, all of them with different strengths and weaknesses and thus appropriate to support particular driver assistance systems. Like for other purposes in road transport, results obtained from data fusion are much better than those derived from a single sensor, as individual errors are compensated (Soriguera and Robusté 2011; Bachmann et al. 2013; Yuan et al. 2014; Martínez-Díaz and Pérez 2015). In a very simple classification, sensors can be short-range or medium/long-range. The first ones are quite elementary and the data they collect is straightforward but very limited. To this group belong ultrasounds, capacitive sensors, and infrared sensors. The remaining are included in the second group. Data from radars and sonars requires no interpretation either: they basically measure the distance and the speed of the nearest object in front of the vehicle, using respectively radio or sound waves. In case danger of collision is perceived, emergency actions are immediately undertaken. For their part, cameras are the basis of artificial vision systems, which support recognizing and tracking tasks. However, they only perform well in light environments with a stable illumination level. Even in these cases, they are computationally demanding.

Equipping vehicles with a *Light Detection and Ranging* (LIDAR) sensor allowed making significant progress toward automation. Born in the 1960s, LIDAR was at first used in meteorology or for mapping in archeology or agriculture. It became known out of these fields when it was used to map the surface of the moon in 1971 during the Apollo 15 mission. In fact, it is this mapping ability that makes it very important for AVs. Powerful LIDARs provide 360º of visibility and measure distances with an exactness of ± 2 cm. This is achieved by continuously shooting laser light beams and measuring the time they need to return to the sensor. Up to distances of 60 m, a LIDAR is able to accurately measure all dimensions of the near objects. With less precision, it can generate 3D maps for distances up to 500 m. Therefore, LIDARs allow mapping and navigating, but also detect and track obstacles, other cars, pedestrians, etc. Finally, a global navigation satellite system, nowadays mainly the American GPS, self-localizes the vehicle. GPS is accurate, but its update time intervals are too long for real-time applications. Therefore, most prototypes of driverless cars also include an inertial measurement unit (IMU) to support navigation. Not only provide IMUs much more frequent estimates, but they also allow working when GPS-signals are unavailable, such as in tunnels, with interferences, etc. Despite their advantages, IMUs could not be standalone because they suffer from cumulative errors. Data fusion from GPS and IMUs stands out as a good solution.

The *Client System* consists of a powerful hardware platform and a highly evolved operating system. This computing framework plays a key role, as it must in real time (i) extract relevant and accurate information from the raw data supplied by the sensors (*perception* task) and (ii) indicate to the vehicle how it must proceed (*decision* task).

Very different types of hardware platforms are being designed: some companies opt for computing boxes containing distinct processors and accelerators. Others are developing system-on-chip (SoC) solutions, which are tiny integrated circuits with a microprocessor and advanced peripherals. The latest consume less energy and have less space requirements, but they do not reach at the moment enough computing capabilities to allow both fast and continuous sequential and parallel data processing. The support of cloud computing will be essential in this regard and will also add robustness to the system, which must continue to work even after a failure. The *perception* task includes three parts: *localization, detection, and tracking*, all of them achieved through data fusion performed at different levels. First of all, raw data from similar detectors is fused to detect outliers and to generate a bigger database. Secondly, the results of the former stage are fused and perceptions already generated. The highest level algorithms fuse action proposals taken according these perceptions. *Localization*, for example, is usually performed by algorithms that fuse data from GPS, IMU, and LIDAR, resulting in a high-resolution ground map. However, vision-based deep-learning technologies are achieving the most accurate results for object detection, as they are able to autonomously handle huge amounts of data. Deep-learning techniques have also demonstrated their suitability for object tracking relative to approaches based on computer vision.

Decision-taking is one of the most challenging tasks that C/AVs must perform, especially in awkward situations. On the one hand, proper and safe actions must be defined in real time. On the other hand, these decisions must be fair/ethical in case some damage is unavoidable. Focusing on the first point of view, decision-making encompasses *prediction, path planning, and obstacle avoidance*, all of them performed on the basis of previous perceptions. Stochastic models and probability distributions are often used to predict the next movements of other vehicles, pedestrians, animals, etc. *Path planning* is being faced with brute-force approaches (i.e., considering all possible paths and using cost functions to choose the best one), deterministic algorithms, and probabilistic planners. These last ones turn out to be the most advisable to work in real time with reasonable computational capabilities. Finally, *obstacle avoidance* must be deployed at least at two levels: a *proactive level* based on traffic predictions and a *reactive level* that immediately takes the control in case the first one fails. This second mechanism is usually based on radar data, because it acts only when obstacles are very close to the vehicle (Liu et al. 2017).

The *Action System* and the HMI are not that burdensome. Although they will be probably improved in line with the remaining parts of the AVs architecture, current prototypes are already acceptable. The action system consists of the mechanical parts of the vehicle (steering system, powertrain, braking system, etc.). HMI handles the interaction driver/passenger/pedestrian-vehicle. HMI are called to be minimalist in SAE5 level vehicles and basically oriented to provide humans in and out of the vehicle with information about the driving. The outer part will be mainly made up of lights, while the internal interface will usually consist in a touch monitor. This screen could also allow passengers to interact with the vehicle, for example, to control the air conditioning or to play certain music. Nevertheless, other designs entrust these tasks to mobile phones or other personal devices (Benderius 2016).

5.2.2 Communications

As explained, an efficient and safe mobility will not be possible if automated vehicles behave individually, but in a cooperative environment. Therefore, they must at least communicate among them (V2V interaction), with the infrastructure (V2I) and with the cloud (V2G, G coming from grid). Besides, communication with pedestrians (V2P), mobile phones, and other personal devices (V2D) are being considered. All these information exchanges are globally known as V2X communications. The establishment of a robust, powerful, safe, and reliable communications network is a main concern for the governments of all developed and developing countries. This network must be able to transmit huge amounts of data at very high speed, with low latency, under all conditions (weather, traffic state, etc.) and without interferences. Additionally, it must prevent hackers or terrorists from entering it and it must be able to work even if a fail occurs. Interoperability between different countries must also be ensured.

Thus, two tendencies are being followed all over the world: the use of evolutions of the wireless standard 802.11p or of mobile networks. In fact, most governments opt for their combination, as both have advantages and disadvantages. Regarding the first ones, the ITS-G5 has been standardized in the EU (C-ITS 2016), while the U.S. relies on DSRC (dedicated short-range communications) (FHWA, 2015). Respectively, they have reserved 75 MHz of the spectrum in the 5.9 GHz band and the 5.875–5.905 frequency for ITS applications. Both communication technologies are short-range, but they are supposed to: (i) be robust and reliable, i.e., to work in all boundary conditions and be able to immediately and independently recover from breakdowns, (ii) have a very low latency, i.e., messages are delivered practically without delay and (iii) be safe, i.e., message veracity and privacy are guaranteed. Notwithstanding, especially the EU is concerned about the possibility of an overload, taking into account that a great number of the entities implied in V2X could try to establish communications at once. In this regard, there are initiatives aimed at avoiding high network loads, like the restriction of the number of information packets generated by each vehicle, the reduction of information lifespan on the network, etc. Besides, the additional support of cellular communications seems advisable. Some experts approve this combination but warn against focusing on these last communications. They claim that the fact of relying on the cellular network causes the information to be indirectly (e.g., vehicle-network-vehicle instead of V2V) and thus slower delivered (ITS-JPO 2014; Filippi et al. 2015, etc.). 5G, 100 times faster than current 4G LTE wireless technology, seems to be the solution for these issues. In fact, some companies and administrations are contemplating it as a standalone system in their designs of a future cooperative environment and are accelerating its implementation. If time demonstrates that 5G suffices, significant savings in the infrastructure equipment could be achieved (Intel 2016; Arriola 2017; Shaheen 2018). Other countries out of the EU and the U.S. follow similar trends.

5.2.3 The Cloud

It is easy to imagine that autonomous driving per se, and to a much greater extent cooperative driving, involve the handling of huge amounts of data. Note that the word "handling" includes the storage, processing (i.e., eliminating erroneous data and outliers and obtaining relevant information from the remaining data), and exchange of these data, all performed in real time, uninterruptedly and with very small update intervals. This task, as mentioned above, will not be possible without a safe, reliable, and powerful communications network that ensures the reception/transmission of such data. However, it must be taken into account that data processing will not take place in the communications network, but it will be shared between the cloud and the vehicle itself. The idea that such processing (and consequent storage) would be carried out exclusively on-board was discarded many years ago: technological development makes it possible to implement in vehicles very small processors with large (and very expensive) capacities, but insufficient for the tasks to be performed. Hence, the cloud appears as a support element. Moreover, its intervention also makes it possible to occupy less space in vehicles, reduces their cost, (which is still expensive) and facilitates the exchange of information. Another important advantage provided by the cloud is that some tasks assigned to the vehicle could be temporarily transferred to it in case of vehicle failure.

However, the term “cloud” must be specified for this context of CAVs. For example, Vehicular ad hoc Networks (VANET) are the basis of V2I and V2V communications, being at first considered to support uncomplicated tasks related to safety, automated toll payment, navigation, etc. In fact, VANETs came up to assist the increasing amount of wireless devices used on-board (mobile phones, control keys, etc.). These networks are created by vehicles within the same area, which act as nodes (Zeadally 2010). Precisely this fact makes them not reliable to support all communications in a CAVs environment: vehicles move and VANETs are thus instable and cover an arbitrary range. To solve this problem, VANETs could be bound to a traditional cloud, which would be responsible for the most important tasks (e.g., data storage, high definition maps updating, distributed processing, etc.). This combination is called Vehicular Cloud Network (VCN) (Gerla et al. 2014; Ahmad et al. 2015). Another much more ambitious approach is that of the Internet of Vehicles (IoV) (Fig. 5.2), which would integrate a lot of different networks, users, means of transport, personal devices, administrations, infrastructures, etc. (Yang et al. 2014). Similar to the concept of the Internet of Things (IoT), bearing in mind how difficult it would be to manage all these data in a reliable and safe way, such configurations are not expected in the medium term.

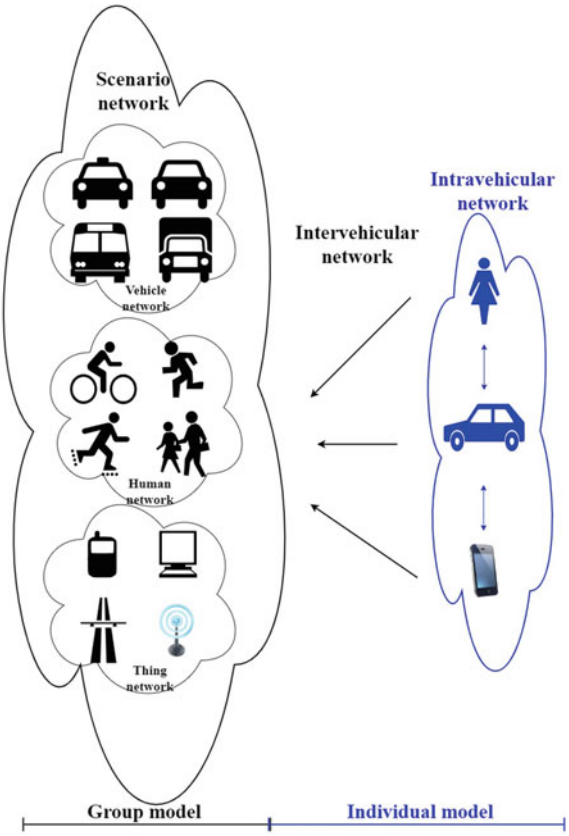

Fig. 5.2 Scheme of Internet of Things (inspired in Yang et al. 2014)

5.2.4 The Infrastructure

Improvements in the infrastructure will be necessary in any case. On the one hand, those aimed at helping AVs to perform the perception tasks: horizontal and vertical road signs must be clear and complete, road layouts should be as smooth as possible, etc. (Bosetti et al. 2015, García et al. 2017). On the other hand, V2I-related technologies must be deployed. As in the case of the communications network, different administrations work together trying to design a system with continuity across borders. In the EU, a multidisciplinary platform comprised of State and local, governments, automakers, cooperative intelligent transportation systems (C-ITS) European associations, user organizations, etc. already proposed in 2016 the initial guidelines for the deployment of C-ITS.

Although especially focused on V2I, some parts of V2V and V2X were also addressed (C-ITS 2016). The platform defined those services that should be implemented in a first phase, according to their significance, the related know-how, the state of the practice and the penetration rate of AVs, and the economic means. For

example, those V2I systems aimed at improving safety take priority. Secondly, those systems for which the cost-social benefit ratio results favorable were considered. This initial document was revised in 2017 (C-ITS 2017). In the second version, the compliance of the former plan was assessed, further needs were included and misstatements rectified. For example, the consideration that full AVs would be quite an important part of the fleet by 2030, what had affected the lead time initially considered for V2I deployment. Other countries out of the EU have also developed their own plans for V2I installation. For example, the U.S. elaborated its "Connected Vehicle Path" for V2I deployment in 2015, which also included some regulations intended for carmakers, trying to ensure the goodness of all communications. The National Highway Traffic Safety Administration (NHTSA) recognized in that year that the U.S. was in its infancy regarding these topics and is determined to make 80% of the national intersections V2I capable by 2040 (U.S. GAO 2016). In spite of general guidelines like the former, final deadlines and probably the quality of the system will depend on particular local or state budgets. The concept of smart roads especially designed for CAVs has also been suggested. However, some drawbacks advise against their construction. First of all, extra expenses would be huge, while current freeways and highways would be misused. In addition, according to full AVs deployment forecasts, they would serve a minority share of the total vehicle fleet during a long period. Furthermore, traditional vehicles would not benefit from sharing roads with self-driving vehicles. These potential benefits are explained in Sect. 5.3.

5.2.5 Other Agents

If IoT or similar schemes really become a reality, any road/street user (e.g., pedestrian, cyclist), any device (e.g., noise meter, retail security camera, radio antenna, etc.) will be able to exchange data voluntarily or involuntarily with the rest of the agents. Whether this will be positive or not is still an open question: having a large amount of data is no guarantee of success if these data are inaccurate, irrelevant, or unmanageable, among other things. There are two current trends in this respect. The first, more grounded in reality, warns of the limitations of communications networks and real-time data processing capabilities, and advocates learning to distinguish which data are important so that they can be the basis for cooperative driving. The second, for the moment more linked to research than to practical applications, envisions AI and data-driven as the basis for the driving of the future (see Chaps. 6 and 7). Therefore, it sees positively the inclusion of any type of available information in the driving and traffic management processes.

5.3 Impact of Cooperative Automated Vehicles on Mobility

Although still difficult to quantify, the main impacts of this new form of driving have already been qualitatively defined. Whether or not and to what extent these impacts become a reality will depend on many factors, including the penetration rate of vehicles with a high degree of automation, the investments made by the different administrations to provide roads and management centers with the necessary equipment and communications and, of course, the correct definition of information and traffic management strategies specifically designed for these scenarios. The huge amount of research underway suggests that at least this last requirement will be met. The following sections summarize very briefly some of the impacts expected from the implementation of cooperative automated mobility, as a bridge to some of the issues that will be addressed in Chaps. 6 and 7. More consequences are anticipated, for example, reductions in consumptions and emissions linked to electrification, changes in legislation, in productivity, etc. A detailed description of these can be found in Martínez-Díaz et al. (2019).

5.3.1 New Approaches for an Improved Traffic Performance

Once the idea of vehicles driving autonomously with the sole objective of maximizing their individual experience has been discarded, the next step is to decide how vehicles should cooperate to optimize the overall flow, even if a particular vehicle is ruled out. This goal involves technical and communications challenges, but also management challenges. Since the 1990s, traffic management has already been progressively improving over time, moving from purely passive guidelines to real-time management. This evolution has also been seen in travel time information systems, which have moved from providing historical information determined offline on to providing instantaneous information or even travel time forecasts (see Chaps. 6 and 7).

However, automated driving opens a window of opportunity to implement more effective dynamic traffic management strategies in a more coordinated and successful way. These strategies should develop in line with vehicle automation and CAVs penetration rate. Therefore, they should be designed for mixed traffic environments in the first term and be gradually adapted according to the increase of the CAVs percentage and of the automation level afterward. This fact is challenging, as CAVs will have to interact with human drivers, whose behaviors are much more aggressive and unexpected. Nevertheless, positive impacts are also envisaged: it has been demonstrated that efficiency improves in line with the increase in CAVs. Enhancements become already noticeable when they are 30% of the flow (Guériau et al. 2016). The main cause for this progress is that cooperation reduces traffic instabilities, like "stop and go" situations. In fact, instabilities are largely responsible for the three major inefficiencies of congestion: capacity reductions, the increase in the accident rate

(due to secondary accidents), and the increase in consumption and, when applicable, emissions (linked to multiple accelerations).

Furthermore, separated strategies for urban areas and freeways or highways will be necessary. Especially the latest should ensure interoperability between regions and countries. In any case, the first scenario to address will be probably that of freeways. The reason is twofold: (i) traffic behavior is more uniform than in cities and thus more manageable to make the first tests and (ii) some freeways already count on part of the needed technology. In fact, some ways in which CAVs should drive on freeways and highways are already being tested. Although it is out of the scope of this book to deepen in these topics, an example is given next in order to give an idea of the great potential of cooperative automated driving.

One very promising form of cooperation is vehicle platooning. That is, vehicles should constitute a sort of road train in which they would be able to drive maintaining very small intervehicle distances (smaller than that of human-driven vehicles), at quite high speeds, but also safely. This idea is not new. Well-known is the experiment performed in 1997 under the PATH (Partners for Advanced Transportation Technology) program at UC Berkeley. However, technological deficiencies hindered the generalization of platoons in that moment. Currently, platoons are already being used by some freight companies in very specific scenarios, with SAE2 level vehicles that usually have additional equipment to this end. Although positive effects are achieved, it is still not possible to make the most of cooperation due to a lack of technology and knowledge. Soon, technology will no longer be a problem. However, there are still a lot of doubts hanging over platoons and there is a lack of traffic strategies especially designed to manage these road trains. Firstly, regarding their formation: minimum vehicle automation level, type of vehicles (cars, vans, trucks, their combinations, etc.), platoon average gap, average speed, maximum length, merging and split (Saeednia and Menéndez 2016; 2017), etc. Secondly, with respect to their interaction with traditional vehicles: via shared lanes, with dedicated lanes, with dedicated roads, etc. Another key issue is which vehicles within the platoon should exchange information and which information (Feng et al. 2019). There is no clear answer to these questions and much more exist. Again, the development of ad hoc dynamic management strategies taken depending on (i) CAVs penetration rate, (ii) traffic conditions and (iii) infrastructure equipment and communications is due. Although insufficient, some initial ideas have already been outlined in this regard: shared lanes seem to be a good option when traffic conditions are good, i.e., when speeds exceed 50 km/h. In this way, the whole traffic stream would take advantage of the mentioned instability reduction linked to cooperation in general. However, when congestion has already appeared and speeds are small, dedicated lanes for platoons would at least allow them to drive efficiently. As said, the best option would be that this "dedication" is only temporary (i.e., dynamic), until congestion disappears. Furthermore, especially with small percentages of CAVs, it could be suitable that platoons would share these dedicated lanes with other vehicles with priorities like high occupancy vehicles (HOV), public busses, etc. The choice of the most favorable lane for platooning is another strategy to define. For example, with trucks, the first idea could be to use the right-most lane. However, if this were proven to be suitable,

platoon management strategies would have to include measures to avoid problems at on- and off-ramps. The left-most lane is usually chosen by researchers.

5.3.2 Expected Impact on Mobility Trends and Figures

The impact of C/AVs on the mobility rate must be analyzed from a comprehensive point of view and taking into account the changes in the way people move. In fact, especially in cities (and without taking into account the special consequences of the COVID-19 pandemic, the long-term impacts of which are not foreseeable at the time of writing), these changes have already started with traditional vehicles: the number of users of car-sharing, ride-hailing, and ride-sharing systems has increased in the last years. More and more users and particularly young people find unnecessary or even unadvisable to own a vehicle for many reasons. For example, private vehicles usually spend more time parked (20–23 h per day according to recent analyses) than in motion, their acquisition and their maintenance imply costs, etc. The increasing climate change awareness plays also a role. This trend toward vehicle use instead of vehicle ownership is expected to significantly intensify in the next years: firstly, the supply of sharing and hailing services grows and, consequently, they become cheaper. Secondly, AVs are ideal to support them. On the one hand, these services have a large technological component. On the other hand, savings in staff costs are appealing for entrepreneurs. Besides, sharing and hailing are called to be integrated in more comprehensive on-demand initiatives like Mobility as a Service (MaaS). In addition, AVs will be clean (primarily electrical): sharing systems based on AVs are expected to be increasingly promoted by administrations and well accepted by the public. In fact, hybrid or electrical vehicles (EVs) with a certain degree of automation are already being used by some companies to serve limited urban areas. Finally, it must be taken into account that the cost of AVs will be very high (primarily at the beginning), what will prevent many people from buying them even if they wanted to. Therefore, many companies are already developing future sharing services based on electric automated vehicles (EAVs). A key idea is that public transportation could and should also be based on EAVs.

The first conclusions that could be drawn from the former considerations are that the vehicle fleet will decrease and, thus, the mobility rate too. However, only the first statement seems unquestionable (Grosse-Ophoff et al. 2017; Litman 2017; etc.,). In fact, the fleet reduction has already been estimated to be 22–25% in Europe and the U.S. by 2030 (Kuhnert and Stürmer 2018). Regarding the mobility rate, there is no agreement: some researchers estimate a reduction if it is expressed in vehicle-Km per passenger and an increase in terms of vehicle-Km. The former is linked to sharing, as vehicles should transport more people in each travel (current car average occupancy is 1.3 pax. and it should increase for shared vehicles), and to the expected increase in users of a more efficient public transportation system. The latter is linked to two main reasons: on the one hand, transport costs are expected to decrease due to a better

amortization of vehicles (e.g., their rest time will be much smaller in a sharing environment). Therefore, freight transport is expected to rise looking for competitiveness and passenger transport too, mainly for leisure/unnecessary purposes. On the other hand, the spectrum of users will spread: non-drivers, very young or very old people, people with special needs, etc. will be able to use C/AVs. However, the most recent studies point out that the increase in car-traveled Km will not be compensated by the number of passengers per travel: i.e., vehicle-Km per passenger will probably increase too (Correia and van Arem 2016; Milakis et al. 2017). The final result will highly depend on the occupancy of the sharing services and of the use of public transportation: a trade-off could be reached if most trips were performed at high occupancy (e.g., with automated vans or buses). If the current tendency persists and sharing systems rely on 2 or 4-seater vehicles shared by 1–2 people per trip, and if public transportation is not boosted, the mobility rate will definitely increase in all its terms. The number of private C/AVs will also play a role in this regard, as they will make empty journeys, for example, to park once transported their owners or to pick them up after work.

As said, the configuration of the public transport will also be critical. Sharing systems will only make mobility more efficient and sustainable if they substitute private journeys but not those made by mass transit, whose occupancy is higher and thus more favorable. In fact, sharing should act as a complement for them, aimed, for example, at covering the last Km of a commuter's journey toward a city center. However, mass transit must adapt to new scenarios. And, of course, benefit from automation and cooperation. In this context, its integration in on-demand services like MaaS is already being considered: first analyses show that the trade-off between the necessary investment in technology and the overall benefits (accessibility, efficiency, congestion release, pollution reduction, etc.) are greatly beneficial (Barceló 2016).

5.3.3 Contribution to Safer Mobility

The number of road deaths has decreased in most developed countries, due to improvements in the vehicles (driver assistance systems, stronger bodyworks, etc.) and to the efforts of traffic administrations in this regard. However, despite reductions, the number of deaths is still huge and very far from the Vision Zero (i.e., no accidents) courted by many administrations. Furthermore, the situation in developing countries is still worrisome.

A cooperative automated driving environment will not be able to avoid all accidents. Nevertheless, taking into account that 90% of accidents derive from human errors, they are expected to be reduced to a minimum (Gong et al. 2016; Koopman and Wagner 2017; Gear 2030, 2017). However, two important conditions must be fulfilled to achieve success in this regard: (i) the penetration rate of cooperation-capable vehicles must be high and (ii) cooperative management strategies must work optimally. On the contrary, the increase in the number of vehicle-Km traveled could

offset the decrease in the number of accidents, being the final number of deaths similar to the current one.

Other types of risks must also be avoided. For example, the possibility that passengers of C/AVs become overconfident and give up using seat belts, or that pedestrians cross streets recklessly assuming that C/AVs will not run over them. It should also be borne in mind that, unless level 5 automation is reached, there will always be situations in which vehicles will require drivers to resume the driving task. Therefore, drivers of automated vehicles will have to maintain a certain attention and know how to react in a proper and timely manner. Numerous studies analyze these reactions, which are very varied and depend on many internal and external factors, usually using driving simulators (e.g., Rahman and Abdel-Aty 2018). Their goal is to at least draw some general conclusions that allow providing the vehicle hardware and software with all the necessary tools and indications to avoid dangerous situations. Also, to train or at least inform drivers and authorities. Note that requirements of future driving licenses, if any, are also on the table. There are also numerous studies that analyze changes in driver behavior when traveling in a vehicle with a high degree of automation or, simply, when interacting with them, even if only visually (e.g., Skottke et al. 2014).

More dangerous situations with implications for safety could arise: CAVs and V2X as a whole could be appealing targets for hackers or terrorists. For example, ransomware aimed at blocking vehicles until owners pay a ransom or malware intended to intercept communications or even to alter them (with the subsequent fatal results) could be easily distributed through the networks (Douma and Aue 2012; Petit and Shladover 2015; Litman 2017; GEAR 2030, 2017). Governments and cyber-experts already analyze these issues and some of them have proposed initial guidelines trying to build an extremely secure system that is resilient to this kind of attacks. Although some probability of communications being hacked will always remain, it should be at least possible to resume the control in a very short time (BMVI 2017; UK 2017; Shaheen 2018).

5.4 The Role and the Evolution of Travel Time Information Systems in Cooperative Driving Environments

Travel time information systems will continue to play a key role in future cooperative driving environments. However, to this end, they will have to gradually adapt to these new scenarios, in which traffic streams will be composed of vehicles with increasingly high automation levels. In this regard, some important considerations must be kept in mind:

- Travel time value is mainly expected to decrease, but doubts exist

Several studies forecast a gradual reduction in the value of travel time in line with the use of more and more intelligent vehicles. According to them, the maximum

reduction would take place with self-driving vehicles (SAE-level 5), as they would be a kind of mobile bedrooms, playrooms, or offices, allowing passengers to rest or be productive while traveling (Litman 2018). Additionally, ideal scenarios that support these reduced values of travel time are envisioned. First, increased comfort, more travel safety, and higher travel time reliability would contribute to lower values of time. Second, less congestion delays, reduced (or null) search time for parking, and/or the increased use of shared vehicles would lead to the need for less travel time to make each particular journey. Finally, enhanced traffic flow efficiency along with electric or more fuel-efficient vehicles would also reduce the monetary cost of travel. People's psychology also plays a role in this regard. Steck et al. (2018) performed a discrete choice experiment, trying to quantify this reduction. They made an online survey among Germans of different ages and gender and from varied income classes, who either walked or used a wide spectrum of transportation modes (private cars, bikes, public transportation, etc.) to make different types of trips (commuting, shopping, and leisure trips). The goal was to analyze the possible changes in the value of travel time savings (VTTS) both for the case of shifting toward AVs or to shared AVs. It was found that autonomous driving in a private car would reduce the VTTS by 31% compared to driving manually. Besides, it was perceived similarly to in-vehicle time in public transportation. Shared AVs resulted less attractive. However, the travel time spent in them was considered less negatively (by 10%) than that spent while driving manually.

Nevertheless, it is important to note that (i) the former studies are case-specific and have no general representativeness and (ii) all of them are based on assumptions or perceptions, as no AVs drive at present on the road. Additionally, other less optimistic studies (also without empirical evidence) considerably limit the reduction in the value of travel time due to some factors already addressed. Among others, the potential increase of travel demand linked to lower transportation costs, the emergence of new users, and urban sprawl could compromise travel time and cost savings (Milakis et al. 2017). Several researchers concluded that passengers' use of their travel time in an AV to work cannot be generalized. In a survey performed by Haboucha et al. (2017), only long-distance commuters seemed willing to effectively work during the time they traditionally spent driving. Cyganski et al. (2015) performed another survey in Germany. Again, only a minor percentage of the respondents perceived as an advantage the ability to work on the move in a self-driving vehicle. Most of those that used other means of transport (e.g., trains) already worked in their commute. The majority of the participants still saw as more valuable the activities that they usually undertook while driving traditional cars (i.e., listening to music). For their part, in a survey in the Netherlands, Yap et al. (2016) analyzed AVs as egress mode of train trips. That is, as the means of transport that travelers that made a trip by train would take at the station in order to reach their final destination. Surprisingly, Yap et al. (2016) found that travelers gave a higher value to time when using self-driving vehicles for this purpose, in comparison to manually driven vehicles. Next, authors compared the value of travel time for the same people in two different situations: (i) the same one as before, i.e., the main travel made by train and a self-driving car taken at the train station to reach the final destination and (ii) the whole trip made by

a private traditional car. Between these alternatives, travelers were somewhat more willing to pay for a certain travel time reduction in the first case. Therefore, results suggested again that travelers did not perceived as an advantage the possibility of doing other things while traveling. Fears in relation to the idea of riding an automated vehicle, the lack of any real-life experience with them and the fact that an egress trip is typically short (thus not allowing travelers to fully experience the potential benefits of automated vehicles, such as en-route work or travel safety), were some of the explanations to these unexpected results given by researchers.

Overall, many doubts exist about how the value of travel time is going to be when autonomous vehicles hit the road. As a clear example, Milakis et al. (2017) reported a possible decrease of the value of time between 1 and 31% for their users in the Netherlands. This is a very wide range that, as said, could even change toward an increase. The impression is that there is no single answer, that is, that the direction and significance of the change in the value of travel time will depend on the particular boundary conditions. In fact, the VTTS is a subjective variable per se (Jara-Díaz 2000). In any case, travel time information systems adapted to the new driving scenarios will continue to contribute to traffic efficiency and safety: travel time information will still be the best indicator of congestion and allow modifying or rescheduling trips for the passengers' sake.

- Like any other information, delivered travel times will have to be very accurate

Taking into account that computers, either on-board, in the cloud or in a traffic agency, are expected to manage traffic autonomously, all exchanged real-time information will have to be extremely precise. Otherwise, not only congestion but, above all, risky situations could arise. In this regard, travel time information, although especially linked to traffic efficiency, also contributes to accident avoidance: traffic inefficiencies clearly increase the possibility of an incident. Pre-trip information about travel time reliability must also be accurate, so that the congestion decrease associated with AVs comes true. Therefore, current travel time information systems that deliver approximate or inaccurate travel time information (see Chaps. 2 and 4) would have more detrimental consequences in cooperative driving environments. Note that in this context, travel time information will directly affect decisions taken for the management of the whole connected system. There is a need, especially, for new methodologies that precisely forecast travel times in real time and with short time updating intervals, like the one proposed in Chap. 4. Besides, these methodologies will benefit from working with more and more complete data than that currently available (see next point), which will undoubtedly contribute to accuracy.

- Vehicles will act as data providers and senders

This fact is crucial and will have great influence on future travel time information systems. Indeed, this section deliberately refers to "travel time information systems" and not to "highway travel time information systems". Although highways and freeways will probably have more surveillance and better communications than other roads also in the future, V2V communications and the own storage and computation capabilities of vehicles and vehicular cloud networks will allow real-time travel time

estimation and exchange also in less equipped scenarios. Obviously, those infrastructures with more technological and communication equipment would also benefit from the information gathered by on-board sensors. AVs are called to be the paradigm of the perfect lagrangian sensors. Their role could be compared to that of probe or floating vehicles (see Chap. 1), but with some important differences: (i) in-vehicle sensors will be varied and (theoretically) standalone when considered jointly, in the sense that they will be able to gather all the necessary data, (ii) these sensors will be extremely accurate and powerful and (iii) the penetration rate of AVs is expected to solve any problem regarding sample sizes or sample bias (e.g., samples composed only of one type of vehicle).

The way in which travel time information is disseminated will also evolve. Current modes by means of apps, websites, or variable message signs are expected to remain, at least partially, and especially in mixed traffic scenarios. However, a key change will take place in a near future, as intelligent vehicles will directly receive the information either from other vehicles or from the management centers. There exists the possibility that this information is shown in the HMI but, in fact, this will not be necessary unless passengers are interested in it: vehicles themselves will act according to the updated information they will continuously receive.

In this context of enormous amounts of available data, considering both raw inputs and the information that is extracted from them, the need for the implementation of protocols that select, limit, and prioritize the most significant ones must again be highlighted (Zhong et al. 2008). Otherwise, communication failures or the late reception of extremely important information could occur, probably with undesirable consequences. Additionally, vehicle-to-all (V2X) communications should be feasible in order to guarantee a good performance of the system. That is, V2V should only be a part of the driving scenario, taking into account that vehicles, although powerful, will have capacity to storage, process, and disseminate data only to some extent. The fact that cooperative networks among them will be structureless and changing (i.e., vehicles that cooperate at a particular moment will just afterward follow different routes) is a key issue in this regard. The data and information exchanged among vehicles must be ranked, including travel time information. Heuristics, artificial neural networks, machine learning, etc., are being applied to find the best procedure for this purpose. An interesting approach is that of Szczurek et al. (2010), developed for the dissemination of peer-to-peer real-time information. Their method is based on the postulate that the most important information is the most recent, the nearest and that allowing drivers to change their path. From this standpoint, they proposed a machine learning algorithm that learns the probability that a travel time report is useful as a function of its attributes. Then, a naïve Bayes learning method is used to find a mapping for the attribute values to the probability of a report being useful.

- Travel time information systems must share basic standards

The information that future travel time information systems deliver must be not only very accurate, but also uniform and coordinated. Actually, this requirement can be applied to any other type of data or message exchanged among the actors of cooperative driving environments. This fact was partially addressed when describing

intelligent vehicles' architecture or communication technologies. Interoperability and coordination must be guaranteed at all levels. To this end, the format of any inputs and outputs as well as the accuracy levels of any calculation should be agreed upon by carmakers and traffic administrations. In the particular case of travel time information systems for AVs, it must be ensured that every vehicle receives the information it needs and that it is able to "understand" it. That is, all vehicles must have access to the communications channel and use the same or compatible information languages. Additionally, a vehicle that is going to undertake a long journey should be able to receive, for example, information about the travel time reliability of the whole path, even if it takes place along roads managed by different administrations. However, if these administrations use different reliability indexes and/or work with different levels of accuracy, the utility of this information will decrease, as it could result too coarse. Regarding real-time information, the need for similar (high) accuracy levels is not linked to interoperability but, as said, to traffic efficiency. That is, when vehicles made a trip, they will receive constant information updates from the corresponding system. This system could vary if, for example, the trip is long. In case more than one system is involved, the vehicle (supposing interoperability is ensured) could receive information with different degrees of accuracy without any technical problem arising. However, this fact would affect the decisions the vehicle would take looking for its own benefit, and that of other vehicles that exchange information with it. That is, it would be impossible to reach the system optimally.

The former considerations about coordination and interoperability do not only affect vehicles, but also traffic agencies. Nevertheless, from the agencies point of view, the consequences of not receiving data from a certain number of AVs would generally not be as severe for them, taking into account that many other moving sensors will be available.

5.5 Conclusions

The former sections tried to explain how AVs and CAVs could contribute to make future mobility more efficient, safer, cleaner, and more inclusive. However, some of the conditions to be fulfilled so that these goals are achieved have been also highlighted. Table 5.1 summarizes the main conclusions that can be drawn from the overview provided in this chapter.

First, it must be borne in mind that AVs behaving "autonomously" in the sense of "independently" would lead to an increase in congestion. They would perform a very conservative and smooth driving seeking for safety and thus lead to capacity reductions. On the contrary, their cooperation could allow their driving fluently and at high speeds while ensuring safety, even with very small intervehicle distances. That is, cooperation would lead to an increase in capacity. Notwithstanding, it does not only involve technological requirements, but also the definition of ad hoc dynamic traffic management strategies. These strategies will be different, for example, on freeways or in urban areas. In fact, in the later, new mobility patterns are called to

Table 5.1 Potential benefits of cooperative driving environments closely linked to mobility and main challenges to be addressed

Automation scope of influence	Potential benefits		Requisites for impact optimization
	AVs	CAVs	
Traffic efficiency	Low or negative	High	V2X. Ad hoc management strategies
Safety	Medium	High	Technological advances. V2X. Responsibility definition. People formation
Mobility/accessibility	Medium or low	Medium	Encouragement of MaaS and/or use of mass/sharing services at high capacity

complement the arrival of CAVs: sharing systems used at high occupancy and seamless synergies among different means of transport will be indispensable to relieve congestion in cities. In this way, the expected growth in the mobility rate could be compensated by a smaller vehicle fleet. Taking into account that traffic inefficiency contributes to air pollution and energy consumption, future mobility is also called to be more environmentally-friendly. The fact that most CAVs will be electric will also be decisive in this regard.

Vehicle automation will also lead to a relative reduction in the number of accidents. However, this reduction will only be significant with quite a big penetration rate of CAVs. Besides, the probability of an accident will always exist. Therefore, ethical behavior rules for these cases are being discussed and they should be agreed among all involved parts. Similarly, a procedure aimed at deriving responsibilities in this new mobility paradigm must be carefully developed. Additionally, unless SAE 5 level vehicles hit the road, the human factor will continue to play a role, either as passenger/driver of CAVs or as road user that interacts with them. Possible changes in human behavior or too late or improper maneuvers by drivers could become new causes of accidents. This fact must be taken into account when configuring vehicle software and designing management strategies. The suitability of specific training to use or coexist with CAVs must also be analyzed.

References

Ahmad F, Kazim M, Adnane A, Awad A (2015) Vehicular cloud networks: architecture, applications and security issues. In: 2015 IEEE/ACM 8th international conference on utility and cloud computing (UCC), Limassol, Cyprus

Arriola JJ (2017) Circulación de vehículos autónomos: retos legislativos. Carreteras 216:18–27

Bachmann C, Abdulhai B, Roorda M, Moshiri B (2013) A comparative assessment of multi-sensor data fusion techniques for freeway traffic speed estimation using microsimulation modeling. Transp Res Part C: Emerg Technol 26:33–48

Barceló J (2016) Paradigm changes: from smart cities to wise cities and the role of transport models. CYTED Workshop 2016, Santiago, Chile

Benderius O, Berger C, Malmstem V (2017) The best rated human-machine interface design for autonomous vehicles in the 2016 grand cooperative driving challenge. IEEE Trans Intell Transp Syst 1(1):20–32

BMVI Ethics Commission (2017) Automated and connected driving. German Federal Ministry of Transport and Digital Infrastructure

Bosetti P, Da Lio M, Saroldi A (2015) On curve negotiation: from driver support to automation. IEEE Trans ITS 16:2082–2093

Correia GH, van Arem B (2016) Solving the user optimum privately owned automated vehicles assignment problem (UO-POAVAP): a model to explore the impacts of self-driving vehicles on urban mobility. Transp Res Part B: Methodol 87:64–88

Cyganski R, Fraedrich E, Lenz B (2015) Travel time valuation for automated driving: a use-case-driven study. In: 94th annual meeting of the transportation research board, January 2015, Washington, D.C., US, 1–19

Diakaki C, Papageorgiou M, Papamichail I, Nikolos I (2015) Overview and analysis of vehicle automation and communication systems from a motorway traffic management perspective. Transp Res Part A: Policy Pract 75:147–165

Douma F, Aue S (2012) Criminal liability issues created by autonomous vehicles. Santa Clara Law Rev 52(4):1157–1169

EU Cooperative Intelligent Transport Systems Platform (2016) 2016 Final Report. C-ITS Platform

EU Cooperative Intelligent Transport Systems Platform (2017) Final report phase II: cooperative intelligent transport systems, towards cooperative, connected and automated mobility. C-ITS Platform

Fagnant DJ, Kockelman K (2015) Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations. Transp Res Part A: Policy Pract 77:167–181

Feng S, Zhang Y, Li SE, Cao Z, Liu HX, Li L (2019) String stability for vehicular platoon control: definitions and analysis methods. Annual Rev Control 47:81–97

Filippi A, Moerman K, Daalderop G, Alexander PD, Schober F, Pfliegl W (2015) Ready to roll: why 802.11p beats LTE and 5G for V2X. Internal Report for Siemens

García A, Camacho FJ, Padovani PV (2017) Influencia de la infraestructura en la velocidad de los vehículos automatizados. Carreteras 216:52–61

GEAR 2030 (2017) Ensuring that Europe has the most competitive, innovative and sustainable automotive industry of the 2030s and beyond. Report of the High Level Group on the Competitiveness and Sustainable Growth of the Automotive Industry in the European Union (GEAR 2030), European Commission

Gerla M, Lee EK, Pau G (2014) Internet of vehicles: From intelligent grid to autonomous cars and vehicular clouds. In: 2014 IEEE World Forum on Internet of Things (WF-IoT), Seoul, South Korea

Gong S, Shen J, Du L (2016) Constrained optimization and distributed computation based car following control of a connected and autonomous vehicle platoon. Transp Res Part B: Methodol 94:314–334

Grosse-Ophoff A, Hausler S, Heineke K, Möller T (2017) How shared mobility will change the automotive industry. McKinsey & Company

Guériau M, Billot R, El Faouzi N-E, Monteil J, Armetta F (2016) How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies. Transp Res Part C: Emerg Technol 67:266–279

Haboucha CJ, Ishaq R, Shiftan Y (2017) User preferences regarding autonomous vehicles. Transp Res Part C: Emerg Technol 78:37–49

Hyde S, Dalton P, Stevens A (2017) Attitudes to autonomous vehicles. Publications of the TRL Academy. Report PPR823, United Kingdom

Intel Go Autonomous driving solutions (2016) Autonomous driving, accelerated. Product brief

ITS-JPO Strategic Plan 2015–2019 (2014) ITS Joint Program Office, U.S. Department of Transportation
J3016:JAN2014 (2014) Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems. Society of Automotive Engineers
J3016A:SEP2016 (2016) Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Society of Automotive Engineers
Jara-Díaz SR (2000) Allocation and valuation of travel-time savings. In: Hensher DA, Button KJ (ed) Handbook of transport modelling. Elsevier Science Ltd., Simi Valley, CA
Koopman P, Wagner M (2017) Autonomous vehicle safety: an interdisciplinary challenge. IEEE Trans Intell Transp Syst 9(1):90–97
Kuhnert F, Stürmer C (2018) Five trends transforming the automotive industry. Report of PriceWaterhouseCoopers
Litman T (2018) Autonomous vehicle implementation predictions: implications for transport planning. Report for the Victoria Transportation Policy Institute
Liu S, Tang J, Zhang Z (2017) Computer architectures for autonomous driving. IEEE Comput Archit Lett 50(8):18–25
Martínez-Díaz M, Pérez I (2015) A simple algorithm for the estimation of road traffic space mean speeds from data available to most management centres. Transp Res Part B: Methodol 75:19–35
Martínez-Díaz M, Soriguera F, Pérez I (2019) Autonomous driving: a bird´s eye view. IET Intel Transport Syst 13(4):563–579
Milakis D, van Arem B, van Wee B (2017) Policy and society related implications of automated driving: a review of literature and direction for future research. J Intell Transp Syst 21(4):324–348
Milakis D, Snelder M, van Arem B, van Wee GP, Homem De Almeida Correia G (2017) Development and transport implications of automated vehicles in the Netherlands: scenarios for 2030 and 2050. Eur J Transp Infrastruct Res 17(1):63–85
Papageorgiou M (2015) Freeway traffic management in the era of VACS (Vehicle Automation and Communication Systems). In: IEEE 18th international conference on intelligent transportation systems "Smart Mobility for Safety and Sustainability", September 15–18 2015, Canary Islands, Spain
Petit J, Shladover SE (2015) Potential cyberattacks on automated vehicles. IEEE Trans Intell Transp Syst 16(2):546–556
Rahman MS, Abdel-Aty M (2018) Longitudinal safety evaluation of connected vehicles' platooning on expressways. Accident Anal Prevention 117:381–391
Saeednia M, Menéndez M (2016) Analysis of strategies for truck platooning. hybrid strategy. Transp Res Record: J Transp Res Board 2547:41–48
Saeednia M, Menéndez M (2017) A consensus-based algorithm for truck platooning. IEEE Trans Intell Transp Syst 18(2):404–415
Shaheen S, Totte H, Stocker A (2018) Future of mobility. The White Paper. A report of UC Berkeley, UCConnect
Shaheen S, Totte H, Stocker A (2018) Future of mobility. The White Paper: Report of UC Berkeley, UCConnect
Shladover SE (2016) The truth about self-driving cars. Sci Am 314(6):52–57
Skottke EM, Debus G, Wang L, Huestegge L (2014) Carryover effects of highly automated convoy driving on subsequent manual driving performance. Hum Factors 56:1272–1283
Soriguera F, Martínez I, Sala M, Menéndez M (2017) Effects of low speed limits on freeway traffic flow. Transp Res Part C: Emerg Technol 77:257–274
Soriguera F, Robusté F (2011) Highway travel time accurate measurement and short-term prediction using multiple data sources. Transportmetrica 7 (1):85–109
Status of the Dedicated Short-Range Communications Technology and Applications (2015) Report FHWA-JPO-15–28. FHWA. U.S. Department of Transportation
Steck F, Kolarova V, Bahamonde-Birke F, Trommer S, Lenz B (2018) How autonomous driving may affect the value of travel time savings for commuting. Transp Res Record: J Transp Res Board

Szczurek P, Xu B, Wolfson O, Lin J, Rishe N (2010) Prioritizing travel time reports in peer-to-peer traffic dissemination. In: 7th international symposium on communication systems, networks & digital signal processing (CSNDSP), 2010, Newcastle upon Tyne, pp 454–458

UK Government, The key principles of vehicle cyber security for connected and automated vehicle. https://www.gov.uk/government/publications/principles-of-cyber-security-forconnected-and-automated-vehicles/the-key-principles-of-vehicle-cyber-securityfor-connected-and-automated-vehicles. Accessed 28 Feb 2020

United States Government Accountability Office (2016) Intelligent transportation systems: vehicle-to-infrastructure technologies expected to offer benefits, but deployment challenges exist. Vehicle-to-infrastructure technologies. Nova Science Pub Inc., New York, pp 1–50

Yang F, Wang S, Li J, Zhihan L, Qibo S (2014) An overview of internet of vehicles. China Commun 11(10):1–15

Yap MD, Correia G, van Arem B (2016) Preferences of travellers for using automated vehicles as last mile public transport of multimodal train trips. Transp Res Part A: Policy Pract 94:1–16

Yuan Y, van Lint H, Hoogendoorn S, van Wageningen-Kessels F (2014) Network-wide traffic state estimation using loop detector and floating car data. J Intell Transp Syst 18(1):41–50

Zeadally S, Hunt R, Chen YS (2012) Vehicular ad hoc networks (VANETs): status, results, and challenges. Telecommun Syst 50:217–241

Zhong T, Xu B, Wolfson O (2008) Disseminating real-time traffic information in vehicular ad-hoc networks. In: IEEE intelligent vehicles symposium, 2008, Eindhoven, pp 1056–1061

Part III
Data Analytics and Models for Dynamic Traffic Management

Chapter 6
Dynamic Traffic Management: A Bird's Eye View

Jaume Barceló and Margarita Martínez-Díaz

Abstract Traffic systems evolved rapidly, becoming soon a specific case of a complex dynamic system, what raised the need for controlling them in order to achieve an efficient performance. One of the main factors of complexity of traffic systems is a consequence of the variable human traveling behavior in time and space. Therefore, traffic control, in the way it had been conceived and implemented, appeared as a restrictive approach just considering one of the control aspects: the time the vehicles are flowing through the network. This raised the need to move a step forward. Thus, traffic management could be seen as an extension of traffic control that simultaneously controls time and space, and is aimed at adjusting the demand and the capacity to avoid mismatching. This chapter summarily reviews the main concepts and approaches in the development of traffic management systems (TMSs) both in terms of managing the supply as well as managing (or influencing) the demand. In this context, travel times become one of the key factors to induce changes in drivers' behavior in terms of making decisions on departure times and route choices. To better achieve such objectives, it would be desirable that TMS have predictive capabilities. The main approaches addressed here support the predictive capabilities of dynamic traffic models, one of whose main components is an estimation of the dynamic mobility patterns in terms of origin to destination (OD) matrices. This chapter summarizes the architecture of such approaches.

