Escola Tècnica Superior d'Enginyers de Camins, Canals i Ports de Barcelona
UNIVERSITAT POLITÈCNICA DE CATALUNYA

PhD Thesis

HIGHWAY TRAVEL TIME ESTIMATION WITH DATA FUSION

Francesc Soriguera Martí

PhD Supervisor:

Prof. Francesc Robusté Antón

PhD Program in Civil Engineering
E.T.S. d'Enginyers de Camins, Canals i Ports de Barcelona
Universitat Politècnica de Catalunya – UPC – Barcelona Tech

Barcelona, October 2010

Escola Tècnica Superior d'Enginyers de Camins, Canals i Ports de Barcelona
UNIVERSITAT POLITÈCNICA DE CATALUNYA

HIGHWAY TRAVEL TIME ESTIMATION WITH DATA FUSION

Francesc Soriguera Martí

Memòria presentada per optar al títol de Doctor
Enginyer de Camins, Canals i Ports
ETSECCPB-UPC

Director de la tesi: Dr. Francesc Robusté Antón

Barcelona, Octubre de 2010

As a general rule the most successful man in life
is the man who has the best information

Benjamin Disraeli

HIGHWAY TRAVEL TIME ESTIMATION WITH DATA FUSION

Francesc Soriguera

ABSTRACT

Travel time information is the key indicator of highway management performance and one of the most appreciated inputs for highway users. Despite this relevance, the interest of highway operators in providing approximate travel time information is quite recent. Besides, highway administrations have also recently begun to request such information as a means to measure the accessibility service provided by the road, in terms of quality and reliability.

In the last century, magnetic loop detectors played a role in providing traffic volume information and also, with less accuracy, information on average speed and vehicle length. New traffic monitoring technologies (intelligent cameras, GPS or cell phone tracking, Bluetooth identification, new MeMS detectors, etc.) have appeared in recent decades which permit considerable improvement in travel time data gathering. Some of the new technologies are cheap (Bluetooth), others are not (cameras); but in any case most of the main highways are still monitored by magnetic loop detectors. It makes sense to use their basic information and enrich it, when needed, with new data sources.

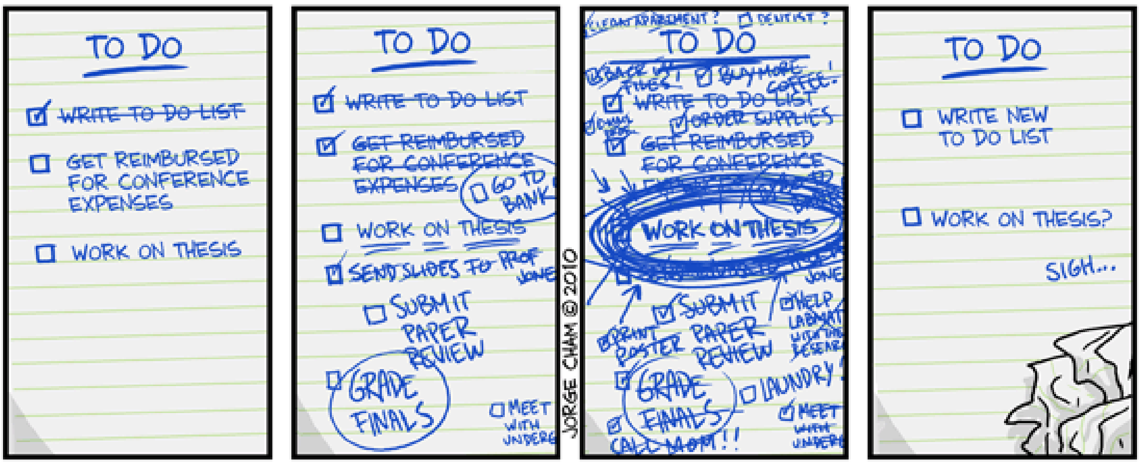
This thesis presents a new and simple approach for the short term prediction of toll highway travel times based on the fusion of inductive loop detector and toll ticket data. The methodology is generic and it is not technologically captive: it could be easily generalized to other equivalent types of data.

Bayesian analysis makes it possible to obtain fused estimates that are more reliable than the original inputs, overcoming some drawbacks of travel time estimations based on unique data sources. The developed methodology adds value and obtains the maximum (in terms of travel time estimation) of the available data, without falling in the recurrent and costly request of additional data needs.

The application of the algorithms to empirical testing in AP-7 toll highway in Barcelona proves our thesis that it is possible to develop an accurate real-time travel time information system on closed toll highways with the existing surveillance equipment. Therefore, from now on highway operators can give this added value to their customers at almost no extra investment. Finally, research extensions are suggested, and some of the proposed lines are currently under development.

Key Words: *Travel time estimation, loop detector data, toll ticket data, space-mean speed, traffic data fusion, Bayesian analysis.*

Francesc Robusté
Professor of Transportation, PhD
Civil Engineering - Barcelona Tech
November, 2010



"*Too due list*" by Jorge Cham (2010) available at www.phdcomics.com

AGRAÏMENTS

M'adono ara de fins a quin punt arriba la perversió de la tesi doctoral, aquesta companya inseparable dels últims anys, aquesta motxilla silenciosa i pesada que no et pots treure del damunt (Miquel aquí tens el treu tribut), aquesta prioritat entre les moltes altres prioritats de la vida. Fins al final. Aquestes línies seran les més llegides d'aquest document. N'estic segur. La seguiran el resum i les conclusions. Totes tres seccions les escric ara. Al final. Sota pressió. Amb poc temps i menys ganes després d'un esprint final de 10 dies. No és just. No m'agraden els esprints (Javi va per tu). Jo soc més de fons.

Només per fer front a aquesta injustícia, ja no escriuria aquestes línies. Li hauria guanyat la última batalla. Però no m'ho puc permetre. No acostumo a expressar gratitud habitualment. Bé, de fet no expresso gran cosa habitualment. Me'n penediria de deixar passar aquesta oportunitat única, així que m'hi llenço.

Aquesta tesi no existiria de no ser d'en Francesc Robusté. Ell m'hi va incitar. Ell es va brindar a dirigir-la. Ell n'ha aconseguit el finançament[1]. Ell em va marcar el camí a Berkeley (those were the most prolific months in my life. Thanks to all the transportation studies community). La tesi tampoc seria el mateix de no ser de la seva manera d'entendre la recerca. Les "*Robus' Guidelines*" em marcaran per sempre. Per tot això només puc dir-te, gràcies.

Les dades... "divino tesoro". En aquest aspecte dec gran part a la gent d'Abertis. Joan, Pepe, Carles, us estic molt agraït. I la idea inicial, va ser teva Leif? Y aquel comentario durante mi primera reunión en el CENIT "*...todo esto no sirve para nada...*" (Magín, eres único), i tots el companys i amics amb els que he treballat colze contra colze des d'aquell ja llunyà 14F 2005 (sí Marta, 14), Dulce, Josep Maria, Manel, Edu, David; o em compartit dinars, curses i preocupacions, Sergi, Deme, César, Ramon... A tots, gràcies per ajudar-me.

Una vegada hi estàs posat això d'agrair t'omple. Ho hauria de fer més sovint. Pare, mare, Fati, padrins, Cèlia sapigueu que tots m'ajudeu moltíssim dia a dia. Ni que no us ho digui mai, Gràcies.

Montbui, 18 d'Octubre de 2010

[1] This work was supported in part by the Spanish Ministry of Science and Innovation under Grant TRA 2008-06470/MODAL.
[1] This work was supported in part by the Abertis Chair – UPC.

TABLE OF CONTENTS

LIST OF FIGURES

THESIS REPORT

APPENDIX A1

APPENDIX A2

APPENDIX A3

APPENDIX A4

LIST OF TABLES

THESIS REPORT

APPENDIX A1

APPENDIX A2

APPENDIX A3

APPENDIX A4

THESIS OVERVIEW AND OBJECTIVES

In the present context of restraint in the construction of new infrastructures due to territorial, environmental and economical restrictions, and given the big bang of mobility in the last decades in all sectors, the transportation system management acquires a fundamental role as an optimizer of the available resources. This requires the application of new policies addressed to achieve two main objectives: sustainability and competitiveness.

Mobility management is the big issue. The information technologies and communications are the tools. It is therefore needed a management system that links these objectives with these available tools. This management system should be based on quantitative traffic information in real time. Travel time and its reliability stand out as key factors in traffic management systems, as they are the best indicators of the level of service in a road link, and perhaps the most important variable for measuring congestion. In addition, travel time is the best and most appreciated traffic information for road users as it plays a fundamental role in the traveler planning process (travel or not travel, best time to travel, route and mode choice…). At the same time, highway travel time measurement and quantitative forecasting in congested conditions pose a striking methodological challenge.

The most elementary method to measure a highway section travel time is by identifying the vehicle at the entrance and exit of the target section, and computing the elapsed time between identifications. However, the necessary automatic vehicle identification (AVI) is not a trivial task. As it will be presented in next sections, it needs somehow advanced technology. All of these technologies require the extensive installation of new hardware in the vast highway network, which cannot be achieved overnight, and possibly it is not profitable in the whole network. To date, these systems have only seen limited demonstration, and the extensive deployment it is not expected beyond the hot spots of the network.

Loop detectors still represent the main source of traffic data in all highways worldwide. And it is expected to remain like this in the medium term [May et al., 2004]. Note that loop detectors are not adequate for link measurements (e.g. like travel times) but are capable of an exhaustive measurement of the punctual data (e.g. traffic counts: the original objective) which is not less valuable and for which the AVI technologies are not so well suited. Given the predominance and preexistence of this surveillance context, lots of traffic agencies worldwide have decided to develop travel time information systems based on a simple and intuitive methodology: they estimate indirectly link travel times from the spatial generalization of the loop punctual measurement of average speed.

As it will be detailed in the next chapters, in both, the direct measurement and the indirect estimation of highway travel times difficulties arise. From the author knowledge, some of these difficulties, which are treated in the present thesis, have not been addressed in the related literature. In addition, different measurement processes lead to conceptually different results (see the part of Appendix A2 devoted to travel time definitions), an issue which is frequently overlooked. Note that a directly measured travel time is a trajectory based measurement in space-time where the vehicle needs to have finished his trip in order to obtain the measurement. In contrast, indirect estimations usually are instantaneously obtained, and do not respond to the trajectory of a particular vehicle. And as if that was not enough, the main objective of a real time highway travel time information system should be to provide the driver with the information of the travel time his trip will undertake once entering the highway. This means that a real time information system needs in reality future information, where the horizon of the forecast is equal to the trip travel time.

This situation, with multiple surveillance equipments, inhomogeneous data with different variables being measured, different travel time estimation algorithms with different accuracies, different spatial coverage, and different temporal implications is the ideal environment for data fusion schemes, where the objective is to use jointly the information provided by different sources in order to infer a more accurate and more robust estimation of the target variable (i.e. the travel time).

THESIS OBJECTIVES

The main objective of the present thesis is to develop a methodology capable to provide a driver entering a highway with accurate information of the travel time it will take the trip he is going to undertake. Note that this information involves two components: the measurement of the current travel times and the estimation of the evolution of the traffic conditions during the time taken for the trip. An additional requirement of the research is that the travel time estimation must be obtained by making the best usage possible of the available multiple highway data sources, neither increasing the highway density of surveillance nor changing the typology of the measurement equipment. Therefore, the objective is to add value to the traffic data which is currently being measured.

Two main research directions appear when facing the issue of real time monitoring the traffic evolution. The first and most intuitive way of quantitatively know what is happening in a highway stretch it is by measuring. In practice this is not an easy task. The measure equipments are limited. They may not be able to measure the most important variables. The amount of measurements may not be representative of the average traffic stream. Their spatial coverage may be limited. The necessary temporal aggregations to reduce the amount of data being transmitted may bias the measurement and may add some delay to the information. The existence of outliers adds additional complexity. And finally, a non negligible amount of measurement units may usually be out of order. The alternative consists in modeling. Highway traffic consists on the interaction of a huge number of drivers; human beings with different ages, different races, different religious creeds, different gender, different political preferences, different way of life options and different psychological stabilities [Vanderbilt, 2008]. Nevertheless (and perhaps surprisingly), on the average they all behave similarly when they face similar conditions. This means that given the characteristics of the drivers' population, the characteristics of the highway environment they are facing and the mobility demand (i.e. the number and characteristics of the trips which are going to be undertaken) it should be possible to know how these vehicles are going to interact and therefore to obtain all the resulting variables of their trips (e.g. their travel times). The forecasting capabilities of this approach are appealing. Again this is not an easy task. It is difficult to know (which means measure) the characteristics of the drivers' population. And it is difficult to accurately model the behavioral laws which steer the relation between the infrastructure and drivers and also between drivers among themselves. In fact some of these behavioral laws are still unknown [Daganzo, 2002]. It is even difficult to know the mobility demand. In conclusion, there are a lot of unknowns.

The dilemma is then to select one option. Lately, modeling has experienced an enormous popularity increase. It seems that modeling is now easier than has ever been before. And the results are better (at least faster – in real time – and more visual). This is due to the quick development of computers and the enhancement of its visual capabilities, which has given rise to traffic simulators (in particular micro simulators, where the performance of each vehicle can be seen in 3D). Although these enhancements, brought up by the digital era, the baseline difficulties remain the same. One must realize that most of these difficulties are solved by case specific over-calibrations which blur the forecasting capabilities of the model. Fortunately, researchers all over the world are working hard in overcoming these problems, and in the future this may come to a happy end. In contrast, traffic monitoring seemed to have fallen out of favor some years ago. This was mainly due to the huge costs of enhancing the traffic surveillances systems given the enormous inertia of the vast highway network. The surveillance equipment rapidly became outdated, with high maintenance cost, and a high rate of malfunctioning. This entire situation discouraged practitioners and researchers of devoting their interests in traffic monitoring. This is now changing with the appearance of high tech low cost traffic detectors. Some technological problems will surely be solved. Costs will also be reduced. However, the conceptual difficulties in the measurements will also remain the same. In addition, all the surveillance system will not be replaced overnight, and different equipments will need to coexist.

Both research directions, measuring and modeling, are appealing. Both have a huge potential. In spite of the modeling "attacks" and the temporary monitoring decline, measuring will not be substituted by modeling. Traffic monitoring will be always necessary. Maybe the axiom should be "measure all you can; model the rest". This is what this thesis is devoted to: providing methodologies to accurately measure highway travel times. In general, the thesis is not technologically captive, as it is more related to the concept of the measurement than to the technological equipment used.

In particular the research is devoted to the specific case of a closed toll highway, where the direct travel time measurement is given by the information contained in the toll tickets (real of virtual by means of electronic toll collection systems), which record the exact time and location where each vehicle enters and leaves the highway. The indirect estimation is obtained from the flow, speed and occupancy measurements of inductive loop detectors. Although this specific environment where the research is developed, some of the proposed methodologies will be easily generalized to other context where both, a direct measurement and an indirect estimation are available.

When facing the travel estimation problem in this closed toll highway environment, with these commonly available sources of data, three main questions arise:

- How travel time can be measured from toll ticket data?
- How travel time can be measured from loop detector data?
- Given these two travel time estimations from different data sources with their intrinsic characteristics, can we combine them to obtain better information?

These questions are going to be answered in the present thesis.

THESIS MAIN CONTRIBUTIONS

In order to improve and facilitate the diffusion of the contents of the thesis, its structure has been conceived as a compendium of journal papers. Four papers are presented as appendixes.

The first paper (Appendix A1) entitled "*Estimation of Traffic Stream Space-Mean Speed from Time Aggregations of Double Loop Detector Data*" presents a method for solving a problem which appears at the very start of trying to estimate travel times using the average speed measured at detector sites. It is a common practice to obtain the average speed of a traffic stream during a short time aggregation period by arithmetic averaging the individual speeds of vehicles. This results in a time-mean speed, which is stored and sent to the traffic management center. Raw data are eliminated. However, in order to compute travel times from average speeds, space-mean speeds are needed. Time-mean speeds and space-mean speeds are related by the individual speed variance, which given the available data is unknown. The paper proposes a statistical method capable of accurately obtain space-mean speed from and only from the commonly time aggregations of loop detector data.

Once space-mean speed is obtained, the main drawback in obtaining travel times from punctually measured average speed is the spatial generalization of this measurement. Several methods are proposed in the literature ranging from the simplest constant interpolated to mathematically complex truncated quadratic interpolations. The research tendency seems lately to follow the direction of continuously increasing the mathematical complexity of the methods overlooking traffic dynamics. This issue is addressed in the second paper (Appendix A2) entitled "*Requiem for Freeway Travel Time Estimation Methods Based on Blind Speed Interpolations between Point Measurements*" This paper, with its iconoclastic title (reminiscent of the title of a paper by Daganzo, 1995), claims that all speed interpolation methods that do not consider traffic dynamics and queue evolution do not contribute to better travel time estimations. Lacking a better approach, and assuming it is naïve alternative, an intelligent smoothing of the noisy loop detector data is proposed. The method is capable of reducing the fluctuations of short time interval aggregations while maintaining the immediacy of the measurements. It must also be pointed out that this paper includes two introductory sections, one where travel time definitions are analytically presented, and another developing a trajectory reconstruction algorithm. The concepts presented in these sections, aimed to create a conceptual framework useful in comparing travel times obtained from different methodologies, should be considered as a baseline knowledge common for all the papers in the compendium.

The third paper (Appendix A3) entitled "*Travel Time Measurement in Closed Toll Highways*" provides a methodology which offers an answer to the first question. The main contribution of this part of the thesis is a method capable of obtaining single section main trunk average travel times (i.e. in between junctions) from specific origin – destination individual vehicle travel times, which include the "exit time" (i.e. the time required to travel along the exit ramp and to pay the fee at the toll plaza. This method allows reducing the intrinsic delay in the information of directly measured travel times, useful for a real time application of the system.

Having obtained travel time estimations from the different available data sources, the last paper in the present compendium (Appendix A4) entitled "*Highway Travel Time Accurate Measurement and Short-term Prediction Using Multiple Data Sources*" proposes a data fusion scheme, partially based on the probabilistic Bayes' Theory, whose objective is to use the potentials of each source of data to overcome the limitations of the others in order to obtain a more accurate and robust travel time estimation. In addition, the proposed method uses the different temporal alignments of travel time estimations to infer a tendency and improve the forecasting capabilities of real time measurements. The source estimation methods used in this last paper are the ones presented in the previous research, plus the introduction of an additional method based on cumulative count curves and conservation of vehicles.

It has been stated that the objective of the thesis is to provide solutions to a global engineering problem. Each one of the papers that conforms the thesis answers a partial question which follows from the original research problem. The unity of the topic treated is therefore granted.