J. Barceló (✉)
Department of Statistics and Operations Research, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Spain
e-mail: jaume.barcelo@upc.edu

M. Martínez-Díaz
Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_6

6.1 Introductory Remarks

The development and evolution of dynamic complex systems in engineering and other processes raised soon an increasing interest in controlling them, in order to ensure that they could behave in an efficient or optimum manner, while holding the stability necessary for such efficiency. In essence, from a conceptual perspective, a basic control system is supported by the idea of a feedback loop (Fig. 6.1), which assumes that the system S object of control is in state $E(t)$ at time t. This state is characterized by the values at that time of a set of variables $\{v_1(t), v_2(t) \ldots v_n(t)\}$, the state variables, which are the output of the system. The main hypothesis is that the state variables are observable, that is, measurable, and they provide the information required to act on the system to drive it into a desired state, more efficient or optimum than the current one.

A key assumption is that the system is equipped with a set of sensors from suitable technologies providing the measurement of the system output, that is, the values of the state variables defining the state of the system $E(t)$ at time t. The objectives of the control process are presumably defined in terms of control policies, formulated as control variables, which provide the references of the expected values of the state variables determining the desired optimal (and efficient) behavior of the system under control. Then, the measured values are compared with the desired ones and their differences, the error measurements, are the input to the controller. This is equipped with the means to implement the corrective control actions, defined in terms of the appropriate control variables, to achieve the expected objectives.

A fundamental underlying concept is that of the observability (Castillo et al. 2008, 2015). Formally, a system is observable if, for any sequence of state and control vectors defined by the values of the state and control variables, the current state of the system can be computed in a finite time using only as input the output of the system. In other words, assuming that the approach to the system representation is based on the *Space State* description, its behavior can be totally determined from the output measurements if these are appropriate, what depends of the sensors layout (Barceló et al. 2012). Castillo et al. (2008, 2015) and Barceló et al. (2012), as many other references, deal with the sensor layout problem from what nowadays could be considered a classical from the perspective of sensor technology, i.e., inductive

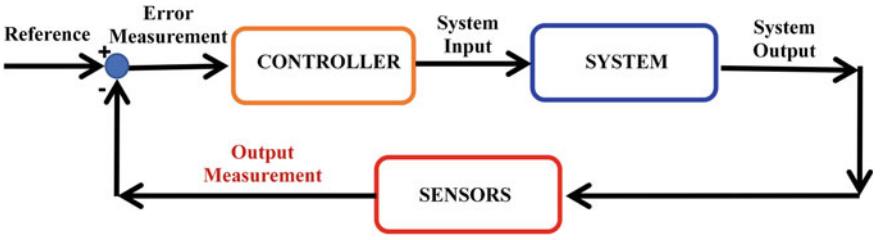

Fig. 6.1 Basic conceptual diagram of a feedback control loop of a system

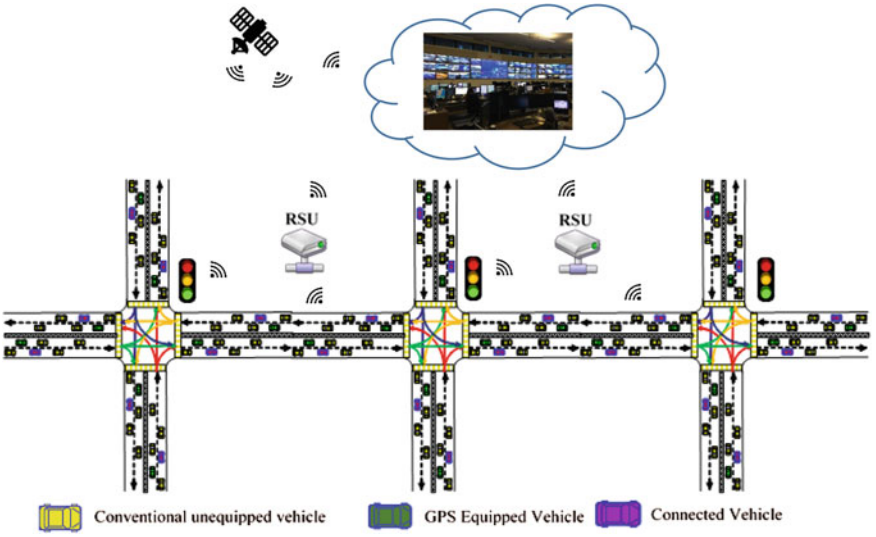

Fig. 6.2 Conceptual example of an adaptive urban traffic control system

loop detectors, radar measurements, license plate recognition, or Bluetooth devices, among others. The irruption of new Information and Communications Technology (ICT) applications, i.e., smartphones or GPS, offers an amazing set of rich possibilities to enhance traffic data collection. Fang et al. (2016), Ibarra-Espinosa et al. (2019), or Antoniou et al. (2019) are good examples of these possibilities. More recently, the forecasted advent of Connected Automated Vehicles (CAVs) has prompted new possibilities (Montero et al. 2016; Xianfeng 2018; or Martínez-Díaz et al. 2019). These new technological scenarios are graphically depicted in Fig. 6.2.

Traffic systems were soon identified as a special case of a dynamic complex system, whose complexity is a consequence of the variable human traveling behavior in time and space. Therefore, the need for an effective control to optimize their performance was also soon detected (Papageorgiou 1983). The first versions of traffic control systems, namely in urban scenarios, were essentially static, based on the average observed traffic behavior, the physical attributes of the road object of control, and the type of available control actions, i.e., the traffic lights settings defined in terms of cycle lengths, splitting of light colors and offsets between successive intersections. These control systems have been continuously evolving toward real-time adaptive control systems aimed at making them as responsive as possible to better fit the time variability of the demand (Gartner 1985). This evolution has been determined by that of the technology, hardware, and software, and its ability to measure and processing in real time the values of the state variables.

Figure 6.2 schematically depicts an example of a hypothetical forthcoming adaptive urban traffic control system. In this system, in addition to the conventional traffic sensors (i.e., inductive loop detectors), GPS-equipped vehicles, and the envisioned CAVs capable of communicating with the traffic controllers become new information

sources, that could be potentially used either locally or in a centralized coordinated way through a "traffic control center" likely operating in the cloud.

For freeway/highway networks, with no signalized intersections, control is based on other approaches. For linear infrastructures, for instance, on freeways, where traffic flows are determined by input flows at on-ramps and exit flows at off-ramps (Martínez-Díaz 2018; Soriguera and Martínez-Díaz 2020), an example of control approaches are those based on ramp metering. This is precisely aimed at controlling these input flows to ensure the fluidity at the main stream.

The general increasing congestion trend, which is especially acute in large metropolitan areas, has prompted the interest in understanding its causes. Also, in looking for solutions, as congestion has clearly negative impacts on the quality of life due to its associated social costs (e.g., waste of time spent in congestion, adverse effects on sustainability in terms of energy consumption, contribution to greenhouse gases and obnoxious emissions endangering human health) (Barceló 2019). These solutions aim at remedying congestion in the long term, and at least alleviating its consequences in the short-medium term. Traffic management has emerged as the most appealing strategy to extend the concept of control in achieving these objectives of fighting congestion and its adverse consequences.

Traffic control, in the way it has been conceived and implemented, as summarily described above, could be considered as a restrictive approach to traffic management, as it only deals with one aspect: the time vehicles are flowing through the network. Traffic control tries to make traffic more fluid, that is, to minimize travel times and delays, while maximizing the use of the available local capacity. However, it does not consider how the space, that is, the total available capacity, is being used. The absence of this global view may lead, and in fact it leads, to unbalances in the use of this available capacity. The other way around, assuming that congestion is a consequence of the timely mismatching between traffic demand and supply (in terms of capacity), traffic management could be seen as an extension of traffic control that simultaneously controls time and space, and is aimed at adjusting the demand and the capacity to avoid this mismatching.

To achieve these objectives, primarily at freeway/highway networks, traffic management has usually combined two type of policies. First, those trying to influence the use of the road network increasing the throughput, regulating traffic inflows, and preventing spillbacks. These are management measures based on control policies that are aimed at preserving the fluidity in the main stream and at avoiding capacity drops, caused by fixed bottlenecks or generated by shock waves. Examples of these policies are the *speed control*, the *lane management,* and the *ramp metering*. Second, those trying to achieve a better distribution of traffic across de network. These policies, commonly referred as *Demand Management*, can be of different types:

- *Travel Time/Route Information Systems*

  They are aimed at influencing the way in which routes are selected and used from travelers' origins to destinations. In other words, they provide reliable information on the current and forecasted traffic conditions, so that drivers reach their

destinations at the desired or expected times. Another goal is to flatten the time distribution, lowering the pick periods of the demand by informing about departure times that minimize travel times while ensuring an acceptable degree of reliability. Thus, these systems shift drivers to off-peak periods.

- *Road pricing, Low Emissions Zones, Wide Area Access Control*, and similar measures

  These policies impose conditions to access a given area in order to reduce inflows to that area, or to ensure that only vehicles fitting certain technological (sustainability-related) conditions can access it.
- Policies that are the consequence of societal and technological changes

  For example those enabling teleworking or virtual access to the realization of any activity without physical displacement. Also, those shifting the current travel modes to other transportation modes (including the emerging micromobility modes), with the support of suitable and reliable information. Those policies that prompt modal chains using different modes for each trip leg (e.g., walking or cycling first, public transport like bus, metro, or commuting trains next, and finishing by using other alternative mode to reach the destination) must be highlighted, as they will become increasingly important.

Figure 6.3 translates these ideas into a conceptual scheme in which it is assumed that the traffic network is suitably equipped with all kind of sensors, both conventional and new technologies, and that an appropriate traffic surveillance system timely collects, stores, and processes the measured data. Traffic data are the main input to:

- A *Network Traffic State Estimation Module*, which determines the current state of the traffic network based on the available measurements and suitable models.
- A *Short-Term Traffic State Prediction Module*, which estimates the forecasted evolution of the traffic state according to the suitable traffic management policies determined by a *Traffic Management Decision Support System* (DSS). The latter is aimed at proposing the control actions to alleviate or prevent the deviations from the desired traffic state, usually defined in terms of a wide set of key performance indicators (KPIs) that translate traffic managers' objectives. These goals range from the conventional ones, i.e., maximizing the throughput, minimizing travel times, reducing congestion, avoiding bottlenecks, etc., to others addressing sustainability objectives like minimizing environmental impacts or energy consumption.
- The *DDS*, which recommends the traffic manager the alternative management actions to activate and indicates their respective potential impacts. Nevertheless, the traffic manager is the ultimate responsible for implementing the policies corresponding to such actions.

The TMS also includes a module to operate the appropriate actuators to impose the decided actions like ramp metering, speed control, lane management, in freeway networks, or gate-in/gate-out zone access policies in wide area control, for example.

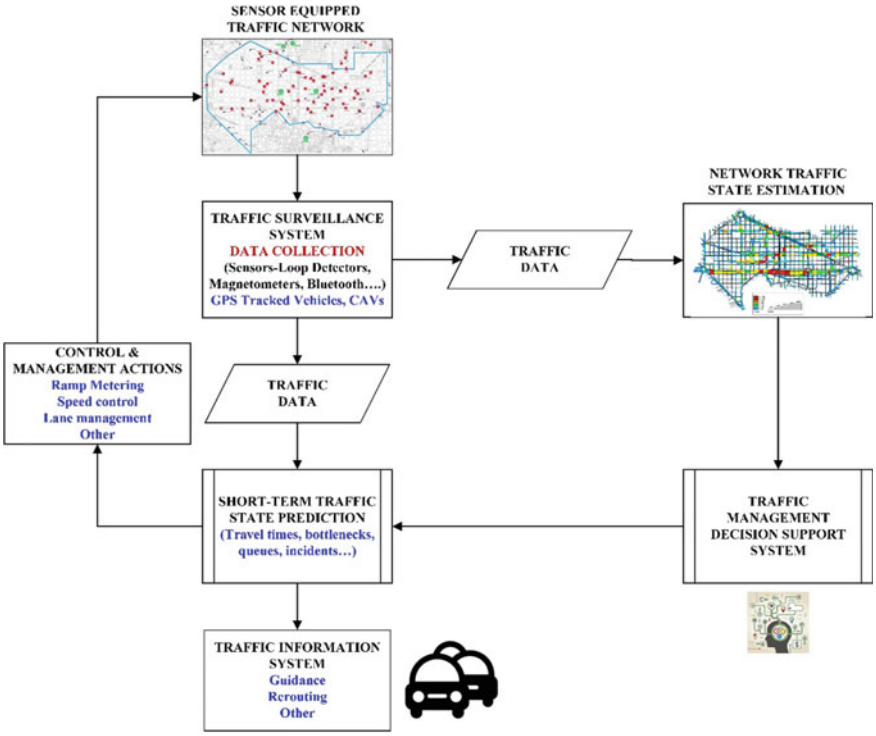

Fig. 6.3 Conceptual scheme of a TMS

Demand management was soon also identified as a set of complementary management actions that, if properly implemented, could contribute to reduce the pressure on the traffic network by flattening the peaks of the time variability of the demand and spreading it over time. Additionally, it can balance the available road capacity with suitable re-routing for a more efficient use of the space. This kind of policies were usually implemented conveying the information to users by displaying the corresponding messages in a set of variable message panels suitably located at key locations in the traffic networks. This dissemination can currently be more pervasive and efficient using the big variety of available mobile devices. That is, conveying the information to drivers in real time, either on-route, i.e., while traveling, dynamically recommending re-routing, or before starting the trip, proposing the most appropriate departure time, routes, or alternative transport modes to use.

6.2 ITS Approaches and Artificial Intelligence

As explained, traffic management was soon identified as a step ahead beyond conventional traffic control, required by the increasing congestion in freeway networks and metropolitan areas, which in turn was a consequence of the growing motorization. Moreover, in the late eighties of the past century, new ICT applications that could potentially enhance traditional management systems started to emerge. The many projects addressing the topic from the early stages of the European Programs are a clear demonstration of the interest that traffic management has aroused and continues to arouse.

Leaving aside the technological aspects, i.e., data collection and management actuators, *grosso modo*, it is possible to differentiate two approaches to traffic management, including the network traffic state estimation, the short-term traffic state prediction, and the core traffic management DDS determining the policies that the traffic manager should implement to avoid or alleviate the conflictive identified situation. The first ones are those approaches inspired in an extension of the control theory conceptually illustrated in Fig. 6.1, and the second ones are those approaches based on Artificial Intelligence (AI). Kirschfink et al. (2000) document this interest in applying the new, and presumably more powerful, techniques of AI to complex traffic problems, expecting that they will be able to overcome the limitations of the current systems when facing critical conditions and congestions. They argue that these traditional systems have been usually conceived in terms of local traffic behavior, while the addressed problems are more global and, therefore, need strategic, high-level approaches.

Two meetings, the *ERUDIT Tutorial on Intelligent Traffic Management Models*, held in Helsinki on 1999, and the *European Symposium on Intelligent Techniques* ESIT'2000, held in Aachen, in 2000, provide a nice panoramic of the state of the art of the developments during more than 10 years (since the beginning of the European Programs) on what at that time was called *Advanced Transport Telematics* (ATM). ATM addressed the applications of the new emergent technologies, Computer Sciences (Informatics) and Telecommunications, to transportation systems and later become ITS. However, from the very beginning, AI was considered one of the main technologies to account for. Rass and Kyamakya (2007) provide a more extended overview of this progress. For their part, Kirschfink et al. (2000) summarize the application of *Advanced Knowledge Modeling Techniques* to *Intelligent Traffic Management Systems* (ITMS), assuming that these systems are conceptually defined to implement two types of measures:

- Direct control measures, that is, measures aimed at managing the infrastructure. For example, control of traffic lights, ramp metering, speed control or variable message signs.
- Indirect control measures, aimed at managing the demand, consisting on recommendations for drivers by means of variable direction signs (VDS), text panels,

or Relational Database Management System-Traffic Message Channel (RDMS-TMC) messages (nowadays replaced by advanced journey planners and navigation systems).

They also explain that TMSs are supported by a global architecture with two main components:

- A *Traffic Surveillance System* that collects and stores traffic and environmental data.
- A *Traffic Control Centre* (TCC) that must be able to suitably process all the detected data by the appropriate algorithms for different purposes. The most important ones are the estimation of the traffic state and the monitoring of the current traffic situation, the estimation of short-term predictions of the potential evolutions of the traffic state, the proposal and coordination of control measures, and the transmission of appropriate recommendations to drivers.

To overcome the above-mentioned limitations of the conventional systems, various approaches based on AI techniques, namely *Knowledge-Based* and *Inference* systems, have been proposed. Examples of these are TRYS (Hernandez et al. 1999), FLUIDS (Hernández 1999) or KITS (Kirschfink et al. 2000). The common idea to all of them is the development of a system that embodies a knowledge model of traffic behavior at a strategic level and is assisted both by knowledge management techniques supporting rule chaining for pattern matching methods, constraint satisfaction, etc., and inference machines able to reason on the acquired knowledge. To achieve such system, they propose adding a functional level in current TCCs, on the top of the existing facilities. This level consists on a *Traffic Knowledge Processing Layer*, whose goal would be to improve the online traffic monitoring and management system. Therefore, this layer would enhance TCCs operations for:

- The estimation of traffic load levels in space and time all over the network.
- The analysis and understanding of traffic demand and routes in the area.
- The qualitative prediction of demand and routes.
- The detection (prediction) of critical traffic situations and bottlenecks.
- The selection and implementation of congestion avoidance/reduction strategies.
- The management of conflictive control objectives and priorities in the different controlled areas.

The experience gained after the first developments and pilots projects reported in KITS (Kirschfink et al. 2000), conceives the knowledge models as a hierarchically structured problem solving, in which a problem domain can be naturally decomposed into substructures. Each of these substructures is specialized in the solution of a class of problem, which in turn can be decomposed into simpler and specialized issues to be addressed. The substructures considered in KITS are:

- *Agents*, which represent a local reasoning process that identify problems, interpret available information, perform specific tasks (i.e., short-term traffic prediction), support local decision processes, and detect inconsistent combinations (i.e., control actions in overlapping areas).

- *Actors*, which are knowledge units specialized in traffic evaluation and management within a particular area. They include reasoning strategies that use agents to solve interpretation and decision problems in the area. They are in correspondence with the topological decomposition of the traffic network.
- *Supervisors*, which are combinations of actors and can be of different types. For example, *Masters* act in a prescriptive way, *Mediators* negotiate conflicts and *Facilitators* act as managers and recognize which Actor has the external knowledge eventually required by each other Actor to perform local tasks.

KITS and TRYS are examples of such architectures whose knowledge modeling approach, quoting from Kirschfink et al. (2000), is supported by two structuring principles:

- A *functional organization*, which functionally decomposes the domain knowledge in specialized units targeting specific types of problems in this domain.
- A *topological organization* that spatially breaks down the traffic network into a set of so-called *Problem Areas*.

Functional and topological organizations should match to ensure that the knowledge and reasoning levels correspond. The envisaged knowledge-based traffic management system would then consist of a "*structured collection of knowledge units, providing specialized knowledge and reasoning mechanisms to deal with the different types of traffic management activities and reflecting the functional and topological knowledge*." The main agents considered in KITS (Boero 1999; Kirschfink et al. 2000), whose architecture is illustrated in Fig. 6.4, were:

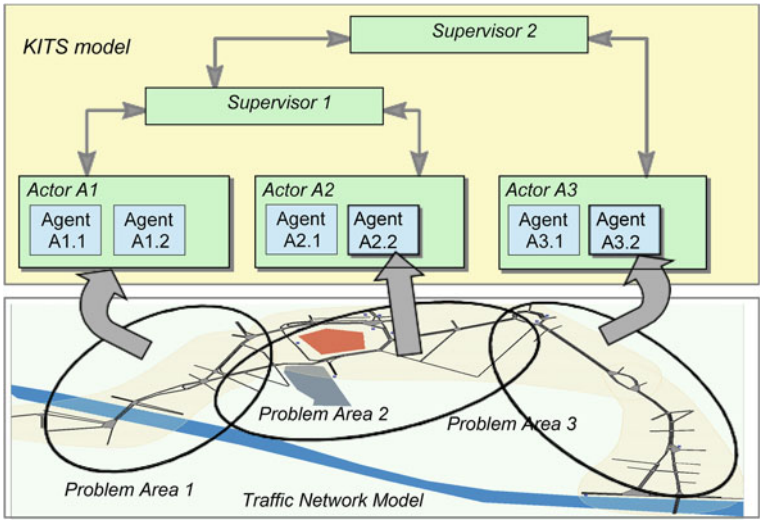

Fig. 6.4 KITS model of AI-based traffic management Boero (1999)

- The *Problem Identification Agent* and the *Traffic Flow Behavior Modeling and Causal Agent*. The first one identifies the problems using heuristic classification criteria based on the historical knowledge of problems in the area and evaluates the information provided by the data collection system. The second one performs the analysis and short-term evolution of traffic flows and looks for causal explanations of the detected problems. The type of agents based on traffic flow models has represented one of the most successful evolutions of these systems, as will be discussed later on.
- *Local Decision Agents* that support the suitable control actions for specific problem areas as consequence of the previous processes, and agents to detect inconsistencies of multiple control actions or actions that could potentially interfere between overlapped areas.
- Finally, a *Strategy Completion Agent* based on a global overview of the controlled network, which synthesizes and harmonizes the local proposals.

For its part, the MOTIC concept (Boero et al. 1997; Kirschfink et al. 1997; Boero and Kirschfink 1999) depicted in Fig. 6.5 could be considered as a hybrid architecture. In this approach, the AI components are combined with traffic simulation models to assist both in the analysis of the situation and in the evaluation of the impacts of the planned control strategies prior to their implementation. Additionally, instead of predefined *Problem Areas* as in KITS, a more flexible concept is introduced. This is that of *Scenario Definition*, enabling a graphic interactive process to dynamically determine the area potentially affected by the identified problem. Thus, once the

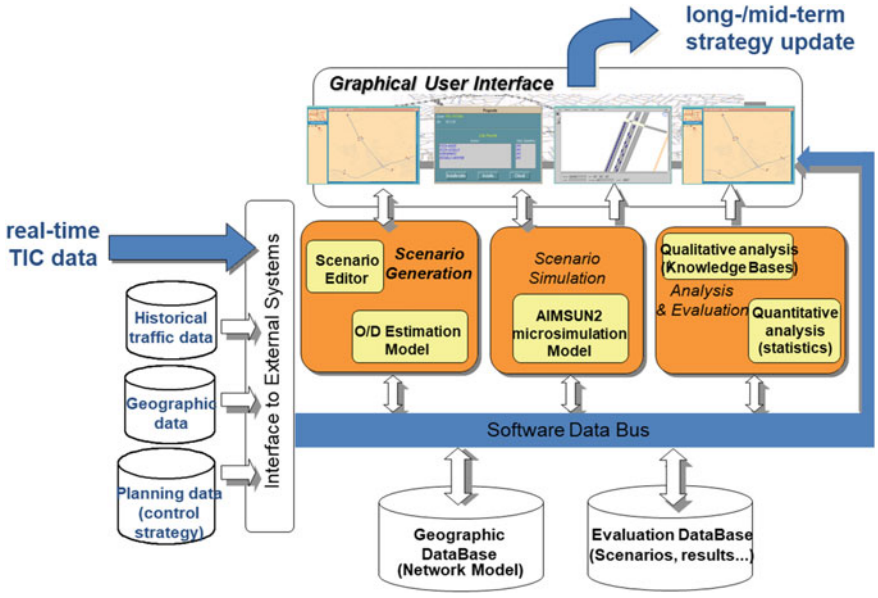

Fig. 6.5 The MOTIC approach to traffic management (Boero and Kirschfink 1999)

scenario is defined, it can be simulated. The simulation results in terms of KPIs can then be used to analyze the scenario before generating the information that will be used to apply the suitable policies and, thus, before disseminating the corresponding suggestions/orders. The use of microscopic simulation in the scenario analysis also enables a simulation-based learning process based on the collected data before and after the actuations. This allows defining and evaluating new management strategies, optimizing the existing ones, assessing the potential impacts of any variations of the strategies, determining the optimal timing to activate actuations, etc. MOTIC was developed and preliminary tested in a pilot test in the project ENTERPRICE, of the ATT European Program of DG XIII of the 4th Framework European Programme 1997–1999 (Boero and Kirschfink 1999).

A differential fact of MOTIC with respect to other approaches, which results evident when analyzing the architecture depicted in Fig. 6.5, is its hybrid structure, which includes AI and traffic models. AI is used to analyze and understand what is going on. In other words, to interpret the identified network state based on the current analysis and past experiences from similar situations. However, traffic models support this network estate estimation as, for example, microsimulation models. Aimsun2 (Boero et al. 1997), an earlier version of Aimsun, was particularly used in the above-mentioned applications. The interactively generated simulation model of the selected scenario, corresponding to a Problem Area, is used to evaluate the alternative strategies to solve the identified problem in terms of specific KPIs, to support the decision-making process.

Another differential aspect of this approach that deserves to be highlighted is that the support of an advanced microsimulation model (Barceló et al. 2004; Barceló 2010), requires an input that is not yet directly observable from traffic measurements: **the model of the mobility patterns in terms of an OD matrix** (see Chap. 7). The dynamic simulation model will describe how the trip makers use the paths connecting origins and destinations depending on traffic conditions. Therefore, it will be able to identify how congestions are generated as well as the spillback across the network and, consequently, to emulate how road users will likely react to the management actuations. A deeper insight into the role of microscopic simulation in ITS applications can be found in Barceló et al. (2004).

A system inspired in MOTIC was developed and applied in the Intermodal Strategy Manager ISM (Barceló et al. 2002; Kirschfink et al. 2003). This was a development within the framework of the Hessian WAYflow-project, with the goal of improving traffic management in the Rhine-Main area by supporting the planning of new strategies, starting with their impact assessment and coordination up to their implementation, as well as by optimizing the already existing strategies. The *Scenario Analysis Module* (Barceló et al. 2004), in Aimsun/ISM uses an Aimsun microscopic traffic simulation model of the traffic network under study to define, verify, and optimize traffic management strategies, evaluate their expected impacts, and determine the triggers for their activation, according to prevailing traffic conditions. A scenario is a microscopic simulation model of a traffic network (or a subnetwork of a large network) in which a traffic problem has been identified. This is consistent with the so-called *Problem Network*.

The model input reproduces the traffic demand in the problem network for the time period for which the traffic problem has been identified with a great degree of accuracy. Also, the current operational conditions in the road network, such as current traffic control at signalized intersections, reductions of capacity at specific parts of the network caused by road works, incidents, etc. The analysis of the scenario consists of a set of simulation experiments whose purpose is to help traffic managers to develop and evaluate the impacts of the single actions or combination of actions. These actions are situation-related measures (i.e., re-routings and/or speed control using variable message signs—VMS—changes in control, etc.), with the objective of alleviating or eliminating the traffic problem identified. This concept of action composed by the various situation-related measures is called a *Strategy*. The evaluation of alternative scenarios, i.e., models of the same problem network with alternative traffic management strategies, compares the values of performance indexes that express saturation levels, quality of service, total travel time, average delays, average queue lengths, or total vehicle-kilometers traveled, among others.

The main objective of Aimsun/ISM is to allow the fast and convenient manipulation of input data to create simulation scenarios and to present result data in a compressible way. It has two main components: the *Simulation Experiment Specification* and the *Result Analysis*. The *Simulation Experiment Specification* includes the setup of a Problem Network (either the network of the whole area or a sub-network), the creation, modification, and adjustment of OD matrices (again global for the whole area as well as local or traversal for the sub-networks), the addition of traffic management policies and their triggers and the simulator tuning. For its part, the *Result Analysis* includes the output data presentation and the comparative study of the performance of a solution, either with previous solutions or with real data. Since a problem can have different solutions and taking into account that these solutions may not be obvious, users can define several experiments combining different policies until the best option is found. During this experimentation, users can also reuse previous solutions and add new ones to, as said, comparing their performance among them and/or with real data. This process can be repeated iteratively until a satisfactory solution is found. The above-mentioned components provide the support for the generation, evaluation, and optimization of traffic management strategies.

The Aimsun/ISM operation is illustrated in Fig. 6.6, where the WAYFLOW Global Network and a potential Problem Network are shown. A Problem Network corresponds to a sub-network of the road network on which a specific traffic problem may arise or is identified by the user. This user can define the target Problem Network graphically by opening a window on the screen on which the WAYFLOW network is displayed (the rectangle in Fig. 6.6 corresponds to the selected Problem Network). Any Problem Network is characterized by the road network within the defined window and an OD database with their distinct demand patterns under various circumstances (season, day of the week, time of the day, special event, etc.). Also, by a strategy database containing the specifications of the potential traffic management strategies to operate on the Problem Network depending on the identified or expected traffic problem and on the demand pattern.

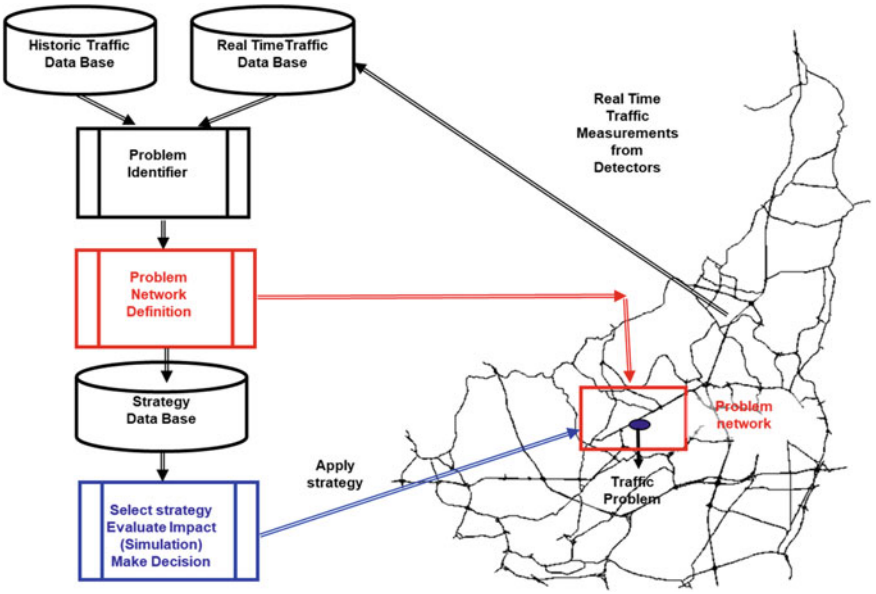

Fig. 6.6 ISM system in the WAYFLOW network in Hessen

Variants of the former approaches can be found in Barceló et al. (2007a, b, c) or Krishnan et al. (2010), among others. Figure 6.7 summarizes the main aspects of this evolution, which are:

- Traffic flow models to estimate the traffic state and its likely short-term evolution can be either basic as in KITS, TRYS, Krishnan et al. (2010) or Hegyi et al. (2009) or complex, usually based on dynamic simulation, as in Barceló et al. (2007a).
- There exists a *Basic Inference* (i.e., set of rules) *System* to infer the primary interventions/actions (i.e., policies) to apply based on the current state.
- These new approaches assume that, during its operational life, the TMS has generated two complementary databases that complement the real-time collected data. First, a historic database of recorded traffic data, identifying specific scenarios (i.e., recurrent congestions and their time evolution at specific locations). Second, a historic database of the control actuations applied to alleviate or solve specific conflicts and all data and related information associated to them.
- New approaches use more complex AI engines, based either a variety of knowledge-based or inference approaches, as explained for KITS, TRYS, or MOTIC, or on a pattern matching process. The latter find similarities between the identified situation and others previously recorded. This looking for similarities is usually a task for specialized clustering approaches, like in AURA (Krishnan et al. 2010).
- The applied advanced traffic models enable an estimate of the likely short-term evolution of traffic conditions in the network, and therefore an evaluation of the

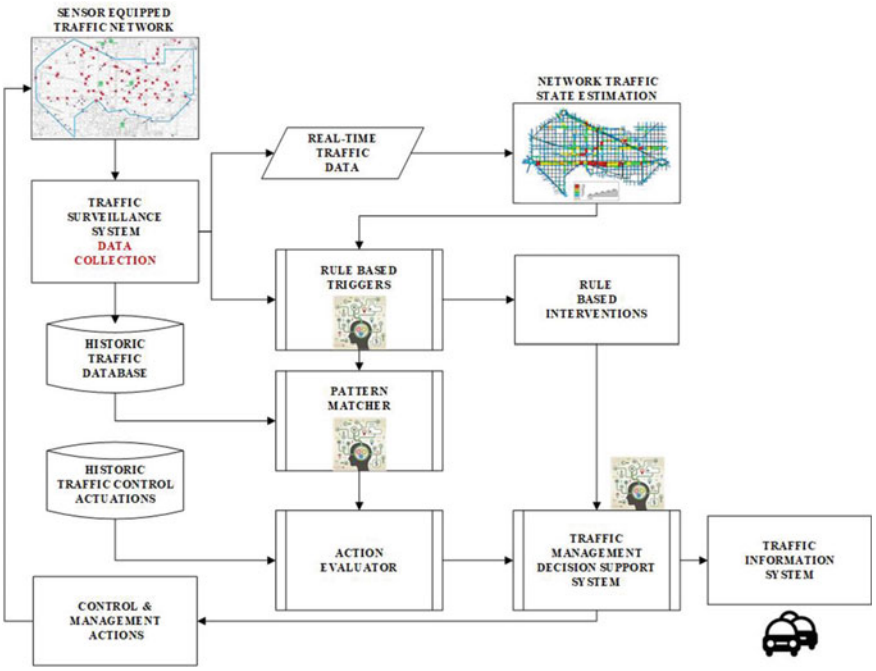

Fig. 6.7 Generic evolution of the AI traffic management-based architectures

potential impacts of the planned actions in terms of a selected set of KPI's, Barceló et al. (2007a).

- These approaches include a *DDS*, i.e., a generalization of the previous inconsistency detection and strategy completion actions, combining the primary rule-based actions and the evaluated actions after the pattern matching. It is responsible for making the final recommendation that the human operator will implement.

6.3 Current Hybrid Approaches

The pursuit of increasingly sophisticated traffic models continues beyond those introduced in the previous sections. However, the goals of these models have not varied. First, they aim at estimating the traffic state and at analyzing its similarity with other already experienced situations. Second, they address the short-term forecasting of its likely evolution under the planed management strategies as well as the evaluation of the potential impacts of these strategies in terms of selected sets of KPIs. This willingness for improvement has prompted the evolution of the generic architecture described in Fig. 6.7 toward some of the commercial systems used in projects that are more recent. Some of their key characteristics are:

- Their ability to dynamically identify the problem area within a la large managed area.
- Their capacity to estimate and adjust the OD matrix by identifying the current traffic patterns in the problem area from both historic records and current traffic measurements.
- Their pattern matching process to find similarities between the identified situation and the historically recorded ones (Mounce et al. 2012).
- They deliver a set of KPIs to evaluate the alternative scenarios according to the possible strategies to apply. Professional platforms usually offer the possibility of customizing the type of KPIs that will support the traffic management decision-making process. However, KPIs based on speeds, travel times (usually path travel times), expected delays, queue lengths, etc., are always present. Therefore, their forecasted values for complex traffic networks in these traffic management platforms are usually based on dynamic traffic models able to estimate path travel times, for example, the dynamic traffic assignment models.

Figure 6.8 adapted from Barceló et al. (2007a), depicts the adaptation of this architecture, to Madrid Calle-30, evolved as described from the ISM project. This became the basic version of Aimsun On-Line, (https://www.aimsun.com/aimsun-live-case-studies/madrid/). The conceptual diagram in Fig. 6.8, highlights this hybridization of AI and analytic components:

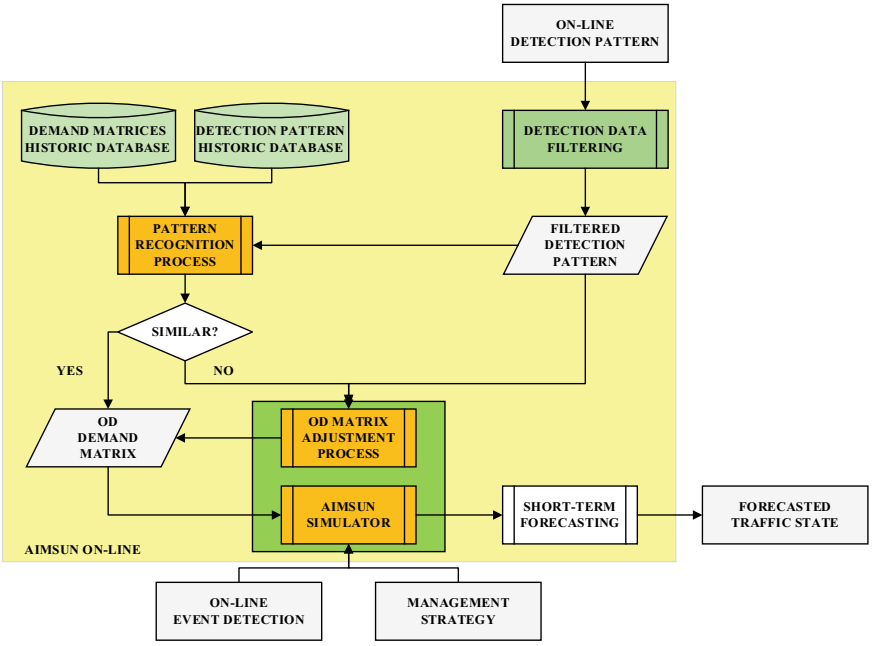

Fig. 6.8 Conceptual diagram of Aimsun on-line platform for real-time traffic management

- The pattern recognition process matches the traffic state identified after the measurement of traffic data with a likely OD pattern, in terms of the demand generating such situation, with some previously historically recorded similar situation.
- In case such similarity does not exit, a new OD is adjusted based on the historical and the current information.
- A traffic simulation model of the target network (Aimsun in this case) is fed with this information and the information from the event detection (incidents or scheduled special events) and the planned strategies, to estimate the forecasted traffic state that will be evaluated in terms of the selected KPIs.

These platforms have been continuously evolving since then. Figure 6.9 depicts the most recent version of Aimsum's platform, called Aimsun Live (Aimsun 2020).

The process has two operational modes: training and prediction. The training mode, as depicted in Fig. 6.10, is responsible for generating the set of mobility patterns. Input data is preprocessed (i.e., data filtering and missing value imputation) and standardized. Next, mobility patterns are extracted using a guided clustering algorithm with the aim of minimizing their number while maximizing their quality, i.e., minimizing cluster variability. The algorithm allows external (or previous) mobility patterns as optional input. This feature is of especial importance for an incremental (or iterative) pattern extraction methodology.

The prediction process is illustrated in Fig. 6.11 and it consists, again, of preprocessing and standardizing input data. Next, the pattern matcher uses traffic data and exogenous variables to estimate the likelihood of each pattern. Note that

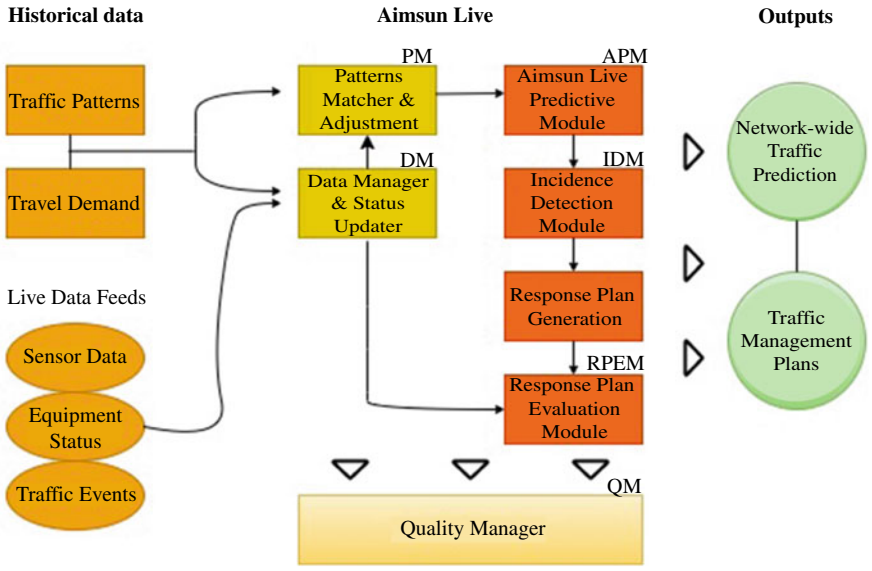

Fig. 6.9 Conceptual structure of Aimsun Live

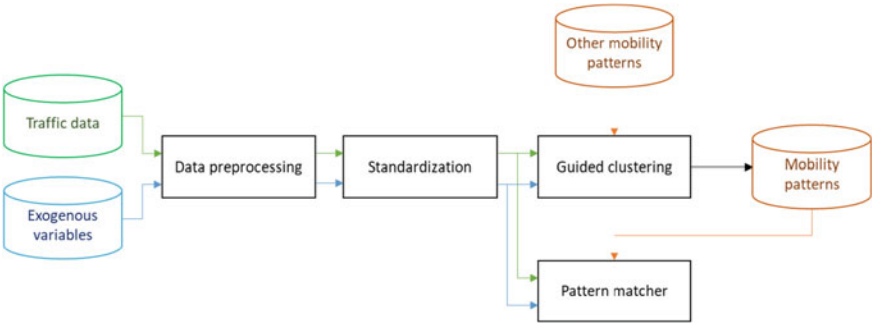

Fig. 6.10 Training workflow of pattern generation and matching process. The colors of the arrows represent the type of data (green for traffic data, blue for exogenous variables, orange for mobility patterns). The solid arrows represent mandatory input/outputs and the dashed arrows represent the optional ones

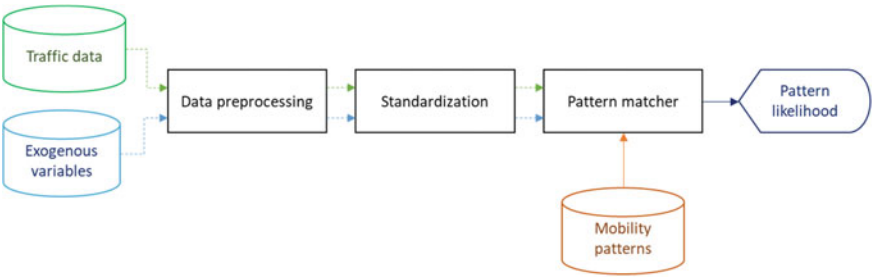

Fig. 6.11 Prediction workflow of the pattern matching process. The colors of the arrows represent the type of data (green for traffic data, blue for exogenous variables, orange for mobility patterns). The solid arrows represent mandatory input/outputs and the dashed arrows represent optional ones

this likelihood is estimated using input data that was either measured or predicted. It is also important to highlight that traffic data and exogenous variables are optional inputs of the pattern matcher. Therefore, depending on the type of input data used to feed the pattern matcher, it will estimate the likelihood of each pattern to occur in the present or in the future. For example, if we feed the pattern matcher with calendar features regarding today and the last N hours of traffic data, it will return the probability of each pattern of happening today. But if we input only calendar features of next Monday, it will return the probability of each pattern to occur next Monday. Therefore, the pattern matcher can be used to predict current and future patterns.

Additional references on traffic management systems based on the use of Aimsun Live can be found in the websites of San Diego Interstate 15 Integrated Corridor Management System, Sydney—M4 Smart Motorway, Wiesbaden DIGI-V, Florida DOT or Aimsun Live Technology Trial (Singapur), among others.

6.4 Other Approaches

At almost the same time, in parallel but independently, approaches different from those addressed in Sect. 6.3. were developed to tackle the problem of traffic management. A good example is that represented by RENAISSANCE (Wang et al. 2006). Conceptually, its architecture is a simplified version of the one depicted in Fig. 6.3, supported by METANET. This is a macroscopic freeway modeling method based on the fundamentals of traffic flow theory (Kotsialos et al. 2002; Papageorgiou et al. 2010), including enhancements from Wang and Papageorgiou (2005). Further extensions can be found in Wang et al. (2008), and Wang et al. (2009).

In this modeling approach, the freeway traffic state is estimated in terms of the traffic flow variables, i.e., flow, mean speed and density for each freeway stretch. All of them are defined with a suitable discretization in both time and space, exploiting the corresponding available real-time measurements. The proposed traffic state estimation combines traffic flow theory with Extended Kalman Filtering (EKF) in an efficient way that jointly estimates the model parameters online, including the fundamental traffic flow variables, the free-flow speeds, the critical densities and the capacities, adding significant adaptive capabilities.

However, the initial version of the RENAISSANCE approach does not consider, in either its modeling or its operation, the potential of control measures like ramp metering, route guidance, or variable speed limits. In case these control measures are applied, RENAISSANCE only allows the estimation of their impacts from the real-time traffic measurements. For example, the impacts of ramp metering can be estimated from the on-ramp inflows and densities at the downstream segments after the on-ramp, route guidance affects the estimates of the turning percentages in junctions, and speed control influences the estimates of the model parameter values. The potential consequences of these control measures in the prediction accuracy is limited by a continuous updating of the prediction horizons in terms of an efficient rolling horizon technique. In this sense, RENAISSANCE is designed to be used as an intermediate layer between the traffic data collection system of the freeway network, which provides the real-time traffic measurements, and the traffic managers, which will take management decisions on guidance and control supported by the provided information. Supported by the same freeway modeling approach in RENAISSANCE, Carlson et al. (2010) developed coordinated control strategies at the network level combining variable speed limits and ramp metering.

The DynaMIT (Dynamic Network Assignment for the Management of Information to Travelers) of Ben-Akiva et al. (2010) is also included in this category of approaches based on traffic flow theory. DynaMIT is "*a simulation-based dynamic traffic assignment (DTA) model system that estimates and predicts traffic conditions.*" However, it has also been designed as a real-time system for the generation of predictive traffic information to support traffic management decisions, in this case primarily intended for route guidance. An example would be the dissemination of travel times by means of any technologies supporting the ATIS, either those placed in-vehicle or that located on the road side (e.g., VMS). A relevant differential feature

of DynaMIT is its objective of avoiding the adverse impacts of improper traffic information. In fact, the proposal and application of management decisions without a previous insight of their potential consequences could be risky, and later corrections would be necessary to deal with unexpected reactions. To achieve such objective, the models in DynaMIT are aimed at providing predictive information consistent with the conditions that drivers will experience in the network, thus accounting for traffic evolution.

A conceptual view of DynaMIT (Ben-Akiva et al. 2010) adapted to the generic framework in Fig. 6.3 is depicted in Fig. 6.12. It integrates models and algorithms designed to fuse data from various sources with two main functions: the network state estimation and the prediction of its short-term evolution. This is achieved by taking advantage of two main modeling components, a demand simulator and a supply simulator, and their interactions. The travel patterns in the network are captured by the demand simulator, whose main input are the time-dependent origin–destination flows, expressed as time-dependent OD matrices. Individual trips are defined in terms of the origin and destination of the trip, the departure time, and the selected

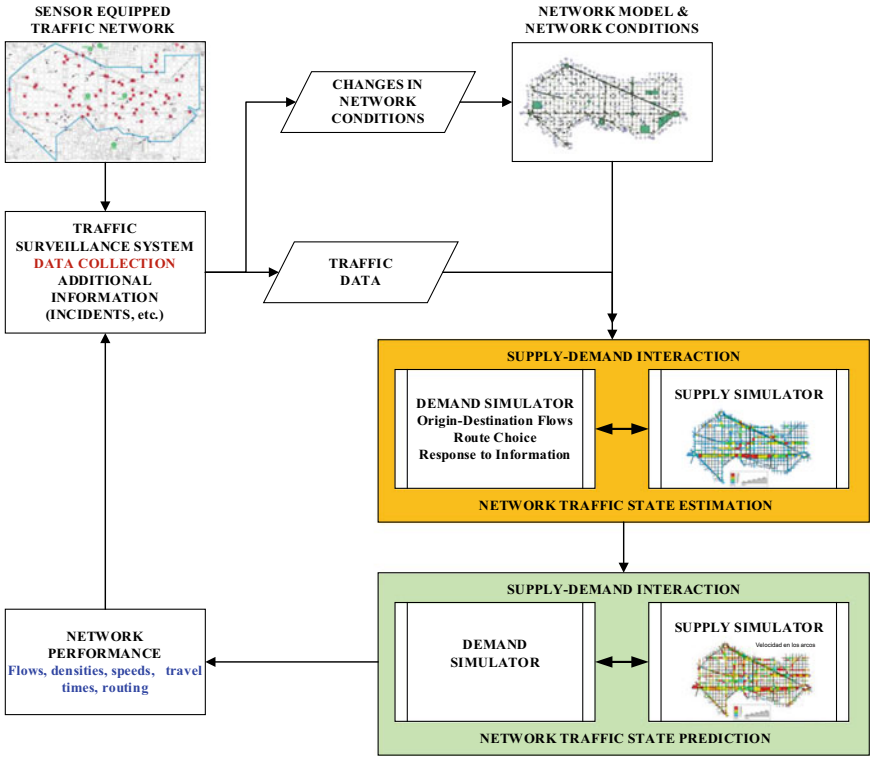

Fig. 6.12 The conceptual architecture of DynaMIT adapted to the general framework of Fig. 6.3

route, considering that these decisions are taken before the trip begins. The aggregated representation of the demand also accounts for individuals' socioeconomic characteristics. To anticipate the response of travelers to the information planned for dissemination (Ben-Akiva et al. 1997), DynaMIT also considers their potential access to such information. The supply simulator, which is mesoscopic, captures traffic dynamics, and evaluates the performance of the network, including the formation and dissipation of queues, spillback effects, or the impacts of incidents and bottlenecks. It represents traffic dynamics using speed–density relationships and queuing theory (Ben-Akiva et al. 2002). The algorithms (Ben-Akiva et al. 2010), used to estimate the current network state, to perform the short-term prediction of its evolution and to support the generation of the anticipatory route guidance and control strategies are suited to properly account for the complex demand–supply interactions. A critical aspect for the quality of the results is to ensure that the models are consistent with the prevailing conditions, which is achieved by means of the dynamic adjustment of the key model inputs and parameters.

Although with conceptual similarities with the framework approaches for traffic management discussed so far, Meschini (2017) proposes a different professional implementation in the platform developed by PTV SISTeMA. In this case, the TMS is the central component of the TCC. Functionalities are split in such way that the TMS is responsible for the collection of data regarding current traffic conditions, the physical implementation of the management decisions (whose final responsible is assumed the human operator) and the dissemination of the information and management decisions to other operators as well as to users. The TCC would be responsible for the continuous monitoring of traffic conditions based on the available measured data. However, the assessment of the current traffic state and its short-term evolution, accounting for the evaluation of alternative management strategies and for incident detection is implemented with the assistance of a DSS. This DDS, suited for real-time management, should support the ATIS and the Optimal Traffic Control (OTC). In agreement with the conceptual frameworks discussed in previous sections of this chapter, Meschini (2017) considers traffic management as a loop involving situation monitoring, identification of disruption, proposal of corresponding mitigation actions, and action implementation (Fig. 6.13). This loop should be fulfilled in time intervals short enough for real-time operations (i.e., within 1 and 10 min).

The real-time data collection envisaged in this approach goes beyond the conventional traffic data measurements provided by the usual technologies, i.e., traffic counts, flows, and occupancies on links, or instantaneous speeds. It accounts for floating car data (FCD), for instance, those provided by GPS devices placed in passenger vehicles, commercial vehicles, or public transport bus fleets, enabling the tracking of vehicle trajectories. It also considers travel time measurements, as those derived from automatic number plate recognition (ANPR) or systems of wireless network sensors. Such rich variety of traffic data coming from different sources is intrinsically heterogeneous. Therefore, there is a need for a data fusion process that merges all of them in homogeneous estimates of traffic variables, i.e., in flows, speeds, densities, and capacities of the links of the underlying transport model supporting the functionalities. This process requires a previous map matching procedure that

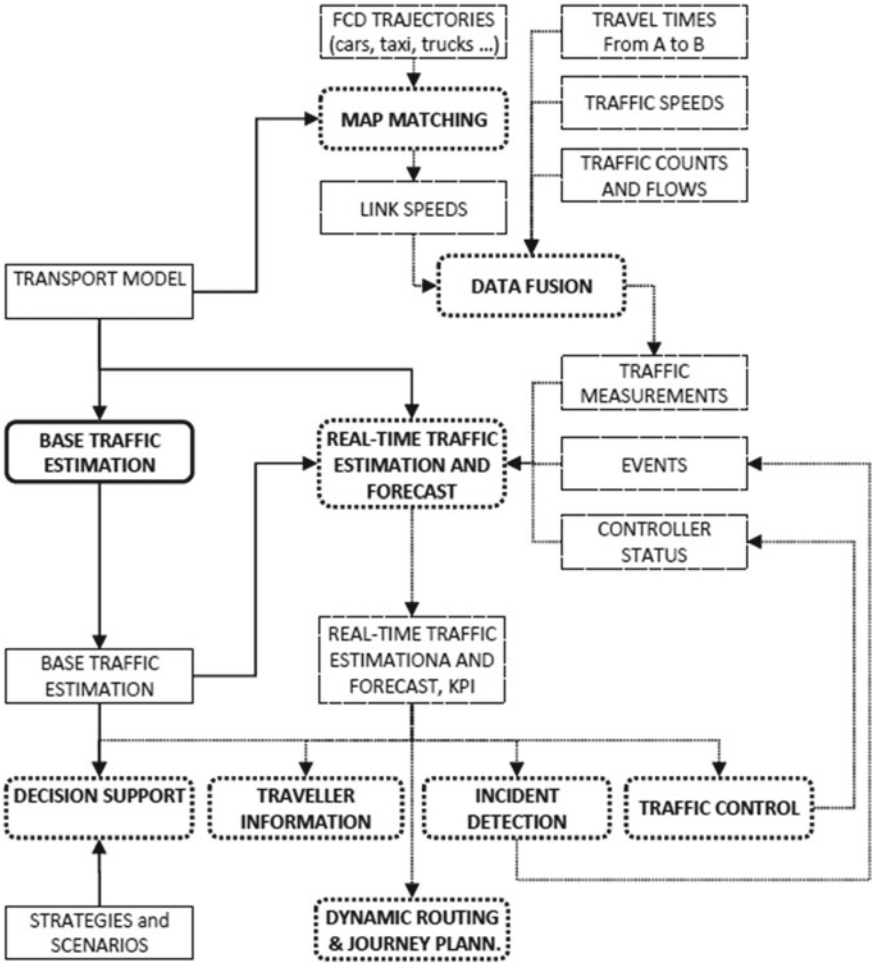

Fig. 6.13 Functional overview of the traffic management system proposed by Meschini (2017). Rounded boxes represent functions or elaboration activities and rectangle represent collected data and inputs produced or exchanged between functions

associates the measured data with the transport model. In the case of OPTIMA, the platform developed by PTV SISTeMA, this consists, first, of the topological representation of the network: links, nodes, turnings, connectors, zones, their attributes (e.g., link id, node id, link maximum speed, number of lanes, capacities, allowed transport modes, etc.). Second, of the functional representation of the road network, including all field related devices such as VMSs, traffic detector locations, traffic signals (and in this case the detailed information of timing and phasing), etc. The model-based traffic prediction also requires the definition of the transport demand defined by OD matrices, their temporal profiles and the modal segmentation.

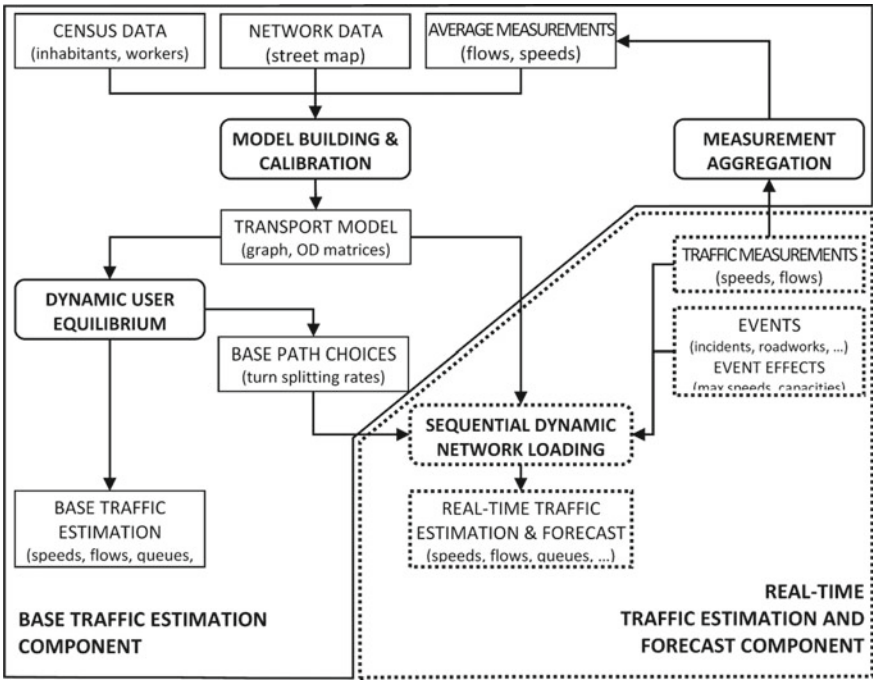

Fig. 6.14 Schematic description of OPTIMA model-based traffic estimation and prediction

In the case of OPTIMA, the proposed simulation and forecast approach is based on a methodology consisting of models and algorithms for dynamic traffic simulation. The logical architecture of the proposed system (Fig. 6.14) is composed of two parts: the offline and the online parts. The offline part is supported by a DTA model, which in turn is based on a dynamic user equilibrium (DUE) model (Bellei et al. 2005; Gentile 2010; Gentile et al. 2007, 2010; Meschini and Gentile 2010). Also, on the transport model that calculates the evolution of link flows, queues, travel times, and path choices over different time intervals within each typical day. The online part uses the model and the base traffic conditions provided by the offline part and combines them with the real-time measurements from the detection system. This is done by means of a traffic model that adjusts the estimations and forecasts to the measured conditions of that particular day. The online part runs automatically on a continuous basis providing a new traffic estimation and forecast in terms of travel times, traffic flows, and queues every few minutes.

OPTIMA was applied, for example, in the Regional Traffic Supervision Centre of the Piedmont region, in northwest Italy, in 2014. Its goal was to provide traffic information over the regional road network of Piedmont, thus enabling traffic management in the region. Another example of OPTIMA's applications is the customer-oriented traffic service www.AnachB.at, in Austria, which provides a comprehensive and effective traffic information system for travelers.

Other references to OPTIMA related traffic management projects are "2015–2020 PTV FR, Direction des routes Ile-de-France (DiRIF)—Supply of traffic data in real time and traffic supervision platform for Traffic operator of Ile-de-France motorways," or "Far EasTone Telecommunications (FET) for Taichung DOT, Taiwan-Development of a real time traffic management and control system for the City of Taichung." The latter provided proactive traffic management, short-term traffic prediction of up 1 h, incident management, alternative route guidance through VMS, and Active Warning using PTV Optima and PTV Balance.

In summary, all these approaches to traffic management are aimed at influencing simultaneously the supply, that is, the capacity of the network and the traffic demand so that they better match in time and space. The usual problems occurring in the road network primarily concern the generation of bottlenecks at specific points at given times, either at peak periods, when the demand exceeds the available capacity (i.e., in a recurrent way) or generated by incident or special events (i.e., in non-recurrent situations). The aim of traffic management is to prevent or alleviate these situations either spreading the traffic demand to different parts of the network or dynamically adapting the available capacity so that it serves the current demand. As explained in Sect. 6.1, ramp metering, dynamic speed limits, lane management, and similar policies are aimed at regulating the inflows, increasing the throughput or preventing spillbacks. They primarily act on the infrastructure maximizing the available capacity given the estimated traffic state and its likely evolution. Therefore, they are usually considered *Supply Management* policies. For their part, transportation *Demand Management* policies are aimed at spreading the demand over time, avoiding high concentrations at peak hours by (i) providing an alternative accessibility to the activities generating the trips (i.e., the case of teleworking), (ii) distributing the demand across the traffic network to use more efficiently the total capacity, or (iii) favoring modal split, particularly shifting the trips to transportation modes other than the private vehicle (i.e., public transport, cycling, walking, or, more recently, micromobility modes). Demand Management is usually implemented in terms of *Travel Time Information Systems* conveying the travelers, as in the case of DynaMIT, reliable information on travel times and their short-term forecasting. This information allows them to make better decisions in choosing the routes from their origins to their destinations. *Advanced Journey Planners* that enable trip makers to make better decisions when choosing the routes and transportation modes (or combinations of them) and, therefore, to achieve their goals more efficiently, also belong to Demand Management.

These approaches to traffic management, usually conceived for freeway networks, can be extended to urban areas adopting the concept of the *Network Fundamental Diagram* (NFD). This seminal concept, developed by Geroliminis and Daganzo (2007) in terms of what they called the *Macro Fundamental Diagram*, is a consequence of their willingness to figure out whether it a fundamental relationship between traffic variables, similar to the fundamental diagram of the basic traffic flow theory for freeways, exists at the network level. That is, if it exists a maximum capacity of a network that can be interpreted as a generalization of the concept of capacity in traffic theory. If it would exist, it would be suitable to define ad hoc management actions by generalizing those strategies extensively used for traffic management on

freeways. Indeed, this relationship was demonstrated (Daganzo 2007; Geroliminis and Daganzo 2007, 2008; Daganzo et al. 2012; Mahmassani et al. 2013) and it can be explained intuitively as follows. The approach assumes that an urban traffic network behaves similarly to a reservoir, with an input flow q_{in} that can be considered a function of time $q(t)$, and an output flow $e(t)$ depending on some system function of the state of the system $n(t)$ at time t, $G[n(t)]$ (Eq. 6.1). The time variation of the accumulation (state) in the system $n(t)$ could then be modeled as in Eq. 6.2:

$$e(t) = G[n(t)] \tag{6.1}$$

$$\frac{dn(t)}{dt} = q(t) - G[n(t)] \tag{6.2}$$

Based on this approach, Geroliminis and Daganzo (2008) show that such NFD exists and can be estimated in terms of the total number of vehicles in the links of the network, measured by traffic detectors, and the total production. The latter should be estimated as the total distance traveled by all vehicles in a link in a given time interval Δt. The resulting diagram is as the one depicted in Fig. 6.15, which was built by simulating microscopically the traffic in a network. This allows estimating the network capacity, that is, the maximum number of vehicles that can be allocated to the network. The possibility of having such measure of maximum capacity of a network and the availability of such an NFD allows exploiting real-time measurements to identify at which point of the diagram (i.e., in which traffic state: free flows, transitions, congested flows, etc.) the network is performing. Moreover, this opens the door to *Active Wide Area Traffic Management Strategies*, as illustrated graphically in Fig. 6.13.

Let us consider a large area to manage and a selected critical subnetwork in this area. For example, the Central Business District (CBD) of Barcelóna depicted in Fig. 6.16. The real-time measurements from sensors allow identifying whether the network is performing at point A, B or C. Point A corresponds to free flow, i.e., the managed subarea has still remaining capacity to allocate more vehicles. At

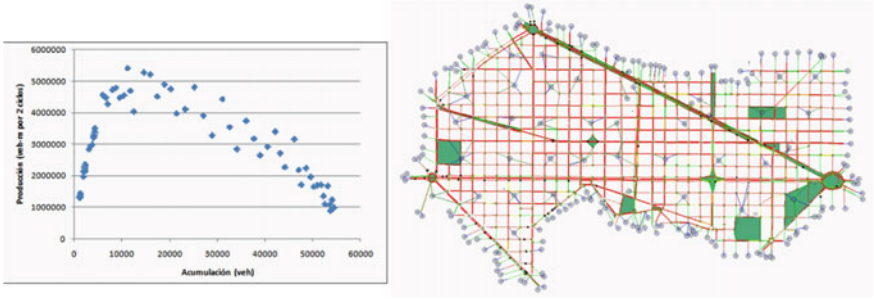

Fig. 6.15 The macro fundamental diagram of Barcelona's central business district built by simulation

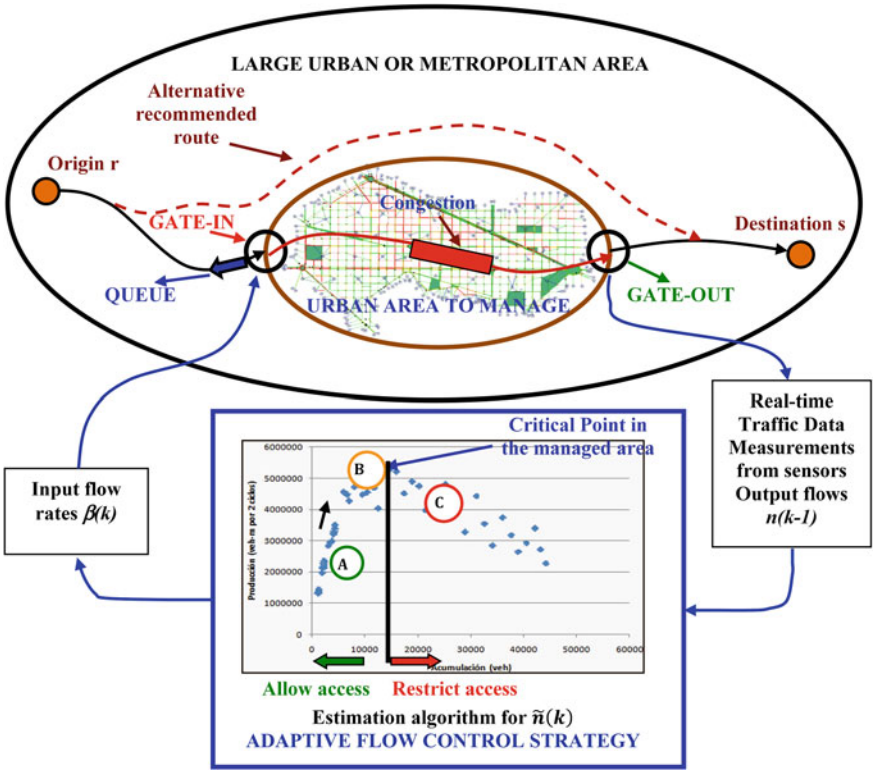

Fig. 6.16 Potential use of the network fundamental diagram to support active traffic management strategies

point B, the subnetwork is reaching the capacity and actions limiting the access to the subnetwork would be desirable to avoid the growing of congestion. Finally, at point C, the network is congested (beyond capacity) and no more vehicles can be allowed to enter the area. Gate-in and gate-out control policies at the boundaries of the area, which are an extension at the network level of the widely used ramp metering policies to manage freeways, can be implemented to, respectively, restrict the access and facilitate the evacuation of the congested area. However, such policies must be applied carefully, as they can generate problems at other parts of the large network due to the queue spillback at entry gates. These gate-in and gate-out policies must be combined with a queue management system and re-routing policies, which in turn require the appropriate dynamic traffic model to account for dynamic traffic patterns between origins and destinations, to balance the situation, (Allström et al. 2017).

The existence of the NFD (or Macro Fundamental Diagram) has been its initial proposal extensively investigated, as well as the procedures to build it from available traffic measurements (Leclercq et al. 2014). At the same time, research has shown

that a proper Macro Fundamental Diagram requires some homogeneity conditions on the network (Geroliminis and Sun 2011), which, in the case of large networks, implies the need for dividing it into homogenous regions (Ji and Geroliminis 2011, 2012). This allows the analysis of the particular phenomenon that takes place in each part and the proposal of an efficient methodology to solve it.

As explained, the existence of a network capacity as a generalization of the key concept of capacity and the availability of practical procedures to estimate it, led very soon to the idea of using it for traffic management in a network, also generalizing the well-known ramp metering strategies. If ramp metering on freeways has the main objective of rating the input flows to the main stream in order to avoid reaching capacity and subsequent consequences as capacity drops, a gate-in–gate-out process could achieve similar objectives at the network level (Aboudolas and Geroliminis 2013). Figure 6.16 illustrates this idea conceptually. Let us assume that output flows $n(k-1)$ are measured in real time by sensors placed at the output gates of a selected subnetwork at time interval $k-1$. Also, that measurements from other sensors placed along the network allow the estimation of the current operational conditions in the network (e.g., A, B or C in Fig. 6.13). Then, an adaptive flow control strategy will determine the input flow rate $\beta(k)$ at time interval k through the input gates (the "gated flows") that keeps the operational condition in the network close to an optimal estimated number of vehicles $\tilde{n}(k)$. This management problem and its variants has generated a rich literature, as Geroliminis et al. (2013), Ampountolas and Kouvelas (2015), Keyvan-Ekbatani et al. (2015, 2016, 2017).

6.5 AMS Approach and ATDM

The cases summarily described so far mostly correspond to a view of traffic management focused on freeways, namely freeway networks, and, in the case of the Wide Area Management discussed in Sect. 6.4, on urban networks. However, there have been also attempts to expand the managed networks, widening the scope to account for mixed networks where freeways and urban roads coexist. An example can be found in Papageorgiou (1995), which considers the concept of corridor as "*a general highway network including both freeways and urban roads.*" That was a first step ahead toward a wider and deeper conception of traffic management, as the traditional one, oriented to single-modal corridor control, was insufficient. Indeed, all traffic management approaches described previously implicitly consider only a transportation mode, i.e., vehicular traffic, ignoring other modes like public transport in all its modal variants, (bus, metro, railways, etc.). This limitation was the natural consequence of freeway networks being the only targets of previous approaches. However, when urban networks or mixed transportation networks spanning large metropolitan areas were considered, it became evident that the other available transportation modes could no longer be disregarded. Responding to this need, Reiss et al. (2006) provided a refined definition of transportation corridor in their Integrated Corridor Management (ICM) initiative report:

> A corridor is a largely linear geographic band defined by existing and forecasted travel patterns involving both people and goods. The corridor serves a particular travel market or markets that are affected by similar transportation needs and mobility issues. The corridor includes various networks (e.g., limited access facility, surface arterial(s), transit, bicycle, pedestrian pathway, waterway) that provide similar or complementary transportation functions. Additionally, the corridor includes cross-network connections that permit the individual networks to be readily accessible from each other.

Shortly before this definition emerged, the US government had envisaged the potential benefits of ICM and started to document the research on this topic. In 2005, the U.S. Department of Transportation's (USDOT) Intelligent Transportation Systems (ITS) Program launched the *ICM Systems Initiative* (FHWA 2005), whose ultimate goal was "*to provide the institutional guidance, operational capabilities, and ITS technology and technical methods needed for effective ICM Systems.*" The initiative further propelled the research on ICM in a regulated way, including the refined definitions for transportation corridor and ICM. Quoting from the TRS 1210 (2012) report:

> In the ICM Program Plan, ICM is defined as the coordination of individual network operations between adjacent facilities that creates an interconnected system capable of cross-network travel management.

To differentiate the ICM from other advanced traffic management systems, Reiss et al. (2006) provided a more detailed definition of ICM:

> ICM consists of the operational coordination of multiple transportation networks and cross-network connections comprising a corridor and the coordination of institutions responsible for corridor mobility. The goal of ICM is to improve mobility, safety, and other transportation objectives for travelers and goods.

Figure 6.17, from Alexiadis (2008a), visualizes this concept of corridor considered as a complex dynamic system composed by a variety of interactive subsystems that need to be considered and managed in an integrated way to achieve the ICM goals. Examples of the subsystems considered in ICM are:

- Different infrastructures, e.g., freeways with on- and off-ramps, arterial streets and local streets.
- Different transportation modes, e.g., passenger car and public transport (bus, metro, light rail).
- Different control systems, e.g., traffic lights at signalized intersections, ramp meters at on-ramps, variable message signs, etc.

Each subsystem may have its own control strategy, such as ramp metering, signal coordination, public transport preemption, or information provision for guidance to foster modal splitting. In the conventional approaches, these strategies would had been developed and integrated independently. The main novelty in ICM is that they must be integrated and coordinated looking at the system as a whole and not only at the individual parts. This is consistently with the view of the transportation corridor in the *ICM Program Plan* (FHWA 2005) as "*a combination of discrete parallel surface transportation networks (e.g., freeway, arterial, transit networks) that link the same*

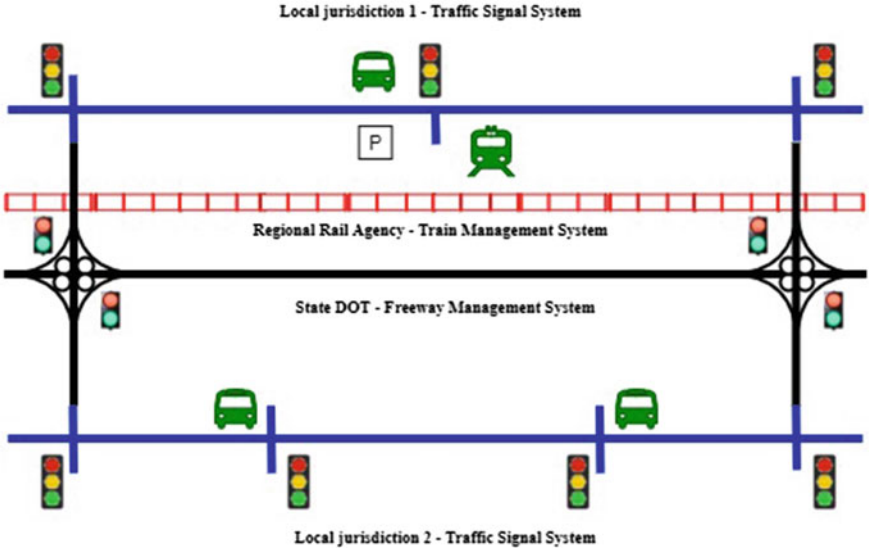

Fig. 6.17 Conceptual scheme of corridor as envisaged by ICM (inspired in Alexiadis 2008a)

major origins and destinations. It is defined operationally rather than geographically or organizationally," which emphases the operational aspects.

ICM is therefore considered in the sphere of the ATMS because of the broad complex operational concepts, the versatility of the integrating subsystems and the variety of management strategies that must be supported, integrated and synchronized by ITS technologies. That is, it must ensure the coordination of the multiple transportation networks that constitute the corridor as well as that of the institutions responsible for each of them. Miller et al. (2008) describe the ICMS in the report on the San Diego ICM project, as a "*system of systems, i.e., a TMS that connects the individual network-based TMS, provides decision support and enables joint operations according to a set of operational procedures agreed to by the network owners.*" They continue explaining that ICMS accounts for the following operations:

- Information share/Distribution: manual information sharing, information clearinghouse (information exchange) between corridor networks and agencies, 511 (pre-trip traveler information), on-route traveler information (smart signage and smart parking), access to corridor information by Internet Service Providers and other value-added entities; automated information sharing (real-time data), common incident reporting system, and asset management system.
- Junctions/Interfaces improvement: signal pre-emption (identifying "best route" for emergency vehicles), multi-modal electronic payment, signal priority for transit, bus priority on arterials, transit hub connection protection, multi-agency/multi-network incident response teams/service patrols, and training exercise.

- Accommodation/Promotion of Network Shifts: modification of ramp metering rates to accommodate traffic (including buses) shifting from arterials; promotion of route shifts between roadway and transit via on-route traveler information devices, promotion of shifts between transit facilities via on-route traveler information devices, congestion pricing for managed lanes, and modification of arterial signal timing to accommodate traffic diverted from the freeway.
- Capacity/Demand Management (short-term): land use control, modification of HOV restrictions, increase of roadway capacity by opening HOV or toll (HOT) lanes/shoulders, scheduled closures for construction, coordination of schedule maintenance and construction activities among corridor networks, planning of temporary addition of transit capacity, and modification of parking fees (smart parking).
- Capacity/Demand Management (long-term): peak spreading, ridesharing programs, expansion of transit capacity, and land use around Bus Rapid Transit (BRT) stations.

Alexiadis (2008a, b) formalizes the Analysis, Modeling and Simulation (AMS), as a methodological proposal for ICM. The proposed AMS methodology combines a variety of traffic models as required. It starts with a trip table manipulation that has its origin in a *Regional Travel Demand Model*, and whose goal is the determination of overall trip patterns, which can be refined and time-sliced as long as the necessary feedback is available. "*In this AMS framework, macroscopic, mesoscopic, and microscopic traffic analysis tools can interface with each other, passing trip tables and travel times back and forth looking for natural stability within the system.*" The elements of this methodological framework were explicitly considered later on by Cronin et al. (2010), who stated that "*conducting analysis, modeling, and simulation tests enables corridor partners to identify the most promising strategies and informs decisions for design of ICM systems.*" They concluded that managers should integrate this methodology with ICM DSS to facilitate the predictive, real-time, scenario-based operational decision-making. The proposed concept of a DSS is in this case "*the interactive, computer-based system that uses historical data and models to identify and solve problems as defined by* Sprague and Watson (1986)." All of them agree that the practice of real-time traffic management must be supported by DSS that assist traffic managers in making sound decisions to avoid, or at least alleviate, conflictive situations in traffic networks occasioned by congestions (recurrent or not), incidents, or any other potential causes. An efficient DSS must help and guide the manager decisions based on a quantitative assessment of the traffic conditions in the network and, if possible, of a short-term prediction of their likely evolution. However, it must be noted that the estimation of the current network state quantified in terms of values of associated indicators (e.g., level of service, queues, delays, travel times, etc.) as well as their short-term forecasting requires the use of suitable dynamic traffic models.

The idea of assisting managers in making decisions with the help of this type of DSS was also explored by Barceló et al. (2005). A first practical implementation was described in Barceló et al. (2007a). Key components in the architecture of these

systems are respectively the traffic models and the trip tables, and the OD matrices and the critical problem of the time updating of these OD matrices. Zhou et al. (2008) provide a seminal exploration on the use of models for the analysis of multimodal urban corridors. The draft report on an AMS Framework (2013) describes the predictive tools as a key component and raises concerns about the current approaches based on travel demand simulators to feed such tools, underlining the limitations of the usual practice of performing a time decomposition of the 24-h trip OD tables. The application of this methodology for the design and development of ICM models has some major challenges:

- The availability of reliable data sources and related data collection and management processes. This an aspect becoming increasingly relevant, considering the availability of new data sources from the pervasive penetration of mobile devices.
- The estimation of the origin–destination time dependent patterns. A key input for most of the existing dynamic models to generate specific control and management decisions, i.e., predictive dynamic re-routing, as discussed for DynaMIT. This input is, however, extremely difficult to estimate, especially under congested scenarios.
- The accuracy and reliability of the information supporting decisions (e.g., travel times, travel time reliability).
- The driver's compliance rate estimation in response to the given control strategies.

Examples of ICM projects implementing this approach using the professional platforms described in Sect. 6.3 are, for instance, the San Diego Interstate 15 Integrated Corridor Management System (SANDAG 2009), led by the San Diego Association of Governments (USA), active since 2013. Also, the Regional ICM System in Florida (USA) for the Florida Department of Transportation (FDOT 2020). Both were powered by Aimsun Live.

Mahmassani et al. (2017) extend the application of the AMS methodology to two programs of the USDOT that represent a step beyond the ICM: The *Active Transportation and Demand Management* (ATDM) and the *Dynamic Mobility Applications* (DMA). A set of effective and reliable tests beds were selected to define a rigorous DMA bundle and to perform a comprehensive ATDM strategy evaluation. These were the San Mateo (US 101), Pasadena, ICM Dallas, Phoenix, and Chicago Testbeds. All tests had the objective of providing valuable mechanisms to refine and integrate research concepts in virtual computer-based simulation environments prior to field deployments. The goals were to (i) "capture a wider range of geographic, environmental and operational conditions under which to examine most appropriate ATDM and DMA strategy bundles, (ii) add robustness to the analysis results and (iii) mitigate the risks posed by a single testbed approach." Mahamassani et al. (2017) particularly report the findings for the Chicago Testbed, in which the strategies tested, all them part of the ATDM strategy bundles applied, were:

- The ATM strategies analyzed were Dynamic Shoulder Lanes, Dynamic Lane Use Control, Dynamic Speed Limits, and Adaptive Traffic Signal Control.

- The ADM Strategies consist of Predictive Traveler Information and Dynamic Routing.
- Weather-related Strategies, including Snow Emergency Parking Management, Traffic Signal Priority for Winter Maintenance Vehicles, Snowplow Routing, and Anti-Icing and Deicing Operations.

These applications were tested considering a proactive network management approach that adopts simulation-based prediction capabilities. The research questions addressed were the impact on any application performance of different facility types under varied operational conditions, the synergies and conflicts among applications, the impact of prediction accuracy and communication latency, and the impact of connected vehicle data versus legacy systems data. The Chicago Testbed was developed using the enhanced, weather-sensitive DYNASMART (Mahmassani et al. 2005) platform in conjunction with a microsimulation tool developed specifically for connected vehicle applications (Talebpour 2016) belonging to the DMA bundle. The authors summarize that the following six algorithmic modules trigger a comprehensive DYNASMART-X simulation:

- A Network State Estimation (RT-DYNA) module, which provides up-to-date estimates of the current state of the network. It has the full simulation functionality as DYNASMART-P, and its execution is synchronized to the real-world clock.
- A Network State Prediction (P-DYNA) module, which provides future network traffic states for a pre-defined horizon, as an extension from the current network state estimated by RT-DYNA.
- An OD Estimation (ODE) module, which uses a Kalman Filtering approach to estimate the coefficients of a time-varying polynomial function that is used to describe the structural deviation of OD demand in addition to a historical regular pattern.
- A OD Prediction (ODP) module, which uses the predicted OD coefficients provided by ODE to calculate the demand that is generated from each origin to each destination at each departure time interval. The predicted time dependent OD matrices are used for both current (RT-DYNA) and future (P-DYNA) stages.
- A Short-Term Consistency Checking (STCC) module, which uses the link densities and speeds of the simulator to evaluate the consistency of the flow propagation with the real-world observations and correct the simulated speeds.
- A Long-Term Consistency Checking (LTCC) module, which compares the simulated and observed link counts to calculate scaling factors that are used to adjust the demand level in both RT-DYNA and P-DYNA.

6.6 Concluding Remarks

From the various approaches and conceptual architectures of traffic management systems discussed in this chapter, a common critical component is the one that provides the system with capabilities for the prediction of the short-term evolution

of the traffic state as a consequence of the envisaged management strategies. This component is usually a dynamic traffic model that, along with the forecasting, evaluates the performance of the system in terms of a selected set of KPIs. One of the key inputs to the KPIs, or a KPI in itself, are the forecasted travel times (sometimes also their reliability) and the subsequent likely route choices.

The dynamic traffic models to be used could be based on various approaches: from pure microscopic, as in certain applications of Aimsun Live depending on the size of the managed scenario, to mesoscopic approaches, based on the concept of Dynamic User Equilibrium (DUE), as in OPTIMA and ATDM, for medium to large scenarios. In any case, all of them share as common input a dynamic OD matrix and its temporal profiles to properly mimic the time dependency of the demand. Nevertheless, dynamic OD matrices are not directly observable yet. Even when direct observations are available, for example, when onboard GPS devices allow vehicle tracking, they only correspond to a sample. The expansion of this sample to the whole population, as required by DUE models, is not always a straightforward exercise. This raises the question of how to accurately estimate such matrices, a question that will be addressed in Chap. 7.

Acknowledgements The authors are very grateful to the collaboration of Dr. Heribert Kirschfink (Momatec Gmbh) and his colleagues Mr. Marco Boero and Dr. Josefa Hernández, for providing access to images and material from the KITS and MOTIC systems. Also, to Professor Guido Gentile and Mr. Lorenzo Meschini, respectively, of SISTeMA S.R.L. and PTV Group, for supplying information and pictures about OPTIMA. Finally, we express our gratitude to Mr. Josep M. Aymamí and Dr. Emmanuel Bert (Aimsun SLU) for the information and pictures regarding Aimsun Live.

References

Aboudolas K, Geroliminis N (2013) Perimeter and boundary flow control in multi-reservoir heterogeneous networks. Trans Res Part B Methodol 55:265–281. https://doi.org/10.1016/j.trb.2013.07.003

Aimsun SLU (2020) Aimsun live overview

Alexiadis V (2008a) Integrated corridor management analysis, modeling, and simulation experimental plan for the test corridor. USDOT Integrated Corridor Management (ICM) Initiative, FHWA-JPO-0-035, EDL 14415

Alexiadis V (2008b) Integrated corridor management analysis, modeling, and simulation results for the test corridor. Technical Report, Federal Highway Administration

Allström A, Barceló J, Ekström J, Grumert E, Gundlegård D, Rydergren C (2017) Traffic management for smart cities. In: Angelakis V, Tragos E, Pöhls HC, Kapovits A, Bassi A (eds) Designing, developing and facilitating smart cities. Springer, Switzerland. ISBN 978-3-319-44922-7

Ampountolas K, Kouvelas A (2015). Real-time estimation of critical values of the macroscopic fundamental diagram for maximum network throughput. In: Transportation research board 94th annual meeting, January 11–15, 2015, Washington, D.C.

AMS Framework for DMA and ATDM Programs. Draft report Version 1.4, May 2013, USDOT

Antoniou C, Dimitriou L, Pereira F (2019) Mobility patterns, big data and transportation analytics. Elsevier, Amsterdam, The Netherlands

Barceló J, García D, Kirschfink H (2002) Scenario analysis a simulation based tool for regional strategic traffic management. In: 9th world conference on intelligent transport systems, 2002, Chicago. Paper 2140

Barceló J, Codina E, Casas J, Ferrer JL, García D (2004) Microscopic traffic simulation: a tool for the design, analysis and evaluation of intelligent transport systems. J Intell Rob Syst 41:173–203

Barceló J, Kirschfink H, Torday A (2005) An integrated software platform to assist advanced traffic management decisions. In: Modelling and simulation workshop, Sedona

Barceló J, Delgado M, Funes G, García D, Torday A (2007a) An on-line approach based on microscopic traffic simulation to assist real time traffic management. In: 14th world congress on intelligent transport systems, Beijing

Barceló J, Casas J, García D, Perarnau J (2007b) A methodological approach combining macro, meso and micro simulation models for transportation analysis. In: 11th world conference on transportation research, Berkeley

Barceló J, Delgado M, Funes G, García D, Torday A (2007c) On-line microscopic traffic simulation to support real time traffic management strategies. In: 6th European congress on intelligent transport systems and services, Aalborg

Barceló J (ed) (2010) Fundamentals of traffic simulation. Springer, Switzerland

Barceló J, Gilliéron F, Linares MP, Serch O, Montero L (2012) Exploring link covering and node covering formulations of detection layout problem. Transp Res Rec J Transp Res Board 2308:17–26

Barceló J (2019) Future trends in sustainable transportation. In: Faulin J, Grasman S, Juan A, Hirsch P (eds) Sustainable transportation and smart logistics. Decision making models and solutions. Elsevier, Amsterdam, The Netherlands. ISBN 978-0-12-814242-4

Bellei G, Gentile G, Papola N (2005) A within-day dynamic traffic assignment model for urban road networks. Transp Res Part B Methodol 39:1–29

Ben-Akiva M, Bierlaire M, Bottom J, Koutsopoulos HN, Mishalani RG (1997) Development of a route guidance generation system for real-time application. In: Proceedings of the 8th IFAC symposium on transportation systems, 16–18 June 1997, Chania, Greece

Ben-Akiva M, Bierlaire M, Koutsopoulos HN, Mishalani R (2002) Real-time simulation of traffic demand-supply interactions within DynaMIT. In: Gendreau M, Marcotte P (eds) Transportation and network analysis: current trends. Miscellenea in honor of Michael Florian. Kluwer Academic Publishers, Boston/Dordrecht/London, pp 19–36

Ben-Akiva M, Koutsopoulos HN, Antoniou C, Balakrishna R (2010) Traffic simulation with DynaMIT. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, Switzerland. ISBN 978-1-4419-6142-6

Boero M, Kirschfink H, Barceló J, Parodi A (1997) Real time traffic management supporting inter-mobility and strategic control. In: 5th European congress on fuzzy and intelligent technologies, EUFIT'97, 1997, Aachen, Germany

Boero M (1999) Case studies of systems: the KITS model. In: ERUDIT tutorial on intelligent traffic management models, Helsinki

Boero M, Kirschfink H (1999) Case studies of systems: the ENTERPRICE model. In: ERUDIT tutorial on intelligent traffic management models, Helsinki

Carlson RC, Papamichail I, Papageorgiou M, Messmer A (2010) Optimal motorway traffic flow control involving variable speed limits and ramp metering. Transp Sci 44(2):238–253

Castillo E, Conejo AJ, Menéndez JM, Jiménez P (2008) The observability problem in traffic network models. Comput Aided Civil Infrastruct Eng 23:208–222

Castillo E, Grande Z, Calviño A, Szeto WY, Lo HK (2015) A state-of-the-art review of the sensor location, flow observability, estimation, and prediction problems in traffic networks. J Sens ID: 903563. https://doi.org/10.1155/2015/903563

Cronin B, Mortensen S, Sheehan R, Thompson D (2010) Integrated corridor management. Public Roads 74(3):6–11

Daganzo CF (2007) Urban gridlock: macroscopic modeling and mitigation approaches. Transp Res Part B Methodol 41(1):49–62. https://doi.org/10.1016/j.trb.2006.03.001

Daganzo CF, Gayah VV, Gonzales EJ (2012) The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. J Transp Logist 1(1–2):49–65

Fang S, Liao H, Fei Y, Chen K, Huang J, Lu Y, Tsao Y (2016) Transportation modes classification using sensors on smartphones. Sensors, 16:1324. https://doi.org/10.3390/s16081324

FDOT (2020) Data and modeling support of off-line and real-time decisions associated with integrated corridor management FDOT project BDV29-977-38. Lehman Center for Transportation Research

FHWA (2005) Integrated corridor management system (ICMS) Work Plan

Gartner NH (1985) Demand-responsive traffic signal control research. Transp Res Part a: Policy Pract 19A(516):369–373

Gentile G, Meschini L, Papola N (2007) Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks. Transp Res Part b: Methodol 41:1114–1138

Gentile G (2010) The general link transmission model for dynamic network loading and a comparison with the DUE algorithm. In: Immers LGH, Tampere CMJ, Viti F (eds) New developments in transport planning: advances in dynamic traffic assignment. Transport economics, management and policy series. Edward Elgar Publishing, MA, USA

Gentile G, Lunardon D, Arenella A, Doninelli T (2010) A new methodology for automatic building of dynamic models for simulation of road transport networks. In: Proceedings of the SIDT scientific seminar 2010 External costs of transport systems: theory and applications, Rome, Italy

Geroliminis N, Daganzo CF (2007) Macroscopic modeling of traffic in cities. In: 86th annual meeting transportation research board, Washington D.C.