In addition, in any engineering thesis, the practical application of the proposed methodology to a pilot test site is desired. In this particular case data from a privileged site was available. The AP-7 highway runs along the whole Spanish Mediterranean coast, from Algeciras to the French border at La Jonquera. On the north eastern stretch of the highway, from La Roca, near Barcelona, to La Jonquera, a closed tolling scheme is in use. Toll ticket data were available to the author. Moreover, in some sections of the highway near Barcelona (in particular from La Roca to Maçanet – see Fig 1.), additional monitoring by means of loop detector data were installed every 5 km approximately. Only a requirement is missing for this stretch being a perfect test site: a congestion episode. Unfortunately for the highway users (but fortunately for the development of the present thesis) every Sunday (among other days) of the summer season (particularly long in the Mediterranean climate) congestion grows in the southbound direction of the highway, due to the high traffic demand towards Barcelona of drivers which have spend a day or the weekend on the coast. Hopefully, the contents in this thesis may help to alleviate this congestion, or at least it will provide information to diminish the drivers' suffering. This was the privileged test site used in all the papers presented here.

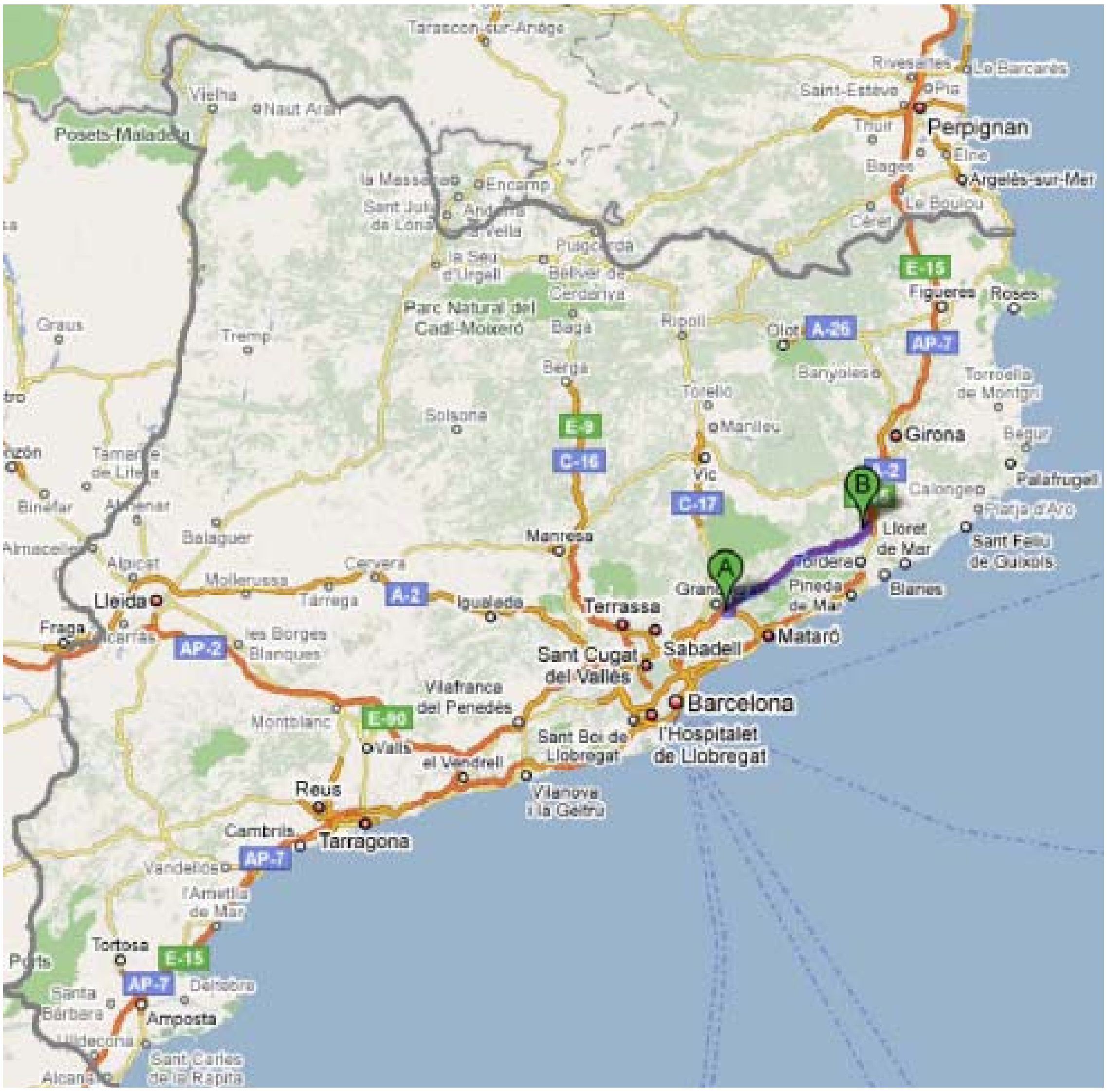


FIGURE 1 Test site location. Source: Google Maps

ROAD MAP

After this overview of the thesis, which gives an introduction to its contents and provides linking arguments between its different parts, the rest of the thesis is structured as follows: two main parts can be differentiated, the thesis report and the appendixes. As stated previously, each appendix corresponds to a paper dealing with some part of the global research question. In these appendixes reside the main research of the thesis and it is where the main contributions are found. The thesis report, structured in several chapters, has the objective of providing a global view of the issues treated and of the results obtained. In addition some baseline concepts, common for all papers and assumed to be known are also introduced. Given that the layout of the thesis is in the form of compendium of papers, the literature review for each topic is provided separately in each paper. Specifically, the chapters that configure the thesis report are: Chapter 1, where the importance of travel time in the mobility management is further analyzed; Chapter 2 devoted to present several methods for the direct and indirect travel time estimation; in

Chapter 3 the relation between data fusion and travel time forecasting concepts are discussed; Chapter 4 is devoted to analyze the dissemination of travel time information among drivers; and Chapter 5 where some issues regarding to the value of travel time information as a traveler oriented reliability measure are discussed. Finally, some overall conclusions and issues for further research are presented in Chapter 6.

To conclude this overview section it is worth mentioning that the present thesis has already seen some of its main contributions been published. Others are in their way. Abridged versions of the papers which conforms the present compendium can be found in the following journals:

- Soriguera, F. and F. Robusté. (2011-a) Requiem for Freeway Travel Time Estimation Methods Based on Blind Speed Interpolations between Point Measurements. *Accepted for publication in IEEE Transactions on Intelligent Transportation Systems.*

- Soriguera, F. and F. Robusté. (2011-b) Estimation of Traffic Stream Space-Mean Speed from Time Aggregations of Double Loop Detector Data. *Transportation Research Part C* **19**(1), 115-129.

- Soriguera, F., D. Rosas and F. Robusté. (2010) Travel Time Measurement in Closed Toll Highways. *Transportation Research Part B* **44**(10), 1242–1267.

- Soriguera, F. and F. Robusté. (2009) Highway Travel Time Accurate Measurement and Short-Term Prediction Using Multiple Data Sources. *Transportmetrica* iFirst 1-25.

1. TRAVEL TIME AND MOBILITY MANAGEMENT

Mobility is synonymous of economic activity and dynamism. It has been vastly proved the relationship between mobility demand and the wealth of a particular region [Robusté et al., 2003]. The mobility increase implies greater competiveness and, if properly panned, territorial cohesion. However, for the mobility to provide these benefits, a good transportation network and a better management of the transportation system is necessary. An infrastructural deficit or the absence of an active management may entail the increase of mobility being counterproductive, transforming the potential benefits to additional costs. These over costs are mainly due to the congestion phenomena.

Congestion is linked to success. It appears when the interaction between transportation demand and the transportation system supply (in terms of infrastructure and organization) generates increasing unitary costs to overcome the same unitary length. Taking into account that the infrastructural supply can hardly go ahead of the demand, given the limitation of resources, congestion has to be considered as an inevitable phenomenon which indicates success and acts as a demand regulator. In spite of this, congestion must be managed and must be sustained as punctual and moderate episodes: it is necessary to maintain a “suitable” level of congestion. The first step is then to know and quantify the level of congestion and to try to limit its damaging variability. This means that for the same trip on two similar days, the travel time should be similar, not the double. This concept is known as travel time reliability.

In most of the metropolitan areas worldwide, the existing levels of congestion are far above from these suitable thresholds. In addition, and despite the actual context of economical recession which has alleviated the growing trends, metropolitan congestion in developed areas is still slightly increasing [Federal Highway Administration, 2010]. This is translated into huge social costs.

When facing this situation of growing congestion in metropolitan areas, two main approaches exist to alleviate the problem: to increase the amount of infrastructures, or to improve the management of the existing ones. Usually, the construction of new metropolitan freeways is only a temporary solution, as involves more induced traffic (and more congestion), plus an increase in the territorial occupation, already severely harmed. It is not possible to maintain a continuous increase in the infrastructural supply, due to the funding limitations, but mainly due to lack of sustainability of this approach, given the difficulties in obtaining a respectful territorial integration. The capacity of territory to absorb new infrastructures is finite. These assertions do not mean that the current infrastructures must remain still. Some regions surely need more construction. And some others may need the reconstruction of the infrastructures, in order to adapt them to more sustainable urban transportation modes. Transportation infrastructures must be capable of continuously adaptation to sustainable society needs.