Geroliminis N, Daganzo CF (2008) Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transp Res Part B: Methodol 42(9):759–770. https://doi.org/10.1016/j.trb.2008.02.002

Geroliminis N, Sun J (2011) Properties of a well-defined macroscopic fundamental diagram for urban traffic. Transp Res Part B: Methodol 45:605–617

Geroliminis N, Haddad J, Ramezani M (2013) Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach. IEEE Trans Intell Transp Syst 14(1):348–359. https://doi.org/10.1109/TITS.2012.2216877

Hegyi A, Bellemans T, De Schutter B (2009) Freeway traffic management and control. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, New York. ISBN 978-0-38730440-3

Hernández J, Cuena J, Molina M (1999) Real-time traffic management through knowledge-based models: the TRYS approach. In: ERUDIT tutorial on intelligent traffic management models (www.erudit.de), Helsinki, Finland

Hernández J (1999) An intelligent model for real-time private traffic management in urban networks: the FLUIDS/CRITIC approach. In: ERUDIT tutorial on intelligent traffic management models (www.erudit.de), Helsinki, Finland

Ibarra-Espinosa S, Ynoue R, Giannotti M, Ropkins K, Dias de Freitas E (2019) Generating traffic flow and speed regional model data using internet GPS vehicle records. MethodsX 6:2065–2075. https://doi.org/10.1016/j.mex.2019.08.018

Ji Y, Geroliminis N (2011) Exploring spatial characteristics of urban transportation networks. In: 14th international IEEE conference on intelligent transportation systems. https://doi.org/10.1109/ITSC.2011.6083062

Ji Y, Geroliminis N (2012) On the spatial partitioning of urban transportation networks. Transp Res Part B: Methodol 46(10):1639–1656. https://doi.org/10.1016/j.trb.2012.08.005

Keyvan-Ekbatani M, Papageorgiou M, Knoop VL (2015) Comparison of on-line time-delayed and non-time-delayed urban traffic control via remote gating. In: 94th transportation research board annual meeting, Paper 15-4289, Washington DC

Keyvan-Ekbatani M, Carlson RC, Knoop VL, Hoogendoorn SP, Papageorgiou M (2016) Queuing under perimeter control: analysis and control strategy. In: 2016 IEEE 19th international conference

on intelligent transportation systems (ITSC), pp 1502–1507. https://doi.org/10.1109/ITSC.2016.7795756
Keyvan-Ekbatani M, Carlson RC, Knoop VL, Papageorgiou M (2017) Balancing delays and relative queues at the urban network periphery under perimeter control. In: Transportation research board 96th annual meeting, Paper 17-05029, Washington DC
Kirschfink H, Boero M, Barceló J (1997) Real-time traffic management supporting intermobility and strategic control. In: Proceedings of the 4th ITS world conference, Berlin, Germany
Kirschfink H, Hernández J, Boero M (2000) Intelligent traffic management models. In: Proceedings of European symposium on intelligent techniques, ESIT'2000, Aachen, Germany
Kirschfink H, Riegelhuth G, Barceló J (2003) Scenario analysis to support strategic traffic management in the region Frankfurt Rhein-Main. In: 10th world conference on intelligent transport systems, 2003, Madrid
Kotsialos A, Papageorgiou M, Diakaki C, Pavlis Y, Middelham F (2002) Traffic flow modeling of large-scale motorway networks using the macroscopic modeling tool METANET. IEEE Trans Intell Transp Syst 3:282–292
Krishnan R, Hodge V, Austin J, Polak J, Jackson T, Smith M, Lee T-C (2010) Intelligent decision support for traffic management. In: Proceedings of 17th ITS world congress, 25–29 October 2011, Busan, South Korea
Leclercq L, Chiabaut N, Trinquier B (2014) Macroscopic fundamental diagrams: a cross-comparison of estimation methods. Transp Res Part b: Methodol 62:1–12. https://doi.org/10.1016/j.trb.2014.01.007
Mahmassani HS, Fei X, Eisenman S, Zhou X, Qin X (2005) DYNASMART-X evaluation for real-time TMC application: chart test bed. Final Report. Maryland Transportation Initiative, University of Maryland College Park, Maryland, US
Mahmassani HS, Saberi M, Ali Zockaie K (2013) Urban network gridlock: theory, characteristics, and dynamics. In: 20th international symposium on transportation and traffic theory. Procedia Soc Behav Sci 80:79–98
Mahmassani HS, Hong Z, Xu X, Mittal A, Yelchuru B, Kamalanathsharma R (2017) Analysis, modeling, and simulation (AMS) testbed development and evaluation to support dynamic mobility applications (DMA) and active transportation and demand management (ATDM) programs evaluation report for the Chicago testbed. FHWA Final ReportApril 2017, FHWA-JPO-16-387
Martínez-Díaz M (2018) Highway travel time information systems: from traditional to cooperative driving environments. PhD Dissertation, Universidade da Coruña, Spain
Martínez-Díaz M, Soriguera F, Pérez I (2019) Autonomous driving: a bird's eye view. IET Intel Transport Syst 13(4):563–579. https://doi.org/10.1049/iet-its.2018.5061
Meschini L, Gentile G (2010) Real-time traffic monitoring and forecast through OPTIMA–optimal path travel information for mobility actions. In: Proceedings of models and technologies for intelligent transportation systems, 2010, Rome, Italy, pp 113–121
Meschini L (2017) Modern traffic control centres and traffic management systems. In: Fusco G (ed) Intelligent transport systems (ITS): past, present and future directions. NOVA Science Publishers, New York, US. ISBN 978-1-53611-815-5
Miller M, Novick L, Li Y, Skabardonis A (2008) San Diego I-15 integrated corridor management (ICM) system: phase I. California PATH Research Report, UCB-ITS-PRR-2008-33
Montero L, Pacheco M, Barceló J, Homoceanu S, Casanovas J (2016) A case study on cooperative car data for traffic state estimation in an urban network. TRR Transp Res Rec J Transp Res Board 2594:127–137
Mounce R, Hollier G, Smith M, Hodge VJ, Jackson T, Austin J (2012) A metric for pattern-matching applications to traffic management. Transp Res Part C. https://doi.org/10.1016/j.trc.2012.04.019
Papageorgiou M (1983) Application of automatic control concepts to traffic flow modeling and control. Springer-Verlag, New York, USA
Papageorgiou M (1995) An integrated control approach for traffic corridors. Transp Res Part c: Emerg Technol 3(1):19–30

Papageorgiou M, Papamichail I, Messmer A, Wang Y (2010) Traffic simulation with METANET. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, New York. ISBN 978-1-4419-6142-6

Rass S, Kyamakaya K (2007) Artificial intelligence techniques in traffic control. Ögai J 25: 5–11

Reiss R, Gordon R, Neudorff L, Harding J (2006) Integrated corridor management phase I concept development and foundational research: task 3.1 develop alternative definitions. Report No. FHWA-JOP-06-034. United States Department of Transportation ITS Joint Program Office, Federal Highway Administration, Federal Transit Administration

SANDAG (2009) https://www.sandag.org/uploads/publicationid/publicationid_4473_23298.pdf. Accessed 10 May 2021

Soriguera F, Martínez-Díaz M (2020) Freeway travel time information from input-output vehicle counts: a drift correction method based on AVI data. IEEE Trans Intell Transp Syst (In Press). https://doi.org/10.1109/TITS.2020.2992300

Sprague RH, Watson HJ (1986) Decision support systems: putting theory into practice. Prentice-Hall, New Jersey, US

Talebpour A, Mahmassani HS, Bustamante FE (2016) Modeling driver behavior in a connected environment: integrated microscopic simulation of traffic and mobile wireless telecommunication systems. Transp Res Rec J Transp Res Board 2560:75–86

The AURA Web Pages, AURA, Advanced Computer Architectures Group. http://www.cs.york.ac.uk/arch/neural-networks/technologies/aura. Accessed 10 May 2021

TRS 1210 (2012) Transportation research synthesis. Minnesota Department of Transportation Office of Policy Analysis, Research & lnnovation Research Services Section

Wang Y, Papageorgiou M (2005) Real-time freeway traffic state estimation based on extended Kalman filter: a general approach. Transp Res Part b: Methodol 39:141–167

Wang Y, Papageorgiou M, Messmer A (2006) RENAISSANCE—a unified macroscopic model-based approach to real-time freeway network traffic surveillance. Transp Res Part c: Methodol 14:190–212

Wang Y, Papageorgiou M, Messmer A (2008) Real-time freeway traffic state estimation based on extended Kalman filter: adaptive capabilities and real data testing. Transportation Research Part a: Policy and Practice 42(10):1340–1358

Xianfeng TY (2018) Vehicle sensor data (VSD)-based traffic control in connected automated vehicle (CAV) environment. Final Report NITC-SS-1175, University of Utah

Zhou X, Mahmassani HS, Zhang K (2008) Dynamic micro-assignment modeling approach for integrated multimodal urban corridor management. Transp Res Part C: Methodol 16(2):167–186

Chapter 7
Data Analytics and Models for Understanding and Predicting Travel Patterns in Urban Scenarios

Jaume Barceló, Xavier Ros-Roca, and Lidia Montero

Abstract The estimation of the network traffic state, its likely short-term evolution and the prediction of the expected travel times in a network are key steps of traffic management and information systems, especially in urban areas and in real-time applications. To perform such functions, most systems have at their core engine specific dynamic traffic models whose main input is a dynamic OD-matrix describing the time dependency of travel patterns in urban scenarios. This chapter provides an overview of the main concepts supporting these dynamic traffic models and their practical implementations in some software platforms, as well as an outline on the main approaches for the estimation of dynamic OD-matrices. Additionally, this chapter provides a basic discussion on one of the main emerging trends: strategies aimed at using the unprecedented amount of new traffic data made available by "new" mobile technologies.

7.1 Dynamic Traffic Assignment Models

Most of the real-time traffic management systems are based on conceptual architectures embedding in their core engines dynamic traffic models, usually a *Dynamic Traffic Assignment* (DTA) or *Dynamic User Equilibrium* (DUE) model. These models are aimed at providing, among others, outputs to predict traffic flows and travel times on road networks, which vary over time because of various factors. One of these factors is particularly relevant: the time variation of the demand. Traffic assignment accounting for these time dependencies are referred to as DTA. When the

J. Barceló (✉) · X. Ros-Roca · L. Montero
Department of Statistics and Operations Research, Polytechnic University of Catalonia, UPC-BarcelonaTech, Catalonia, Spain
e-mail: jaume.barcelo@upc.edu

X. Ros-Roca
e-mail: xavier.ros.roca@upc.edu

L. Montero
e-mail: lidia.montero@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_7

predicted flows are such that no user can unilaterally reduce their travel times, the resulting assignment is said to be a DUE. In any case, their main input is an OD-matrix, that is, the matrix representing the time dependencies of the demand (e.g., Barceló et al. 2004; Allström et al. 2017). Dynamic Traffic Models, either DTA or DUE, are the key tool to estimate traffic states, understanding traffic patterns. And, as already mentioned, to be able to provide a predictive information consistent with the conditions that drivers will experience in the network, thus accounting for traffic evolution. Both important functionalities become more relevant in the case or complex urban networks. This is explained in detail in Ben-Akiva et al. (2010), which describes the approaches on which DynaMIT is based as well as its objectives. Descriptions of other similar systems can be found in Barceló et al. (2007), Heygi et al. (2009), Meschini (2017), and Aimsun (2020). The last two references illustrate these approaches through their implementation in two worldwide used professional systems based on these applications: OPTIMA and Aimsun. The role of DTM becomes even more critical in recent real-time traffic management systems like the *Active Transportation and Demand Management* (ATDM) and the *Dynamic Mobility Applications* (DMA), two programs of the United States Department of Transportation (USDOT) (Mahmassani et al. 2017).

The DTA problem can be considered an extension of the well-known *Static Traffic Assignment* (STA) problem, widely used in transport planning. The dynamic version must be able to determine how link and path flows evolve with time in the traffic network because of a time-dependent demand defined in terms of a time-varying OD-matrix. In other words, the dynamic approach to traffic assignment must describe how traffic flow patterns evolve in time and space on the network (Mahmassani 2001). Subsequently, it must provide the estimations of the link and path travel times and their short-term expected evolution. These are the main inputs to derive the KPIs that lead to specific traffic management policies, namely, those concerning information to travelers, alternative dynamic re-routing, etc.

From this standpoint, the DUE problem can be defined as the dynamic version of Wardrop's Principle (Wardrop 1952; Friesz et al. 1993; Smith 1993; Ran and Boyce 1996): "*If, for each OD pair at each instant of time, the actual travel times experienced by travelers departing at the same time are equal and minimal, the dynamic traffic flow over the network is in a travel-time-based dynamic user equilibrium state*". In other words, the DUE formulation stipulates that the experienced travel cost, including travel time and early/late arrival penalties, is identical for those route and departure time choices selected by travelers between a given OD pair. There are several attempts to translate this formulation into a suitable model.

In a recent paper, Han et al. (2019) review the various formulations of the models and the associated algorithms used to compute DUE, starting from the seminal proposal of Friesz et al. (1993), which formulates it as an open-loop, non-atomic Nask-like game. "Open-loop" means in this context that the selection of routes by the travelers after leaving the origin does not vary in response to changes in the dynamic network conditions. For its part, the term "non-atomic" implies the assumption of techniques based on aggregated traffic flow dynamics instead of techniques based on individual vehicle dynamics. This hypothesis ensures that DUE suitably accounts

for two main aspects of travel behavior: the departure time choice and the route choice. Therefore, the modeling hypothesis implies that travel times are identical for all trips departing at the same time interval using the same route. Following with the contribution of Han et al. (2019), the two main components of DUE modeling approaches are highlighted:

- The mathematical expression of the equilibrium condition.
- The network performance model, which mimics flow propagation through the network. This is usually referred to as *Dynamic Network Loading* (DNL).

DTA/DUE have been the subject of intensive research and developments both from the theoretical point of view and as key components of most software platforms used for the practical implementation of traffic management systems. Consequently, as Han et al. (2019) report, the concept of dynamic equilibrium has been implemented in various ways, as, for example, variational inequalities, nonlinear complementarity problems, differential variational inequalities, etc. In this Chapter, we limit our discussion to the formulation in terms of variational inequalities (Friesz et al. 2013; Smith and Wisten 1995), which is the most frequent in practical implementations. It is based on the mathematical model (Eqs. 7.1. and 7.2) proposed by Wu (1998):

$$\left[tt_{rsp}(t) - \theta_{rs}(t)\right] * x_{rsp}(t) = 0 \forall p \in K_{rs}(t), \forall (r,s) \in, t \in [0,T] \tag{7.1}$$

$$s.t.\ \ tt_{rsp}(t) - \theta_{rs}(t) \geq 0 \forall p \in K_{rs}(t), \forall (r,s) \in, t \in [0,T]$$

$$tt_{rsp}(t), \theta_{rs}(t,), x_{rsp}(t) > 0 \forall p \in K_{rs}(t), \forall (r,s) \in, t \in [0,T]$$

and the flow balancing equations:

$$\sum_{\forall p \in K_{rs}(t)} x_{rsp}(t) = X_{rs}(t)\ \forall (r,s) \in I, t \in [0,T] \tag{7.2}$$

where $x_{rsp}(t)$ is the flow on path p departing from origin r to destination s, $tt_{rsp}(t)$ is the actual path cost from r to s on route p, $\theta_{rs}(t)$ is the cost of the shortest path from r to s, $K_{rs}(t)$ is the set of all available paths from r to s and $X_{rs}(t)$ is the demand (number of trips) from r to s, all of them at time interval t. For their part, I is the set of all origin–destination pairs (r, s) in the network and T the overall time period considered. It can be demonstrated that this is equivalent to solve a finite-dimensional vibrational inequality problem consisting of finding a vector $\boldsymbol{x}*$ of path flows and a vector τ of path travels times, such that

$$\left[x - x^*\right]^T * \tau \geq 0, \forall x \in \aleph \tag{7.3}$$

where $\aleph$ is the set of feasible flows defined by

$$\aleph = \left\{ x_{rsp}(t) \left| \sum_{\forall p \in K_{rs}(t)} x_{rsp}(t) = X_{rs}(t) \forall (r,s) \in I, t \in [0,T], x_{rsp}(t) > 0 \right. \right\} \quad (7.4)$$

Wu et al. (1991,1998a; b) probe that this is equivalent to solve the discretized variational inequality:

$$\sum_{t \in [0,T]} \sum_{p \in \Re} tt_{rsp}(t) * \left[x_{rsp}(t) - x^*_{rsp}(t) \right] \geq 0 \quad (7.5)$$

where $\Re = \bigcup_{(r,s) \in I} * K_{rs}$ is the set of all available paths from origins to destinations.

Reviews of DTA models can be found in Boyce et al. (2001), Peeta and Ziliaskopoulos (2001), Szeto and Lo (2005), Szeto and Wong (2012), Jeihani (2007), and Bliemer et al. (2017).

Algorithms to deal with DTA or DUE problems usually involve solving this variational inequality formulation. A wide variety of algorithms has been proposed: from projection algorithms (Wu et al. 1991,1998a; b; Florian et al. 2001) or methods of alternating directions (Lo and Szeto 2002) to various versions of the *Method of Successive Averages* (MSA) (Tong and Wong 2000; Florian et al. 2002; Mahut et al. 2003a, b; Mahut et al. 2004; Varia and Dhingra 2004).

The computational approaches proposed to solve the DTA problem can be broadly classified into two classes: mathematical formulations, looking for analytical solutions, and traffic simulation-based approaches, looking for approximate heuristic solutions. Both fit the conceptual framework proposed by Florian et al. (2001) and Florian et al. (2002), formalizing the relationships and dependencies between the two main components identified (Fig. 7.1):

- A method to determine the path-dependent flow rates on the paths on the network, usually applying any of the approaches mentioned above (MSA, projection methods, etc.).
- A DNL method, which determines how these path flows give raise to time-dependent arc volumes, arc travel times and path travel times.

Quite frequently, and basically in all practical implementations mentioned above, DNL method is based on a mesoscopic simulation model (Barceló 2010a) emulating the flow propagation through the network in the current conditions. Depending on how the convergence criterion and the iterative process implemented, the resulting assignment is a DTA or a DUE (see Chiu et al. 2011 for more details).

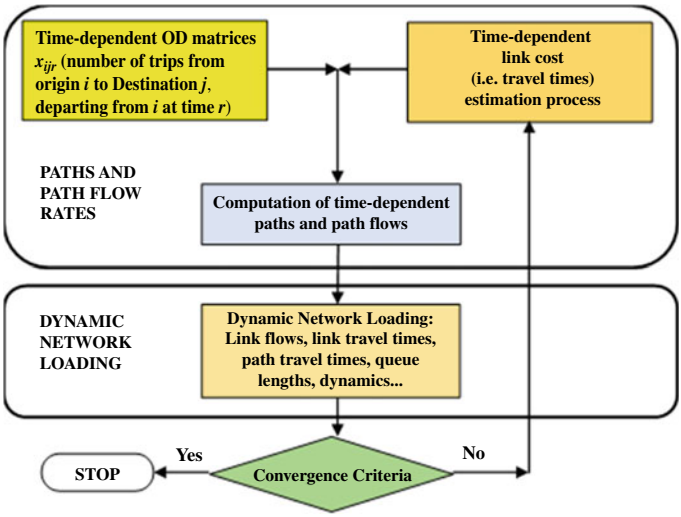

Fig. 7.1 Conceptual algorithmic scheme for DTA

7.1.1 Determining the Path-Dependent Flow Rates by MSA: Convergence Criterion to Equilibrium

If the convergence criteria are not met after one particular iteration of the conceptual algorithmic scheme in Fig. 7.1., a new one is performed. In this new iteration, after computing the new potential paths once the link costs have been updated, the key point is the determination of how the demand will be split among these paths, producing the corresponding path flows. Carey and Ge (2012) or Han et al. (2019) provide a comprehensive panoramic view of the many computational alternatives.

To illustrate these concepts, in this chapter we address the MSA method, one of the most frequently used in practice to estimate the path-dependent flow rates to solve (Eq. 7.5). MSA is a procedure that redistributes the flows among the available paths in an iterative procedure that, at any iteration n, computes a new shortest path, $c_{rs}(t)$, from origin r to destination s at time interval t. Depending on $c_{rs}(t)$ the path flows update process is as follows:

If $c_{rs}(t) \notin K_{rs}^{n}(t)$

$$x_{rsp}^{n+1}(t) = \begin{cases} \alpha_n * x_{rsp}^{n}(t) & if \ \ p \in K_{rs}^{n}(t) \\ (1-\alpha_n) * X_{rs}(t) & if \ \ p = c_{rs}(t) \end{cases} \quad \forall r, s, t \tag{7.6a}$$

$$K_{rs}^{n+1}(t) = K_{rs}^{n}(t) \cup c_{rs}(t) \tag{7.6b}$$

Otherwise if $c_{rs}(t) \in K_{rs}^{n}(t)$

$$x_{rsp}^{n+1}(t) = \begin{cases} \alpha_n * x_{rsp}^n(t) & if\ p \neq c_{rs}(t) \\ \alpha_n * x_{rsp}^n(t) + (1-\alpha_n) * X_{rs}(t) & if\ p = c_{rs}(t) \end{cases} \quad \forall r, s, t \tag{7.7a}$$

$$K_{rs}^{n+1}(t) = K_{rs}^n(t) \tag{7.7b}$$

Depending on the values of the weighting coefficients α_n, different MSA schemes can be implemented (Carey and Ge 2012), probably being the most typical value $\alpha_n = \frac{n}{n+1}$. Many variants have been suggested. For example, Varia and Dhingra (2004) propose a modified MSA algorithm where the weighting coefficient takes into account a variable step length that depends on the current path travel times (Eq. 7.8):

$$\alpha_n = \frac{\lambda_k * \left[exp\left(-tt_{rsp}(t)\right)\right]}{(n+1) * \left[\sum_p * \left[exp\left(-tt_{rsp}(t)\right)\right]\right]} \tag{7.8}$$

One of the potential computational drawbacks of these implementations of MSA is the growing number of paths when dealing with large networks. To avoid this in the case of DTA assignments, an alternative is to specify the maximum number K of paths to keep for each OD pair. Several modified implementations have been suggested to keep control of the number of paths in MSA algorithms (Peeta and Mahmassani 1995; Sbayti et al. 2007). Interesting proposals are those in Mahut et al. (2003a,2004; b). Possibly, one of the most computationally efficient is the one proposed by Florian et al. (2002). This variant of the algorithm initializes the process based on an incremental loading scheme that distributes the demand among the available shortest paths. The process is repeated for a predetermined number of iterations, after which no new paths are added and the corresponding fraction of the demand is redistributed according to the MSA scheme. This modified MSA works as follows:

Let K be the maximum number of iterations to compute new paths.

If $n \leq K$

a new shortest path $c_{rs}(t) \notin K_{rs}^n(t)$ is found. Then,

$$x_{rsp}^{n+1}(t) = \frac{1}{n+1} * X_{rs}(t) \quad \forall p \in K_{rs}^n(t), \forall (r,s) \in I, t \in [0,T] \tag{7.9a}$$

$$K_{rs}^{n+1}(t) = K_{rs}^n(t) \cup c_{rs}(t) \tag{7.9b}$$

If $n > K$

the new shortest path is computed among the existing paths $c_{rs}(t) \in K_{rs}^n(t)$. Then, the set $K_{rs}^n(t)$ does not change, $K_{rs}^{n+1}(t) = K_{rs}^n(t)$, and

$$x_{rsp}^{n+1}(t) = \begin{cases} \frac{1}{n+1} * X_{rs}(t) & if\ p \neq c_{rs}(t) \\ \\ \frac{n}{n+1} * x_{rsp}^{n}(t) + \frac{1}{n+1} * X_{rs}(t) & if\ p = c_{rs}(t) \end{cases} \quad \forall p \in K_{rs}^{n}(t), \forall (r,s) \in I, t \in [0,T] \tag{7.10}$$

However, the possibility of repeating shortest paths from one iteration to the next to keep a maximum K of different shortest paths in a proper implementation of the algorithm implies a requirement: that the number of iterations n is defined for any OD pair and time interval.

All the approaches for DUE based on simulation procedures for the network loading process are, therefore, heuristic in nature. Thus, no formal proof of convergence can be provided. However, a convergence criterion is necessary. In this context, a way to empirically determine if the solution reached can be interpreted in terms of a DUE, in the mentioned sense that "*the actual travel time experienced by travelers departing at the same time are equal and minimal*", can be based on an ad hoc version of the *Relative Gap Function* proposed by Janson (1991):

$$Rgap(n) = \frac{\sum_t \sum_{(r,s)\in I} \sum_{p \in K_{rs}(t)} x_{rsp}^{n}(t) * \left[tt_{rsp}^{n}(t) - \theta_{rs}^{n}(t) \right]}{\sum_t \sum_{(r,s)\in I} X_{rs}(t) * \theta_{rs}^{n}(t)} \tag{7.11}$$

where $x_{rsp}^{n}(t)$ is the flow on path p from r to s departing at time t at iteration n. The difference $tt_{rsp}^{n}(t) - \theta_{rs}^{n}(t)$ measures the excess cost experienced because of using a path of cost $tt_{rsp}^{n}(t)$ instead of the shortest path, with cost $\theta_{rs}^{n}(t)$, at iteration n. The ratio measures the total excess cost with respect to the total minimum cost if all travelers would have used the shortest paths.

7.1.2 Dynamic Network Loading

Once the path flows have been estimated, the next step in the DTA determines how these flows propagate across the network along the assigned paths. Thus, it yields travel times as a function of flows and accounting for their temporal profiles (Xu et al. 1999). The procedures to achieve this goal are precisely the DNL methods, which have been, and still are, a fertile research domain. In fact, a wide variety of DNL have been already proposed. Carey and Ge (2012) or Han et al. (2019) provide comprehensive overviews about them. Some of these methods, for example, those in Friesz et al. (1993), Wu et al. (1998b), or Xu et al. (1999), assume travel time functions of the form $tt_{ij}\left(x_{ij}^{t}\right) = f_{ij}\left(x_{ij}^{t}\right)$, where $f_{ij}\left(x_{ij}^{t}\right)$ is the travel time function for link (ij) that provides the travel time tt_{ij} to traverse the link as a function of x_{ij}^{t}, i.e., the flow in link (ij) at time t.

However, most of the DNL currently used both in research as well as in the professional practice are based on a mesoscopic modeling of traffic flow dynamics.

This is a simplification that, while capturing the essentials of the dynamics, is less data demanding and computationally more efficient than microscopic models, which emulate the dynamics of traffic flows from the detailed dynamics of each vehicle. Mesoscopic approaches sometimes combine microscopic aspects in a simplified way (basically, they can deal with individual vehicles) with macroscopic aspects (e.g., those directly concerning the flow dynamics). There are two main approaches to mesoscopic traffic simulation. First, those in which individual vehicles are not taken into account, and vehicles are grouped into packages or multivehicle platoons that move along the links. This is, for example, the case in CONTRAM (Leonard et al. 1989). Second, those in which flow dynamics are determined by simplified dynamics of individual vehicles. DYNASMART (Jayakrisham et al. 1994), DYNAMIT (Ben-Akiva et al. 1997, 2001,2002,2010), Dynameq (Mahut 2000; Florian et al. 2001, 2002; Mahut et al. 2003a, b, 2004; Mahut and Florian 2010), MEZZO (Burghout 2004; Burghout et al. 2005), or Aimsun (Casas et al. 2010) are well-known examples.

From a methodological point of view, the simulation approach of mesoscopic modeling lays in the way it deals with time. The most common approaches are based on synchronous timing, that is, time-oriented simulations in which time in the model progresses according to an appropriately chosen time unit Δt, also known as the simulation step. This is the case of DYNASMART and DynaMIT. Other approaches are asynchronous or event-based. That is, the state of the model changes when some events occur. Thus, time advances in variable amounts. Dynameq and MEZZO are examples of event-based mesoscopic traffic simulators.

One of the main phenomena determining the time evolution of traffic flows across the network are vehicle queues and their backward propagation (or spillback). As the finite-difference approximations to the fluid flow models in terms of the theory of kinematic waves (LWR, Lighthill and Whitham 1955; Richards 1956), satisfactorily reproduces that dynamics, it has been quite natural to use it to develop DNL models. One of the first was the Cell Transmission Model (CTM) proposed by Daganzo (1994, 1995a), which has been extensively used by other authors (e.g., Lo and Szeto 2002; Szeto and Lo 2004). This model assumes a triangular or trapezoidal flow-density function. Daganzo (1995b) developed a second model similar to the CTM, in this case a Finite-Difference Approximation Method (FDAM), which assumes a general nonlinear flow-density function. This FDAM can be used for network loading in the same way as the Cell Transmission Model for networks in Daganzo (1995a). These basic models exhibit limitations, namely in the case of urban networks, since they only account for flow dynamics in links. This means that they do not explicitly deal with intersections and more in particular with signalized intersections, quite usual in urban networks. In this context, Bellei et al. (2005) propose a DUE approach that is an extension of the CTM. This approach, described theoretically in detail in Gentile et al. (2007), is the basis for the General Link Transmission Model (GLTM), which can deal with any concave fundamental diagram and node topology. The road network is modeled in terms of an oriented graph $G = (N, A)$, where N is the set of nodes, each one representing an intersection and where links A, connecting two intersections, converge or diverge. The forward and backward stars of each node identify the set of links converging or diverging to/from it. The GTLM link model

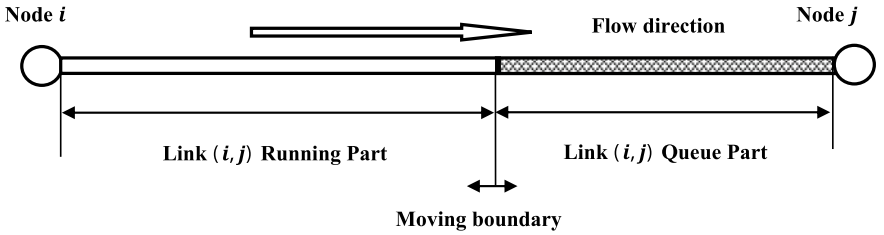

Fig. 7.2 Link model

provides the main input to the node model in terms of the incoming flows. The output from the node model is the outflows that constitute the main input for the link model (Gentile et al. 2010, 2015).

This modeling approach has also been used in many other developments that model the link, explicitly or implicitly, splitting it into two parts: the running part and the queueing part (Fig. 7.2). The running part is that where vehicles are not yet delayed by the queue spillback at the downstream node, where the capacity is limited by stop or give way signs, or traffic lights.

Nodes are modeled according to the interactions between traffic flows at intersections, as node transfer modules, or according to a queue server approach, explicitly accounting for traffic lights and the delays that they cause (Mahmassani et al. 1994). In this case, a simplified car-following model compatible with the macroscopic speed-density relationship on the link approximates the individual vehicle dynamics in the running part. This speed is used to estimate the earliest time at which the target vehicle could exit the link, unless it is affected by the queue spillback when reaching the border between the running part and the queueing part. Vehicle dynamics are then ruled by the queue discharging process. The boundary between the running part and the queueing part is dynamic, according to the queue spillback and queue discharge processes.

Various solutions have been proposed for simulating flow dynamics in the link running part in a simplified way. In essence, they solve the continuity equation of traffic flow:

$$\frac{\partial q}{\partial x} + \frac{\partial k}{\partial t} = g(x, t) \tag{7.12}$$

$$q(x, t) = k(x, t) * u(x, t) \tag{7.13}$$

Link densities are determined by solving the finite differences form of the continuity Eq. (7.12). This can be done using a suitable approach, as, for example, CTM or GTLM, and a functional form (7.13) of the fundamental diagram, where $q(x, t)$ is the flow, $k(x, t)$ the density, $u(x, t)$ the spatial speed and $g(x, t)$ a flow generation term, all of them at time t in x. Jayakrisham et al. (1994) solve these equations in DYNASMART given the densities and the in- and outflows for each section at each

time step and assuming that section speeds are calculated from the densities using the modified Greenshields (1934) speed-density relationship (Eq. 7.14):

$$u_i^t = \left(u_f - u_0\right) * \left(1 - \frac{k_i^t}{k_{jam}}\right)^{\alpha} + u_0 \tag{7.14}$$

where u_i^t and k_i^t are, respectively, the mean speed and density in section i at time step t, u_f and u_0 are the mean free speed and the minimum speed, k_{jam} is the jam density and α is a parameter that captures speed sensitivity to density. DYNAMIT (Ben-Akiva et al. 2001, 2010) includes a speed-density relationship (7.15) that generalizes the one proposed by May and Keller (1967) including a lower bound limiting density, k_{min}, and a second parameter β to capture speed sensitivity to concentration:

$$u = \begin{cases} u_f & if\ k < k_{\min} \\ u_f * \left[1 - \left(\frac{k - k_{min}}{k_{jam}}\right)^{\alpha}\right]^{\beta} & otherwise \end{cases} \tag{7.15}$$

More in particular, the link speed is modeled assuming that it is constant on the upstream section of the link, changes along a deceleration zone covering a downstream section, and varies linearly as a function of the position in this section. According to this assumption, v_u is the speed at the upstream end of the link and the one that is a function of the average density on its running part. That is, v_u is determined by Eq. 7.15. For their part, v_d is the speed at the downstream end of the segment and L_s is the length of the deceleration zone. L_s depends on the geometry of the segment and on traffic conditions. Ben-Akiva et al. (2001) propose a way to determine L_s that is consistent with the empirical evidence that the majority of delays are related to queuing. Finally, assuming that the target link starts at position 0 and ends at position L (i.e., L is the length of the segment), the speed function at an intermediate point x in the segment can be written as follows (Eqs. 7.16 and 7.17):

$$\boldsymbol{v}(x) = \begin{cases} v_d & if\ 0 \le x \le L - L_s \\ \lambda * (x * L) + v_d & if\ L - L_s \le x \le L \end{cases} \tag{7.16}$$

where

$$\lambda = \frac{v_d - v_u}{L_s} \tag{7.17}$$

Other models like MEZZO (Burghout 2004; Burghout et al. 2005) complement this approach according to empirical evidence establishing that there are two limiting densities $k_{\min}$ and $k_{\max}$, which delimit the range in which speed is still a function of the density (del Castillo and Benitez 1995; Eq. 7.18):

$$u = \begin{cases} u_f & \text{if } k < k_{\min} \\ u_0 + (u_f - u_0) * \left[1 - \left(\frac{k - k_{\min}}{k_{\max} - k_{\min}}\right)^{\alpha}\right]^{\beta} & \text{if } k \in [k_{\min}, k_{\max}] \\ u_{\min} & \text{if } k > k_{k\max} \end{cases} \tag{7.18}$$

$u_{\min}$ is the minimum speed in congested conditions. Various queueing models have been proposed to calculate the waiting times in the queuing part of the link. That is, the delays incurred by vehicles because of the output and acceptance capacities of the links. These, respectively, determine the rate at which vehicles can leave the link and how many vehicles can enter it depending on the available space. Obviously, when the acceptance capacity of a link is zero no more vehicles can enter the segment and spillbacks occur. A good example that illustrates this idea is the simplified model in DynaMIT (Ben-Akiva et al. 2001), which considers that the delay of the $i-$ th vehicle in the queue is given by Eq. 7.19:

$$\frac{i}{\rho} \tag{7.19}$$

where ρ is the output capacity of the link. Then, during a time period of length t, $\rho * t$ vehicles will leave the queue. A vehicle in the running part that at time t reaches the end of the queue will find it at $lq(t)$, length of queue at time t, given by

$$lq(t) = lq_0 + l_{\text{eff}} * (\rho * t - m) \tag{7.20}$$

In Eq. 7.20, lq_0 is the position of the end of the queue at time $t = 0$, l_{eff} is the effective length of the queue (i.e., the physical length plus headways), and m is the number of vehicles that reached the queue before the considered vehicle. Obviously, the model is relevant only when $0 < lq(t) < L$.

A completely different approach is taken in Dynameq (Mahut and Florian 2010). It is based on a simulation model that moves vehicles individually, according to a simplified car-following model. In this model, given two consecutive vehicles, the leader vehicle n and the follower $n + 1$, the position $x_{n+1}(t)$ of the follower at time t relative to the position of the leader at $x_n(t - T)$ is estimated according to Eq. 7.21:

$$x_{n+1}(t) = \text{Min}\left[x_{n+1}(t - \varepsilon) + \varepsilon u_f, x_n(t - T) - l_{\text{eff}}\right] \tag{7.21}$$

where T is the reaction time, u_f the free-flow speed, l_{eff}, as before, the effective vehicle length and ε an arbitrary short time interval. The first term inside the minimizing operator represents the farthest position downstream that can be attained at time t based on the follower's position at time $(t - \varepsilon)$, as constrained by the maximum speed of the vehicle, u_f. The second term inside this operator represents the farthest position downstream that can be attained based on the trajectory of the next vehicle downstream in the same lane, according to a simple collision-avoidance rule (Mahut

1999, 2001; Newell 2002). It is a simplified model that only depends on the free-flow speed and does not account for accelerations. It can be considered a lower-order model, since it only defines the position of each vehicle in time, rather than vehicle speed or acceleration.

The solution of the car-following relationship (Eq. 7.21) for time results in (Eq. 7.22):

$$t_{n+1}(x) = \text{Max}\left[t_{n+1}(x-\delta) + \frac{\delta}{u_f}, t_n(x + l_{\text{eff}}) + T\right] \tag{7.22}$$

This relationship in Eq. 7.22 enables the event-based simulation approach used in Dynameq, because it is possible to derive the following expression in Eq. 7.23. It calculates the link entrance and exit times for each vehicle:

$$t_{n+1}(L_1) = \text{Max}\left[t_{n+1}(0) + \frac{L_1}{u_f^1}, t_n(L_1) + T + \frac{l_{\text{eff}}}{\min\left[u_f^1, u_f^2\right]}, t_{n+L_2/l_{eff}}(L_2) + \frac{L_2}{l_{\text{eff}}} * T\right] \tag{7.23}$$

where L_1 and L_2 are the lengths of two sequential links with speeds u_f^1 and u_f^2, respectively. The vehicle attributes represented by l_{eff} and T are considered identical over the entire traffic stream, and each vehicle adopts the link-specific free-flow speed when traversing a given link. The link lengths are assumed to be integer multiples of the vehicle length, l_{eff}. It can be shown (Mahut 2000) that this model yields the triangular fundamental flow-density diagram (Daganzo 1994). The main events changing the state of the model are the arrivals of vehicles to links, their link departures or transfers from one link to the next, according to the turning movements at intersections.

This one-lane link model can be extended to multilane links, including lane changing decisions and additional terms to (7.23) to account for conflicts at nodes with multiple outgoing links. Details can be found in Florian et al. (2008), Mahut and Florian (2010).

The summary description of the most common DTA and DUE included in this section has shown how they can provide TMS with useful information. On the one hand, with the inputs allowing them to estimate the network traffic state. On the other hand (and what is even more relevant), with the necessary outputs to predict traffic flows and travel times on road networks. Moreover, this prediction accounts for their evolution over time because of various factors, being one of them particularly relevant: the time variation of the demand. The main pending question at this point is how to provide this time variation of the traffic demand that constitutes the main input to DTA or DUE. In other words, how to estimate OD-matrices.

7.2 The Static Formulation of the OD-Estimation Problem

Traffic assignment models aim at estimating traffic flows in the network assigning a trip OD-matrix to it, in terms of a route choice mechanism. Therefore, OD trip matrices become their major data input to describe the patterns of traffic behavior across this network. All formulations of static traffic assignment models (e.g., Florian and Hearn 1995), as well as the dynamic ones (e.g., Ben-Akiva et al. 2001), assume that a reliable estimate of an OD is available. However, OD-matrices are not directly observable yet, especially in the case of the time-dependent OD-matrices that are necessary for DTA models. Consequently, it has been natural to resort to indirect estimation methods. These are the matrix adjustment methods, whose main modeling hypothesis can be stated as follows: "*if traffic flows in the links of a network are the consequence of the assignment of an OD matrix to a network, and if we are capable of measuring link flows, the problem of estimating the OD matrix that generates such flows can be considered as the inverse of the assignment problem*" (Cascetta 2001). In other words, the traffic assignment problem is defined as the direct problem, i.e., "*given the O/D matrix X and the cost conditions for using links on the road network, the user equilibrium assignment problem estimates the user equilibrium flows Y on the links of the road network*" (Eq. 7.24):

$$Y = \text{Assignmt}(X) \tag{7.24}$$

where Y is the set of all link flows, X is the OD-matrix, and *Assignmt* is an equilibrium assignment algorithm assigning the OD-matrix X to the network. The reciprocal problem would be that of estimating, from the observed link flows y_l, the OD-matrix X that originated them. In other words, the reciprocal problem of traffic assignment, as described by Cascetta (2001), consists in "*assuming that the observed flows* y_l *on a subset* $\hat{L} \subseteq L$ *of links in the network (or in all links) constitute an user equilibrium flow pattern as defined by* Wardrop (1952)*, determining the OD matrix X whose assignment would produce the observed flows* y_l*"*. Formally, this implies that (Eq. 7.25)

$$X = \text{Assignmt}^{-1}(Y) \tag{7.25}$$

Since the earlier formulation proposed by van Zuylen and Willumsen (1980), the matrix adjustment problem has been a relevant research and practical problem. Given a road network $G = \{L, N\}$, with a set of links L, a set of nodes N, and a set I of OD pairs, the OD-matrix estimation problem consists in finding a feasible vector (OD-matrix) $X \in \Omega$, where Ω is the set of all feasible OD-matrices. For their part $X = \{X_i\}, i \in I$, are the demands for all OD pairs, being $I = \{\text{setofall}OD\text{pairsinthenetwork}\}$. (r, s), as introduced in Sect. 7.1, stands for the $i - th$ OD pair. The assignment of the OD-matrix explains the observed flows y_l on a subset $\hat{L} \subseteq L$ of links equipped with counting stations. It is usually accepted

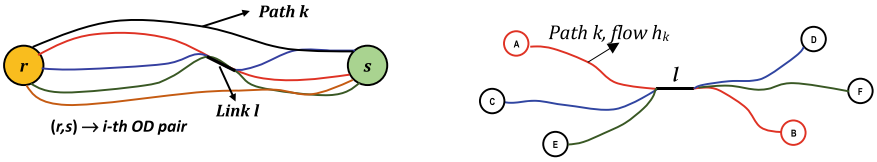

Fig. 7.3 Possible link-path relationships

that the assignment of the OD-matrix to the links of the network is made according to an assignment proportion matrix $P = \{p_{il}\}, \forall i \in I, \forall l \in L$, where each element p_{il} in the matrix is defined as the proportion of the OD demand X_i that uses link l. The notation $P = P(X)$ remarks that, in general, these proportions depend on the demand.

The hypotheses supporting the approach are illustrated in Fig. 7.3, which depicts possible positions of a hypothetical detector at a link l.

Let y_l be the flow measured by one detector and h_k the flow on path k to which this link belongs. If φ_{ik} is the fraction of the demand of the ith OD pair X_i, the flow h_k is given by Eq. 7.26:

$$h_k = \varphi_{ik} * X_i \tag{7.26}$$

δ_{lk} is the link-path assignment matrix, taking the following values (Eq. 7.27):

$$\delta_{lk} = \begin{cases} 1 & \text{if link } l \text{ belongs to path } k : l \in \text{Path } k \\ 0 & \text{otherwise} \end{cases} \quad \forall l \in L, k \in K_i, i \in I \tag{7.27}$$

where K_i = {Set of all paths connecting the ith OD pair}. The relationship between the measured flow y_l on link l and the flows h_k on the paths using link l is given by Eqs. 7.28 and 7.29:

$$y_l = \sum_{i \in I} \sum_{k \in K_i} h_k * \delta_{lk} = \sum_{i \in I} \sum_{k \in K_i} \varphi_{ik} * \delta_{lk} * X_i = \sum_{i \in I} p_{il} * X_i \tag{7.28}$$

$$p_{il} = \sum_{k \in K_i} \varphi_{ik} * \delta_{lk} \tag{7.29}$$

When assigned to the network, the OD-matrix induces a flow $Y = \{y_l\}, \forall l \in L$ in its links. If we assume that observed flows $\widehat{Y} = \{y_l\}$ are available for a subset $\widehat{L}$ of the links, $l \in \widehat{L} \subseteq L$, and that a target matrix $X^H \in \Omega$ is also available, the generic OD-matrix estimation problem can be formulated (Lundgren and Peterson 2008) as (Eq. 7.30):

$$\text{Min}_{XY}(X, Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right)$$

$$s.t. \sum_{i \in I} p_{il}\left(X^H\right) * X_i = \hat{y}_l, \forall l \in \widehat{L} \tag{7.30}$$

$$X \in \Omega$$

The functions $F_1\left(X, X^H\right)$ and $F_2\left(Y, \widehat{Y}\right)$, respectively, represent generalized distance measures. The first one that between the estimated OD-matrix X and the given target matrix X^H, and the second one between the estimated link flows Y and the observed link flows $\widehat{Y}$. The parameters w_1 and w_2 reflect the relative belief (or uncertainty) in the information contained in X^H and $\widehat{Y}$. The problem expressed in Eq. 7.30 can be interpreted as a two-objective optimization problem, being precisely these objectives F_1 and F_2, whereas w_1 and w_2 are the corresponding weighting factors.

The set Ω of feasible OD-matrices normally includes the non-negative OD-matrices. However, it can also be limited to those matrices within a certain deviation from the target values (Eq. 7.31), i.e.,

$$\Omega = \left\{X \geq 0 \middle| (1-\alpha) * X^H \leq X \leq (1+\alpha) * X^H\right\} \tag{7.31}$$

for some suitable parameter $\alpha > 0$ stating the tolerance level. An analogous formulation can be used to state, instead, a maximum deviation from the link flow observations with a tolerance parameter $\beta > 0$ (Eq. 7.32):

$$\Omega = \left\{X \geq 0 | (1-\beta) * \hat{y}_l \leq y_l \leq (1+\beta) * \hat{y}_l\right\} \tag{7.32}$$

Another possibility is to restrict the total travel demand in all OD pairs originating or terminating at a certain node. This is the four-step demand model (Ortúzar and Willunsen 2011), which makes an adjustment of the trip distribution with respect to the trip generation. In any case, all these constraints on Ω are linear or convex and can be easily handled from the optimization point of view.

Obviously, the resulting OD-matrix is dependent on the objective function minimized in (7.30), that is, on the distance measure chosen. One of the distances initially proposed, probably as an analogy with the trip distribution problem, was the maximum entropy function. It was derived from the principle of minimum information (van Zuylen and Willumsen 1980) and is expressed as in Eq. 7.33:

$$F_1\left(X, X^H\right) = \sum_{i \in I} X_i * \left\{\log \frac{X_i}{X_i^H} - 1\right\} \tag{7.33}$$

The function F_2 in (7.30) can be formulated in a similar way.

A type of objective function that is becoming very used in these models is the one based on the least squares formulation. This is equivalent to assume a Euclidean distance function between observed and estimated variables. For example, the function F_2 for the observed volumes would correspond to Eq. 7.34:

$$F_2\left(Y, \widehat{Y}\right) = \sum_{l \in \widehat{L}} \left(y_l - \hat{y}_l\right)^2 \tag{7.34}$$

and could be weighted using the information on the significance of each observation. For instance, when the measurements contained in y are computed as means from a set of observations for each link, the variance σ_l^2 can be used as a measure on how important each link observation is. Equation 7.34 would be then reformulated as Eq. 7.35:

$$F_2\left(Y, \widehat{Y}\right) = \sum_{l \in \widehat{L}} \frac{1}{\sigma_l^2} * \left(y_l - \hat{y}_l\right)^2 \tag{7.35}$$

One disadvantage of the entropy maximizing approaches as formulated in Eq. 7.33 lies in the treatment of link flow observations as error-free constraints (Bell and Iida 1997). An attempt to overcome this disadvantage consists in using a generalized least squares approach to provide a framework accounting for errors from various sources. The method, first proposed by Cascetta (1984), also yields standard errors for the trip table, thereby indicating the relative robustness of the fitted values. The equivalent problem, assuming that the weighting factors w_1 and w_2 have the same value, takes the following form (Eq. 7.36):

$$\begin{aligned} \mathrm{Min}_X F(X) = & \frac{1}{2} * \left[\left(X - X^H\right)^T * \left(X_C^H\right)^{-1} * \left(X - X^H\right)\right] \\ & + \frac{1}{2} * \left[\left(\widehat{Y} - P\left(X^H\right) * X\right)^T * (Y_C)^{-1} * \left(\widehat{Y} - P\left(X^H\right) * X\right)\right] \end{aligned} \tag{7.36}$$

The inputs are prior estimates of OD flows, X^H, link flow measurements, $\widehat{Y}$, variance–covariance matrices of the prior estimates and link flow measurements, respectively, X_C^H and Y_C and the matrix of link choice proportions $P(X^H)$. As the variance–covariance matrices are positive definite and the objective function is convex, the minimum is uniquely given by (Eq. 7.37):

$$\begin{aligned} \nabla F\left(X^*, Y^*\right) = & \left(X_C^H\right)^{-1} * \left(X^* - X^H\right) \\ & - P^T\left(X^H\right) * (Y_C)^{-1} * \left(\widehat{Y} - P\left(X^H\right) * X^*\right) = 0 \end{aligned} \tag{7.37}$$

This yields the following linear estimator (Eq. 7.38):

$$X^* = \left[\left(X_C^H\right)^{-1} + P^T\left(X^H\right) * (Y_C)^{-1} * P\left(X^H\right)\right]^{-1} * \left[\left(X_C^H\right)^{-1} * X^H + P^T\left(X^H\right) * (Y_C)^{-1} * \hat{Y}\right] \tag{7.38}$$

For their part, the sensitivities of this factor are given by Eq. 7.39:

$$\Delta X^* = \left[\left(X_C^H\right)^{-1} + P^T\left(X^H\right) * (Y_C)^{-1} * P\left(X^H\right)\right]^{-1} * \left[\left(X_C^H\right)^{-1} * \Delta X^H + P^T\left(X^H\right) * (Y_C)^{-1} * \Delta\hat{Y}\right] \tag{7.39}$$

Additionally, taking into account that X^H and y are uncorrelated and assuming that $E = \left[\left(X_C^H\right)^{-1} + P^T\left(X^H\right) * (Y_C)^{-1} * P\left(X^H\right)\right]^{-1}$, the variance of X^* is given by Eq. 7.40:

$$\mathrm{Var}\left\{X^*\right\} = E\left(X_C^H\right)^{-1} * E + E * P^T\left(X^H\right) * (Y_C)^{-1} * P\left(X^H\right) * E \tag{7.40}$$

Unlike the maximum entropy model, there is nothing to prevent negative fitted values for the OD flows being produced by the generalized least squares estimator. While negative values would reflect small real values, they are nonetheless counterintuitive. Bell (1991) has also considered the introduction of non-negativity constraints for the fitted OD-matrix.

7.3 Bi-level Optimization Models for OD Adjustment

The estimation of OD-matrices from observed flows as the reciprocal of the assignment problem is a highly undetermined problem. That is, there are in general many OD-matrices, which, when assigned to the network, induce equivalent link flows. The objective function and the set of constraints in the formulation of the problem are aimed at reducing this indetermination. However, this simple formulation can still have some drawbacks. The set of constraints in the generic problem formulation (Eq. 7.30) to determine X is expressed by Eq. 7.41:

$$\sum_{i \in I} p_{il}\left(X^H\right) * X_i = \hat{y}_l, \forall l \in \hat{L} \tag{7.41}$$

$$s.t.\ X \in \Omega$$

It consists of one equation for every link flow observation. Thus, it is an undetermined equation system, as long as the number of OD pairs $|I|$ is greater than the number of link flow observations $\left|\hat{L}\right|$. This fact is especially true for large real-world networks. Additionally, the information transferred through the equation system is delimited by topological dependencies. A basic principle in network flows is that, for consistent flows, the balance equations must hold. In other words, the sum of

incoming and outgoing flows at any intermediate node must be zero. This principle, which can also be interpreted in physical terms using Kirchhoff's law, means that, for each intersection, at least one link flow is linearly dependent from the others. This results in a row-wise dependency for the equation system.

On the other hand, the elements $p_{il}\left(X^H\right)$ are non-zero because they are part of one or more shortest paths for OD pairs $i \in I$. However, since every subpath of a shortest path is a shortest path, every pair of nodes along a certain shortest path is connected through parts of this shortest path. This results in a column-wise dependency for the equation system. Thus, we can conclude that the equation system (Eq. 7.31) is most likely not fully ranked, which further increases the freedom of choice for the OD-estimation problem. Therefore, the way of determining $p_{il}\left(X^H\right)$ is crucial for the quality of the OD-matrix estimation model. This is usually done depending on how the assignment matrix $P(X^H)$ is calculated, and whether it is dependent of X or not. In other words, if the route choices are made depending on the congestion or not. If the assignment of the OD-matrix to the network is independent of the link flows, that is, if we have an uncongested network,$P\left(X^H\right) = P$ is a constant matrix. In that case, the first set of constraints in Eq. 7.41 can be reformulated as in Eq. 7.42:

$$\sum_{i \in I} p_{il} * X_i = \hat{y}_i \ \forall l \in \hat{L} \tag{7.42}$$

$$s.t.\ X \in \Omega$$

In addition, this substitution can be directly performed in the objective, i.e., in the function $F_2\left(Y, \widehat{Y}\right)$, which reduces the OD-matrix estimation to a problem only in the variable X. Assuming that the deviation measures F_1 and F_2 are convex and that the set of feasible OD-matrices Ω is linear or, at least, convex, the OD-estimation problem can be easily solved with some suitable standard algorithms for nonlinear programming. This is the usual approach in most cases (van Zuylen and Willumsen 1980). However:

> The assumption that the assignment, i.e., the route choice, is independent of the load on the links is only realistic in a network with a very low congestion rate or in networks where, in practice, only one route can be used.

If we assume that the network is congested and that the routes are chosen depending on the current travel times, the route proportions are in turn dependent on the existing traffic situation. For its part, this situation depends on the OD-matrix. Thus, the relationship between the route proportions P and the OD-matrix X can only be defined implicitly. In this case, a plausible hypothesis is to assume that the choice proportions can be derived from a traffic assignment model. Then, the set of feasible solutions to the estimation problem (Eq. 7.30) is defined as all points (X, Y) in which Y is the link flow solution satisfying an assignment of the corresponding demand $X \in \Omega$. In this case, the generic OD-matrix estimation problem (Eq. 7.30) can be reformulated as a bi-level optimization problem. Bell and Iida (1997) propose

an approach based on the hypothesis that a traffic assignment model can be represented by a function whose input is the OD-matrix X and whose outputs are the link flows Y (Eq. 7.43)

$$Y = A(X) * X \tag{7.43}$$

That is simply a reformulation of the direct assignment problem as defined in Sect. 7.2, in which, given the OD-matrix X, it is possible to find the link flows $\hat{y}$. The reciprocal problem of finding X given y (Eq. 7.25) is not possible, since the inverse of this function does not exist. However, a way of accounting for this functional relationship in the OD-estimation process could be to reformulate the least squares formulation including it explicitly in the model (Eq. 7.44):

$$\begin{aligned}\text{Minimize}_X F(X) = &\frac{1}{2} * \left(X - X^H\right)^T * \left(X_C^H\right)^{-1} * \left(X - X^H\right) \\ &+ \frac{1}{2} * \left[y - A(X) * X\right]^T * (Y_C)^{-1} * \left[\hat{Y} - A(X) * X\right]\end{aligned} \tag{7.44}$$

If the assignment function (Eq. 7.73) is differentiable, then (Eq. 7.45):

$$\nabla F(X) = \left(X_C^H\right)^{-1} * \left(X - X^H\right) - \nabla A(X)^T * (Y_C)^{-1} * \left[\widehat{Y} - A(X) * X\right] \tag{7.45}$$

And if the Jacobian of the assignment function $\nabla A(X)$ is independent of X^H, then (Eq. 7.46):

$$\nabla^2 F(X) = \left(X_C^H\right)^{-1} + \nabla A(X) * (Y_C)^{-1} * \nabla A(X) \tag{7.46}$$

$\nabla^2 F(X)$ is positive definite, since X_C^H and Y_C are variance–covariance matrices, and there is a unique solution to the equivalent optimization problem. Yang (1995) proposes an efficient heuristic approach to solve this bi-level problem.

However, as Florian and Chen (1995) probe, the assignment function is usually not differentiable. Therefore, analytical approaches are of limited usefulness, since they are constrained to simple uncongested cases. Consequently, other formulations have been proposed. The most common formulation of the bi-level OD-matrix estimation problem for the general case is that Eqs. 7.47 and 7.48, respectively, referred to the upper level and to the lower level problem. Equation 7.47 is as follows:

$$\text{Min}_X F(X, Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) \tag{7.47}$$

$$\text{s.t. } X \in \Omega$$

We want to find the X that minimizes $F(X, Y)$ subject to $X \in \Omega$ under the hypothesis that the induced link flow $\hat{y}$ satisfies the equilibrium assignment conditions

obtained by solving Eq. 7.48, that is, the lower level problem:

$$
\begin{aligned}
Y(X) &= \operatorname{argmin} \sum_{l \in L} \int_0^{y_l} s_l(x)dx \\
s.t. \sum_{k \in K_i} h_k &= X_i, \forall i \in I \\
h_k &\geq 0 \forall k \in K_i, \forall i \in I \\
y_l &= \sum_{i \in I} \sum_{k \in K_i} \delta_{lk} * h_k, \forall l \in L
\end{aligned}
\tag{7.48}
$$

The algorithm for this OD adjustment method based on a bi-level optimization process can be viewed as the calculation of a sequence of OD-matrices, so that the least squares error between traffic counts coming from detectors and traffic flows obtained by a traffic assignment is increasingly reduced. The estimation of the OD-matrix requires information about the routes used for the trips contained in the OD-matrix, X_{rs}. Particularly, it requires the definition of the route and the trip proportions relative to the total trips X_{rs} originated at zone r and ending at zone s. This information is difficult both to handle and to store in traffic databases, considering that the number of routes connecting all OD pairs on a connected network can grow exponentially with the size of the network. This is the reason to use a mathematical programming approach based on a traffic assignment algorithm, which is solved at each iteration without requiring the explicit route definition. The algorithmic scheme to numerically solve the bi-level formulation of the OD-matrix adjustment problem is illustrated in the conceptual diagram in Fig. 7.4. The solution at the $k-$ th iteration of the upper level nonlinear optimization problem for the current estimates of the link flows, $\widehat{Y}^k$, provides a new estimate X^k of the OD-matrix, which is the input to the lower level equilibrium assignment problem. In turn, the solution to this lower level problem updates the link flow estimates. The iterative process continues until certain convergence criterion is satisfied.

One of the first operational approaches of the bi-level algorithm was the one proposed by Spiess (1990), whose bi-level optimization adjustment procedure solves the following bi-level nonlinear optimization problem (Eqs. 7.49 and 7.50):

$$
\text{Min } F[Y(X)] = \frac{1}{2} * \left\{ \sum_{l \in \widehat{L}} \left[y_l(X) - \hat{y}_l \right]^2 \right\} \tag{7.49}
$$

$$
Y(X) = \operatorname{argmin} \sum_{l \in L} \int_0^{y_l} s_l(x)dx \tag{7.50}
$$

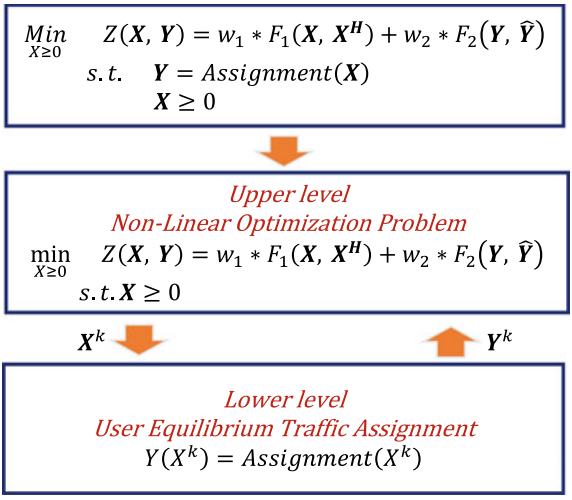

Fig. 7.4 Algorithmic scheme for the bi-level approach to the OD adjustment problem

$$s.t. \sum_{k \in K_i} h_k = X_i, \forall i \in I$$

$$h_k \geq 0 \forall k \in K_i, \forall i \in I$$

$$y_l = \sum_{i \in I} \sum_{k \in K_i} \delta_{lk} * h_k = \sum_{i \in I} X_i \sum_{k \in K_i} \delta_{lk} * p_k, p_k = \frac{h_k}{X_i} \forall l \in L$$

where $y_l(X)$ is the flow on link l estimated by the lower level traffic assignment problem with the adjusted trip matrix X, h_k is the flow on the $k-$th path for the $i-$th O-D pair and $\hat{y}_l$ is the measured flow on link l. I is the set of all OD pairs in the network, and K_i is the set of paths connecting the $i-$th O-D pair. $s_l(y_l)$ is the volume-delay function for link $l \in L$. The algorithm used to solve the problem is heuristic in nature, of steepest descent type, and does not guarantee that a global optimum of the problem will be found. The iterative process for a generic iteration k is as follows:

- Given a solution X_i^k, an equilibrium assignment is solved, yielding link flows y_l^k and proportions $\{p_{il}^k\}$ satisfying the relationship in Eq. 7.51:

$$y_l^k = \sum_{i \in I} p_{il}^k * X_i^k \quad \forall l \in L \tag{7.51}$$

  The target matrix is used in the first iteration (i.e., $X_i^1 = X_i^H, \forall i \in I$).
- The estimate of the OD-matrix at iteration $k+1$ is calculated in terms of the gradient of the objective function $F[Y(X)]$ with Eq. 7.52:

$$X_i^{k+1} = \begin{cases} X_i & \text{for } k = 0 \\ X_i^k * \left[1 - \lambda^k * \left(\frac{\partial F[Y(X)]}{\partial X_i}\right)_{X_i^k}\right] & \text{for } k = 1, 2, 3 \ldots \end{cases} \tag{7.52}$$

That is, a change in the demand is proportional to the demand in the initial matrix and zeroes are preserved in the process.

- The gradient is approximated as in Eq. 7.53:

$$\frac{\partial F[Y(X)]}{\partial X_i} = \sum_{k \in K_i} p_k \sum_{l \in \widehat{L}} \delta_{lk} * \left(\hat{y}_l - y_l\right) \; \forall i \in I \tag{7.53}$$

where $\widehat{L} \subset$ is the subset of links with flow counts and $p_k = \frac{h_k}{X_i}$.

- The step length is approximated as in Eq. 7.54 and 7.55:

$$\lambda^* = \frac{\sum_{l \in \widehat{L}} y_l^{'} * \left(\hat{y}_l - y_l\right)}{\sum_{l \in \widehat{L}} \left(y_l^{'}\right)^2} \tag{7.54}$$

where

$$y_l^{'} = -\sum_{i \in I} X_i * \left(\sum_{k \in K_i} p_k \sum_{l \in \widehat{L}} \delta_{lk} * \left(\hat{y}_l - y_l\right)\right) * \left(\sum_{k \in K_i} \delta_{lk} * p_k\right) \tag{7.55}$$

To ensure the convergence the step length must satisfy the condition in Eq. 7.56:

$$\lambda^* \frac{\partial F[Y(X)]}{\partial X_i} < 1 \; \forall i \in I \tag{7.56}$$

If the condition is violated for some I, the step length must be bounded accordingly (Eq. 7.57):

$$\lambda^* = \frac{1}{\max\limits_i \left\{\frac{\partial F[Y(X)]}{\partial X_i}\right\}} + \varepsilon \tag{7.57}$$

where ε is added to avoid numerical errors.

Further details on the algorithmic properties of this approach are available in Florian and Chen (1995). Alternative approaches improving the simplified gradient approach can be found in Codina and Barceló (2004) and Lundgren and Peterson (2008), among others.

In summary, the most common practices consist in using an initial OD estimate, the OD seed X^H as input, and adjusting it. This adjustment is done based on the

available link counts y provided by an existing layout of traffic counting stations and on other additional information, whenever it is available. Adjustments can be considered as indirect estimation methods based on optimization approaches. All of them share two fundamental modeling hypotheses:

- A mapping scheme of OD flows-link flow counts is available
- If L is the set of links in the network, flow detectors are only located in a subset $\hat{L} \subset L$, from which link flow measurements $\hat{y}_l$, $l \in \hat{L}$ are available.

Assuming these hypotheses, a bi-level optimization model can be proposed, which is usually solved by computational schemes like the one conceptually depicted in Fig. 7.4. That is, iterating between an upper and a lower level. Again, the upper level solves a nonlinear optimization problem that minimizes the distance between available empirical evidence (i.e., a target OD-matrix X^H and observed flows $\widehat{Y}$ in a subset of links) and the estimations provided by the algorithm, while the lower level solves a *User Equilibrium Traffic Assignment* (UETA). The solution to the upper level nonlinear optimization problem provides new estimates of the OD-matrix, which constitute the input to the lower level assignment problem. In turn, the solution to this latter problem provides new estimates of link flows. This computational scheme is in fact a computational framework from which multiple algorithmic variants to solve the problem, both at the upper and at the lower level, can be derived.

The second modeling hypothesis strongly depends on the detection layout available in the network. Unfortunately, they are usually designed and implemented with the primary purpose of providing the data required by traffic control applications. Therefore, current detection layouts in traffic networks are not appropriate for the reconstruction of OD-matrices, as they do not take into account the OD pattern structure explicitly. This could represent a serious drawback regarding the quality of the OD reconstruction, since it has been observed in practice that the adjustment procedure can act implicitly as a metaregression model. That is, it would fit quite well those parts of the network with a relatively rich detection infrastructure (in fact overfit them is most cases), while completely distorting other parts of the network where detection is sparse. This would generate an unbalancing process moving trips between parts of the network, depending on the numerical requirements of the process, but completely unrelated to the underlying transportation phenomena modeled by the OD pattern. In this context, the objective of identifying a detection layout that optimizes the coverage of origin–destination demand on the road network while minimizing the uncertainties of the estimated OD is a subsidiary prior requirement. Since the seminal work of Yang and Zhou (1998), the problem has received substantial attention in recent years, being Ehlert et al. (2006), Fei et al. (2007) just example references. Castillo et al. (2008), who formulate the problem from the perspective of the observability of systems being a *sine qua non* condition for their state estimation and forecasting, must be highlighted. Larsson et al. (2010) provide an overview of the pros and cons of various approaches, and Barceló et al. (2012) complement the detection layout models with a sensitivity analysis, enabling the analyst to establish a relationship between the quality of the layout and the quality of the OD pattern reconstruction.

In consequence, for the practice of the matrix adjustment, it is not only relevant the mathematical modeling approach to be used but it is also highly recommendable to pay attention to the detection layout whose measurements are going to be used for the adjustment of an OD matrix.

7.4 Analytical Formulations for the Dynamic OD Matrix Estimation (DODME) Problem

The static bi-level optimization OD adjustment problem can be reformulated as (Eqs. 7.58 and 7.59):

$$\text{Min } Z(X, Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) \tag{7.58}$$

$$s.t.\ Y = \text{Assignmt}(X) \tag{7.59}$$

$$X \geq 0$$

where F_1 and F_2, as before, are suitable distance functions between estimated and observed values, while w_1 and w_2 are weighting factors reflecting the uncertainty of the information contained in X^H and $\widehat{Y}$, respectively. The underlying hypothesis is that $Y(X)$ are the link flows predicted by assigning the demand matrix X to the network, which can be expressed by a proportion of the OD demand flows passing through the count location at a certain link. In terms of the assignment matrix $A(X)$, the proportion of OD flow that contributes to a certain link traffic count is (Eq. 7.60):

$$Y = A(X) * X \tag{7.60}$$

This is a bi-level optimization problem that solves (at the upper level) the nonlinear optimization problem by substituting the estimated flows Y in the objective function (Eq. 7.59) using the relationship in Eq. 7.60. Thus, it results in (Eq. 7.61):

$$\text{Min } Z(X, Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(A(X) * X, \widehat{Y}\right) \tag{7.61}$$

$$s.t.\ X \geq 0$$

To estimate a new assignment matrix X while at the lower level, a Static User Equilibrium Assignment is used to solve the assignment problem $Y = \text{Assignmt}(X)$, i.e., to estimate the assignment matrix $A(X)$ induced by the new X. Spiess (1990) is a good example of a seminal model based on this approach. Static models have made wide use of the analytical approaches that include flow counts as complementary information to reduce indeterminacy when solving the minimization problem

(Eq. 7.61), as in Codina and Montero (2006), Lundgren and Peterson (2008), and Spiess (1990). The various algorithmic approaches to numerically solve the problem look for algorithmic efficiency, convergence properties, and stability. However, since they are static, they are supported by static assignment models.

In this context, some researchers as Frederix et al. (2011), Lundgren and Peterson (2008), Toledo and Kolechkina (2013), or Yang et al. (2017) drew attention to the role played by the quality of the assignment matrix, which results from the lower level assignment process when estimating the flows used in the upper level. Therefore, they proposed either analytical or empirical approaches for improving it. The analytical approaches assume a functional dependency that allows for a Taylor expansion around the current solution. While some authors like Lundgren and Peterson (2008) still derive the expansion from a static traffic assignment, others like Frederix et al. (2013) or Toledo and Kolechkina (2013) propose a dynamic traffic assignment to account for time dependencies. The approaches based on the hypothesis of linear relationships may be invalid when congestions build up in the network, resulting in non-linearities. The dynamic assignment would be more appropriate for working with congestion building processes that would be captured by the analytical expansion of the dynamic assignment matrix. Frederix et al. (2013) offer a relevant theoretical contribution, while Toledo and Kolechkina (2013) provide more insights to apply it to large networks.

A simpler approach is the modification of the Spiess procedure performed by Ros-Roca et al. (2020). They used, on the one hand, a first-order approach to the assignment matrix that is provided by replacing the static assignment at the lower level by a dynamic traffic assignment. On the other hand, an ad hoc reformulation of the analytical calculation of the gradient that is suitable for a straightforward calculation of the step length at each iteration.

The following notation is used for the dynamic analytical extension from this point until the end of the chapter:

- I is the set of OD pairs.
- $\mathcal{T} = \{1, \ldots, T\}$ is the set of time intervals.
- L is the set of links in the network. $\widehat{L} \subseteq L$ is the subset of links that have sensors.
- $\hat{y}_{l,t}$ are the measured flow counts at link l during time period t. $y_{l,t}$ are the corresponding simulated flow counts, $\forall l \in \widehat{L} \subseteq L$ and $\forall t \in \mathcal{T}$. $Y = \left(y_{l,t}\right)$ and $\widehat{Y} = \left(\hat{y}_{l,t}\right)$ are the link flow counts in vector form.
- $x_{n,r}$ are the OD flows for $n-$ th OD pairs departing during time period r, $\forall n \in I$ and $\forall r \in \mathcal{T}$.$X = \left(x_{n,r}\right)$ are the OD flows in vector form.
- $a_{n,r}^{l,t}$ is the flow proportion of the $n-$ th OD pair, $n \in I$, departing at time period $r \in \mathcal{T}$ and captured by link $l \in \widehat{L}$ at time period $t \in \mathcal{T}$. $A = \left[a_{n,r}^{l,t}\right]$ is the assignment matrix.

Given a network with a set of links L, a set I of OD pairs, and the set of time periods $\mathcal{T}$, the goal of the dynamic OD-matrix estimation problem is to find a feasible vector (OD-matrix) $X^* \in \mathrm{G} \subseteq \mathbb{R}_+^{I \times \mathcal{T}}$, where $X^* = \left(x_{n,r}^*\right)$, $n \in I, r \in \mathcal{T}$ consists of the demands for all OD pairs. It can be assumed that the assignment of the time-sliced

OD-matrices to the links of the network should be done according to an assignment proportion matrix $A = \left[a_{n,r}^{l,t}\right], \forall l \in L, \forall n \in I, \forall r, t \in \mathcal{T}$, where each element in the matrix is defined as the proportion of the OD demand $x_{n,r}$ that uses link l at time period t. The notation $A = A(X)$ is used to indicate that, in general, these proportions depend on the demand. The linear relationship between the flow count on a link and the given OD pair has a matrix form, which thus sets the vector of detected flows as $Y = (Y_1, \ldots, Y_T) = (y_{1,1}, \ldots, y_{L,1}, \ldots, y_{1,T}, \ldots y_{L,T})$ and the vector of OD flows as $X = (X_1, \ldots, X_T) = (x_{1,1}, \ldots, x_{N,1}, \ldots, x_{1,T}, \ldots, x_{N,T})$. Expressing this relationship as the matrix product (Eq. 7.42), $A(X)$ is now (Eq. 7.62):

$$A(X) = \begin{pmatrix} A^{1,1} & 0 & \cdots & 0 \\ A^{1,2} & A^{2,2} & 0 & \vdots \\ \vdots & \ddots & \ddots & 0 \\ A^{1,\mathrm{T}} & \cdots & A^{T-1,T} & A^{T,T} \end{pmatrix} \text{where} A^{r,t} = \begin{pmatrix} a_{1,r}^{1,t} & \cdots & a_{N,r}^{1,t} \\ \vdots & \ddots & \vdots \\ a_{1,r}^{L,t} & \cdots & a_{N,r}^{L,t} \end{pmatrix} \tag{7.62}$$

$a_{n,r}^{l,t}$ represents the proportion of OD flow departing at time $r, x_{n,r}$, passing through link l at time t, $y_{l,t}$. $A^{r,t}$ represents the assignment matrix for the departing flows at time window r detected at time window t. Therefore, A is a lower-diagonal matrix, because OD flows departing at time r cannot pass through link l at time $t < r$.

This linear mapping between the link flows and the OD flows is indeed the first term in the Taylor expansion of the relationship between link flows and OD flows, where additional terms capture the assignment matrix's sensitivity to changes in the OD flows, path choice, and congestion propagation effects (Frederix et al. 2011, 2013; Toledo and Kolechkina 2013). Let X' be in the neighborhood of X. Then, the Taylor expansion is (Eq. 7.63):

$$\begin{aligned} y_{l,t} &= \sum_{n\in I}\sum_{r=1}^{t} a_{n,r}^{l,t}(X') * x'_{n,r} + \sum_{n\in I}\sum_{r=1}^{t} \frac{\partial y_{l,t}(X')}{\partial x_{n,r}} * \left(x_{n,r} - x'_{n,r}\right) = \\ &= \sum_{n\in I}\sum_{r=1}^{t} a_{n,r}^{l,t}(X') * x'_{n,r} \\ &\quad + \sum_{n\in I}\sum_{r=1}^{t} \left.\frac{\partial\left[\sum_{n\in I}\sum_{r=1}^{t} a_{n,r}^{l,t}(X') * x_{n,r}\right]}{\partial x_{n,r}}\right|_{X'} * \left(x_{n,r} - x'_{n,r}\right) = \\ &= \sum_{n\in I}\sum_{r=1}^{t} a_{n,r}^{l,t}(X') * x'_{n,r} \\ &\quad + \sum_{n\in I}\sum_{r=1}^{t} \left(x_{n,r} - x'_{n,r}\right) * \left[\sum_{n'\in I}\sum_{r'=1}^{t} \left.\frac{\partial a_{n',r'}^{l,t}(X')}{\partial x_{n,r}}\right|_{X'} * x'_{n',r'}\right] \end{aligned} \tag{7.63}$$

This enables redefining Spiess' approach to the dynamic case by simply using the first term in the above Taylor expansion. It does not account for the propagation effects, but it explicitly considers time dependencies. The traffic assignment problem at the lower level must now be a dynamic traffic assignment (DTA). Then, the time

periods for the entire formulation must be considered as follows (Eqs. 7.64 and 7.65):

$$\text{Min } Z(X) = \frac{1}{2} * \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} \left(\left(\sum_{n\in I} \sum_{r=1}^{t} a_{n,r}^{l,t} * x_{n,r} \right) - \hat{y}_{l,t} \right)^2 \tag{7.64}$$

$$s.t.\ a_{n,r}^{l,t} = \text{Assignment}(X) \tag{7.65}$$

$$x_{n,r} \geq 0$$

where $a_{n,r}^{l,t}$ is the assignment matrix described before. Therefore, the linear combination inside the brackets is the simulated flow $y_{l,t}$, applying (Eq. 7.66):

$$\frac{\partial y_{l,t}}{\partial x_{n,r}} = a_{n,r}^{l,t} \tag{7.66}$$

As in Spiess (1990), the chain rule can be used to obtain the gradient of the objective function (Eq. 7.67):

$$\frac{\partial Z}{\partial x_{n,r}} = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} \frac{\partial y_{l,t}}{\partial x_{n,r}} * \left(y_{l,t} - \hat{y}_{l,t} \right) = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} a_{n,r}^{l,t} * \left(y_{l,t} - \hat{y}_{l,t} \right) \tag{7.67}$$

We obtain similar equations finding the optimal step size by using the same procedure (Eq. 7.68):

$$y_{l,t}^{'} = \frac{dy_{l,t}}{d\lambda} = \sum_{r=1}^{t} \sum_{n\in I} \frac{dx_{n,r}}{d\lambda} * \frac{\partial y_{l,t}}{\partial x_{n,r}} = \sum_{r=1}^{t} \sum_{n\in I} -x_{n,r} * \frac{\partial Z}{\partial x_{n,r}} * \frac{\partial y_{l,t}}{\partial x_{n,r}} \tag{7.68}$$

The optimal step length λ can be calculated solving the 1-dimensional optimization problem in Eq. 7.69 and whose solution is given by Eq. 7.70:

$$Z^{'}(\lambda) = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} y_{l,t}^{'} * \left(\tilde{y}_{l,t} - \hat{y}_{l,t} + \lambda * y_{l,t}^{'} \right) = 0 \tag{7.69}$$

$$\lambda^{*} = \frac{-\sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} y_{l,t}^{'} * \left(y_{l,t} - \hat{y}_{l,t} \right)}{\sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} {y_{l,t}^{'}}^{2}} \tag{7.70}$$

Then, the iterative procedure described by Spiess (1990) can be used in DTA using these new equations, which are expanded with the time windows. In addition, this procedure can be improved by adding a second term in the objective function to compare it with a historical OD-matrix. If the quadratic function is used, and replacing w_1 and w_2 by $w = w_2 / w_1$ for simplification, Eq. 7.71 arises

$$\text{Min } Z = \frac{1}{2} * \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} \left(\left(\sum_{n\in I} \sum_{r=1}^{t} a_{n,r}^{l,t} * x_{n,r} \right) - \hat{y}_{l,t} \right)^2 + \frac{w}{2} * \sum_{r\in\mathcal{T}} \sum_{n\in I} \left(x_{n,r} - x_{n,r}^{H} \right)^2 \tag{7.71}$$

In this case, Eq. 7.47 is updated, resulting in Eq. 7.72:

$$\begin{aligned} \frac{\partial Z}{\partial x_{n,r}} &= \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} \frac{\partial y_{l,t}}{\partial x_{n,r}} * \left(y_{l,t} - \hat{y}_{l,t} \right) + \frac{w}{2} * x_{n,r} \\ &= \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} a_{n,r}^{l,t} * \left(y_{l,t} - \hat{y}_{l,t} \right) + \frac{w}{2} * x_{n,r} \end{aligned} \tag{7.72}$$

Therefore, the *Iterative Dynamic Spiess Procedure* would be (Eq. 7.73):

$$X_i^{(k+1)} = \begin{cases} X_i^H & \text{for } k = 0 \\ X_i^{(k)} * \left(1 - \lambda^{(k)} * \left[\frac{\partial Z(X)}{\partial X_i} \right]_{X_i^{(k)}} \right) & \text{for } k > 0 \end{cases} \tag{7.73}$$

The use of Euclidean distances to measure the distance between the estimated OD, X, and the historical X^H has been discussed critically in Frederix et al. (2013). For example, Djukic (2014) shows that using a Euclidean distance term can result in two matrices that have very different structures but maintain the same distance value with respect to the reference matrix. Other distance measures have been suggested, for example, in Ros-Roca et al. (2020). Although additional measurements are expected to improve the outcome of the OD-estimation in terms of structural similarity, the analytic approaches do not seem capable of adding measurements different from link counts.

The resort to the classical entropy function, as in the original analytical formulations, is an appealing option because of its structural meaning. With this approach, Eqs. 7.71 and 7.72, respectively, become Eqs. 7.74 and 7.75:

$$\text{Min } Z = \frac{1}{2} * \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} \left(\left(\sum_{n\in I} \sum_{r=1}^{t} a_{n,r}^{l,t} * x_{n,r} \right) - \hat{y}_{l,t} \right)^2 + \frac{w}{2} * \sum_{r\in\mathcal{T}} \sum_{n\in I} x_{n,r} * \log\left(\frac{x_{n,r}}{x_{n,r}^H} \right) \tag{7.74}$$

$$\frac{\partial Z}{\partial x_{n,r}} = \sum_{t\in\mathcal{T}} \sum_{l\in\hat{L}} a_{n,r}^{l,t} * \left(y_{l,t} - \hat{y}_{l,t} \right) + \frac{w}{2} * \left(\log\left(\frac{x_{n,r}}{x_{n,r}^H} \right) + 1 \right) \tag{7.75}$$

7.5 Practical Applications for Traffic Management

Because DTA is a core component of most Dynamic Traffic Management Systems and the Dynamic Origin–Destination Matrices are the main input to DTA, algorithms to numerically implement DODME approaches become a basic procedure in all of them. The main approaches are:

- The strict analytical dynamic approaches based on State-Space Modeling (Ashok and Ben-Akiva 1993, 2002), which are the basis of DynaMIT (Ben-Akiva et al. 2020).
- The numerical approximations of analytical optimization approaches, as the ones proposed by Frederix et al. (2011), Frederix et al. (2013), Toledo and Kolechkina (2013), or Ros-Roca et al. (2020). Other variants are those studied by Djukic et al. (2017,2018,2019), currently implemented in Aimsun Live, Aimsun (2020), or OPTIMA.
- Simulation-based approaches: Stochastic Perturbation Stochastic Approximation (SPSA).

7.5.1 Analytical Approaches Based on State-Space Modeling

The approach taken in DynaMIT to estimate dynamic OD-matrices, aimed at providing support to real-time management decisions, is different from the bi-level optimization considered so far. DynaMIT formulates the real-time dynamic OD-estimation based on the Kalman Filtering framework proposed by Ashok and Ben-Akiva (1993). The basic information, as in all other approaches, is that contained in the historical OD-matrix, which is combined with traffic count data from the counting stations along the network. Other differential aspects of the estimation proposed in DynaMIT are the use it makes of each day's estimate to update the original historical OD estimate in a learning process. These updated historical OD-matrices contain rich information about the latent factors that affect travel demand and its daily variations, which the approach tries to capture. To achieve this goal, this approach uses as state variables the deviations of the OD flows from the historical OD estimates, instead of the actual flows themselves.

The underlying hypothesis states that (Antoniou et al. 2007) modern surveillance systems generate data and historical information that can be used for the estimation and prediction of the time evolving demand patterns represented by OD-matrices. The wealth of information contained in these off-line values, which affects trip making and traffic dynamics, as well as their temporal and spatial evolution, can be incorporated into the DODME process as a priori estimates.

The approach based on Kalman filtering assumes an autoregressive procedure that provides a prediction tool consistent with the estimation process. That autoregressive procedure models the temporal relationships among deviations in OD flows, also accounting for unobserved factors that are correlated over time, as, for instance,

weather effects. A proper approach that incorporates this information and its associated errors in the estimation process considers transport systems as dynamic systems and resorts to the state-space modeling approach. The formulation of the DODME problem discussed so far shows that the most critical issue is the calculation of the assignment matrix, a_{ijr}^{lt}, mapping the observed link flows, y_{lt}, and the unobserved OD flows, x_{ijr}. This matrix must be estimated at each step of the iterative processes by solving numerically the corresponding mathematical model (Eq. 7.76):

$$y_{lt} = \sum\nolimits_{(i,j)\in I} \sum\nolimits_{r=1}^{t} a_{ijr}^{lt} * x_{ijr} \ \forall l \in \hat{L}, t \in T \tag{7.76}$$

The dynamic problem formulation assumes that the assignment matrix depends on link and path travel times and on traveler route choice factors, being all of them time-varying. Precisely, time variations are captured by the time indices in Eq. 7.76. The mapping can be interpreted as the contribution, i.e., the fraction, of the OD flow of pair (i, j) departing origin i with destination j, at time interval r, that flows across detectors located at link l, during time interval t.

Ashok and Ben-Akiva (2002), in an extension to their previous seminal work in Ashok and Ben-Akiva (1993), make the observation that "*all quantities are imperfectly observed, thereby they introduce errors into the OD estimation process, erroneous travel times and/or route choice fractions resulting in an imperfect assignment matrix*". Therefore, they propose reformulating Eq. 7.76 as Eq. 7.77:

$$y_{lt} = \sum\nolimits_{(i,j)\in I} \sum\nolimits_{r=1}^{t} a_{ijr}^{lt} * x_{ijr} + v_{lt} \ \forall l \in \hat{L}, t \in T \tag{7.77}$$

where v_{lt} is the measurement error. The reformulation of the DODME as a state-space model involves two types of equations:

- Transition equations that capture the evolution of the state vector over time.
- Measurement equations that, according to Antoniou et al. (2007), "*capture a mapping of the state vector on the measurements:* a priori*values of the model parameters provide direct measurements of the unknown parameters*".

Let X_k be the vector of state variables whose values define the state of the system at time interval k. A Kalman filter iterates between an *updating* (prediction) of the system's state at time k, obtained from the system's state at time $k-1$, and a *correction* based on an update of the measurements of the system. This corresponds to a process model that models the transformation of the system's state in terms of a linear stochastic difference equation (Eq. 7.78):

$$X_k = \Phi X_{k-1} + w_{k-1} \tag{7.78}$$

where Φ is the *transition matrix* from system's state at time $k-1$ to system's state at time k, and w_{k-1} is the process error term. Additionally, a measurements model describes the relationship between the process changing the system's state and the system measurements (Eq. 7.79):

$$Y_k = A * X_k + v_k \tag{7.79}$$

Assuming initial estimates of the state vector $\widehat{X}_{k-1}$ and of the error covariance P_{k-1} at time interval $k-1$, the *prediction phase* consists of two steps: (i) a state projection step (Eq. 7.80) and (ii) a covariance projection step (Eq. 7.81), respectively, projecting forward the state estimate or the covariance from time step $k-1$ to step k:

$$\widehat{X}_k^{k-1} = \phi * \widehat{X}_{k-1}^{k-1} + w_{k-1} \tag{7.80}$$

$$P_k^{k-1} = \Phi * P_{k-1}^{k-1} * \Phi^T + Q \tag{7.81}$$

The correction regarding the measurements update consists of three steps: (i) the computation of the Kalman Gain (Eq. 7.82), (ii) the update of the error covariance (Eq. 7.83) and (iii) the update of the state estimates with the measurements Z_k (Eq. 7.84):

$$K_k = P_k^{k-1} * A^T * \left(A * P_k^{k-1} * A^T + R\right)^{-1} \tag{7.82}$$

$$P_k^k = (I - K_k * A) * P_k^{k-1} \tag{7.83}$$

$$\widehat{X}_k^k = \widehat{X}_k^{k-1} + K_k * \left(Y_k - A * \widehat{X}_k^{k-1}\right) \tag{7.84}$$

where w_k and v_k, the process and measurement errors, are independent, white noise, and normally distributed (Eqs. 7.85 and 7.86):

$$p(w) \sim N(0, Q) \tag{7.85}$$

$$p(v) \sim N(0, R) \tag{7.86}$$

Q and R are, respectively, the covariance matrices of the process and the measurement errors.

When applying Kalman filtering to DODME, the state vector is the vector X of unknown OD flows, and the transition equation represents an autoregressive process. However, Ashok and Ben-Akiva (1993) state that "*an autoregressive process can only capture interdependencies among OD flows. It does not include structural information about trip patterns, which are a function of spatial and temporal distribution of activities, as well as of the characteristics of the transportation system*". Therefore, it is desirable to modify the model in such a way that it also incorporates structural information. This information could be, for example, that contained in a prior estimate. For instance, a historical OD-matrix X^H provided by a reliable surveillance system. It can be accommodated in the model by reformulating the state vector in terms of the deviations from that historical OD flows. The transition equation would then be as follows (Eq. 7.87):

$$X_{ij(t+1)} - X^H_{ij(t+1)} = \sum_{r=t-s}^{t} \sum_{(p,q)\in I} f_{ijt}^{pqr} * \left(X_{pqr} - X^H_{pqr}\right) + w_{ijt} \tag{7.87}$$

where f_{ijt}^{pqt} describes the effect of the deviation $\left(X_{pqr} - X^H_{pqr}\right)$ on the deviation $\left(X_{ij(t+1)} - X^H_{ij(t+1)}\right)$. The first one is the deviation of the OD flow from origin p to destination q and departing at time r. Equivalently, the second one is the deviation of the OD flow from origin i to destination j and departing at time $t+1$. In this second deviation, w_{ijt} is a random term error for OD pair (i,j) at time t and s is the order of the autoregressive process, that is, the number of lagged OD flow deviations assumed to affect the OD deviation in interval $t+1$. Equation 7.87 "*models the temporal relationship among deviations in OD flows, capturing the correlation over time among deviations which arise from unobserved factors that correlated over time. It assumes dependency of deviations corresponding to one OD pair on deviations corresponding to other OD pairs in prior periods*" (Ashok and Ben-Akiva 1993). It can be rewritten in matrix form (Eq. 7.88):

$$\Delta X_{t+1} = X_{t+1} - X^H_{t+1} = \sum_{r=t-s}^{t} \Phi_t^r * \left(X_{pqr} - X^H_{pqr}\right) + w_t \tag{7.88}$$

In the general case, the computation of the transition matrix Φ_t^r involves estimating linear regression models for each OD pair and for each time interval. However, depending on the network topology, some of these correspondences may

be ignored and thus the matrix is simplified. There are also some other hypotheses enabling further simplifications, as, for example, the assumption that the autoregressive process remains constant with respect to t. This implies that it depends only on the difference $(t-s)$ and not on the individual values of t and s. Equation 7.77 can be rewritten accordingly to get the measurements equation in terms of deviations with respect to historical values y_{lt}^{H}, as in Eq. 7.89:

$$y_{lt} - y_{lt}^{H} = \sum\nolimits_{(i,j)\in I} \sum\nolimits_{r=\mathrm{t-s}}^{t} a_{ijr}^{lt} * \left(x_{ijr} - x_{ijr}^{H}\right) + v_{lt} \ \forall l \in \widehat{L}, t \in T \tag{7.89}$$

It can also be expressed in matrix form (Eq. 7.90):

$$\Delta Y_t = Y_t - Y_t^{H} = \sum\nolimits_{r=\mathrm{t-s}}^{t} A_r^t * \left(X_r - X_r^{H}\right) + v_t \tag{7.90}$$

where v_t is the measurements random error vector at time$t$. Error terms w_t and v_t are uncorrelated, which means that$E[w_t] = E[v_t] = 0$. The variance–covariance matrices are Q_t and R_t, respectively.