The alternative is the improvement in traffic management. This may imply actions to modulate the demand (e.g. increase of vehicles' occupation, smooth peak hour demands, derive demand to other transportation modes) usually by means of taxation or restriction. And, in addition, a better management of the supply (e.g. improving the lane usage, avoiding traffic instabilities by imposing variable speed limits, avoiding the capacity drop by imposing ramp metering, …). Is in this context where a common baseline requirement appears: traffic information. Traffic information is needed by traffic managers in order to set their operational policies. It is also needed by drivers in order to take their own decisions.

Travel time information appears as the key element. Travel time is the fundamental variable to provide traffic information, because is the best indicator of the level of service in a road stretch and it is completely understandable by all users. In fact, several surveys have shown that travel time is the worthiest information from the user point of view [Palen, 1997], as it allows him to decide in advance when is the best time to start a trip and the best routing option, or to modify this initial planning once on route. Travel time information is not only useful to the driver, but also to the road system operator as it is a basic knowledge to assess the operational management and planning of the network. Travel time forecasting allows the operator/manager to beat the incidents and operational problems in the system, while the real time information allows monitoring the evolution of these incidents. The network manager must not only provide travel time information to drivers, but also look after its variation and achieve high reliability of the infrastructure.

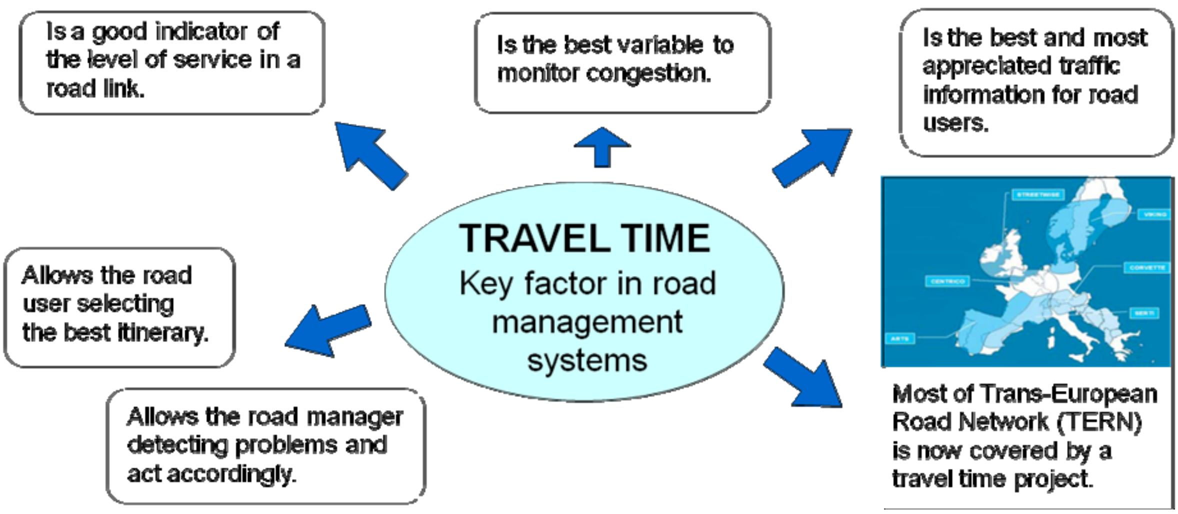


FIGURE 2 Travel time importance in the mobility management context.

In this context, travel time measurement and forecasting must be a priority objective for road network managers and operators.

2. TRAVEL TIME MEASUREMENT

The need for traffic data appeared as soon as motor vehicles became popular and the road network started to develop, almost 100 years ago. Originally, the objective was limited to measuring traffic volumes in order to know the usage of the network for planning purposes. Soon thereafter average speeds became necessary to estimate the level of service of the different links and enhance the planning process [Highway Research Board, 1950]. More recently, with the development of the information and communication systems, real time traffic data play the main role in the so called "Highway Advanced Traffic Management and Information Systems" (ATMS/ATIS) [Palen, 1997]. At every step of this evolution, the requirements in terms of equipment, communications, processors … in short, cost, to fulfill the traffic data needs increase enormously. In addition the vast extension of the highway network adds a huge inertia to the surveillance system already installed, so that it cannot make the most of the continuous technological and economical improvement of equipments. These factors lead to very heterogeneous levels of surveillance within a highway network. On the one hand the hot spots of the network (e.g. metropolitan freeways) may be densely monitored, with various types of equipments, including high-tech. On the other hand, some parts of the network may remain with some isolated out-dated detectors. All surveillance levels could be found in between. Obviously, metropolitan freeways concentrate most of the traffic of the highway system (and therefore most of the operational problems) and the intensive monitoring is completely justified [OECD/JTRC, 2010].

Highway travel time estimation reflects this reality. Although being the most primitive variable to be measured in any trip [Berechman, 2003], the systematic measurement of travel times in the network is quite recent. Travel time information is considered to be the key factor in the ATMS/ATIS, as it is widely accepted that the real time knowledge of highway travel times is the most informative traffic variable for both, drivers (it is easily understood, allows supporting trip decisions like changing the time of departure, switching routes or transportation modes…) and traffic agencies (it is a clear indicator of the level of service provided). This belated blossoming of highway travel time information systems may be result of the difficulties in the systematic direct measurement of travel times.

Basically there are two methodologies to measure travel time in a road link: the direct measurement and the indirect estimation. The direct travel time measurement is based on measuring the time interval that a particular vehicle takes to travel from one point to another. The alternative is the indirect travel time estimation from traffic flow characteristics (density, flow and speed), obtained for example from inductive loop detectors. To obtain travel time estimations from these last measurements some type of algorithm must be applied. Indirect measurement is especially interesting when direct measurement is extremely difficult or costly, or when all the monitoring equipment for indirect measurement is already available, while for direct measurement it is not.

Direct travel time measurement has only been carried out under regional travel time research projects, mainly in USA and Western Europe in limited corridors. Catalonia, and Spain are not an exception, and the unavailability of valid travel time databases is a reality. This lack of ground truth data has been a recurrent problem for practitioners in developing their road management schemes.

Taking into account that travel time is the preferred information for all the stakeholders (managers and users), filling the gap of ground truth data should be a main objective. Although being aware of the problem, efforts in direct measuring travel times have been very rare until recent times. This results from the traditional point of view in relation to the traffic data, considered only useful for pavement maintenance and planning objectives. As a consequence, all operational usage of traffic data relies on data gathered with very different objectives and with unsuitable accuracy requirements.

2.1. DIRECT TRAVEL TIME MEASUREMENT

Travel time can be directly measured from the vehicles travelling on the highway. One of the main properties of directly measured travel times is their spatial implication: the measurements include all the effects suffered by the vehicle while traveling along space. This is a main difference from indirect estimations, which, as will be seen next, are generally based on punctual measurements and spatial extrapolations.

In the direct measurement, the travel time is an individual property of each vehicle. This means that in order to obtain a representative average of a particular section travel time, a significant number of vehicles must be measured within the traffic stream. This usually represents a drawback in this type of measurements.

There are two main procedures to obtain travel time as a direct measure: identifying the vehicle in at least two control points or following the vehicle along with its trajectory. Both are analyzed in the following sections.

2.1.1. Vehicle Identification at Control Points

In the identification based techniques, the vehicle is identified at the entrance and at the exit of the stretch, and its passing time is stored. By pairing both registers travel time is directly obtained. Obviously, clock synchronization at control points is a major issue in order to warrant the accuracy of measurements. A collateral benefit of this travel time measurement method is that the individual identification of vehicles allows for constructing origin – destination matrices a key input for simulation models which usually is very difficult to obtain.

Travel time measurement in this case responds to a particular trip of a vehicle. Therefore, it has to be finished for being measured. This time alignment implication, analyzed in more detail in Appendix A2, involves some delay in the real time application of travel time information systems. This type of measurement is generally named Arrival Based Travel Time (ATT).

Another drawback of reidentification methods is that travel times are spatially captive of the control point locations. Travel times can only be obtained between two control points. No partial measurements can be obtained. It is evident that the number and location of control points plays an important role. For high control point densities, all the drawbacks are less dramatic: information delay is slight and sections are so short that no partial information is desired. However installation and maintenance costs increase. A trade-off must be reached. In this optimization process, not only the number of control points matter, also its location, in relation to the mobility patterns, has implications. Some particular locations may add more added value to the system, while others may be irrelevant [Sherali et al., 2006]. As an order of magnitude control points are located approximately every 2 km in metropolitan freeways with a high density of junctions while in interurban freeways with fewer junctions they are located up to 8 km apart [Turner et al. 1998].

In practice, more difficulties arise. For instance a common difficulty encountered when directly measuring travel times between control points is the elimination of "outliers". Only travel times related to the traffic conditions in the section should be considered. Other factors, not related to traffic, may introduce false delay to some vehicles (e.g. stopping for refueling or to have a break, or motorbikes dodging congestion). If the amount of measurements is high, it is not difficult to identify these outliers using standard statistical algorithms. However, if the identification rate is low, and given the high variance on section travel times introduced in case of congestion, it is particularly difficult to discriminate, for example an episode of growing congestion from a vehicle which has stopped. This issue is analyzed in Appendix A3.

The number of identifications, crucial to obtain a representative sample, depends hardly on the identification technology. Nowadays, all systems with the objective of a systematic application must rely on the AVI (Automated Vehicle Identification) systems. Manual identification should only be considered in small specific analysis in order to avoid the implementation costs of an automatic system. Some common AVI technologies include the license plate video recognition [Buisson, 2006; NYSI&SI, 1970] see Fig. 3, the reidentification of vehicle signatures from video cameras [Huang and Russell, 1997; MacCarley, 2001], the identification of toll tags in the case of turnpikes (traditional toll tickets in case of closed turnpikes – see Appendix A3 – or equipped with an electronic toll collection system – ETC – [Nishiuchi et al., 2006] see Fig 5.) or the innovative Bluetooth signature identification of on-board devices [Barceló et al., 2010]. Take into account that in some of these identification methods (e.g. license plates) it is possible to link the information with a particular person. This may imply additional legal difficulties in relation to privacy issues.

Note that the amount of travel time measurements depends on the technology but also on the configuration of the control points. A control point in the main highway trunk can be exhaustive in case it tries to identify all vehicles crossing the section, or partial where for example only some lanes are monitored. The number of identifications in each type of control point depends on the technology. Despite of technological malfunctioning, all vehicles could be identified in case of the license plate reading or closed toll highways scenarios. Only some of them in case of using the rest of the technologies described (depending on its penetration rate). However, note that the amount of measurements is not directly the amount of identifications, but the amount of pairings between control points. In case of an exhaustive control point, the number of pairings should be almost the same as the number of identifications, despite the technology used, and if there is no on/off ramp in between. In case there is one junction, the differences respond to the originated or finished trips in the junction, which provides the data for the origin – destination matrix construction. In case of partial control points, the same cannot be asserted. The number of pairings could be significantly lower than the identifications due to the amount of "leaks" in the system. In addition, in case of an in between junction, nothing can be said about the amount of input/output vehicles. The limitations of partial control points are therefore evident. If the control points are located many kilometers apart with a considerable number of in between junctions, the origin – destination properties of the method are lost, and the amount of pairings, even in the case of exhaustive control points, will depend on the number of junctions and the number of vehicles which travel the whole itinerary. This may be a very small part of the identifications, questioning the ability of the method for providing a continuous and significant average of travel times. As a general rule if long trips are predominant a smaller number of control points may suffice.

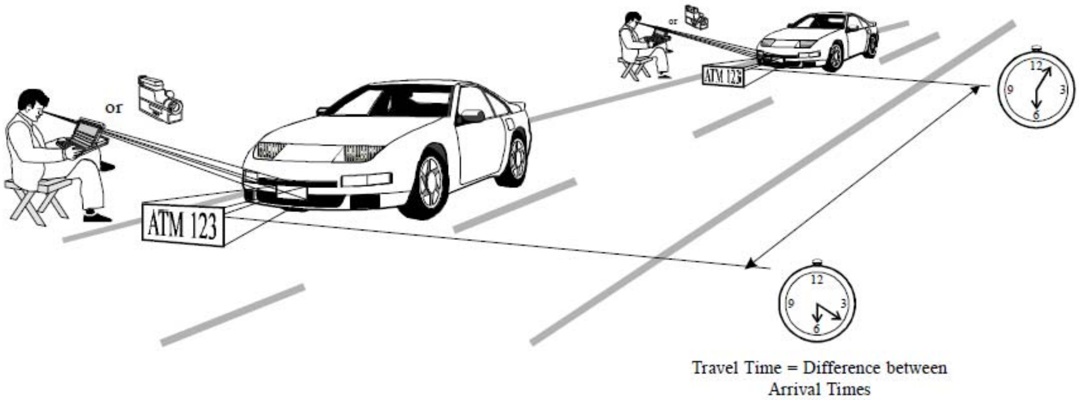


FIGURE 3 Travel time estimation from license plate recognition.
Source: [Turner et al., 1998].

As stated previously, and despite the technological feasibility of automated vehicle identification, highway traffic monitoring is, and will be for the next years, based on inductive loop detectors. This results from a reminiscence of the past, where the technological options were by far more limited, and the objectives to be fulfilled by the obtained traffic data more elementary. In addition, the huge inertia implied by the vast extension of the highway network, prevents from an extensive and fast technological update. Therefore, if a highway travel time information system aims to be generally implemented in the next years, it has to be based on loop detector data.

Considering this situation, researchers have attempted to improve the travel time measurement capabilities of loop detectors by trying to reidentify vehicles at the detector spot. This allows for the direct travel time measurement. The reidentification by means of the vehicles' electromagnetic signature [Abdulhai, 2003; Kuhne and Immes, 1993; Kwon, 2006] see Fig. 4, needs retrofitting loop detector hardware. An alternative is using the vehicles' distinctive length [Coifman and Cassidy, 2002; Coifman and Ergueta, 2003; Coifman and Krishnamurthya, 2007]. However, only rare vehicles are being reidentified using these methods when lane changing and in/out flows at ramps between detectors are considered. This may add some bias to the results in free flowing situations, but it may not imply a serious flaw in congested conditions where FIFO prevails. Other approaches [Lucas et al., 2004; Dailey, 1993; Petty et al., 1998], try to reidentify the platoon structure of a traffic stream, which is lost in congested periods, when travel time information is more valuable. Despite these limitations, the use of inductive loop detectors as an AVI equipment stands out as an active research field.

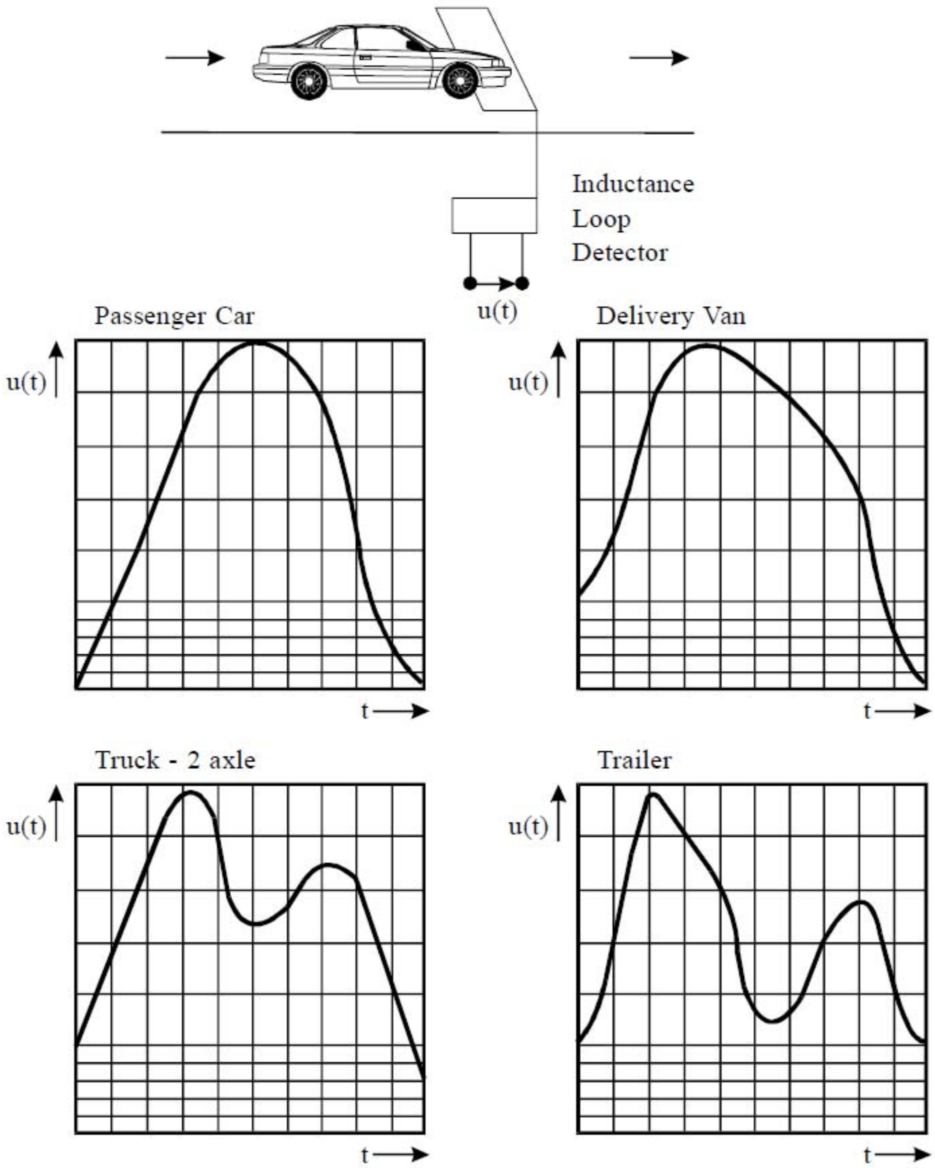


FIGURE 4 Electromagnetic signature of different vehicle types over an inductive loop detector. Source: [Turner et al., 1998].

2.1.1.1. Direct travel time measurement in toll highways

As it is one of the main objectives in the present thesis, a little more attention will be given at the particular case of using the equipment originally designed for collecting the toll at turnpikes for the direct travel time measurement. Travel time measurement here, like in the rest of technologies belonging to the present category, is obtained by means of vehicle reidentification.