There is an additional advantage in reformulating the Kalman filtering in terms of deviations as state variables and measurements, since the traffic flow variables have skewed distributions (Antoniou et al. 2007). However, the deviations from these variables from available estimates have symmetric distributions and, hence, are more amendable to approximations to normal distributions. This is a useful property in terms of Kalman filtering (Kalman 1960; Gelb 1974). Then, assuming an initial state of the system with ΔX_0, with mean $\Delta \overline{X}_0$, and variance–covariance P_0, the Kalman filtering algorithm for DODME, for a time horizon T divided into N intervals of equal length, is

Initialization

$$\Delta X_0^0 = \Delta X_0$$
$$P_0^0 = P_0$$

For k=1 to N do

Time update (Transition)

$$\Delta \hat{X}_k^{k-1} = \phi * \Delta \hat{X}_{k-1}^{k-1} + w_{k-1}$$

$$P_k^{k-1} = \Phi * P_{k-1}^{k-1} * \Phi^T + Q$$

Measurement update

$$K_k = P_k^{k-1} * A^T * \left(A * P_k^{k-1} * A^T + R\right)^{-1}$$

$$\Delta \hat{X}_k^k = \Delta \hat{X}_k^{k-1} + K_k * \left(\Delta Y_k - A * \Delta \hat{X}_k^{k-1}\right)$$

$$P_k^k = (I - K_k * A) * P_k^{k-1}$$

End

Many alternative versions of these basic algorithms resorting to variants of Kalman filtering have been proposed, as those in Ashok and Ben-Akiva (2002), Hu et al. (2001), Antoniu et al. (2007), Lin and Chang (2007). In essence, many of the most appealing ones deal with the calculation of matrices Φ and A. That is, with the characteristics of the autoregressive model, the mapping OD paths and the links flows, being these latter the most critical. Antoniu et al. (2007) propose nonlinear relationships for the measurement equations, generically defined as (Eq. 7.91):

$$\Delta M_t = M_t - M_t^H = \mathcal{S}(\Delta X_t) - M_t^H + v_t \tag{7.91}$$

where M_t is the vector of measurements at time t, $\mathcal{S}(\Delta X_t)$ is a simulation model and $M_t^H = \mathcal{S}(\Delta X_t^H)$. When traffic flow models are used to simulate the time progression of traffic flows through the network, they can be approximated by continuous functions $h(x)$(Antoniou 2004). These functions can be linearized to approximate the measurement equation as in Eq. 7.92:

$$H_t = \left.\frac{\partial h(x^*)}{\partial x^*}\right|_{x^*=X_t^{t-1}} \tag{7.92}$$

An example based on this linearization included in Antoniou (2004) and Antoniou et al. (2007) is the following Extended Kalman Filter (EKF):

Initialization

$$\Delta X_0^0 = \Delta X_0$$
$$P_0^0 = P_0$$

For k=1 to N do

Time update (Transition)

$$\Delta \hat{X}_k^{k-1} = \phi * \Delta \hat{X}_{k-1}^{k-1} + w_{k-1}$$

$$P_k^{k-1} = \Phi * P_{k-1}^{k-1} * \Phi^T + Q$$

Linearization step

$$H_k = \left.\frac{\partial h(x^*)}{\partial x^*}\right|_{x^*=X_k^{k-1}}$$

Measurement update

$$K_k = P_k^{k-1} * H_k^T * \left(H_k * P_k^{k-1} * H_k^T + R\right)^{-1}$$

$$\Delta \hat{X}_k^k = \Delta \hat{X}_k^{k-1} + K_k * \left(\Delta Y_k - H_k * \Delta \hat{X}_k^{k-1}\right)$$

$$P_k^k = (I - K_k * H_k) * P_k^{k-1}$$

End

Equation 7.91 also opens the door to the consideration of additional measurements in Kalman filters other than traffic variables like the link flow counts from fixed counting stations (e.g., inductive loop detectors, magnetometers…). For example, the travel times between pairs of points in the network, as measured by ICT applications (e.g., Bluetooth, GPS…).

7.5.2 Aimsun Live

A professional software platform for traffic management with a DTA as core engine and that haves as main input dynamic OD-matrices is Aimsun Live (Aimsun 2020). The DODME process implemented in Aimsun Live is a variant of the numerical approximations of analytical optimization approaches discussed Sect. 7.4. Djukic et al. (2017,2018,2019) reformulate the bi-level approach (Eq. 7.61) as in Eq. 7.93:

$$\text{Min } Z(X) = \alpha * \|X - X^H\|^2 + (1-\alpha) * \|A(X) * X - \widehat{Y}\|^2 \tag{7.93}$$

$$s.t.\ X \geq 0$$

Assuming that the flow estimates are provided by the DTA at the lower level, i.e., at the algorithmic framework in Fig. 7.4. implemented in Aimsun (2020), then $Y = DTA(X)$. This allows a Taylor expansion as in Eq. 7.63. Then, Djukic et al. (2018) propose a modified bi-level approach that, at iteration k, replaces at the upper level the objective function in Eq. 7.95 by the approximation in Eq. 7.94:

$$Z_k(X) = \alpha * \|X - X^H\|^2 + \|\widehat{Y} - Y_k - A_k * (X - X_k)\|^2 \tag{7.94}$$

where at iteration k, X_k is the estimated OD demand vector, A_k the assignment matrix estimated from Aimsun's DTA using Eq. 7.63 and Y_k the vector of estimated link flow counts in the subset of links with counting stations. Aimsun's DTA estimates A_k by stopping the Taylor expansion at either the first or the second term, depending on the desired degree of accuracy or on the affordable computing cost. Djukic et al. (2018) propose to solve the approximated upper level optimization problem (Eq. 7.94) with non-negative variable constraints, using a gradient descent method. Particularly, one using as descent direction the one defined by the following gradient (Eq. 7.95):

$$d_k = -\nabla Z_k(X) \tag{7.95}$$

This gradient can be calculated from (7.94) as

$$\nabla Z_k(X) = 2 * \alpha * \left(X - X^H\right) + 2 * \left(A_k^T * A_k * X - A_k^T * \hat{Y} + A_k^T * Y_k - A_k^T * A_k * X_k\right) \quad (7.96)$$

Then, the new OD-matrix for the lower level iteration $k+1$ is given by Eq. 7.97:

$$X_{k+1} = X_k + \lambda_k * d_k \quad (7.97)$$

where λ_k is the optimal step length in the gradient movement along the descent direction. The gradient procedure to optimize Eq. 7.94 is also iterative. It recalculates the step size at each iteration until either a convergence criterion is met or a maximum number M of iterations is reached, whatever occurs first. At gradient iteration m, the estimated demand is X_k^m, the search direction at this iteration is given by $\nabla Z\left(X_k^m\right)$ (calculated from Eq. 7.96) and the step size calculation can be calculated solving Eq. 7.98, using any of the available line search procedures (Bazaraa et al. 1993):

$$\lambda_k^m = Min_\lambda Z\left[X_k^m - \lambda * \nabla Z\left(X_k^m\right)\right] \quad (7.98)$$

However, since $Z(X)$ is quadratic, the optimal step can be computed analytically using Eq. 7.99:

$$\lambda_k^m = \frac{\left\|\nabla Z\left(X_k^m\right)\right\|^2}{\left\|\nabla Z\left(X_k^m\right)\right\|^2 + \left\|A_k * \nabla Z\left(X_k^m\right)\right\|^2} \quad (7.99)$$

The proposed algorithm iteratively updates the demand at iteration $k+1$ from the demand at the previous iteration k, until some convergence criteria are satisfied. The algorithm is modified with respect to the usual approaches to better fit the requirements for congested large-scale networks. The proposed modification relaxes the assumption on link flow proportions provided by the DTA assignment matrix by computing the marginal effects of the demand deviations on link flows given by Eq. 7.63. Therefore, it reduces the number of OD variables in this Eq. 7.63 by including only those OD pairs whose change in demand values causes significant deviations in the link flows. The modified algorithm is, according to Djukic et al. (2018), as follows:

1) Initialization

Initiate prior OD demand matrix. Set $k = 0$, $I' = X_0$ (seed matrix).

2) Assignment

Assign the demand to the network to obtain the assignment matrix, A_k and to estimate the link traffic counts with traffic observations with Equation 7.63 (1st or 1st & 2nd terms of Taylor's expansion, depending on the choice).

3) Convergence test

Check the value convergence of the objective function. If it has converged, stop and accept the current demand. Otherwise, proceed to 4).

4) Objective Function performance test

Check the performance of the objective function value. If the objective function decreases, proceed to 5). Otherwise, proceed to 6) with $k = k - 1$.

5) Update the OD demand

Estimate the OD demand with the link flows obtained from DTA, as given by Equation 7.63. Otherwise, proceed to 2) with $k = k + 1$.

6) Select OD pairs

Determine the OD pairs whose variation has a considerable impact on the link flows variation in the previous iteration, and insert them in *I'*.

7) Update assignment

For the selected OD pairs in *I',* update the link-flow proportions in the assignment matrix $A_{k-1}$with values obtained from the chosen version of Equation 7.63.

8) Update the OD demand

Estimate the OD demand with the link flows obtained from Equation 7.63. Proceed to 2) with $k = k + 1$.

End

The computational testing of this proposed modified bi-level optimization framework, which solves the high-dimensionality of nonlinear OD-estimation problems by computing the marginal effects only for the most significant OD pairs with respect to traffic observations, allows the modeler to control the trade-off between the simplicity of the model and the level of realism. It is thus very efficient for practical purposes.

7.5.3 Simulation-Based Approaches: Stochastic Perturbation Stochastic Approximation (SPSA)

The optimization problem in Eqs. 7.58 and 7.59, as already mentioned, is highly underdetermined because there are many more variables than equations in the system. In other words, $X \in \mathbb{R}^{|I| \times T}$, $Y \in \mathbb{R}^{\left|\widehat{L}\right| \times T}$ and $|I| \gg |\widehat{L}|$. Therefore, the problem is very sensitive to the quantity of data and the detection layout in the real network. As the availability of new measurements like those provided by smartphone and GPS localization allows calculating travel times between arbitrary pairs of points, the use of these data seems to be a promising approach for reducing the aforementioned underdetermination. An apparently straightforward extension of the bi-level formulation in Eqs. 7.58 and 7.59 accounting for measured, $\widehat{tt}$, and estimated travel times, tt, would be the expansion of the objective function adding a third term, $F_3\left(tt, \widehat{tt}\right)$. This term would be aimed at minimizing the distance between measured and estimated travel times between arbitrary pairs of points in the network, assuming that trips are most likely made via the shortest paths. The hypothetical formulation (Ros-Roca et al., 2021a) would be (Eqs. 7.100–7.102):

$$\text{Min } Z(X) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) + w_3 * F_3\left(tt, \widehat{tt}\right) \tag{7.100}$$

$$s.t.\ Y(X) = \text{Assignmt}(X) \tag{7.101}$$

$$tt(X) = \mathcal{F}(X) \tag{7.102}$$

$$X \in \Omega$$

Assuming that $Y(X) = \text{Assignmt}(X) = A(X) * X$, that is, the relationship between the estimated link flows and the estimated OD-matrix defined by the assignment, the problem can be reformulated as follows (Eqs. 7.103 and 7.104):

$$\text{Min } Z(X) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(A(X)X, \widehat{Y}\right) + w_3 * F_3\left(tt, \widehat{tt}\right) \tag{7.103}$$

$$s.t.\ tt(X) = \mathcal{F}(X) \tag{7.104}$$

$$X \in \Omega$$

The analytical relationship in Eq. 7.104 either does not exist or is unclear. However, in practice, travel times can be estimated from it if the assignment is a DTA. Therefore, it can be accepted that some kind of relationship exists and the relationship $tt(X) \sim \text{Assignmt}(X)$ is assumed. The problem to be solved is again reformulated

as (Eq. 7.105):

$$\text{Min } Z(X) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) + w_3 * F_3\left(tt, \widehat{tt}\right) \tag{7.105}$$

$$s.t.\ (Y, tt) = \text{Assignmt}(X)$$

$$X \in \Omega$$

As mentioned before, it is unclear how these new measurements can be included in the analytical formulations. Nevertheless, it seems rather easy to deal with them by using approaches based on derivative-free optimization methods that approximate the descent direction based on simulation. Among them, simulation optimization techniques are especially suited to deal with optimization problems that cannot be solved with the usual analytical algorithms. Some reasons are:

- The objective function cannot be analytically expressed as a function of parameters because its evaluation requires a simulation. Therefore, it is not differentiable in terms of the parameters.
- The time cost of evaluating the objective function is expensive, as it requires having simulated data for each evaluation of the function.

Simulation-based optimization techniques can be generically formulated assuming that there is a mathematical model $\mathcal{M}$ with a set of parameters $P = \{p_1, p_2, \ldots, p_N\}$ and an objective function $\mathcal{F}(\mathcal{R}, S)$ defined as the sum of error functions between real observations $\mathcal{R}$ and the corresponding simulated data S. The purpose of $\mathcal{M}$ is then to provide (Eq. 7.106):

$$\text{Min } \mathcal{F}(\mathcal{R}, S) \tag{7.106}$$

$$s.t.\ P \in \Omega \subseteq \mathbb{R}^N$$

When $\mathcal{F}(\mathcal{R}, S)$ (i) is on-convex, nonlinear, (ii) cannot be represented analytically as a function of the set of parameters P and (iii) has to be evaluated by simulation.

There is a wide range of different simulation optimization techniques to solve Eq. 7.106. For example, Nelder-Mead, SNOBFIT, and SPSA are optimization techniques, either derivative free or approximating the gradient, that evaluate it using simulation. Osorio and Linsen (2015) make an approximation of the upper level function by building a metamodel that can be solved analytically. Its conceptual diagram is depicted in Fig. 7.5.

Simultaneous Perturbation Stochastic Approximation (SPSA) (Spall 1992) is commonly used in OD-matrix estimation (Cipriani et al. 2011; Cantelmo et al. 2014; Antoniou et al. 2015; Lu et al. 2015; Ros-Roca et al. 2020) and it can easily account for additional measurements (Bullejos et al. 2014; Antoniou et al. 2016; Carrese

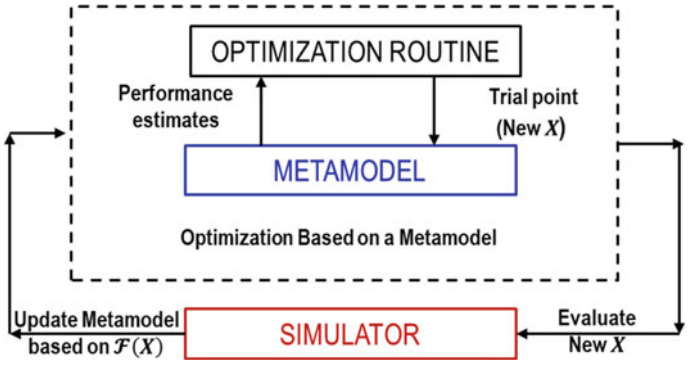

Fig. 7.5 Conceptual diagram of the simulation-based optimization approach of Osorio and Linsen (2015)

et al. 2017; Nigro et al. 2018). SPSA preserves the original upper level formulation and is easy to implement for simulation optimization problems.

SPSA is a simulation-based optimization algorithm, and it only requires two evaluations of the objective function to approximate the gradient instead of N, as in the case of a finite-difference gradient approach. Like in many iterative procedures, it begins with an initial OD-matrix (usually a historical OD-matrix). The next matrix at iteration $k+1$ is computed from the matrix at iteration k, moving a distance a_k along the descent direction provided by the following gradient (Eq. 7.107):

$$X_{k+1} = X_k - a_k * \hat{g}_k(X_k) \tag{7.107}$$

Two particularities distinguish this method from the conventional gradient descent method:

- The estimated gradient $\hat{g}_k(X_k)$, is calculated according to Eq. 7.108:

$$\hat{g}_k(X_k) = \frac{Z(X_k + c_k * \Delta_k) - Z(X_k)}{c_k} * \begin{pmatrix} \Delta_{k,1}^{-1} \\ \vdots \\ \Delta_{k,N}^{-1} \end{pmatrix} = \begin{pmatrix} \frac{Z(X_k + c_k * \Delta_k) - Z(X_k)}{c_k * \Delta_{k,1}} \\ \vdots \\ \frac{Z(X_k + c_k * \Delta_k) - Z(X_k)}{c_k * \Delta_{k,N}} \end{pmatrix} \tag{7.108}$$

  where Δ_k is a random perturbation N-dimensional vector with $\Delta_i, \forall i$ independent identically distributed random variables that satisfy $\mathbb{E}(\Delta_i) = 0$ and $\left|\mathbb{E}\left(\left(\Delta_i^{-1}\right)^n\right)\right| < \infty, \forall n$. One commonly used perturbation is $\Delta_i \sim Be(1/2, \pm 1)$, which is a Bernoulli distribution with a probability of ½ for each ± 1. This is the asymmetric design, although a symmetric design using $Z(X_k + c_k * \Delta_k)$ and $Z(X_k - c_k * \Delta_k)$ can also be considered.

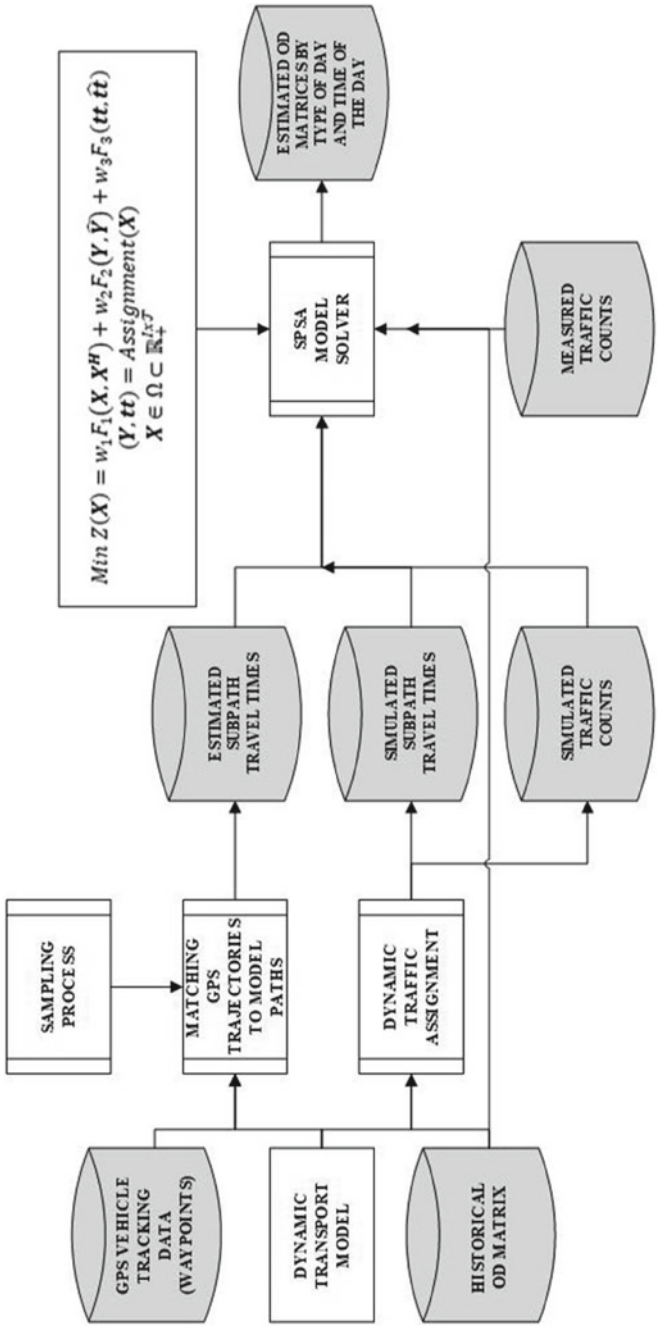

Fig. 7.6 Conceptual diagram of a SPSA approach adding travel times

- The spacing coefficient c_k and the step size a_k are decreasing sequences of positive real values, and they satisfy some regularity conditions in order to ensure the convergence of the method, as detailed in Spall (1992). Typically, the sequences used are (Eqs. 7.109 and 7.110):

$$a_k = \frac{a}{(A + k + 1)^{\alpha}} \tag{7.109}$$

$$c_k = \frac{c}{(k + 1)^{\gamma}} \tag{7.110}$$

  where a, A and c are chosen depending on the problem, while $\alpha = 0.602$ and $\gamma = 0.101$.

Averaging many independent estimates of the gradient of Eq. 7.108 contributes to a more stable and quicker convergence of the SPSA method (Spall 1992). Therefore, the gradient estimation is finally calculated as (Eq. 7.111):

$$\hat{g}(X_k) = \frac{1}{n_g} * \sum_{j=1}^{n_g} \hat{g}_k^j(X_k) \tag{7.111}$$

where $\hat{g}_k^j(X_k)$ is precisely calculated as in Eq. 7.108. The asymmetric design for the gradient saves a large number of assignments, since all $\hat{g}_k^j(X_k)$, $\forall j$ share the mid-point X_k evaluation.

The versatility of simulation optimization techniques, especially when using SPSA, allows including additional information in a new form, such as the constraints in the OD-estimation problem. Ros-Roca et al. (2017) tried adding constraints to simulation optimization problems when dealing with the calibration of microsimulation models.

A potential improvement with respect to the original formulation (Bullejos et al. 2014; Cantelmo et al. 2014) replaces the gradient by the Conjugate Gradient (CG) (Luenberger and Ye 2008), a descent method for the optimization algorithm of the OD-estimation problem. This modifies the descent direction in the iterative procedure by using the previous iteration gradient. It can be incorporated into SPSA by replacing Eq. 7.107 with Eqs. 7.112–7.114:

$$X_k = X_{k-1} + a_k * d_k \tag{7.112}$$

$$d_k = -\hat{g}(X_k) + \beta_k * \hat{g}(X_{k-1}) \tag{7.113}$$

$$\beta_k = \frac{\hat{g}(X_k)^T * d_{k-1}}{\|d_{k-1}\|^2} \tag{7.114}$$

SPSA's main drawback for the OD-estimation problem is that all different OD flows receive the same perturbation magnitude (Eq. 7.108). As OD flows usually have very different magnitudes, this implies very different changes to each flow, which can lead to several problems of convergence. Tympakianaki et al. (2015) approached this phenomenon by clustering the variables according to their magnitude. A different alternative can be normalizing to the interval [0, 1] all variables using some particular reasonable bounds $[a_i, b_i]$. For example, Ros-Roca et al. (2018) performed a classical linear transformation from $[a_i, b_i]$ to [0, 1], where a_i and b_i were based on additional information from the network, particularly socioeconomic or past reliable OD-matrices. The normalization was performed using the following linear application (Eq. 7.115):

$$\begin{array}{ccc} \varphi_i : [a_i, b_i] & \rightarrow & [0, 1] \\ X_i & \mapsto & \frac{X_i - a_i}{b_i - a_i} \end{array} \tag{7.115}$$

Using the normalized variables in SPSA procedure, each variable will be perturbed according to its magnitude.

Experience with similar problems shows that the selection of SPSA gain sequences a_k and c_k is crucial for the convergence and performance of the algorithm. The sequences in the form of Eq. 7.109 and 7.110 are widely used, as they satisfy the conditions of convergence that were proved in Spall (1992). This reduces the problem of selecting appropriate values for a, A, α, c and γ. Kostic et al. (2017b) showed the sensitivity of SPSA with respect to these parameters. Based on the guidelines in Spall (2003), an automated selection of the parameters a, A and c, can be based on the objective function's variability that results from the simulation, and on the desired perturbation steps in the early iterations. The selection would be done according to the following schema:

- First, those values stated as optimal for convergence in Spall (1998) are fixed. That is, $\alpha = 0.602$, $\gamma = 0.101$.
- Several evaluations of $Z(X^H)$ to capture the variability of the objective function are computed. Since the variables have been normalized, it seems natural to use the coefficient of variation ($CoV(Z) = \sigma_Z / \mu_Z$) for this purpose. The parameter c is set at $c = CoV$.
- A is set as 10% of the maximum number of iterations ($A = 0.1 \cdot \text{iter}_{\max}$).
- n_g experiments are simulated using the SPSA logic $X_i = X^H + c\Delta_N$. This allows finding the respective gradients g_k as in the SPSA procedure.
- The desired iterative modification of the first iteration must be determined with Eq. 7.116:

$$X_{k+1} = X_k - a_k * g_k \rightarrow X_{k+1} - X_k = |a_k * g_k| \tag{7.116}$$

- The corresponding a for the desired change in the initial iteration must be computed using Eq. 7.117:

$$\left|a_k * g_k\right| = \frac{a}{(1+A+k)^{\alpha}} * |g_k| \rightarrow a = \frac{\left|a_k * g_k\right| * (1+A+k)^{\alpha}}{|g_k|} \tag{7.117}$$

- The minimum of the n_g performed experiments must be finally chosen. That is (Eq. 7.118):

$$a = \min\left\{a_{\{i=1\}}, \ldots, a_{\{i=N_g\}}\right\} \tag{7.118}$$

As already mentioned, the underdetermination of the OD-estimation problem can lead to different adjusted OD-matrices that show the same traffic counts at the sensor locations even though they are different. Furthermore, the adjusted OD-matrix can also be non-consistent with the socioeconomic factors of the area under study. In traffic analyses, practitioners usually have access to historical data in the form of an OD-matrix X^H which, with a certain degree of uncertainty, provides prior information about the mobility patterns of the target area. Therefore, including constraints in the SPSA formulation that accounts for this information can lead to more realistic results. A possible approach is to add bounding values to the OD values, which is not easy to do in analytical formulations (Codina and Montero 2006) but is relatively easy to manage in SPSA. In Cipriani et al. (2011), a single generation constraint is added to the minimization problem (Eq. 7.119):

$$\sum_{i=1}^{n_h} G_o^i \leq G_o^* \quad \forall o \in \{\text{origins}\} \tag{7.119}$$

with G_o^* being the a priori generation value for the origin zone 0 and n_h the number of time periods. Other approaches, that of Ros-Roca et al. (2020), specify upper and lower bounds for each OD flow, defined in terms of a percentage β of this flow's historical value, according to its degree of uncertainty. With the constraints, the minimization problem is updated as follows (Eqs. 7.120 and 7.121):

$$\text{Min } Z(X,Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) \tag{7.120}$$

$$s.t.\ Y = \text{Assignmt}(X) \tag{7.121}$$

$$X \in G = \left\{(1-\beta) * x_{n,r}^H \leq x_{n,r} \leq (1+\beta) * x_{n,r}^H, \forall x_{n,r} \in X\right\} \subset \mathbb{R}_+^{I \times T}$$

$$X \geq 0$$

This single constraint in Eq. 7.119 results from summing for each origin all the upper bounds in the former minimization problem. The addition of all constraints makes the feasible region bigger. Greater values are therefore allowed for some

variables, but this is compensated by others having low values. On the contrary, the proposal for constrained SPSA in Ros-Roca et al. (2020) defines a smaller feasible region that accounts for further information of each OD pair.

These constraints added to the problem also have an effect on the originally presented SPSA algorithm. Sadegh and Spall (1998) proposed to add a projection to the set G during the iterative procedure shown in Eq. 7.107. The projection would be applied only to the iterative procedure as $X_{k+1} = \pi_G\left(X_k - a_k * \hat{g}_k(X_k)\right)$, while $Z(X_k + c_k * \Delta_k)$ could be computed subject to non-negative OD values. This method, in which some strict constraints are added to the procedure, is called Constrained SPSA.

Inspired in Wang and Spall (1999), other formulations equivalent to Eqs. 7.120 and 7.121 add penalty functions to the objective function (Eqs. 7.122 and 7.123):

$$\text{Min } Z(X, Y) = w_1 * F_1\left(X, X^H\right) + w_2 * F_2\left(Y, \widehat{Y}\right) + r_k * P\left(X, X^H\right) \tag{7.122}$$

$$s.t.\ Y = \text{Assignmt}(X) \tag{7.123}$$

$$X \geq 0$$

where r_k is an increasing sequence of the form $r_k = r*(1+k)^{\rho}$ and $P\left(X, X^H\right)$ is a set of penalization functions for the set of constraints that delimit the constraints of set G. Formally (Eq. 7.124):

$$G \triangleq \left\{q_{n,r}\left(X, X^H\right) \leq 0, \forall n \in I, r \in T\right\} =$$

$$= \left\{x_{n,r} - (1+\beta) * x_{n,r}^H \leq 0, (1+\beta) * x_{n,r}^H - x_{n,r} \leq 0\ \forall n \in I, r \in T\right\} \tag{7.124}$$

The penalty function $P\left(X, X^H\right)$ must be differentiable, non-negative, and an increasing function. Wang and Spall (1999) propose a sum for each constraint of penalizing functions that satisfy $p(x) = 0$ if and only if $x \geq 0$. That is (Eq. 7.125):

$$\begin{aligned} P\left(X, X^H\right) &= \sum_{n \in I} \sum_{r=1}^{T} w_{n,r} * p\left(q_{n,r}\left(X, X^H\right)\right) \\ &= \sum_{n \in I} \sum_{r=1}^{T} w_{n,r} * \max\left\{0, q_{n,r}\left(X, X^H\right)\right\}^2 \end{aligned} \tag{7.125}$$

As in the previous variant, the iterative procedure is also modified to incorporate the gradient of the penalization function (Eq. 7.126):

$$X_{k+1} = X_k - a_k * \hat{g}_k(X_k) - a_k * r_k + \nabla P\left(X_k, X^H\right) \tag{7.126}$$

When additional information from ICT measurements is available, it can be included in the SPSA formulation (Eq. 7.107) as long as it can be estimated from the current OD-matrix X by means of a DTA. This is, for example, the case of subpaths travel times $\hat{tt}$ measured either by Bluetooth (Bullejos et al. 2014; Antoniou et al. 2016) or by GPS tracking (Ros-Roca et al. 2021a). The logical diagram of this process is described in Fig. 7.6. The calculation of these observed subpaths travel times $\hat{tt}$ requires the identification of the most used paths from the available measurements and their map matching to the transport model supporting the DTA. This allows computing the corresponding estimated travel times tt from the current OD, which will be added in the additional term to the objective function in Eq. 7.107. The processing of the GPS data to calculate $\hat{tt}$ is described in Sect. 7.6.

In Kostic et al. (2017a), the additional term of the objective function in Eq. 7.107 is formulated as a function of the measured speeds at detection stations equipped with conventional technologies (i.e., inductive loops), and the DTA used is TRE (Gentile et al. 2007; Gentile 2010), supporting OPTIMA.

7.6 Data-Driven Approaches

The availability of new traffic data supplied by ICT applications, i.e., mobile phones, image processing techniques for license plate recognition, Bluetooth devices, FCD from onboard tracking mobile devices vehicles like GPS, etc., prompted the research interest in finding which could be the advantages of including these data explicitly in the OD-estimation methods. In this context, probe (or equipped) vehicles can be grouped into two generic classes (Nanthawichit et al. 2003; Eiseman and List 2004), according to the explanations in Chap. 1. First, those vehicles equipped with devices that can only be detected at specific locations (i.e., where the detection technology is located), as, for example, those equipped with a tag-reader or with a Bluetooth or Wi-Fi device. Known as "space-based" probe vehicles, their true origin and destination are not known, and their approximate estimates can only be inferred, being this inference strongly dependent on the layout of the detection devices (e.g., tag-readers, Bluetooth antennas). Second, those vehicles equipped with wireless communication mobile devices that are fully visible in the areas covered by the corresponding telecommunications system. Therefore, these systems can provide seamless data about their location, speed, travel direction, etc., depending on the device. These are known as "time-based" probe vehicles.

Methodologies related to space-based probe vehicles that have received significant attention are those based on the identification and reidentification of the license plate of all vehicles passing the area covered by a TV camera with a LPR technology (Mo et al. 2020). Also, those based on the identification of Bluetooth devices between coupled pairs of Bluetooth antennas (Barceló et al. 2013; Behara et al. 2021). However, as already mentioned, results of these methodologies have a strong dependency on the layout of TV Cameras or Bluetooth antennas in the network, this layout becomes a critical aspect for the observability of the system (Castillo et al. 2008) and thus determines the capability of the methods to estimate and predict its state.

As regards time-based probe vehicles, the pervasive penetration of mobile phones has allowed a better understanding of human mobility patterns from their traces, that is, by means of their digital footprints. As mobility patterns include information about where people are and how they got there, mobile phones were soon identified as an important data source for urban modeling. They attracted the interests of researchers and practitioners, as they were seen as a powerful data source that would allow overcoming the well-known drawbacks and limitations of conventional methods in transportation analysis (i.e., household survey). Analyses are usually conducted using datasets, the so-called Call Detail Records (CDR), previously recorded by a mobile provider for communication and billing purposes, after an anonymization process. A seminal example of this process can be found in González et al. (2008), where each individual calling activity is characterized to allow monitoring the user's movement over time. Calabrese et al. (2013) provide an example of techniques aimed at extracting useful mobility information from mobile phone traces of millions of users from which to infer individual mobility patterns in large urban areas, especially OD-matrices (Zhang et al. 2010; Calabrese et al. 2011). Since CDR are time tagged and locations can be identified after suitable processing, added value information for a variety of mobility analyses can be extracted from the (Çolak et al. 2015). Additionally, OD-matrices can be differentiated by purpose and time of the day (Alexander et al. 2015). However, this requires resorting to very specific Data Analytics techniques, given the huge amount of data frequently recorded from millions of users. Gundlegård et al. (2015) or Jianga et al. (2016) are good examples of this data processing to extract the OD-matrices.

However, the type of OD-matrices that dynamic traffic models used in traffic management systems require as input is rather different from the matrices directly extracted from DCR. Indeed, the mobility patterns modeled by these latter OD-matrices are global, that is, they include all types of trips without distinguishing the transportation mode used. Conversely, the OD-matrices of interest for traffic management purposes are usually those modeling the passenger cars patterns. Additional work is necessary to estimate these specific OD-matrices. For example, DCR OD-matrices can be combined with simulation models like MITSIM (Iqbal et al. 2014) or they can be fused with other data sources (Montero et al. 2019). Bassolas et al. (2019) propose also a fusion variant to generate inputs to activity-based travel demand models using MATSIM.

Among the time-based probe vehicles, the better suited to generate OD-matrices that can be exploited by dynamic traffic models seem to be those allowing the tracking of the equipped individual vehicles and the reconstruction of their trajectories. Assuming that the collected data from the tracking technologies include geolocation and time stamps, i.e., waypoints in the terminology of commercial GPS providers, map matching and path inference procedures could provide comprehensive information about origins, destinations, taken paths, and path travel times. This was essentially the assumption in an early paper of van Aerde et al. (1993), accepting that probe vehicles were fully visible. The mentioned seminal papers of Nanthawichit et al. (2003), and Eisenman and List (2004) later accepted this hypothesis. Therefore, assuming that these sampled trajectory data are available, the question is whether

and how they can be used to find sound estimates of dynamic OD-matrices, that is, OD-matrices discretized in time, exploiting for that purpose the time tag recorded data.

Research on the potential use of these mobile data for transport analysis has also prompted a key question with relevant practical applications. Most of the DODME approaches discussed in the previous sections usually assume that one of the inputs is provided by an available historical matrix. The reliability and quality of such historical OD-matrices has been questioned in practical transport planning practice, as it could be largely outdated or even not exist. However, this is not the case in most practical traffic management applications due to the amount and quality of data supplied by modern surveillance systems. Therefore, the improvement of the seed matrices used in DODME by means of sample data from probe vehicles is a relevant contribution. However, the previous discussion on DODME approaches makes evident that all of them rely on the estimate of a dynamic assignment matrix. The fact that this assignment matrix must be estimated by a DTA or a DUE and that the approach implies an iterative process, this could represent a heavy computational burden not affordable in real-time applications. Therefore, the key question is: can the dynamic traffic assignment matrix be empirically estimated from probe vehicle data? And, if possible, how can it be used to improve DODME approaches? A positive answer to the first question opens the door to build models some of whose components are directly derived from an empirical procedure, which would be based on the observed data, instead of from an analytical procedure. In other words, this approach paves the way to build data-driven models.

7.6.1 A Conceptual Proposal on Data-Driven Modeling

From this latter perspective, an interesting proposal is that made by Yang et al. (2017). They wanted to determine whether the availability of such trajectory data could be used to develop an approach to DODME independent of the reliability of an historical OD. That is, to make a good empirical estimation of the assignment matrix, making it unnecessary to resort to DTA. According to the above-mentioned statements, it is assumed that each probe vehicle reports its position in the form of GPS coordinates after a preprocessing procedure performed with map-matching techniques. In summary the approach is as follows:

- It is assumed that vehicle trajectories from origins to destinations are traceable for each probe vehicle, and that the supplied GPS locations have been suitable preprocessed by data cleansing and map-matching procedures. Therefore, GPS locations in the approach are assumed to be exact.
- If $\widehat{L} \subseteq L$ is the subset of links with counting stations, two link flow measurements are available for each time period r. There are $\hat{y}_{lr}, l \in \widehat{L}, r \in T$ flows from the counting stations at links $l \in \widehat{L}$, and $\widehat{h}_{lr}, l \in \widehat{L}, r \in T$ flows of probe vehicles crossing that link l at time interval r.

- The OD probe ratios, that is, the average number of vehicles observed across the entire network during a time interval are given by Eq. 7.127:

$$\gamma_r = \frac{\sum_{l \in \hat{L}} \hat{h}_{lr}}{\sum_{l \in \hat{L}} \hat{y}_{lr}} \forall r \in T \tag{7.127}$$

- Thus, the seed OD-matrix $\hat{x}_{nr}$ can be estimated with Eq. 7.128:

$$\hat{x}_{nr} = \frac{\hat{z}_{nr}}{\gamma_r} \forall n \in I, \forall r \in T \tag{7.128}$$

  where I is the set of all OD pairs, and $\hat{z}_{nr}$ is the number of identified probe vehicles traveling from the origin to the destination of the $n-th$ OD pair.
- The assumption of the identification of locations of probe vehicles allows, in a similar way, directly estimating the assignment matrix (Eq. 7.129):

$$a_{ln}^t = \frac{1}{|T|} * \sum_{r \in T} \left(\frac{\hat{z}_{ln}^{r.r+t}}{\hat{z}_{nr}} \right) t \in T, n \in I, l \in \hat{L} \tag{7.129}$$

Assuming this data-driven approach, a variant of the model in Eq. 7.71 is proposed (Eqs. 7.130):

$$\underset{x_{nr}}{Min} \left[\sum_{r \in T} \sum_{n \in I} \frac{\left(x_{nr} - \hat{x}_{nr}\right)^2}{w_{nr}^2} + \sum_{r \in T} \sum_{l \in \hat{L}} \frac{\left(y_{lr} - \hat{y}_{lr}\right)^2}{q_{rl}^2} \right] \tag{7.130}$$

s.t.

$$y_{lr} = \sum_{t \in T} \sum_{n \in I} a_{ln}^t * x_{n,r-t} \forall l \in \hat{L}, \forall r \in T$$

$$-\beta x_{nr} \leq x_{n,r+1} - x_{nr} \leq \beta x_{nr} \forall n \in I, \forall r \in T$$

$$x_{nr} \geq 0 \forall n \in I, \forall r \in T$$

where the first constraint expresses, as in the analytical models, the relationships set up by the empirical assignment matrix a_{ln}^t between y_{lr}, the estimated flows at the links l with traffic detection stations $l \in \hat{L}$ for each time interval r, and $x_{n,r-t}$, the OD flows leaving the origin at time $r-t$, observed in link l at time r. The objective function in this case is formulated in terms of a quadratic distance function. Its metrics are defined, as in Eq. 7.35, respectively, by the matrix of variances, w_{nr}^2 and q_{rl}^2, of the empirical OD-matrix, $\hat{x}_{nr,}$ and the link flow measurements, $\hat{y}_{lr}$. The coefficient β in

the bounding constraints of Eq. 7.130 is the maximum change percentage of OD flows between two consecutive intervals.

Yang et al. (2017) also propose a more general variant of the model assuming that there is a correlation between the OD probe ratios, γ_{nr}, and the link probe ratios, $\theta_{lr.}$. In other words, that there exists a function $\theta_{lr} = P(\gamma_{nr})$, for which they postulate the following form (Eq. 7.131):

$$\theta_{lr} = \sum_{t\in T}\sum_{i\in I}\rho_{ln}^{t} * \gamma_{nr} = \sum_{t\in T}\sum_{i\in I}\rho_{ln}^{t} * \left(\frac{\hat{z}_{n,r-t}}{x_{n,r-t}}\right) \forall l \in \hat{L}, \forall r \in T \tag{7.131}$$

where ρ_{ln}^{t} define the assignment matrix of probe ratios, which is assumed to be computed empirically from the collected data according to the main hypothesis of the method. These link probe ratios depend on the estimated OD-matrix, x_{nr} as expressed in Eq. 7.131, being therefore new variables of the model. Assuming that the available GPS data allow estimating the empirical values $\hat{\theta}_{lr,}$ they can be added to the objective function (Eq. 7.130) yielding the enhanced model in Eq. 7.132:

$$Min_{x_{nr}}\left[\sum_{r\in T}\sum_{n\in I}\frac{\left(x_{nr}-\hat{x}_{nr}\right)^2}{w_{nr}^2} + \sum_{r\in T}\sum_{l\in\hat{L}}\frac{\left(y_{lr}-\hat{y}_{lr}\right)^2}{q_{rl}^2} + \sum_{r\in T}\sum_{l\in\hat{L}}\frac{\left(\theta_{lr}-\hat{\theta}_{lr}\right)^2}{v_{rl}^2}\right] \tag{7.132}$$

s.t.

$$y_{lr} = \sum_{t\in T}\sum_{n\in I} a_{ln}^{t} * x_{n,r-t} \forall l \in \hat{L}, \forall r \in T$$

$$\theta_{lr} = \sum_{t\in T}\sum_{i\in I}\rho_{ln}^{t} * \left(\frac{\hat{z}_{n,r-t}}{x_{n,r-t}}\right) \forall l \in \hat{L}, \forall r \in T$$

$$-\beta x_{nr} \le x_{n,r+1} - x_{nr} \le \beta x_{nr} \forall n \in I, \forall r \in T$$

$$x_{nr} \ge 0 \forall n \in I, \forall r \in T$$

where, as before, v_{rl}^2 is the variance of the observed ratios. Since the optimization model is quadratic, the gradient can be easily calculated and a gradient algorithm is proposed to numerically solve the problem.

7.6.2 Accounting for Mobility Learning from ICT Data Collection

Cascetta et al. (2013) formulate the hypothesis that "*an OD estimator can be based on the assumption of constant distribution shares across larger time horizons with respect to the within-day variation of the production profiles, leading to an estimator that dramatically improves the unknowns/equations ratio*". Krishnakumari et al. (2019) propose to go a step further. They assume that all realized travel times are available over all (shortest) paths. Also, that it is only necessary to specify how many of the shortest paths are actually used for each OD pair and the proportions of each OD flows over these used shortest paths. These proportions are a behavioral assumption at the macroscopic scale (a path flow proportion), and not in the form of a detailed route choice model with (elaborated) trade-offs.

Nevertheless, the assumptions in Krishnakumari et al. (2019) about the distribution of traffic over the network are not sufficient to estimate the underlying OD-matrix, They must be complemented with additional information that, for instance, can be provided by measured link flow counts $\hat{y}_{lt}$ from counting stations, measured at links $l \in \hat{L}$ at time $t \in T$. Link flow counts y_{lt} that, as shown in Eq. 7.76, can be estimated in terms of the relationships between flows and OD flows x_{ijr} departing from origin i with destination j at time interval r, arriving at link l at time t, and the assignment matrix a_{ijr}^{lt}.

However, to be valid, these relationships must be set up considering that counts in or downstream congestion are not informative of demand, but of (discharge) capacity, as shown in Frederix et al. (2011). Information on demand is only provided if y_{lt} are estimated in uncongested conditions, and no path flows for OD pair $(i, j) \in I$ using a path k to which l belongs, experience a bottleneck upstream before crossing link l. Therefore, in order to overcome these limitations, the computation of the assignment matrix, or that of any related terms, must be done in a way that explicitly accounts for congestion effects.