In closed toll highways, where the toll paid by each vehicle depends on its particular origin and destination and on the application of a kilometric fee, the vehicle reidentification is straightforward. Note that in order to compute the fee, the vehicle must be reidentified. This is achieved by means of a toll ticket, real (i.e. a piece of paper) or virtual (i.e. a register in an electronic tag) where the precise time and location the vehicle enters and exits the highway is stored. The real time exploitation of these data, which is easy as the ticket travels with the vehicle like a baton, provides the desired travel time measurements and the origin – destination matrix for all vehicles. It is a valuable and exhaustive source of information usually not fully exploited. In addition to the general problems stated before for all methods based on vehicle reidentification, the method presented here suffers for a specific problem. The control points (i.e. the toll booths where the toll tickets are processed) are not located on the main highway trunk, but at the very far end of the on / off ramps, at junctions. This means that the measured travel time from an origin to a destination includes the main trunk travel time, but also the time required to travel along the on ramp, the time required to travel along the off ramp, plus the time required to pay the toll. In short trips this additional time is not negligible. Furthermore, if one constructs itinerary travel times by adding up different single section travel times, it will add up as many entrance and exit times as sections contains the itinerary. This process may result in a completely overestimated itinerary travel time. An interesting method for solving these complexities is presented in Appendix A3.

In contrast, open toll highways do not need to reidentify the vehicle to charge the toll. Every vehicle is charged the same toll, resulting from an average trip in the highway. In this case, toll plazas are strategically located in the main highway trunk every now and then. The entrance and exit time complications do not appear in this case, but the exhaustive travel time measurement is lost. Note that in order to reidentify vehicles in this open configuration it is needed that the vehicles travel across, at least, two toll plazas, and leave a trace at each payment. This is the achieved in case of electronic payment (e.g. credit card or electronic toll collection tag - ETC). Short trips or cash users will travel unidentified. This will not be a major drawback as 80% of toll highway users in Catalan turnpikes are actually using electronic payment methods, which should be enough to obtain a representative average travel time measurement. In addition, in order to obtain the measurements in real time, the communications requirements are more challenging, as there is no baton travelling with the vehicle. Finally note that main toll plazas are usually located several kilometers apart. Probably, a finer discretization of travel time measurements will be desired. This can be achieved by installing ad hoc overhead gantries capable of reading the electronic toll collection tags “on route”, see Fig 5. This technology has been proved capable of identifying all ETC equipped vehicles even if they are travelling simultaneously in different lanes at speeds as high as 180 km/h. The main issue here is the penetration rate of these electronic tags, which nowadays is around 20%, as a result of marketing policies (e.g. free dissemination of tags for frequent users) and priority benefits at toll plazas, where usually equipped vehicles cross undisturbed.

TABLE 1 Direct Travel Time Measurement in Toll Highways: Open vs Closed Toll Configurations

Closed Toll Configuration	Open Toll Configuration
Vehicle reidentification by means of Toll Ticket (real or virtual).	Vehicle reidentification by means of credit card number or ETC identification.
Exhaustive sample.	Sample is made up only by electronic payment users who travel across two or more toll plazas.
Toll ticket travels with the vehicle.	Communication system between toll plazas needed for pairing.
All on / off ramps are control points.	Main trunk toll plazas are the control points. Additional ad hoc identification gantries may be needed.
Travel times affected by entrance / exit times.	Main trunk travel times are measured.

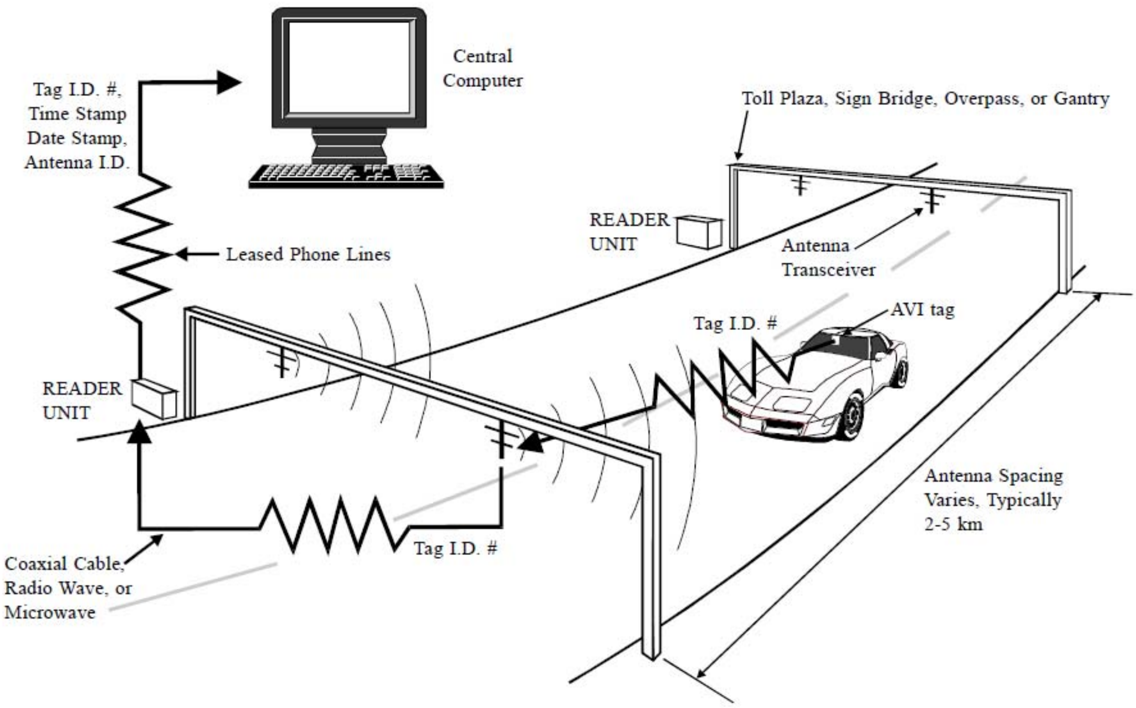


FIGURE 5 Travel time estimation from ETC reidentification.
Source: [Turner et al., 1998].

2.1.2. Vehicle Tracking

The second group of techniques for the direct travel time measurement is related to the vehicle tracking concept. In this case the vehicles act as probes and record their position every defined time interval. There are not control points or infrastructure related monitoring equipment. The vehicles become active sensors, instead of being passive as in the previous case, and compute travel times by continuously tracking its trajectory.

Historically the vehicles used as probes have been dedicated cars. These probe cars traveled with the only purpose of gathering travel time data. This is the case of traditional probe car data. In order to obtain a continuous flow of travel time measurements to be used as a real time information system in a highway corridor, the amount of ad hoc probe cars would be huge (e. g. probe cars at 3 minute headways), and not sustainable in the long term. This traditional method is restricted to case specific studies.

The development of ITS (Intelligent Transportation Systems) and the popularization of GPS (Global Positioning System) technologies has favored that each vehicle which travels in a particular road could be a potential probe vehicle. These GPS equipped vehicles are nowadays regular transportation fleets (like buses, parcel companies vans, roadside assistance vehicles, patrol service vehicles, taxi cabs, ...) which travel regularly over a selected route and who have at their disposal an active management center to elaborate the necessary data treatment. Take into account that the specificities of these fleets may bias the sample. Currently, the weak spot of the system is the data location transmission from the vehicle to the control center, usually using radio channels (e.g. GPRS system). Surely these schemes (see Fig 6.) will be expanded in next future to every particular car who volunteers (this will solve the privacy issues). This extensive and automatic version of traditional probe cars needs a high penetration of on board GPS devices plus the collaboration of the driver in order to transmit the data. This may be possible with the future popularization of GPS-enabled smartphones [Herrera et al., 2010] or by directly geo-locating the phone signal [Yim and Crayford, 2006], although this last option does not seem to provide the necessary location accuracy.

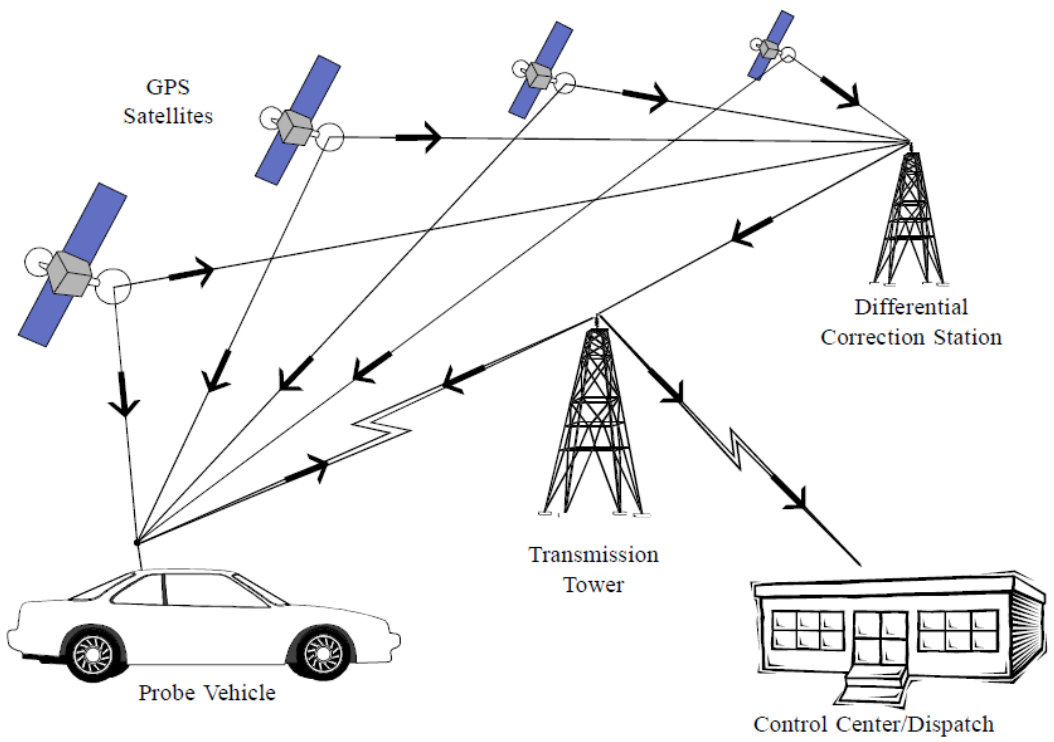


FIGURE 6 Travel time estimation from GPS tracking. Source: [Turner et al., 1998].