The approach proposed by Krishnakumari et al. (2019) assumes that, in addition to the availability of OD travel times, also the productions P_{ir}, i.e., the total outgoing flows from each origin i, during period r, as well as the attractions A_{jr}, i.e., the total incoming flows to each zone j during period r, are observable and, therefore, available.

The availability of these inputs from the observed data is based on a methodology proposed by López et al. (2017b) that is based on specific Data Analytics techniques suited to build consensual 3D speed maps by clustering techniques from link speeds. These speeds are estimated from field data by a heuristic procedure defined in López et al. (2017a). According to the authors, this procedure can exploit classical data (e.g., from inductive loops or cameras) as well as those from more modern data sources (e.g., mobile phone records, GPS tracking, etc.). The methodology is illustrated in the referenced papers for a case in which link speeds were estimated from individual travel times recorded by TV cameras equipped with LPR technology. However, it is extensible to other technologies as long as the requirements are met.

In any case, field data must be appropriately cleansed out and the outliers removed. Krishnakumari et al. (2019) discuss several procedures for this purpose, among them a moving average process where, if τ_n is the $n-$ th realized travel time for a given OD pair, Eq. 7.133 gives the moving average:

$$\overline{\tau}_n = \frac{1}{k}\sum_{i=0}^{k-1}\tau_{n-i} \tag{7.133}$$

Outliers are defined by $\overline{\tau}_n + \Delta\tau$, where $\Delta\tau$ is a time window empirically determined, for instance, as the standard deviation times recorded during the peak demand. The mean of the observed travel times for a given OD pair (i, j) at a given period is considered the travel time from i to j at time t,$tt_{ij}(t)$. Additionally, the considered k-shortest paths as the most likely used between each OD pair. For a particular one(i, j), a path L_p is characterized by a sequence of links $L_p = \left(l_{p1}, l_{p2}, \ldots, l_{pn}\right)$. Then, the path speed is (Eq. 7.134):

$$s_p = \frac{\text{dist}(L_i)}{tt_{ij}(t)} \tag{7.134}$$

Krishnakumari et al. (2019) also consider various approaches to impute link speeds when no data are available.

The main assumption behind the approach proposed by Lopez et al. (2017b) is that the availability of the data provided by these more modern data sources allows finding empirically driving insights of human mobility, namely, those concerning their dynamic aspects, and thus enables their use in mathematical models aimed at predicting that dynamic mobility. This means to investigate the regularity of macroscopic mobility patterns, how they vary within days and from day to day. For that purpose, Lopez et al. (2017b) propose a methodology based on what they call 3D maps, in essence spatial–temporal speed cluster maps, which are a joined partition of space (i.e. the road network) and time into homogeneous clusters characterized by constant mean speeds. The proposed approach considers that link speed data can be reconstructed from trip travel time observations with Eq. 7.134, as in Lopez and al. (2017b), and that the network is coded in Open Street Map Geographical Information System (OSM GIS) Database, also used to compute all shortest paths. The cluster building process is based on the following partitioning criteria:

- All clusters should contain a single connected component. In other words, all links in the cluster are reachable within the cluster.
- An intra-cluster homogeneity criterion, formulated in terms of the minimization of the internal speed variance for all clusters. If n is the number of clusters, the total within cluster variance TV_n is given by (Eq. 7.135):

$$TV_n = \frac{1}{\sum_{i=1}^{n} n_i} * \left(\frac{\sum_{i=1}^{n} n_i * s_i^2}{s^2}\right) \tag{7.135}$$

where n_i is the number of links in cluster i, s_i the standard deviation of links speeds in cluster i, and s the standard deviation for the whole network. It is assumed that link speeds have been estimated from Eq. 7.134.

- An inter-cluster dissimilarity criterion that maximizes the difference in speed between neighbor clusters, where the inter-cluster dissimilarity is given by Eqs. 7.136 and 7.137:

$$CCD_n = \frac{\sum_{i=1}^{n} \sum_{k=i+1}^{n} \delta_{ik} * \sqrt{n_i * n_k} * |\overline{v}_i - \overline{v}_k|}{\sum_{i=1}^{n} \sum_{k=i+1}^{n} \delta_{ik} * \sqrt{n_i * n_k}} \tag{7.136}$$

$$\delta_{ik} = \begin{cases} 1 & \text{if clusters } i \text{ and } k \text{ have a common border} \\ 0 & \text{otherwise} \end{cases} \tag{7.137}$$

where $\overline{v}_i$ is the mean speed in cluster i.

Lopez et al. (2017b) test three different clustering approaches, *k-means*, DBSCAN, and S-cut and conclude that, at least in the case study reported in the paper, *k-means* is the most economical in terms of computational cost to obtain the envisaged 3D speed maps. Furthermore, assuming that the observational data cover a period of M days, they add a new process to find commonalities in these days' congestion patterns, the so-called "consensual" patterns, by means of *Consensus Learning Techniques* (Filkov and Skiena 2004).

The approach proposed by Krishnakumari et al. (2019) uses these results for different purposes:

- To estimate or predict the production and attraction patterns using the identified 3D speed and flow patterns (possibly augmented with other data) using machine learning techniques (especially Neural Network techniques, although other techniques could also be used).
- To compute N weighted (by travel time) shortest paths, where N is an assumption on how many alternative routes are used on average for each OD flow on these paths.
- To estimate path flows on the used paths assuming that are inversely proportional to the realized travel times on these paths, considering path overlap, and under the additional constraint that the path flow solution space is determined by all admissible link flow counts.

Let's assume that x_{ijr}^k is the path flow from origin $i \in O$ (where O = {set of all origins}) to destination $j \in D$ (where D = {set of all destinations}), departing from origin i at time period $r \in T$ (where T is the time horizon) on path $k \in N_{ij}^k$ (where N_{ij}^k is the set of all paths between origin i and destination j at time period k}; x_{ijr} is the OD flow from origin $i \in O$ to destination $j \in D$, departing from origin i at time period $r \in T$; P_{ir} is the production of origin i during period r, A_{jr} is the attraction of destination j during period r; TT_{ijr}^k is the travel time for vehicles traversing path k from origin i to destination j departing from i in time period r; P_{ijr}^k is

the proportion of vehicles traveling on path k from origin i to destination j departing from i in time period r and $\hat{y}_{lr}$ is the measured flow count in link l at time period r. P_{ir} is the sum of all outgoing flows from i at this time period along all paths $k \in N_{ij}^{r}$ from i to all destinations $j \in D$ (Eq. 7.138):

$$P_{ir} = \sum_{j \in D} \sum_{k \in N_{ij}^{r}} x_{ijr}^{k} \tag{7.138}$$

In a similar way, the attraction A_{jr} of destination j during period r is the sum of all incoming flows to destination j from all origins $i \in O$ along all paths $k \in N_{ij}^{k}$ (Eq. 7.139):

$$A_{jr} = \sum_{i \in O} \sum_{k \in N_{ij}^{r}} x_{ijr}^{k} \tag{7.139}$$

Since links speeds are available, path travel times TT_{ijr}^{k} can be calculated. From them, a behavioral assumption can be made on the proportion of trips using each available path in terms of each utility, which is defined by the path travel time. Krishnakumari et al. (2019) estimate this path proportion with the modified logit-based model proposed by Ben-Akiva and Bierlaire (1999) (Eq. 7.140):

$$P_{ijr}^{k} = \frac{e^{TT_{ijr}^{k} * (1 - PS^{k})}}{\sum_{p \in N_{ij}^{r}} e^{TT_{ijr}^{p} * (1 - PS^{p})}} \tag{7.140}$$

In this Eq. 7.140 a correction term PS^{k} is added to the deterministic component of the discrete-choice mode. It is the path size factor defined by Eqs. 7.141 and 7.142:

$$PS^{k} = \sum_{a \in \text{Path } k} \left(\frac{l_a}{L^{k}} \right) * \frac{1}{\sum_{p \in N_{ij}^{r}} \delta_{ap}} \tag{7.141}$$

$$\delta_{ap} = \begin{cases} 1 & \text{if link } a \text{ belongs to path } p \\ 0 & \text{otherwise} \end{cases} \tag{7.142}$$

where l_a is the length of link a, L^{k} is the length of paths k, and δ_{ap} is the link-path incidence matrix. The path size factor tries to capture the correlations between alternative paths measuring the dependencies in terms of a certain degree of similarity among the shared links. The calculation of the path proportions allows setting up the relationships between the OD flows, $x_{ijr,}$ and the path flows, x_{ijr}^{k} (Eq. 7.143):

$$x_{ijr}^{k} = P_{ijr}^{k} * x_{ijr} \forall i \in O_i, \forall j \in D_j, \forall r \in T, \text{and } k \in N_{ij}^{r} \tag{7.143}$$

The number of paths N_{ij}^{r} can be exponentially large but, in practice, as not all of them are significantly used, this number can be reduced to a smaller set $N_{ij}^{r*} \leq N_{ij}^{r}$. This smaller set can be identified as part of the data analytics procedures to estimate the values of the model components. This leads to the approximation of the estimated OD-matrix as in Eq. 7.144:

$$x_{ijr} = \sum_{k \in N_{ij}^{r*}} x_{ijr}^{k} \tag{7.144}$$

This approximation is sufficiently good if N_{ij}^{r*} has been properly defined.

The relationship between link flows and path flows can be reformulated explicitly considering the effects of congestion in order to account for the conditions discussed above. That is, that flows y_{lr} measured in link l at time r are informative of path flows crossing the link only if they are not congested at that time and if none of the links upstream of it experiences a bottleneck. The approach chosen by Krishnakumari et al. (2019) considering the subset of paths to which link l belongs and satisfying these conditions can be formulated as follows. If $\wp_r^l$ is the set of paths to which link l belongs at time r, the subset of paths satisfying the conditions is given by (Eq. 7.145):

$$\wp_{ijt}^{k} \in \wp_r^l, t \leq r - \overline{TT}_{ijr}^{k \setminus l} \ \forall i, j, k \quad k \text{ all paths traversing } l \text{ during period } r \tag{7.145}$$

where $\overline{TT}_{ijr}^{k \setminus l}$ estimates the partial arrival travel times to link l along the paths in $\wp_{ijt}^{k}$. This implies that (Eq. 7.146):

$$\sum_{\wp_{ijt}^{k} \in \wp_r^l} P_{ijt}^{k} * x_{ijr} = \begin{cases} 0 & \text{if link supstream of } l \in \wp_{ijt}^{k} \text{ are congested} \\ P_{ijt}^{k} * x_{ijr} & \text{otherwise} \end{cases} \tag{7.146}$$

Thus, if $\hat{y}_{lr}$ are the link flows measured at links $l \in \hat{L} \subseteq L$ equipped with detection stations, their relationships with the OD flows x_{ijr} can be stated with Eq. 7.147:

$$\hat{y}_{lr} = \sum_{(i,j) \in I} P_{ijr}^{k} * x_{ijr} \ \forall l \in \hat{L}, \forall r \in T \tag{7.147}$$

Together with the corresponding reformulations of Eqs. 7.134 and 7.135 and in terms of Eq. 7.146, a system of equations (Eqs. 7.148 and 7.149) is defined:

$$P_{ir} = \sum_{j \in D} x_{ijr} \ \forall i \in O, \forall r \in T \tag{7.148}$$

$$A_{jr} = \sum_{i \in O} x_{ijr} \ \forall j \in D, \forall r \in T \tag{7.149}$$

As highlighted in Krishnakumari et al. (2019) "*this system is underdetermined or overdetermined (or rare cases full rank) depending on the available link counts and the choice and number of link paths for each OD pair*". To solve the system, the authors propose to use the constrained least squares algorithm of Altman and Gondzio (1999), either with lower bounds set to 0 to ensure non-negative solutions, or without bounds when no solution exist and ignoring the negative values in computing the solution error.

A potential limitation of the proposed approach arises for large networks. That is, when the number of origins and destinations grows and, then, the number of OD flows grows quadratically. However, the number of Eqs. 7.148 and 7.149 in the system only grows linearly, as link flow equations do (Eq. 7.147), assuming also an increase in the number of detection stations. The authors propose to use in this case the dimensionality reduction techniques studied in Djukic et al. (2012), which are based on the application of the *Principal Components Analysis* (Jolliffe 2002).

To end this section, it should be noticed that this data-driven approach is the planned forthcoming OD-estimation method in future versions of the corresponding modules of Aimsun Next and Live software platforms for traffic analysis and management.

7.6.3 Estimating Assignment Matrices from FCD Data

As mentioned, the computational burden associated with the DTA required in the analytical approaches to the DODME problem, which is necessary to estimate the assignment matrix, and the existing doubts on how to integrate the additional information that can be available, have fostered research on these issues not only among researchers, but also among practitioners and developers of professional software platforms. An example of this motivation can be found in a recent work of the team supporting the OPTIMA traffic management platform (Mitra et al. 2020). This platform is aimed at estimating base OD demand matrices for large-scale networks using the information that can be extracted from large amounts of FCD data and link flow counts. The main assumption, similar than that of previous approaches, is that a detailed analysis of FCD trajectories, if properly and accurately done, enables the estimation of the two main required inputs: (i) a revealed OD-matrix X^0 extracted from the FCD trajectories, playing the role of seed matrix and (ii) information to build from FCD data a reliable assignment matrix that can replace the one provided by DTA in analytical approaches.

A critical point is that of the quality of the FCD data, since they can be poor, not homogeneous, or biased. However, Mitra et al. (2020) claim that, even in these cases, it is possible to take advantage of these data. Their suitable cleansing and filtering and their clustering according to similar average behaviors are useful techniques to apply. Also, the use of specialized map-matching algorithms matching each individual raw GPS trajectory on the transportation graph in order to reconstruct the most likely paths in this graph (Hart et al. 1968; Marchal et al. 2004; Quddus et al. 2007; Kubicka et al.

2018; Millard-Ball et al. 2019). The map-matched trajectories can be associated with origin and destination zones, departing from origin zones at specific times of the day.

Let's assume that X is the estimated OD vector of size $|I|*|T|$ (where I is the set of OD pairs and T the set of time intervals), that $\widehat{Y}$ is the vector of traffic counts (of size $\left|\widehat{L}\right| * |T|$, being $\widehat{L}$ the set of links with counting stations), and that A is the estimated assignment matrix from FCD data. Then, mapping the estimated OD flows to the estimated link flow counts Y (with $Y = A * X$) and simplifying the formulation for a simple fixed time interval (no interdependencies between time intervals are assumed in this approach. See Mitra et al. (2020) for additional details), the DODME problem can be formulated as in Eq. 7.150:

$$\text{Min } \varphi(X) = \frac{1}{2} * \|A * X - \widehat{Y}\|^2 + \frac{\lambda}{2} * \|X - \widehat{X}\|^2 \tag{7.150}$$

where λ is the relative weight of the demand term, and $\widehat{X}$ is the reference demand vector, whose $ij-$ th element is given by Eq. 7.151:

$$\widehat{X}_{ij} = \gamma * \alpha_i * \beta_j * X_{ij}^0 \ \forall i \in O, \forall j \in D \tag{7.151}$$

being O the set of origins, D the set of destinations and X^0 the observed seed OD-matrix from FCD trajectories. γ is a constant factor that homogenously scales all OD pairs, and $\alpha_i, \forall i \in O$ and $\beta_j, \forall j \in D$, respectively, are the generation and attraction factors for each origin and each destination.

The solution to Eq. 7.150 is found by an iterative process that generates a sequence of feasible solutions $\{X^k\}$. This is done in such way that a new solution is found at iteration $k+1$ from the solution at iteration k by moving a step of length $\theta^k \in (0,\ 1]$ along a feasible descent direction ΔX^k (Eq. 7.152):

$$X^{k+1} = X^k + \theta^k * \Delta X^k \tag{7.152}$$

Since $\varphi(X)$ is a quadratic problem, the descent direction can be found by a Newton method solving with Eq. 7.153:

$$\Delta X^k = \left[\nabla^2 \varphi\left(X^k\right)\right]^{-1} * \nabla \varphi\left(X^k\right) \tag{7.153}$$

where $\nabla \varphi\left(X^k\right)$ is the gradient of $\varphi(X)$, and $\nabla^2 \varphi\left(X^k\right)$ the Hessian at X^k. In practice, Eq. 7.153 can be solved efficiently without inverting the Hessian and, since $\varphi(X)$ is quadratic, the solution can be exactly found in one step if the Hessian is definite positive.

Several alternatives have been proposed (Mitra et al. 2020) to estimate the values of factors γ, α_O and β_D, where α_O and β_D are the vectors of attraction and generating factors. An example procedure that simultaneously optimizes α_O and β_D could consist

in (i) calculating the optimal value of γ the common global factor by solving the quadratic problem in Eq. 7.154 and (ii) calculating the optimal values of α_O and β_D by solving Eq. 7.155:

$$\begin{aligned} &\text{Min } \varphi(X) \\ &\gamma \geq 0 \end{aligned} \tag{7.154}$$

s.t. $\alpha_O = \beta_D$=1

$$\begin{aligned} &\text{Min } \varphi(X) \\ &\alpha_O, \beta_D \geq 0 \end{aligned} \tag{7.155}$$

Mitra et al. (2020) present promising results of this approach applied to the large-scale network of Turin, with 438 zones, 96,420 links, 6,352 nodes, 1203 counting locations and GPS data for 1 year.

The potential problems of dealing with GPS data reported when discussing previous approaches fostered the search for other practical solutions. Most of these problems concern the unbiased reconstruction of vehicle trajectories and the estimation of the observed seed OD-matrix X^0 and are usually linked to the fact that most of the available commercial GPS data are obtained from non-homogeneous vehicle fleets (e.g., indiscriminate mix of commercial vehicles and passenger cars). Another source of issues is trajectories being split by random identity changes due to privacy reasons. However, once these data properly cleansed and filtered out, the waypoints or POIs (Points of Interest) supplied by GPS data can be considered reliable. These are usually given as an ordered sequence of waypoints containing the information (*IDk, date, ts(kl), latkl, longkl*), as illustrated in Table 7.1. *IDk* is the identity of each trip k, the date stands for the recording date, *ts*(k,l) is the time tag for the $l-$ th observation of trip k and *latkl* and *longkl,* respectively, are its latitude and longitude.

However, these geographically referenced data do not usually correspond to the analyzed road network. Therefore, as already mentioned, they must be properly map-matched to transform these sequences of waypoints in points corresponding to

Table 7.1 Example of GPS recorded waypoints

ID	Date	Time stamp	Latitude	Longitude
4,261,353	2019–11-30	22:43:58	45.445988	9.1244048
4,261,353	2019–11-30	22:44:27	45.445496	9.1241952
……	……	……	……	……
4,261,353	2019–11-30	22:50:57	45.444767	9.1192517
4,261,355	2019–11-30	22:43:58	45.445980	9.1247048
4,261,355	2019–11-30	22:44:27	45.445574	9.1192821
……	……	……	……	……
4,261,355	2019–11-30	22:50:57	45.444767	9.1197541
……	……	……	……	……

paths on that network. The most used procedures (Marchal et al. 2004; Schuessler and Axhausen 2009; Pereira et al. 2009; Rahmani and Koutsopoulos 2013; Kubicka et al. 2018) assign each waypoint to a point in the nearest link of the network. There are available tools provided by software platforms to perform this operation, as, for instance, OpenLR (OpenLR 2020), or GPX (PTV Visum 2020). An example on how this works is depicted in Fig. 7.7, in which the red stars are the waypoints and the red numbers near the links are the relative position of the waypoint projection over the target link. Timestamps for waypoints are depicted in green.

Link travel times can be heuristically estimated from waypoint timestamps according to their sequence (Ros-Roca et al. 2021b). In this example, for all links in the sequence, the interpolated travel time for a link is the sum of the timestamp differences of two consecutive waypoints mapped in the target link. In the case of two consecutive waypoints that are not wholly projected within one link, the distance-based fraction within the link is taken (lk is the length of link k in Fig. 7.7). For instance, the travel time for link l_3 can be estimated taking into account that the travel time for the trip between the 3rd and 4th waypoints is 20 s, and that it is the estimated travel time of the whole link l_3 plus a 0.2 fraction of l_2 and a 0.7 fraction of l_4(Eq. 7.156, with the result in s):

$$tt_3 = \frac{l_3}{0.2{*}l_2 + l_3 + 0.7{*}l_4} * 20 \tag{7.156}$$

The estimated travel time in link l_4 is obtained by adding two parts, the first part is the travel time proportion between the 3rd and 4th timestamps in link l_4 (adding 0.7 of l_4 to 0.2 of the length of link l_2 plus the entire length of link l_3). The second part is estimated directly from the proportion of link l_4 lying between 4 and 5th timestamps (a fraction of 7 s, which is the travel time between waypoints, calculated as 0.3 of the l_4 distance over the total distance between the 4th and 5th waypoints, that is 0.3 l_4+0.2 l_5). Overall, the travel time in link l_4 is given by Eq. 7.157 (in s):

$$tt_4 = \frac{0.7 * l_4}{0.2 * l_2 + l_3 + 0.7 * l_4} * 20 + \frac{0.3{*}l_4}{0.3{*}l_4 + 0.2 * l_5} * 7 \tag{7.157}$$

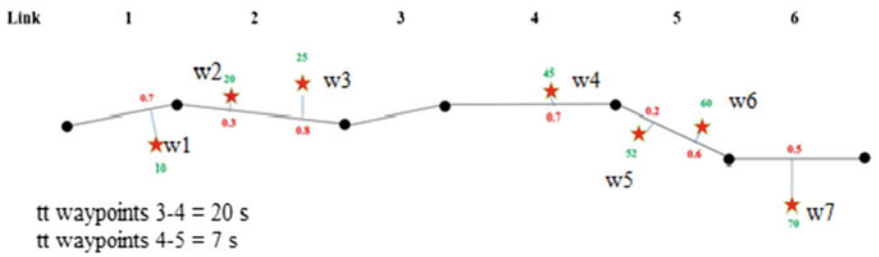

Fig. 7.7 Schematic overview of the map-matching process

Finally, once all the waypoint sequences are converted to several paths with full details at the link level, the link travel times are averaged. The outcome of this process is the set of observed link travel times at each time period t, $\widehat{tt}_{lt} \forall l \in L, \forall t \in T$, for all links in the network that are monitored by GPS tracking. That is, the dataset of estimated link travel times. Despite possibly being huge the quantity of trajectories available for the target network, which will depend on the penetration rate of devices with GPS among the population, the final sample may uncover links. It is also possible that some of them are not fully covered by time information, as, for instance, the first and last links in each sequence (e.g., links 1 and 6 in the example in Fig. 7.7). Moreover, the procedure that infers link travel times can produce non-feasible values when they are below the free-flow link travel time. In these situations, scaled travel times are used (Eqs. 7.158 and 7.159):

$$\hat{tt}_{l't} = R * tt_{0l'} \tag{7.158}$$

$$R = \text{mean}_{l \in GPS} \left(\frac{\hat{tt}_{lt}}{tt_{0l'}} \right) \tag{7.159}$$

where $tt_{0l'}$ is the free-flow travel time at each link, and R is computed using all observed link travel times and their corresponding free-flow travel times. That is, R is the arithmetic mean of the expanding factors found for each link and can be understood as a global expanding factor linked to the congestion effect. The methodological process for generating the observed link travel times dataset is summarized in Fig. 7.8.

The estimated average link travel times $\widehat{tt}_{lt}$ for each link $l \in L$, for each time interval $t \in T$ can be used to generate a plausible *Route Choice Set* $\mathcal{K} = \left\{K_{ijr}, \forall i \in O, \forall j \in D, \forall r \in \mathcal{T}\right\}$ of the most likely used paths between each origin and each destination at each departure time. This can be done by applying variants of Dijkstra-based algorithms explicitly accounting for commonalities between paths in terms of shared links, as in Krishnakumari et al. (2019). However, as we are in this case considering link travel times, other alternatives like those proposed by Chabini (1998), dealing directly with time-dependent shortest paths, can be more appropriate. Nassir et al. (2014), Janmyr and Wadell (2018), use the penalization of overlapping in

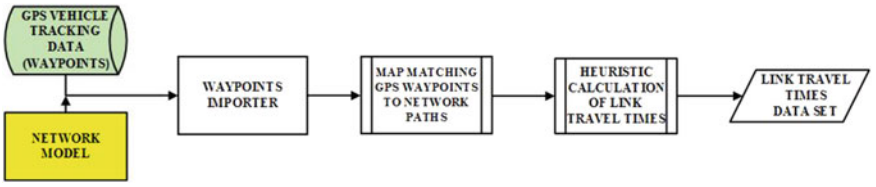

Fig. 7.8 Conceptual methodological approach to the process of importing waypoints into a network model and their use to estimate link travel times

terms of "commonality factors" proposed by Cascetta et al. (1996), Cascetta (2001) as a measure of similarity between alternatives. This allows overcoming the problems derived from the basic hypothesis of irrelevant alternatives with discrete-choice models reducing the systematic utility of paths, being this utility measured in terms of travel time, in proportion to its level of overlapping with other alternative paths. Such procedures can be additionally strengthened by applying the modification of the variant of Bovy et al. (2008) proposed by Janmyr and Wadell (2018). According to this modification, paths in K_{ijr} are denoted here as $k(i,j,r) \in K_{ijr}$ in order to explicitly show the dependence on (i,j,r). Let's assume that the sequence of links that compound a certain path $k(i,j,r)$ is $\Gamma_{k(i,j,r)} = \{e_1, \ldots, e_m\}$. Then, the proportion of paths choice for each path in the set K_{ijr} is calculated in terms of the following modified discrete logit-based choice model that uses the commonality factor (CF) for each OD pair and time period (Eqs. 7.160 and 7.161):

$$CF_{k(i.j.r)} = \frac{1}{\mu_{CF_k}} * \sum_{a \in \Gamma_{k(,i,j,r)}} \left(\frac{l_a}{L_{k(i,j,r)}} * \log \left(\sum_{h \in K_{ijr}} (\delta_{ahr} + 1) \right) \right) \tag{7.160}$$

$$P_{k(i,j,r)} = \frac{\exp[\mu_{P_k}(-\hat{tt}_{k(i,j,r)} - CF_{k(i,j,r)})]}{\sum_{h \in K_{ijr}} \exp[\mu_{P_k}(-\hat{tt}_{h(i,j,r)} - CF_{h(i,j,r)})]} \tag{7.161}$$

where $\delta_{ahr} = 1$ if path $h \in K_{ijr}$ uses link a at time r and 0 otherwise, l_a is the length of link a and $L_{k(i,j,r)}$ is the total length of path $k \in K_{ijr}$. In order to adapt magnitudes for the discrete-choice summation, μ_{P_k} and μ_{CF_k} are parameters fixed as in Eq. 7.162:

$$\mu_{P_k} = \mu_{CF_k} = \frac{1}{\text{mean}_{k \in K_{ijr}} \left(\hat{tt}_{k(i,j,r)} \right)} \tag{7.162}$$

These calculations provide the flow distribution for each path on the basis of observed path travel times, which are the summation of the observed time-dependent link travel times. That is, they consider the arrival time, $\hat{tt}_{at(k)}$, at each link a belonging to the path $k(i,j,r)$ (Eq. 7.163):

$$\hat{tt}_{k(i,j,r)} = \sum_{a \in \Gamma_{k(i,j,r)}} \hat{tt}_{at(k)} \tag{7.163}$$

Once $P_k = \left\{ P_{k(i,j,r)} \right\}$ is determined from the k shortest paths obtained from the estimated travel times, the estimated time-dependent assignment matrix $\overline{A} = \left[\overline{a}_{ijr}^{lt} \right]$ can be calculated with Eq. 7.164 and 7.165:

$$\overline{a}_{ijr}^{lt} = \sum_{k \in K_{ijr}} \delta_{k(i,j,r)}^{lt} * P_{k(i,j,r)} \quad \forall i,j,r,l,t \tag{7.164}$$

$$\delta_{k(i,j,r)}^{lt} = \begin{cases} 1 \text{ if path } k(i,j,r) \text{ uses link } l \text{ at time } t \\ 0 \text{ otherwise} \end{cases} \tag{7.165}$$

where $\delta_{k(i,j,r)}^{lt}$ is the estimated incidence indicator.

This is the estimated assignment matrix that can replace the calculated assignment matrix from DTA in an alternative formulation of DODME. Therefore, the relationship in Eq. 7.76 that the assignment matrix establishes between estimated link flows y_{lt} and estimated OD flows x_{ijr} can now be rewritten as Eq. 7.166:

$$y_{lt} = \sum_{i \in O} \sum_{j \in D} \sum_{r=1}^{t} \bar{a}_{ijr}^{lt} * x_{ijr} \tag{7.166}$$

If data collected from a sample of GPS-tracked vehicles is available and if it is possible to create a discrete time estimate of a seed OD-matrix from it, that is, the observed OD-matrix $X^0 = \left[x_{ijr}^0\right]$, this last matrix could be expanded to estimate the OD-matrix in terms of the scaling factors per origins, $\alpha_i, \forall i \in O$, and per destinations $\beta_j, \forall j \in D$, such that (Eq. 7.167):

$$x_{ijr} = \alpha_i * \beta_j * x_{ijr}^0 \tag{7.167}$$

It can be assumed, as in all previous formulations, that a reliable historical OD-matrix X^H is available. As already mentioned, this assumption would be questionable in long-term planning applications, as this matrix could be either largely outdated or simply not exist. However, its existence is a reasonable hypothesis in traffic management applications, where a surveillance system is already in operation and provides rich structural information (Ashok and Ben-Akiva 1993; Ben-Akiva et al. 2001; Djukic et al. 2018; Aimsun 2020). Once the existence of a historical OD-matrix accepted, the DODME problem can be reformulated in terms of the estimation of the scaling factors α_i, and β_j, in the following way (Eq. 7.168):

$$\begin{aligned} \underset{\alpha_i, \beta_j}{\text{Min}} \Bigg[& w \left(\sum_{i \in O} \sum_{j \in D} \sum_{r=1}^{t} \left(x_{ijr}^H - \alpha_i * \beta_j * x_{ijr}^0 \right)^2 \right) \\ & + \sum_{l \in \hat{L}} \sum_{t \in \mathcal{T}} \left(\hat{y}_{lt} - \sum_{i \in O} \sum_{j \in D} \sum_{r=1}^{t} \alpha_i * \beta_j * \bar{a}_{ijr}^{lt} * x_{ijr}^0 \right)^2 \Bigg] \\ & s.t.\ \alpha_i, \beta_j \geq LB \quad \forall i \in O, \forall j \in D \end{aligned} \tag{7.168}$$

The problem variables are multiplicative scaling factors for each origin α_i and each destination β_j, which significantly reduces the number of variables from $|I| * |J| * |T|$ to $|I| + |J|$. Moreover, the fact of using the scaling factors as variables aims at

preserving the structure of the seed OD-matrix, as gravity models do. Since the model is no longer quadratic and is bounded from below, other optimization procedures could be advisable. Ros-Roca et al. (2021a, b) report good results using the L-BFGS-B method (Morales and Nocedal 2011). It is a quasi-Newton method suitable for constrained nonlinear problems with a high number of variables, and it efficiently reduces the memory requirements and the computational burden.

Theoretically, the lower bound (LB) should be a non-negativity constraint for all the scaling factors α_i, β_j. However, from a practical point of view, $\alpha_i = 0$ or $\beta_j = 0$ implies that a positive OD flow of the seed OD-matrix from a certain origin or to certain destination would become null. Therefore, considering that the seed OD-matrix in Eq. 7.167 comes from reliable information on mobility, the scaling factors cannot be null and the lower bound should therefore be larger than zero.

If the quality of the observed seed matrix X^0 is questionable due to the conditions in which GPS data have been collected, (this could be the case for some commercial GPS data, as mentioned) but the historical matrix X^H is very reliable, both matrices could be fused to generate an improved seed matrix (Ros-Roca et al. 2021b).

7.7 Measuring the Quality of the OD Estimates

A critical question when estimating an OD is how the quality of the resulting estimated matrix can be assessed. This quality has been usually assessed in terms of the convergence of the objective function and the R^2 fit between measured and simulated traffic counts at links with counting stations. From the optimization point of view, these measures are a good selection because they can show explicitly that the used method works specifically for the purpose of minimizing the objective function designed as an OD-matrix estimation problem. Furthermore, it verifies that the estimated OD acceptably replicates the observed flows. However, despite R^2 being a good indicator of how the optimization problem is performing, it can produce misleading results. For example, it is possible that a high regression is achieved but the resulting estimated OD-matrix does not match the reality of the demand pattern and the internal mobility of the study area. Therefore, some other indicators that evaluate the mobility patterns in the OD-matrices are needed.

These indicators do not pay any attention to the quality of the results from a structural point of view. In other words, they do not distinguish whether the traffic OD patterns resulting from the adjustment approach exhibit an acceptable degree of structural similarity to the historical OD-matrix (when a reliable one is available), or whether the used approach provides a perturbed matrix that, even fitting the observed link flows, is structurally different. If this last is the case, it could be doubtful that such a structural change could be physically interpretable in terms of the underlying transportation system. Particularly when considering increases or decreases in the total number of trips between transportation zones that cannot be consistent with the socioeconomic attributes of the zone generating or attracting them. Looking at the link-path relationships visualized on the right-hand side of Fig. 7.3, it may happen

that the optimization process used to solve the DODME problem locally behaves as a retrogression model. This model could pull forth and back the OD flows in paths crossing the link with the counting station in order to fit measured and simulated flows as well as possible. This would just be a consequence of a numerical procedure, ignoring the underlying structure of the modeled reality.

A widely used proposal has been to resort to other goodness of fit indicators, like the Mean Square Error (MSE) and other similar ones (Hollander and Liu 2008). Other approaches consider alternative formulations of the objective function in terms of the distance function between the historical and the estimated OD-matrices. Classical distances between vectors can be applied to matrices by considering these matrices $X^H, X \in \mathcal{M}_n(\mathbb{R})$ as vectors of $X^H, X \in \mathbb{R}^{n \times n}$. Euclidean, Manhattan, and other vector distances can be used in the objective function of the OD-estimation problem aimed at minimizing the distance between matrices. However, these metrics, although comparing the OD-matrices cell by cell, do not have the ability to capture the differences and similarities of many aspects, such as their structure. Therefore, the spatio-temporal similarities of OD-matrices are not captured by these measures (Djukic 2014) and it seems clear that alternatives to these vector measures must be used. Djukic (2014) or Behara (2019) present a reference matrix M_R, which could be considered as a hypothetical ground truth matrix, X^{GT}, and two additional matrices M_1 and M_2 generated by perturbations of that reference matrices, such that they clearly have different structures but are indistinguishable in terms of measures like MSE or similar. The example in Fig. 7.9. illustrates this situation. Let us consider the three matrices,M_R, M_1 and M_2, the reference and perturbed matrices, respectively, generated following the guidelines of Djukic (2014).

Comparing M_R, M_1 and M_2 in terms of MSE, the results are that MSE (M_R, M_1) = MSE (M_R, M_2) = 16. Therefore, MSE does not help to discriminate which of the two matrices M_1 and M_2 is closer to M_R.

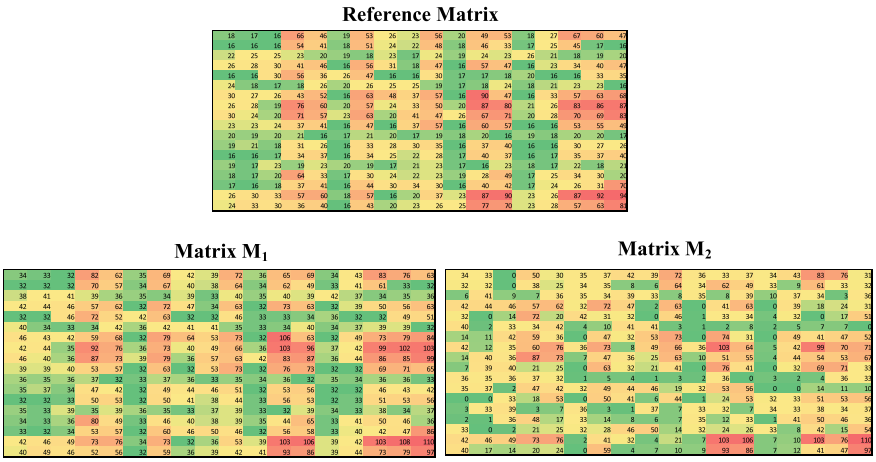

Fig. 7.9 Comparing matrices with the same MSE and different structures

Djukic (2014) proposes a measure of structural similarity based on the *Image Quality Assessment* process for comparing two different images (Wang et al. 2004). This measure is the Structural SIMilarity index (SSIM) for a matrix of pixels, that is, the product of three different comparison components: luminance, contrast, and structure. Luminance corresponds to the intensity of illumination, which is indeed the mean of the different pixels in a sub-matrix. Contrast is the squared average between pixels once the luminance is removed, thus making it the standard deviation. Finally, the structure is compared by using the covariance between the two matrices. These three factors are firstly transformed with the aim of adjusting them to the interval [0, 1], where 1 means perfect match and 0 means no match. SSIM is therefore a similarity measure that is independent of the magnitude of the values in the matrix. Equation 7.169 gives the formula summarizing this explanation:

$$SSIM\,(x, y) = l(x, y)^{\alpha} * c(x, y)^{\beta} * s(x, y)^{\gamma} \tag{7.169}$$

where luminance, contrast, and structure are, respectively, defined by Eqs. 7.170–7.172:

$$l(x, y) = \frac{2{*}\mu_x * \mu_y + C_1}{\mu_x^2 + u_y^2 + C_1} \tag{7.170}$$

$$c(x, y) = \frac{2{*}\sigma_x * \sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{7.171}$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x \sigma_y + C_3} \tag{7.172}$$

and μ_x, σ_x, μ_y, σ_y, σ_{xy} are the mean, standard deviation, and covariance of the vectors x and y, respectively. C_1, C_2, C_3 are stability constants aimed at avoiding numerical problems and are typically set to $C_1 = C_2 = 2 * C_3 = 1$. For their part, α, β, γ are weighting coefficients typically set to 1 (Wang et al. 2004). In image comparison, because pixel proximity is crucial in image pattern recognition, Wang et al. (2004) propose to first generate sliding submatrices of dimension N entirely covering the image, then compute the SSIM index for each of them and, finally, calculate the MSSIM as the mean of the SSIM of all submatrices of dimension N. Djukic (2014) assimilates the OD-matrix to an image whose pixels would be the OD cells and explores various alternatives for generating these sliding windows in terms of proximities. Behara (2019) and Behara et al. (2020) propose a procedure to generate them based on the geographical structure of the area spanned by the transport system object of study. Ros-Roca et al. (2020) propose to use rectangular sliding windows as submatrices corresponding to either rows or columns in the OD-matrix. In any case, SSIM will capture the similarity between these distributions by considering the mean, the variance, and the structure of departing and arriving distributions, all

of which correspond to the structural property of the trip patterns described by the OD-matrix.

Furthermore, let us assume that the number of generated submatrices is N_s that a and b are, respectively, the corresponding windows of the matrices A and B to compare and that $SSIM(a,b)$ is their similarity value. Then, if MSSIM is $SSIM\,(a, b)$ averaged over N_s sliding windows, a key question arises. Particularly, whether all windows have the same weight or whether their role in the total demand requires that they have different weights. In the case of OD-matrices, it is obvious that not all origins or destinations are equivalent in a transport network. Therefore, a weighted MSSIM as in Wang and Simoncelli (2008) prioritizes those origins and destinations with more impact on the network. This proposed weighting average is defined as in Eq. 7.173:

$$MSSIM\,(A, B) = \frac{\sum_{i=1}^{N_s} W\,(a_i, b_i) * SSIM\,(a_i, b_i)}{\sum_{i=1}^{N_s} W\,(a_i, b_i)} \tag{7.173}$$

where a_i, b_i are, respectively, the $i - th$ windows of A and B, while the weight $w(a_i, b_i)$ is given by Eq. 7.174:

$$w(a_i, b_i) = \log\left[\left(1 + \frac{\sigma_{a_i}^2}{C_2}\right) * \left(1 + \frac{\sigma_{b_i}^2}{C_2}\right)\right] \tag{7.174}$$

The weighting factors for the sliding windows, in the case of OD-matrices, account for variances of the selected windows that, given how they are defined, represent the variance of trips from an origin to all destinations or from all origins to one destination. The use of MSSIM in addition to the conventional performance indicators has demonstrated that the usual R^2 goodness of fit between observed and simulated links flows must be carefully complemented (e.g., Djukic 2014; Behara et al. 2020; Ros-Roca et al. 2020, 2021b). Particularly, it must be complemented with a MSSIM analysis in order to check the structural quality of the estimated OD-matrix X when an acceptable historical X^H that conveys reliable structural information on the OD patterns is available.

Comparing again M_R, M_1, and M_2 in terms of MSSIM, the results are, MSSIM $(M_R, M_1) = 0.914882$ and MSSIM $(M_R, M_2) = 0.510276$, which clearly shows that M_2 is structurally different from M_R.

The relevance of this structural similarity measure (Behara et al. 2020; Behara et al. 2021) led to explicitly include it in the objective function of the mathematical model for DODME, reformulating it as follows (Eqs. 7.175–7.179):

$$\begin{aligned}\text{Min } Z(X) = \frac{1}{2} * &\left[\left(c_1 + \left(Y - \hat{Y}\right)^T * \left(Y - \hat{Y}\right)\right)\right] \\ &* \left[\left(c_2 + f\left(s, \hat{s}\right)\right)^T * \left(c_2 + f\left(s, \hat{s}\right)\right)\right]\end{aligned} \tag{7.175}$$

$$Y = A * X \tag{7.176}$$

$$s = Q * X \tag{7.177}$$

$$f\left(s, \hat{s}\right) = \frac{1 - \rho\left(s, \hat{s}\right)}{2} \tag{7.178}$$

$$\rho\left(s, \hat{s}\right) = \frac{\left(\hat{s} - \mu_{\hat{s}}\right)^T * (s - \mu_s)}{\sqrt{\left(\hat{s} - \mu_{\hat{s}}\right)^T * \left(\hat{s} - \mu_{\hat{s}}\right)} * \sqrt{(s - \mu_s)^T * (s - \mu_s)}} \tag{7.179}$$

where A is the assignment matrix, Y and $\widehat{Y}$ are, respectively, the estimated and the observed link flows at links with counting stations and s and $\hat{s}$ denote the observed and simulated flows at subpaths detected by Bluetooth (or Wi-Fi) antennas. For their part, Q is the corresponding subpath assignment matrix, while c_1 and c_2 are stabilizing constants. The algorithmic approach assumes that t A and Q are locally constant.

7.8 Concluding Remarks

The main objective of this Chapter has been to highlight the role of two key components of the engine of most traffic management and information systems. First, a *Dynamic Traffic Model*, usually a DTA or DUE, which is quite frequently supported by a *Network Loading* process based on a mesoscopic traffic simulation approach. Second, a *Dynamic Origin–Destination Matrix Estimator* (DODME) that suitably models the time-dependent mobility patterns. The main goals of these components are the estimation of the traffic state in the managed road network and its short-term prediction, accounting for impacts of external events like traffic incidents that would change the operational conditions in the network. Travel times are one of the main outputs describing these states for both managers and travelers in the network. Figure 7.10 conceptually summarizes a generic architecture of a traffic management and information system highlighting the role of these two key components and their interactions since, as it has been discussed in the chapter, the main input to a DTA or DUE is a Dynamic OD, and DOME procedures usually rely on information generated by a DTA.

This chapter has also provided an overview of the main approaches to both models, DODME and DTA/DUE, and their relationships. The role of one critical component, the dynamic assignment matrix, has been extensively discussed. This matrix describes the structure of the dynamic of the use of the links of the network by the traffic flows in the paths from origin to destinations. The possibility of exploiting the huge amount of traffic data supplied by ICT applications, which allows empirically reproducing the assignment matrix from data instead of from models in the direction

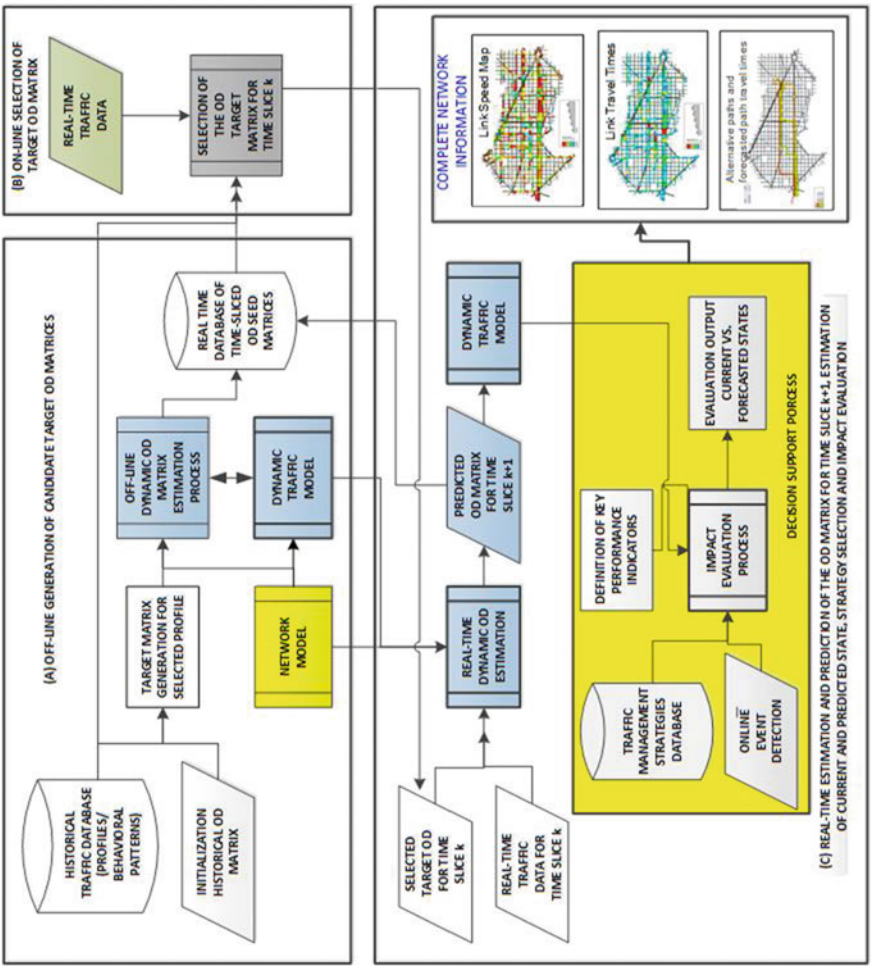

Fig. 7.10 Generic architecture of a traffic management and information system highlighting the role of the two main engine components: a DOMME and a DTA/DUE

of the data-driven modeling, has also been addressed. This trend is intellectually very appealing and, in fact, it is currently leaving the Academia realm to enter the domain of real-life applications, as it can be deduced from the last versions of some professional software platforms.

Acknowledgements The authors are very grateful to Professor Guido Gentile and Mr. Lorenzo Meschini, respectively, of SISTeMA S.R.L. and PTV Group, for supplying information OPTIMA. We also express our gratitude to Mr. Josep M. Aymamí and Dr. Emmanuel Bert (Aimsun SLU), for the information regarding Aimsun Next and Aimsun Live.

References

Aimsun SLU (2020) *Aimsun Next*. https://www.aimsun.com/es/aimsun-next/. Accessed 5 May 2021

Alexander L, Jiang S, Murga M, González MC (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transp Res Part C: Emerg Technol 58(B):240–250

Allström A, Barceló J, Ekström J, Grumert E, Gundlegård D, Rydergren C (2017) Traffic management for smart cities. In: Angelakis V, Tragos E, Pöhls HC, Kapovits A, Bassi A (ed) Designing, developing and facilitating smart cities. Springer, Switzerland. ISBN 978-3-319-44922-7

Altman A, Gondzio J (1999) Regularized symmetric indefinite systems in interior point methods for linear and quadratic optimization. Optim Meth Softw 11(1–4) (interior Point Methods)

Antoniou C (2004) On-line calibration for dynamic traffic assignment. PhD dissertation, Mass. Inst. Technol, Cambridge, MA

Antoniou C, Ben-Akiva ME, Koutsopoulos HN (2007). Nonlinear Kalman filtering algorithms for on-line calibration of dynamic traffic assignment models. IEEE Trans Intell Transp Syst 8(4):661–670

Antoniou C, Azevedo CL, Lu L, Pereira F, Ben-Akiva M (2015) W-SPSA in practice: approximation of weight matrices and calibration of traffic simulation models. Transp Res Part C: Emerg Technol 59:129–146

Antoniou C, Barceló J, Breen M, Bullejos M, Casas J, Cipriani E, Ciuffo B, Djukic T, Hoogendoorn S, Marzano V, Montero L, Nigro M, Perarnau J, Punzo V, Toledo T, van Lint H (2016) Towards a generic benchmarking platform for origin-destination flows estimation/updating algorithms: design, demonstration and validation. Transp Res Part C: Emerg Technol 66:79–98

Ashok K, Ben-Akiva M (1993) Dynamic origin-destination matrix estimation and prediction for real-time traffic management systems. In: Daganzo C (ed) Transportation and traffic theory. Elsevier Science Publishing Company. Inc. Proceedings of the 12th ISTTT.

Ashok K, Ben-Akiva ME (2002) Estimation and prediction of time-dependent origin-destination flows with a stochatic mapping to path flows and link flows. Transp Sci 36(2):184–198

Barceló J, Codina E, Casas J, Ferrer JL, García D (2004) Microscopic traffic simulation: a tool for the design, analysis and evaluation of intelligent transport systems. J Intell Rob Syst 41:173–203

Barceló J, Delgado M, Funes G, García D, Torday A (2007) An on-line approach based on microscopic traffic simulation to assist real time traffic management. In: 14th World congress on intelligent transport systems, 2007. Beijing

Barceló J (2010a) Models, traffic models, simulation and traffic simulation. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, Switzerland. ISBN 978-1-4419-6142-6

Barceló J, Gilliéron F, Linares MP, Serch O, Montero L (2012) Exploring link covering and node covering formulations of detection layout problem. Transp Res Records: J Transp Res Board 2308:17–26

Barceló J, Montero L, Bullejos M, Serch O, Carmona C (2013) A Kalman filter approach for the estimation of time dependent OD matrices exploiting bluetooth traffic data collection. JITS J Intell Transp Syst: Technol, Plan Oper 17(2):1–19

Bassolas A, Ramasco JJ, Herranz R, Cantú-Ros OG (2019) Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transp Res Part A: Policy Pract 121:56–74

Bazaraa MS, Sherali HD, Shetty CM (1993) Nonlinear programming. theory and algorithms. Wiley, USA

Behara K (2019) Origin-Destination matrix estimation using big traffic data: a structural perspective. PhD Thesis, School of Civil Engineering and Built Environment Science and Engineering Faculty Queensland University of Technology

Behara KN, Bhaskar A, Chung E (2020) Geographical window based structural similarity index for origin-destination matrices comparison. J Intell Transp Syst 1–22

Behara KNS, Bhaskar A, Chung E (2021) A novel methodology to assimilate sub-path flows in bi-level OD matrix estimation process. IEEE Trans Intell Transp Syst (in press)

Bell MGH (1991) The estimation of origin-destination matrices by constrained generalized least squares. Transp Res B: Methodol 25B:115–125

Bell MGH, Iida Y (1997) Transportation network analysis. Wiley, USA

Bellei G, Gentile G, Papola N (2005) A within-day dynamic traffic assignment model for urban road networks. Transp Res Part B: Methodol 39:1–29

Ben-Akiva M, Bierlaire M, Bottom J, Koutsopoulos HN, Mishalani RG (1997) Development of a route guidance generation system for real-time application. In: Proceedings of the 8th IFAC symposium on transportation systems, 1997. Chania, Crete

Ben-Akiva M, Bierlaire M (1999) Discrete choice models and their application to short term travel decisions. In: Hall RW (ed) Handbook of transportation science. Springer, Switzerland. ISBN: 0-7923-8587-X

Ben-Akiva M, Bierlaire M, Burton D, Koutsopoulos HN, Mishalani R (2001) Network state estimation and prediction for real-time traffic management. Netw Spatial Econ 1:293–318

Ben-Akiva M, Bierlaire M, Koutsopoulos HN, Mishalani R (2002) Real-time simulation of traffic demand-supply interactions within DynaMIT. In: Gendreau M, Marcotte P (ed) Transportation and network analysis: current trends. Miscellanea in honour of Michael Florian. Kluwer Academic Publishers, Boston/Dordrecht/London

Ben-Akiva M, Koutsopoulos HN, Antoniou C, Balakrishna R (2010) Traffic simulation with DynaMIT. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, Switzerland. ISBN 978-1-4419-6142-6

Bliemer MCJ, Raadsen MPH, Brederode LJN, Bell MGH, Wismans LJJ, Smith MJ (2017) Genetics of traffic assignment models for strategic transport planning. Transp Rev 37(1):56–78

Bovy P, Bekhor S, Prato C (2008) The factor of revisited path size. Transp Res Board 2076:132–140

Boyce D, Lee DH, Ran B (2001) Analytical models of the dynamic traffic assignment problem. Netw Spatial Econ 1:377–390

Bullejos M, Barceló J, Montero L (2014) A DUE based bi-level optimization approach for the estimation of time sliced OD matrices. International symposium of transport simulation, 2014. France, pp 1–19

Burghout W (2004) Hybrid microscopic-mesoscopic traffic simulation. Doctoral Thesis, Royal Institute of Technology, Stockholm, Sweden

Burghout W, Koutsopoulos H, Andréasson I (2005) Hybrid mesoscopic-microscopic traffic simulation. In: Proceedings of the 83rd TRB annual meeting, 2005. Washington, DC.

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using mobile phone location data. IEEE Pervasive Comput 10

Calabrese F, Diao M, Di Lorenzo G, Ferreira J Jr, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res Part C: Emerg Technol 26:301–313

Cantelmo G, Cipriani E, Gemma A, Nigro M (2014) An adaptive bi-level gradient procedure for the estimation of dynamic traffic demand. IEEE Trans Intell Transp Syst 15(3):1348–1361
Carrese S, Cipriani E, Mannini L, Nigro M (2017) Dynamic demand estimation and prediction for traffic urban networks adopting new data sources. Transp Res Part C: Emerg Technol 81:83–98
Carey M, Ge YE (2012) Comparison of methods for path flow reassignment for dynamic user equilibrium. Netw Spatial Econ 12:337–376
Casas J, Ferrer J, García D, Perarnau J, Torday A (2010) Traffic simulation with Aimsun. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, Switzerland. ISBN 978-1-4419-6142-6
Cascetta E (1984) Estimation of origin-destination matrices from traffic counts and survey data: a generalised least squares estimator. Transp Res Part B: Methodol 18(B):289–299
Cascetta E, Nuzzolo A, Russo F, Vitetta A (1996) A modified logit route choice model overcoming path overlapping problems. In: Proceedings of the 13th international symposium on the theory of road traffic flow, 1996. France
Cascetta E (2001) Transportation systems engineering theory and methods. Springer, Switzerland
Cascetta E, Papola A, Marzano V, Simonelli F, Vitiello I (2013) Quasi-dynamic estimation of OD flows from traffic counts: formulation, statistical validation and performance analysis on real data. Transp Res Part B: Methodol 55:171–187
Castillo E, Conejo AJ, Menéndez JM, Jiménez P (2008) The observability problem in traffic network models. Comput-Aided Civil Infrastruct Eng 23:208–222
Chabini I (1998) Discrete dynamic shortest path problems in transportation applications: complexity and algorithms with optimal run time. Transp Res Records 1997
Chiu YC, Bottom J, Mahut M, Paz A, Balakrishna R, Waller T, Hicks J (2011) Dynamic traffic assignment: a primer. Transp Res E-Circular (E-C153)
Cipriani E, Florian M, Mahut M, Nigro M (2011) A gradient approximation approach for adjusting temporal origin-destination matrices. Transp Res Part C: Emerg Technol 19(2):270–282
Codina E, Barceló J (2004) Adjustment of O-D matrices from observed volumes: an algorithmic approach based on conjugate gradients. Eur J Oper Res 155:535–557
Codina E, Montero L (2006) Approximation of the steepest descent direction for the O-D matrix adjustment problem. Ann Oper Res 114:329–362
Çolak S, Lima A, González MC (2015) Understanding congested travel in urban areas. Nat Commun 7:10793
Daganzo CF (1994) The cell-transmission model: a simple dynamic representation of highway traffic. Transp Res Part B: Methodol 28(4):269–287
Daganzo CF (1995) The cell transmission model part II: network traffic. Transp Res Part b: Methodol 29:79–93
Daganzo CF (1995) A finite difference approximation of the kinematic wave model of traffic flow. Transp Res Part B: Methodol 29(4):261–276
Del Castillo JM, Benitez FG (1995) On the functional form of the speed-density relationship I: general theory. Transp Res Part B: Methodol 29(5):373–389
Djukic T, van Lint JWC, Hoogendoorn SP (2012) Application of principal component analysis to predict dynamic origin-destination matrices. Transp Res Record: J Transp Res Board 2283(1):81–89
Djukic T (2014) Dynamic OD demand estimation and prediction for dynamic traffic management. PhD Thesis, TU Delft
Djukic T, Breen M, Masip D, Perarnau J, Budin J, Casas J (2017) Marginal effects evaluation with respect to changes in OD demand for dynamic OD demand estimation. In: Proceedings of the international conference on intelligent transport systems in theory and practice. TUM'17, 2017, Munich
Djukic T, Masip D, Breen M, Perarnau J, Casas J (2018) Heuristic-based framework for dynamic OD demand estimation in the congested networks. Transportation research board 97th annual meeting transportation research board, 18, 03283

Djukic T, Masip D, Breen M, Casas J (2019) Efficient metamodel framework for nonlinear OD matrix estimation problem. Transportation research board 98th annual meeting transportation research board, 19, 05188

Ehlert A, Bell MGH, Grosso S (2006) The optimisation of traffic count locations in road networks. Transp Res Part B: Methodol 40:460–479

Eisenman SM, List GF (2004). Using probe data to estimate OD matrices. In: Proceedings of the 7th international IEEE conference on intelligent transportation systems (ITSC '04), October 2004. Washington, DC, USA, pp 291–296

Fei X, Eisenman SM, Mahmassani H (2007) Sensor coverage and location for real-time traffic prediction in large-scale networks. In: 86th annual meeting of the transportation research board, January 2007. Washington, DC, USA

Filkov V, Skiena S (2004) Integrating microarray data by consensus clustering. Int J Artif Intell Tools 13:863–880

Florian M, Chen Y (1995) A coordinate descent method for the bi-level OD matrix adjustment problem. Int Trans Oper Res 2(2):165–175

Florian M, Hearn D (1995) Network equilibrium models and algorithms. In: Ball MO et al (ed) Handbooks in operations research and management science, 8. Elsevier Science B.V., The Netherlands

Florian M, Mahut M, Tremblay N (2001) A hybrid optimization-mesoscopic simulation dynamic traffic assignment model. In: Proceedings of the 2001 IEEE intelligent transport systems conference, 2001. Oakland, pp 118–123

Florian M, Mahut M, Tremblay N (2002) Application of a simulation-based dynamic traffic assignment model. In: Kitamura R, Kuwahara M (eds) International symposium on transport simulation, 2002, Yokohama (also in: Simulation approaches in transportation analysis, 2005. Kluwer, US

Florian M, Mahut M, Tremblay N (2008) Application of a simulation-based dynamic traffic assignment model. Eur J Oper Res 189(3):1381–1392

Frederix R, Viti F, Corthout R, Tampère C (2011) New gradient approximation method for dynamic origin-destination matrix estimation on congested networks. Transp Res Record: J Transp Res Board 2263(1):19–25

Frederix R, Viti F, Tampère C (2013) Dynamic origin-destination estimation in congested networks: theoretical findings and implications in practice. Transportmetrica a: Transport Science 9(6):494–513

Friesz TL, Bernstein D, Smith TE, Tobin RL, Wie BW (1993) A variational inequality formulation of the dynamic network user equilibrium problem. Oper Res 41(1):179–191

Gelb A (1974) Applied optimal estimation. MIT Press, Cambridge, MA

Gentile G, Meschini L, Papola N (2007) Spillback congestion in dynamic traffic assignment: a macroscopic flow model with time-varying bottlenecks. Transp Res Part B: Methodol 41:1114–1138

Gentile G (2010) The general link transmission model for dynamic network loading and a comparison with the DUE algorithm. In: Immers LGH, Tampere CMJ, Viti F (eds) New developments in transport planning: advances in dynamic traffic assignment. Transport Economics, Management and Policy Series, Edward Elgar Publishing, MA, USA

Gentile G (2015) Using the general link transmission model in a dynamic traffic assignment to simulate congestion on urban networks. Transp Res Procedia 5:66–81

González MC, Hidalgo A, Barabasi A-L (2008) Understanding human mobility patterns. Nature 453(7196):779–782

Greenshields BD (1934) A study of traffic capacity. In: Proceedings of the fourteenth annual meeting of the highway research board, held at Washington, D.C. December 6–7, 1934, Part I, 14, 448–477

Gundlegård D, Rydergren C, Barcelo J, Dokoohaki N, Görnerup O, Hess A (2015) Travel demand analysis with differentially private releases. D4D challenge Senegal 2014, Netmob 2015, November 2015, MIT, Boston

Han K, Eve G, Friesz TL (2019) Computing dynamic user equilibria on large-scale networks with software implementation. Netw Spatial Econ 19:869–902

Hart PE, Nilsson NJ, Raphael B (1968) A formal basis for the heuristic determination of minimum cost paths. IEEE Trans Syst Sci Cybern 4:100–107

Hegyi A, Bellemans T, De Schutter B (2009) Freeway traffic management and control. In: Meyers RA (ed) Encyclopedia of complexity and systems science. Springer, New York. ISBN 978-0-38730440-3

Hollander Y, Liu R (2008) The principles of calibrating traffic microsimulation models. Transportation 35:347–362

Hu S, Madanat SM, Krogmeier J, Peeta S (2001) Estimation of dynamic assignment matrices and OD demands using adaptative Kalman filtering. Intell Transp Syst J 6:281–300

Iqbal MS, Choudhury CF, Wang P, González MC (2014) Development of origin–destination matrices using mobile phone call data. Transp Res Part C: Emerg Technol 40:63–74

Janmyr J, Wadell D (2018) Analysis of vehicle route choice during incidents. MSc Thesis, University of Linkoping, Department of Science and Technology

Janson BN (1991) Dynamic traffic assignment for urban road networks. Transp Res Part B: Methodol 25(2):143–161

Jayakrisham R, Mahmassani HS, Yu TY (1994) An evaluation tool for advanced traffic information and management systems in urban networks. Transp Res Part C: Emerg Technol 2C(3):129–147

Jianga S, Yanga Y, Gupta S, Veneziano D, Athavale S, González MC (2016) The TimeGeo modeling framework for urban mobility without travel surveys. Proc Natl Acad Sci USA 113:37

Jeihani M (2007) A review of dynamic traffic assignment computer packages. J Transp Res Forum 46:35–46

Jolliffe IT (2002) Principal component analysis. Springer, Switzerland

Kalman RE (1960) A new approach to linear filtering and prediction problems. J Basic Eng (ASME) 82D:35–45

Kostic B, Gentile G, Antoniou C (2017a) Techniques for improving the effectiveness of the SPSA algorithm in dynamic demand calibration. In: 5th IEEE international conference on models and technologies for intelligent transportation systems, MT-ITS 2017. Napoli, Italy

Kostic B, Annunziata A, Gentile G, Meschini L (2017b) A sequential approach to time-dependent demand calibration: application, validation and practical implications for large-scale networks. In: 5th IEEE international conference on models and technologies for intelligent transportation systems, MT-ITS 2017. Napoli, Italy

Krishnakumari P, van Lint H, Djukic T, Cats O (2019) A data driven method for OD matrix estimation. Transp Res Part C: Emerg Technol 113:38–56

Kubicka M, Mounier H, Niculescu SI, Cela A (2018) Comparative study and application-oriented classification of vehicular map-matching methods. IEEE Intell Transp Syst Mag 10(2):150–166

Larsson T, Lundgren JT, Peterson A (2010) Allocation of link flow detectors for origin-destination matrix estimation: a comparative study. Comput-Aided Civil Infrastruct Eng 25:116–131

Leonard DP, Gower P, Taylor N (1989) CONTRAM. Structure of the model, transport and road research laboratory. Research Report 178, Department of Transport, Crowthorne

Lighthill M, Whitham G (1955) On kinematic waves. II. A theory of traffic flow on long crowded roads. In: Proceedings of the royal society of London. Series A, mathematical and physical sciences, vol 229, no 1178, pp 317–345

Lin P, Chang G (2007) A generalized model and solution algorithm for estimation of the dynamic freeway origin-destination matrix. Transp Res Part b: Methodol 41:554–572

Lo HK, Szeto WY (2002) A cell-based variational inequality formulation of the dynamic user optimal assignment problem. Transp Res Part b: Methodol 36:421–443

López C, Krishnakumari P, Leclercq L, Chiabaut N, van Lint H (2017) Spatio-temporal partitioning of the transportation network using travel time data. Transp Res Record: J Transp Res Board 2623(1):98–107

López C, Leclercq L, Krishnakumari P, Chiabaut N, van Lint H (2017) Revealing the day-to-day regularity of urban congestion patterns with 3D speed maps. Sci Rep 7:14029

Lu L, Xu Y, Antoniou C, Ben-Akiva M (2015) An enhanced SPSA algorithm for the calibration of dynamic traffic assignment models. Transp Res Part C: Emerg Technol 51:149–166

Luenberger DG, Ye Y (2008) Linear and nonlinear programming. Springer, Switzerland
Lundgren JT, Peterson A (2008) A heuristic for the bilevel origin–destination-matrix estimation problem. Transp Res Part b: Methodol 42:339–354
Mahmassani HS, Hu TY, Peeta S, Ziliaskopoulos A (1994) Development and testing of dynamic traffic assignment and simulation procedures for ATIS/ATMS applications. Technical Report DTFH61–90-R00074-FG, Center for Transportation research, The University of Texas at Austin
Mahmassani H (2001) Dynamic network traffic assignment and simulation methodology for advanced system management applications. Netw Spatial Econ 1:267–292
Mahmassani H, Hong Z, Xu X, Mittal A, Yelchuru B, Kamalanathsharma R (2017) Analysis, modeling, and simulation (AMS) testbed development and evaluation to support dynamic mobility applications (DMA) and active transportation and demand management (ATDM) programs evaluation report for the Chicago testbed. Final Report—April 2017, FHWA-JPO-16–387
Mahut M (1999) Behavioural car following models. Report CRT-99–31. Centre for Research on Transportation. University of Montreal. Montreal, Canada
Mahut M (2001) Discrete flow model for dynamic network loading. PhD Thesis, Département d'informatique et de recherhe opérationelle, Université de Montréal. Published by the Center for Research on Transportation (CRT), University of Montreal
Mahut M, Florian M, Tremblay N (2003a) Space-time queues and dynamic traffic assignment: a model, algorithm and applications. Transportation research board, 82nd annual meeting, 2002. Washington DC, USA
Mahut M, Florian M, Tremblay N (2003b) Traffic simulation and dynamic assignment for off-line applications. In: 10th world congress on intelligent transportation systems, 2003. Madrid, Spain
Mahut M, Florian M, Tremblay N, Campbell M, Patman D, McDaniel ZK (2004) Calibration and application of a simulation based dynamic traffic assignment model. Transp Res Record: J Transp Res Board 1876:101–111
Mahut M, Florian M (2010) Traffic simulation with dynameq. In: Barceló J (ed) Fundamentals of traffic simulation. Springer, Switzerland. ISBN 978-1-4419-6142-6
Marchal F, Hackney JK, Axhausen KW (2004) Efficient map-matching of large GPS data sets—tests on a speed monitoring experiment in Zurich. Arbeitsbericht Verkehrs und Raumplanung. Technical report, UNAM, p 244
May AD, Keller HEM (1967) Non-integer car-following models. Highway Res Rec 199:19–32
Meschini L (2017) Modern traffic control centres and traffic management systems. In: Fusco G (ed) Intelligent transport systems (ITS): past, present and future directions. NOVA Science Publishers. ISBN 978-1-53611-815-5
Millard-Ball A, Hampshire RC, Weinberger RR (2019) Map-matching poor-quality GPS data in urban environments: the pgMapMatch package. Transp Plan Technol 42(6):539–553
Mitra A, Attanasi A, Meschini L, Gentile G (2020) Methodology for O-D matrix estimation using the revealed paths of floating car data on large-scale networks. In: IET intelligent transport systems special issue: the scientific seminar of the Italian association of transport academicians 2019 (SIDT 2019), vol 14, pp 1704–1711
Mo B, Li R, Dai J (2020) Estimating dynamic origin–destination demand: a hybrid framework using license plate recognition data. Comput Aided Civil Infrastruct Eng 35(7):1–19
Montero L, Ros-Roca X, Herranz R, Barceló J (2019) Fusing mobile phone data with other data sources to generate input OD matrices for transport models. Transp Res Procedia 37:417–424
Morales JL, Nocedal J (2011) Remark on algorithm 778: L-BFGS-B: fortran subroutines for large-scale bound constrained optimization. ACM Trans Math Softw 38(1):7
Nanthawichit C, Nakatsuji T, Suzuki H (2003) Application of probe vehicle data for real-time traffic state estimation and short-term travel time prediction in a freeway. Transp Res Record: J Transp Res Board 1855(1):49–59
Nassir N, Ziebarth J, Sall E, Zorn L (2014) Choice set generation algorithm suitable for measuring route choice accessibility. Transp Res Record 2430(1):170–171
Newell GF (2002) A simplified car-following theory: a lower order model. Transp Res Part B: Methodol 36B(3):195–205

Nigro M, Abdelfatah A, Cipriani E, Colombaroni C, Fusco G, Gemma A (2018) Dynamic O-D demand estimation: application of SPSA AD-PI method in conjunction with different assignment strategies. J Adv Transp 2018:1–18

OpenLR (2020. OpenLR White Paper. Version 1.5, revision 2. https://ww.openlr-association.com/fileadmin/user_upload/openlr-whitepaper_v1.5.pdf

Ortúzar JD, Willumsen LG (2011) Modelling transport. Wiley, USA

Osorio C, Linsen C (2015) A computationally efficient simulation-based optimization algorithm for large-scale urban transportation problems. Transp Sci 49(3):623–636

Peeta S, Mahmassani HS (1995) System optimal and user equilibrium time-dependent traffic assignment in congested networks. Ann Oper Res 60:81–113

Peeta S, Ziliaskopoulos AK (2001) Foundations of dynamic traffic assignment: the past, the present and the future. Netw Spatial Econ 1:233–265

Pereira FC, Costa H, Pereira NM (2009) An off-line map-matching algorithm for incomplete map databases. Eur Transp Res Rev 1:107–124

PTV AG Visum (2020) PTV Visum 2020—user's manual. PTV Group, Karlsruhe, Germany

Quddus MA, Ochieng WY, Noland RB (2007) Current map-matching algorithms for transport applications: state-of-the art and future research directions. Transp Res Part C: Emerg Technol 15:312–328

Rahmani M, Koutsopoulos HN (2013) Path inference from sparse floating car data for urban networks. Transp Res Part C: Emerg Technol 30:41–54

Ran B, Boyce D (1996) Modeling dynamic transportation networks. Springer, Switzerland

Richards PI (1956) Shockwaves on the highway. Oper Res 4(1):42–51

Ros-Roca X, Montero L, Barceló J (2017) Notes on using simulation-optimization techniques in traffic simulation. Transp Res Procedia 27:881–888

Ros-Roca X, Montero L, Schneck A, Barceló J (2018) Investigating the performance of SPSA in simulation-optimization approaches to transportation problems. Transp Res Procedia 34:83–90

Ros-Roca X, Montero L, Barceló J (2020) Investigating the quality of Spiess-like and SPSA approaches for dynamic OD matrix estimation. Transportmetrica 17(3):235–257

Ros-Roca X, Montero L, Barceló J, Nökel K (2021a) Dynamic origin-destination matrix estimation with ICT traffic measurements using SPSA. Accepted for presentation at *MTITS2021*, to appear in Scopus-indexed IEEE Xplore Digital Library conference proceedings (conference number 49943)

Ros-Roca X, Montero L, Barceló J, Nökel K, Gentil G (2021b) A practical approach to assignment-free dynamic origin-destination matrix estimation problem. Accepted for publication in Transportation Research C: Emerging Technologies

Sadegh P, Spall JC (1998) Optimal random perturbations for stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans Autom Control 43(10):1480–1484

Sbayti H, Lu C, Mahmassani HS (2007) Efficient implementations of the method of successive averages in simulation-based DTA models for large-scale network applications. TRB 2007 Annual Meeting, 2007. Washington DC, USA

Schuessler N, Axhausen KW (2009) Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique (MHT). Working paper 568 Institute for Transport Planning and Systems, Swiss Federal Institute of Technology Zürich

Smith MJ (1993) A new dynamic traffic model and the existence and calculation of dynamic user equilibria on congested capacity-constrained road networks. Transp Res Part B: Methodol 27:49–63

Smith MJ, Wisten MB (1995) A continuous day-to-day traffic assignment model and the existence of a continuous dynamic user equilibrium. Ann Oper Res 60(1):59–79

Spall JC (1992) Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans Autom Control 37(3):332–341

Spall JC (1998) An overview of the simultaneous perturbation method for efficient optimization. Johns Hopkins APL Tech Digest (applied Physics Laboratory) 19(4):482–492

Spall JC (2003) Introduction to stochastic search and optimization: estimation, simulation, and control. Wiley-Interscience
Spiess H (1990) A gradient approach for the OD matrix adjustment problem. Publication No. 693, Centre de Recherche sur les Transports, Université de Montréal
Szeto WY, Wong SC (2012) Dynamic traffic assignment: model classifications and recent advances in travel choice principles. Open Eng 2(1):1–18
Szeto WY, Lo HK (2005) Dynamic traffic assignment: review and future. Inf Technol 5:85–100
Szeto WY, Lo HK (2004) A cell-based simultaneous route and departure time choice model with elastic demand. Transp Res Part B: Methodol 38:593–612
Tympakianaki A, Koutsopoulos HN, Jenelius E (2015) C-SPSA: cluster-wise simultaneous perturbation stochastic approximation algorithm and its application to dynamic origin-destination matrix estimation. Transp Res Part C: Emerg Technol 55:231–245
Toledo T, Kolechkina T (2013) Estimation of dynamic origin-destination matrices using linear assignment matrix approximations. IEEE Trans Intell Transp Syst 14(2):618–626
Tong CO, Wong SC (2000) A predictive dynamic traffic assignment model in congested capacity-constrained road networks. Transp Res Part b: Methodol 34:625–644
van Aerde M, Hellinga B, Yu L, Rakha H (1993) Vehicle probes as real-time ATMS sources of dynamic OD and travel time data. Queen's University, Department of Civil Engineering
van Zuylen HJ, Willumsen LG (1980) The most likely trip matrix estimated from traffic counts. Transp Res Part B: Methodol 14:281–293
Varia HR, Dhingra SL (2004) Dynamic user equilibrium traffic assignment on congested multidestination network. J Transp Eng 130(2):211–221
Wang IJ, Spall JC (1999) A constrained simultaneous perturbation stochastic approximation algorithm based on penalty functions. In: IEEE Proceedings of the 1999 American control conference (Cat.No.99CH36251), 1999. USA
Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
Wang Z, Simoncelli EP (2008) Maximum differentiation (MAD) competition: a methodology for comparing computational model of perceptual quantities. J vis 8(12):1–13
Wardrop JG (1952) Some theoretical aspects of road traffic research. Proc Inst Civil Engineers II:325–378
Wu JH (1991) A study of monotone variational inequalities and their application to network equilibrium problems. Ph. D. Thesis, Centre de Recherche sur les Transports, Université de Montréal, Publication #801
Wu JH, Chen Y, Florian M (1998a) The continuous dynamic network loading problem: a mathematical formulation and solution method. Transp Res Part B: Methodol 32(3):173–187
Wu JH, Florian M, Xu YW, Rubio-Ardanaz JM (1998b) A projection algorithm for the dynamic network equilibrium problem. In: Yang Z, Wang KCP, Baohua M (eds) Traffic and transportation studies, ASCE proceedings of the ICTTS'98
Xu YW, Wu JH, Florian M, Marcotte P, Zhu DL (1999) Advances in the continuous dynamic network loading problem. Transp Sci 33(4):341–353
Yang H (1995) Heuristic algorithms for the bi-level origin-destination matrix estimation problem. Transp Res Part B: Methodol 29:231–242
Yang H, Zhou J (1998) Optimal traffic counting locations for origin-destination matrix estimation. Transp Res Part B: Methodol 32B(2):109–126
Yang X, Lu Y, Hao W (2017) Origin-destination estimation using probe vehicle trajectory and link counts. J Adv Transp 2017:4341532
Zhang Y, Qin X, Dong S, Ran B (2010) Daily O-D matrix estimation using cellular probe data, Paper 10–2472. In: 89th TRB annual meeting, 2010. Washington DC

Part IV
Overall Conclusions and Further Research

Chapter 8
Overall Conclusions and Further Research

Margarita Martínez-Díaz

Abstract This chapter summarizes very succinctly the main contributions included in this book and suggests some challenges for new research that, among others and from the traffic management point of view, would gradually contribute to achieving the optimal mobility we look for.

8.1 Overall Conclusions

Two main conclusions can be drawn from the lecture of this book. The first one can be stated as follows:

> It is **already** possible to gradually adapt current travel time information systems to new driving environments, so that they are able to provide accurate travel time predictions in real time, making the most of the available surveillance.

Travel time information is and will be very valuable both for drivers/passengers and traffic agencies. However, current schemes do not provide accurate information under all traffic conditions, especially in congestion, due to two reasons: (i) lack of surveillance equipment and/or (ii) lack of adequate estimation (prediction) methodologies. It has been demonstrated in this book that travel time information systems can be significantly improved, immediately, and with no or limited monetary expenditures. Moreover, it is possible to implement methodologies that will also be valid and even more effective in future driving scenarios, for example, exploiting the accessibility to huge amount of data supplied by the pervasive penetration of mobile devices.

In those highways that, still, exclusively rely on loop detector data for traffic management, travel time estimations are obtained from spot speed methods. These

M. Martínez-Díaz (✉)
Department of Civil and Environmental Engineering, Area of Transport and Territorial Infrastructures, Barcelona Innovative Transportation (BIT) Research Group, Polytechnic University of Catalonia, UPC-BarcelonaTech, Barcelona, Spain
e-mail: margarita.martinez.diaz@upc.edu

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*,
Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0_8

methods calculate the travel time in a link between loop detectors by diving the length of this link by an average speed. This average speed is the result of different mathematical interpolations of the time mean speeds provided by the loops every some time interval. Although spot speed methods have additional inconveniences, the first inaccuracy is precisely introduced in this step: space mean speeds and not time mean speeds should be used in order to estimate average travel times. Taking into account that time means are higher than space means, current procedures lead to travel time underestimations. In case no other surveillance is available, an algorithm to obtain space mean speeds from the data commonly supplied by loop detectors has been derived. For this purpose, it was assumed a log-normal distribution of speeds over small space–time regions. Using real data from the AP-9 freeway in Spain, the method has been proven to perform better than today's schemes. Additionally, it's worth mentioning that the outputs of the algorithm (i.e., space mean speeds) will be useful for many traffic studies other than those related to travel time estimation. In fact, most of them rely on space mean speeds and not on time mean speeds. In spite of the proposed method, it must be highlighted that simple modifications to the standard loop detector data treatment process would suffice to directly obtain space mean speeds. However, this has not been undertaken so far. Additionally, the increasing presence of high-tech devices in vehicles and on the roads will, for sure, gradually lead to the withdrawal of spot speed methods, at least in their current forms.

In fact, future travel time information systems will not only rely on these traditional methods. Even with the proposed improvement, none of them performs well in transients (i.e., stop and go, shock wave on- and offsets, etc.). Moreover, the spot speed methods used in practice only provide good estimates when free flow prevails. The reason is that they overlook traffic dynamics and the nature of queue evolution when giving spot speeds a spatial consideration via blind mathematical interpolations. Therefore, spot speed methods are useless precisely when travel time information is more valuable, that is, in congested episodes. Several attempts have tried to go over this issue, but none of them with complete success. Finally, it is important to note that, even if this problem could also be solved, spot speed methods provide at best instantaneous travel times (ITT), but not travel time predictions. These last outputs should be the goal of any system at present, and even more in the near future. In fact, travel time predictions are called to be essential in future cooperative scenarios, in which any information will have even a higher impact on traffic performance.

Therefore, travel time information systems must take advantage of technological progress to be increasingly precise. More in particular, to be able to provide accurate travel time forecasts, at least for the short term. Cooperative driving environments will gradually and heterogeneously settle in the road network. From this standpoint, the need for travel time prediction methodologies that adapt to and benefit from the available equipment is clear. That is, methodologies that benefit from new data sources but perform well even if these are scarce. And this is the case of the method presented in Chap. 4 of this book. It rises with a vocation of continuity in the sense that it can be immediately put into practice even in low surveillance scenarios, but it could also form part (and take advantage) of future travel time information systems and complement/support, for example, data-driven methods. In fact, the method is

not technologically captive. The proposed methodology is aimed at fusing the information provided by input–output diagrams, obtained from loop detectors, with direct measurements of travel times obtained from either automatic vehicle identification (AVI) or tracking technologies. This fusion allows exploiting the accuracy of the direct measurements, which, however, are somehow outdated, in order to correct the count drift in loop detectors. Then, corrected input–output curves can be used to obtain reliable short-term predictions of travel time, using the predictive capabilities of the vehicle accumulation. The proposed data fusion method has been applied to a test site in the AP7 freeway near Barcelona using real and simulated data. Results show that the method is able to provide predicted travel times that anticipate changes in traffic conditions much faster than when simply disseminating measured travel times. This implies lower average and maximum errors of real-time information systems, as it has been demonstrated. For the particular cases analyzed, the mean and maximum errors as well as the MAPE corresponding to the predicted travel times represented significantly smaller percentages of the ground truth travel times when compared to those errors linked to the dissemination of direct travel time measurements. Another advantage of the method is that the real-time information provided can be more frequently updated. The proposed methodology is especially suited for moderate to severe congestion episodes, in contrast to current procedures. That is, for the context in which travel time information is more valuable and more difficult to obtain.

The second key conclusion drawn from the present book can be formulated in the following way:

> A good (dynamic) traffic management system must have two main components, namely a dynamic traffic model with predictive capabilities and a dynamic origin to destination (OD) matrix predictor. Especially in urban environments, the real-time dynamic OD estimation becomes a critical step for the successful performance of the system.

A common critical component of any traffic management system is the one that provides the system with capabilities for the prediction of the short-term evolution of the traffic state depending on the real-time situation and on the application of different management strategies. This component is, usually, a dynamic traffic model. Among other inputs, the accurate estimation of travel times (both their prediction and their reliability) is key, as they have a clear relationship with the subsequent route choices.

No matter if these dynamic traffic models are microscopic or mesoscopic, they require the dynamic estimation of origin-to-destination matrices accounting for the evolution of the demand over time. The accurate estimation of these matrices continues to be a challenge in itself. Data-driven methods that try to take advantage of the ever-increasing amount of traffic (or traffic-related) data available are the subject of much research and of some practical applications, and could represent a considerable breakthrough in this regard. In any case, at least at these early stages, the application of these methodologies should not be done lightly, but from a deep knowledge of traffic engineering. Ideally, they should be introduced in conjunction or contrasted with other methods based on more traditional (and more applied) models, in order to detect possible inconsistencies.

8.2 Further Research

Some ideas for further research have already been detailed in particular chapters of this book. Much more could have been mentioned and much more will arise, taking into account the enthralling evolution of transportation, which is expected to continue in the next years. Especially regarding cooperative driving environments, more issues to overcome will appear and, thus, more challenges for research. Notwithstanding, some interesting possibilities for future research work are summarized next.

For example, it was already explained that spot speed methods, because of their inability to predict travel times, are not called to be the basis of future travel time information systems. Therefore, it would not be worth devoting too many efforts or investments to them. However, as some time will go by until they are substituted in many networks, further quite simple improvements could be introduced. For instance, the proposed algorithm for the estimation of space mean speeds could be combined with similar ones that assume different speed distributions. All of them for which more than one output were possible (like for the one proposed) should include a module to calculate the confidence interval for the means, to enable the quick choice of the most accurate one. With this configuration, a complementary algorithm should previously detect the most suitable distribution for each particular period and, afterward, the proper algorithm for the calculation of space mean speeds would turn on. Additionally, a simple but smart smoothing process could account for the noisy loop detector speed measurements, reducing the fluctuations typical of short time interval aggregations. This smoothing method should be directly applicable with the existing loop detector hardware. Additionally, any proposed improvement should be straightforward enough so as not to imply delay in the dissemination of the information. Current schemes aimed at identifying stationary periods could be integrated in the proposed methodology for this purpose.

Regarding the data fusion short-term travel time prediction methodology introduced in this book, it would be interesting to perform a sensitivity analysis taking into account, among others, different layouts (e.g., more on- or off-ramps and different section lengths) and different traffic conditions (e.g., different levels of congestion). Although already addressed, a deeper assessment of the impact of the different time intervals of aggregation for the direct and indirect travel time measurements under these distinct scenarios would also be desirable. Especially for the fusion of ITTs, different percentages of tracked vehicles, different frequencies for the GPS update or even their sourcing via geolocalization would be interesting variations to test. Additionally, despite the proven goodness of the data fusion procedure, it was noticed that it was sensitive to loop detector failures. Although this fact is less prejudicial than the detector drift (note that drift nearly affects all loops, whereas detector failure is less frequent and more disseminated in the infrastructure), the implementation of a previous algorithm accounting for detector malfunctioning or data loss would enhance the proposed methodology.

By extending the vision to traffic management systems as a whole, the research possibilities are as broad as attractive. It would be impossible to summarize them,

given the many existing variants of the boundary conditions, the different possible approaches of these systems, the varied data available, etc. In any case, what can be said is that further work is needed to make these systems increasingly accurate and adaptable to future cooperative driving environments. As already indicated, the correct use of the vast amounts of data that management centers have at their disposal is a clear goal to achieve. The development of increasingly advanced dynamic models, the introduction of artificial intelligence, data-driven methodologies, etc., that take advantage of these data is crucial. Nevertheless, it should not be forgotten that, even though some of these advanced techniques (e.g., deep learning) are intellectually appealing and promising, they face an important challenge: knowledge about the nature of traffic flows and their dependencies on behavioral aspects must be suitably embedded in these models so that their predictions can be considered reliable.

However, while working toward these ambitious goals, it is also necessary to put the feet on the ground: all these disruptive advances will be seen sooner in the sphere of research than on roads and in cities. And society cannot wait for their large-scale application to become feasible. Therefore, work must be done at the same time on implementations that, even not becoming perfect solutions, can improve existing traffic management schemes and thus mobility in the short term. They must therefore be proposals that take advantage of the data and equipment commonly found on roads and in management centers, and that optimize their use without requiring significant expenses.

At present and in the future, on roads and in cities, dynamic traffic management is not only essential to ensure efficient, safe, sustainable, and inclusive mobility but also an exciting and challenging area of research.

Glossary

AI Artificial Intelligence
ALPR Automatic License Plate Reidentification
AMS Analysis, Modeling and Simulation
ANPR Automatic Number Plate Reidentification
ATIS Advanced Traveler Information Systems
ATDM Active Transportation and Demand Management
ATT Advanced Traffic Telematics
ATTMS Advanced Traffic/Transport Management Systems
ATT Arrival-based Travel Times
ATTL Advanced Transport Telematics
AV Autonomous/Automated vehicle
AVCS Advanced Vehicle Control System
AVI Automatic Vehicle Identification
BRT Bus Rapid Transit
BTI Buffer Time Index
CAV Cooperative/connected Autonomous/Automated Vehicle
CBD Central Business District
CDR Call Detail Records
CF Commonality Factor
CG Conjugate Gradient
C-ITS Cooperative Intelligent Transportation Systems
CTM Cell Transmission Model
CV Coefficient of Variation
DMA Dynamic Mobility Applications
DMS Dynamic Message Sign
DNL Dynamic Network Loading
DODME Dynamic Origin–Destination Matrix Estimation
DSRC Dedicated Short-Range Communications
DSS Decision Support System

M. Martínez-Díaz (ed.), *The Evolution of Travel Time Information Systems*, Springer Tracts on Transportation and Traffic 19,
https://doi.org/10.1007/978-3-030-89672-0

DTA Dynamic Traffic Assignment
DTT Departure-based Travel Times
DUE Dynamic User Equilibrium
EAV Electric Automated Vehicles
EKF Extended Kalman Filter/ing
ETC Electronic Toll Collection Systems
EU European Union
EV Electric Vehicle
FCD Floating Car Data
FDAM Finite Difference Approximation Method
FIFO First in–First out
GDP Gross Domestic Product
GIS Geographical Information System
GLTM General Link Transmission Model
GPRS General Packet Radio Service
GPS Global Positioning System
HF High Frequency
HMI Human–Machine Interface
HOT High Occupancy Toll
HOV High Occupancy Vehicle
I2X Infrastructure-to-All Comm./Interactions
ICM Integrated Corridor Management
ICT Information and Communications Technology
ID/id Identifier
IMS Incident Management Systems
IMU Inertial Measurement Unit
IoT Internet of Things
IoV Internet of Vehicles
ITMS Intelligent Traffic Management Systems
ITS Intelligent Transportation Systems
ITT Instantaneous Travel Times
IVHS Intelligent Vehicle Highway Systems
IVS Intelligent Vehicle Systems
KPIs Key Performance Indicators
KS Kolgomorov–Smirnov
KWT Kinematic Wave Theory
LB Lower Bound
LF Low Frequency
LIDAR Light Detection and Ranging
LOS Level of Service
LWR Lighthill, Whitham and Richards Theory
MaaS Mobility as a Service
MI Misery Index
MMS Multimedia Message Service
MSA Method of Successive Averages

MSE Mean Square Error
MTT Measured Travel Times
NFD Network Fundamental Diagram
OD Origin/Destination
OCR Optical Character Recognition
OSM Open Street Map
OTC Optimal Traffic Control
P2X Person-to-All Comm./Interactions
PATH Partners for Advanced Transit and Highways
PDA Personal Digital Assistant
PTI Planning Time Index
PTT Predicted Travel Times
RDMS-TMC Relational Database Management System-Traffic Message Channel
RFI Radio Frequency Identification
SMS Short Message Service
SoC System-on-Chip
SS Structural Similarity
STA Static Traffic Assignment
TCC Traffic Control Centre
TMS Traffic Management System
TT Travel time
UAV Unmanned Aerial Vehicle
UETA User Equilibrium Traffic Assignment
UHF Ultra High Frequency
US The United States
USDOT United States Department of Transportation
V2G Vehicle-to-Grid Communications/Interactions
V2I Vehicle-to-Infrastructure Communications/Interactions
V2N Vehicle-To-Network Communications/Interactions
V2P Vehicle-To-Person Communications/Interactions
V2V Vehicle-to-Vehicle Communications/Interactions
V2X Vehicle-to-All Communications/Interactions
VANET Vehicular ad hoc Networks
VC Vehicular Cloud
VCN Vehicular Cloud Network
VDS Variable Direction Sign
VMS Variable Message Sign
VTTS Value of Travel Time Savings