TABLE 2 Direct Travel Time Measurement Methods: Benefits and Drawbacks

		Benefits	Drawbacks
AVI methods	License plate, toll ticket, digital imaging, electromagnetic signature, Bluetooth devices	• Potential exhaustive sample (license plates, toll tickets) or high penetration rates (ETC devices). • Continuous flow of travel time measurements. • Promising reidentification methods based on loop detector data.	• Site specific and costly. The highway must be equipped with technologically advanced detectors. • Clock sincronization between control points. • Partial travel measurements in between control points are not available. • Privacy issues. • Difficulties in outlier detection in case of few measurements.
Vehicle tracking	Traditional probe cars, GPS tracking, cell phones geo-location	• Not infrastructure related. • Not spatial captive. The travel time measurement can be obtained between any two desired points. • GPS equipped smart phones imply new opportunities for the method.	• Traditional probe cars only useful for specific studies, but not for a systematic implementation. • Currently based on specific fleets. This implies small and biased samples. • Needs high penetration of GPS equipped vehicles. • Large amount of data must be transmitted from vehicles to a data management center. • Cell phone geo-location does not still provide enough accuracy.

2.2. INDIRECT TRAVEL TIME ESTIMATION

Indirect travel time estimation is based in the measurement of fundamental traffic flow variables (flow, speed and density) in a particular spot of a highway and the extrapolation, using some type of algorithm, of these point measurements to the spatial implications of travel times. These fundamental variables capture the whole physical traffic process, and so it should be possible to derive any other variable from them, in special travel time. Loop detectors are, by far, the most widely spread technology to collect flow, speed and occupancy (the proxy for the traffic density) of a traffic stream. Take into account that single loop detectors only collect flow, and occupancy, while speed must be approximated by usually assuming an average constant vehicle length. Besides, dual loop detectors are capable of measuring all traffic variables (i.e. flow, speed and occupancy). This issue is addressed in detail in Appendix A1.

In general, using loop detector data always imply the same problem: data quality. Flow, speed and occupancy measured on a highway spot over short aggregation periods (of the order of some few minutes) suffer from important fluctuations, particularly when measuring instable stop and go traffic. Huge variations are possible in short time periods, and still the measurements are correct. This makes extremely difficult to detect erroneous loop detector data on real time, unless very abnormal data is measured. In contrast, smoothing or aggregating data over longer time periods (e.g. one day) makes the detector malfunctioning a lot more evident and easily detectable [Chen et al., 2003]. Fortunately this is enough in most cases (although several hours of undetectable malfunctioning may be unavoidable), because loop errors do not arise randomly in between correct measurements. Usually, detector failures respond to the breaking down of some part of the detector, and not to a circumstantial malfunctioning. This means that erroneous data usually respond to "stuck" measurements during long periods of time (days, weeks, months or even years) until the detector is repaired.

Loop detectors are an old technology, which require intensive and costly maintenance as they are exposed to severe conditions (i.e. traffic, extreme hot, extreme cold, water, road works …). If this maintenance work is neglected, frequent breakdowns occur. This implies great holes in the database, with the consequent implications in the algorithms which rely on these data. These algorithms should be prepared to deal with this frequent missing data. Fortunately, travel time algorithms rely on sectional measures. This means that, in general, no lane specific data is used, but data aggregated over all lanes. Detectors in different lanes can be considered as redundant (this is a simplistic approach, because traffic each lane has its own characteristics, and some analysis or applications need to measure this specific lane behavior). This implies that, in case some detector in the section is not functioning, it is easier to reconstruct the data of the whole section by using neighboring detectors, if they are properly working. Therefore, a complete failure only happens when all the detectors in the measurement section fail all together. This is not as rare as it may seem, because it is only necessary the failure of the roadside unit which steers the measurements of all the detectors in the section.

From the previous paragraphs must be concluded that data quality assessment and data reconstruction processes are a necessary first step in the utilization of loop detector data for whichever desired objective.

For the particular objective of indirect travel time estimation from standard loop detector data, two basic methodologies can be distinguished: the estimation from point speed measurements and the estimation from cumulative count curves. Both are analyzed in the following sections.

2.2.1. Indirect Travel Time Estimation from Point Speed Measurement

The first and most widely used approach for estimating travel times from loop detector data is the spot speed algorithm. This method is based in the extrapolation of the point speed measurement at the loop location to a complete freeway section The hypothesis considered in the application of this algorithm are that punctually measured traffic stream characteristics are representative of the whole assigned section.

First of all, this method relies on a speed measurement, which therefore needs to be accurate. Single loop detectors' approximations to speed are not enough, as the assumption of constant average vehicle length does not provide the necessary accuracy.

In order to solve this problem, traffic agencies have tended to install detectors in a double loop configuration (i.e. speed traps) to accurately measure the vehicles' speed. In addition, in order to estimate travel times from averaged speeds, resulting from measured individual speeds over a time period in a particular spot of the highway, this averaging must report the space – mean speeds. The problem here is that the standard loop detector data treatment reports the time – mean speeds, and the raw data useful to compute space – mean speeds is eliminated. In this situation the obtained travel times with the proposed method would be generally underestimated. In order to solve this common drawback a methodology to obtain space – mean speeds from commonly used loop detector data aggregations is presented in Appendix A1.

In relation to the spatial representativeness of punctual measurements, different agencies use different speed interpolation methods between detectors (e.g. constant, linear, quadratic …) trying to better approximate the traffic conditions in the stretch, but without taking into account traffic dynamics. As it is proved in Appendix A2, all of them are simplistic and inaccurate in congested conditions. In view of these limitations, and in order to obtain meaningful travel time estimation using these methods detector density must be extremely high. This has forced a process of continuous increase in the loop detector density. While one single detector in between junctions was enough to measure average daily traffic (ADT) volumes, at least one double detector every 500m is necessary to compute accurate travel times using these methods (see Fig. 7). Obviously this is not economically feasible for the whole highway network, and can only be achieved in some privileged stretches of metropolitan freeways. In conclusion, there are lots of kilometers of interurban highways with low surveillance density (e.g. typically one detector per section between junctions to fulfill the ADT requirements) where the systematic travel time measurement using the existing equipment is devoid of an adequate method.

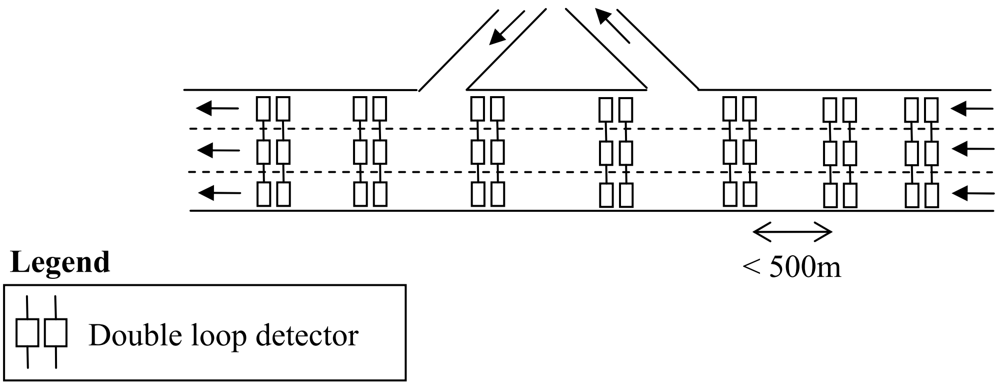


FIGURE 7 Spot speed algorithm required surveillance configuration.

Finally, it is worth mentioning that itinerary travel times are frequently obtained by the addition of several section travel times (where each of these sections is defined by two loop detectors). Each section travel time is obtained from the average speed measured over the last few minutes (as a frequent update is desired). This means that the obtained itinerary travel time is not trajectory related. It is a like a picture of the actual travel times on the stretch. It is possible that any vehicle trajectory responds to this travel time. This temporal alignment concept of travel time is frequently known as ITT (Instantaneous Travel Time) and it may be considered to be the best approximation to the desired real time "future" information, without falling under the uncertainties of forecasting.

2.2.2. Indirect Travel Time Estimation from Cumulative Count Curves

The alternative to avoid the required high surveillance density and the lack of accuracy of the spot speed algorithm in congested situations relies on a cumulative count balance algorithm, which estimates travel times directly from loop detector count measurements. The algorithm uses the entrance and exit flows in the highway stretch to calculate the travel time using the conservation of vehicles' equation. In order to apply the proposed method, the monitoring of the section under analysis has to be "closed", in the sense that all the on/off ramps must be monitored, in addition to some main trunk loop detectors (e.g. typically one on every section between junctions) (see Fig. 8). Under these conditions the vehicle accumulation in the section can be computed.

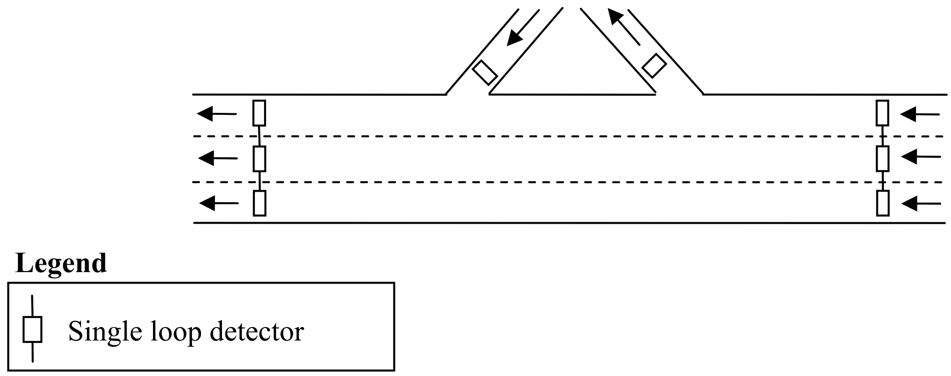


FIGURE 8 Cumulative flow balance algorithm required surveillance configuration.

Despite the apparent potential and simplicity of the method, it is hardly used in practice. This may result from the oversight of the researchers' community to the practical problems which appear in the implementation of the method. For instance, from the author knowledge, contributions are not found in the literature analyzing in detail the problematic detector drift phenomenon, which accumulate in the input / output curves until they become meaningless. The effects of inner section input / output flows at junctions are not treated either. These issues remain for further research.

Unlike some other methods presented in this thesis report which are analyzed in detail in the correspondent appendix, the cumulative count curve detailed description is not provided in any appendix, despite a brief introduction in Appendix A4 where it is applied. Therefore it is found convenient to include here a detailed description of the method, although it may seem excessively detailed and analytical within the broad analysis performed in this report. In this case, next subsections can be skipped without loss of continuity.

2.2.2.1. Basic concepts for the travel time estimation from N-curves

Given a location "*x*" on the highway, one can consecutively count and accumulate the vehicles passing the location. This process defines a function "*N(x,t)*" that gives the cumulative number of vehicles to have passed location "*x*" by time "*t*", starting from an arbitrary initial reference vehicle, which has passed at "*t=0*". By cumulative curve (or equivalently N-curve) it is meant the graph of such a function, which will be always non-decreasing with "t".

N-curves are a convenient way of analyzing traffic data [Makigami et al. 1971]. The works of Newell (1982, 1993) in queuing theory and in traffic flow theory demonstrated their full potential and simple geometric interpretations. Note that if one draws the curves "$N(x_u,t)$" and "$N(x_d,t)$" for two locations "x_u" and "x_d" (e.g. upstream and downstream detector locations) on the same graph as in Fig. 9, the vertical distance between the curves at time "t^*", "$N(x_u,t^*) - N(x_d,t^*)$", represents the number of vehicles between "x_u" and "x_d" (i.e. the vehicle accumulation), provided that vehicles do not enter or leave the intervening space (i.e. vehicle conservation). At the same time, the horizontal distance between the curves at height "j" represents the trip time between "x_u" and "x_d" of the j^{th} vehicle, if the vehicles do not pass each other (i.e. FIFO system).

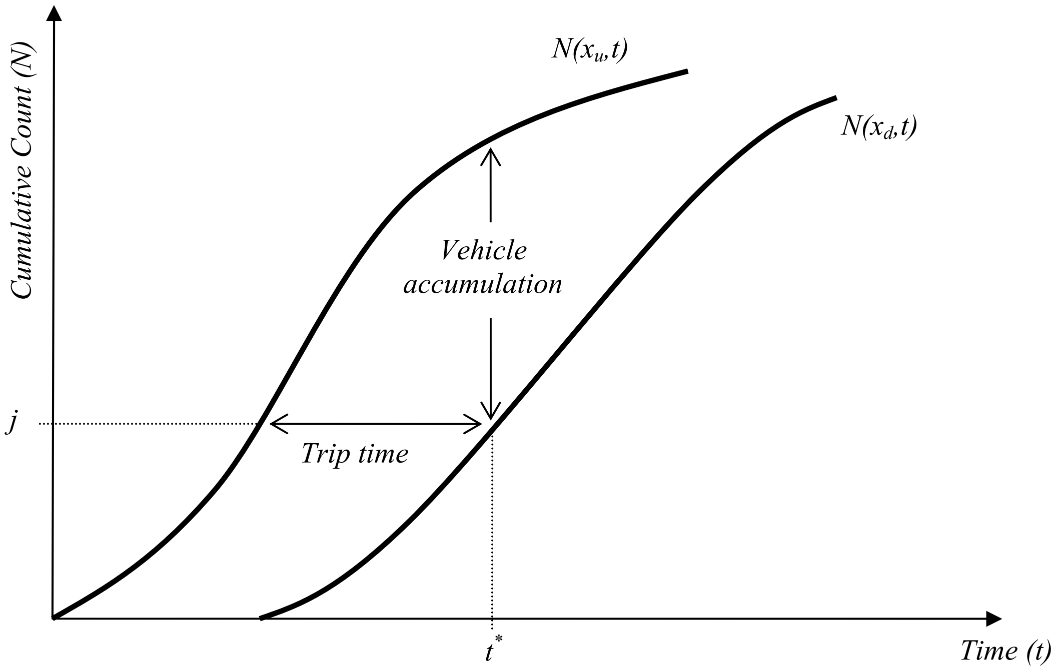


FIGURE 9 Graphical interpretation of cumulative curves at two locations.

A slight modification of the previous concepts can be introduced by considering "T_f" to be the free flow travel time between "x_u" and "x_d", which can be assumed to be approximately the same for all the vehicles. In this case, the "virtual" downstream cumulative curve, "$V(x_d,t)$", can be defined as the number of vehicles that would have been seen at "x_d" by time "t" if all vehicles would have travelled undisrupted. This curve can be constructed by simply translating to the right "$N(x_u,t)$" by an amount "T_f", because "$V(x_d,t) = N(x_u,t-T_f)$". Obviously, if there is no delay "$V(x_d,t) = N(x_d,t)$". The inclusion of "$V(x_d,t)$" allows obtaining the part of the "j^{th}" vehicle trip time corresponding to the delay (i.e. the difference between actual travel time and free flow travel time) as the horizontal distance between "$V(x_d,t)$" and "$N(x_d,t)$" in case of a FIFO system (see Fig. 10). Note that delay is not the same as time spend in the queue [Daganzo, 1983]. The latter is always bigger, because it includes the delay plus the time vehicles would take to travel along the physical length of the queue at free flow speed. Similarly, the vertical separation between "$V(x_d,t^*)$" and "$N(x_d,t^*)$", namely the vehicle excess accumulation, is smaller than the number of vehicles in the physical queue, because queues take up space. Abstractly, the excess accumulation can be seen as the number of vehicles that would form the queue in case vehicles queue one on the top of the other (i.e. without taking physical space). Daganzo (1997) further elaborates these concepts.

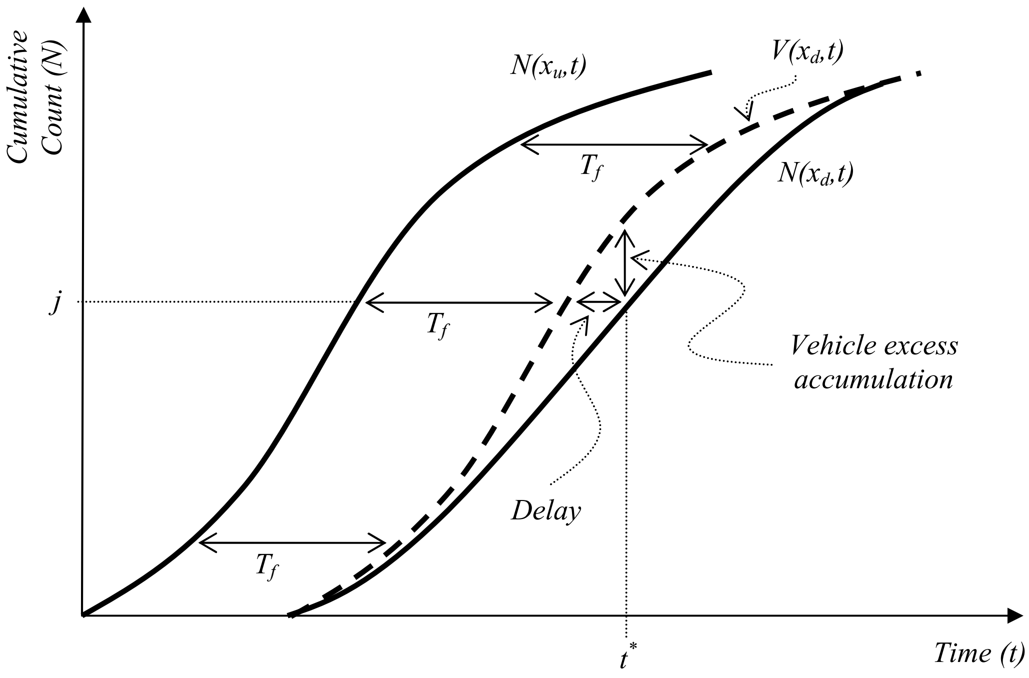


FIGURE 10 Graphical derivation of delay.

2.2.2.2. Different estimation processes lead to different average travel time definitions

Generally, average magnitudes of the traffic stream measured and averaged across a time-space region are more informative than the behavior of a particular vehicle. At the same time, loop detector measurements are not usually available in a per vehicle basis, but aggregated or averaged over short time periods of duration "Δt", which may range from the 30sec common in North America to several minutes in Europe (see Appendix A1). Therefore, cumulative count curves are constructed by linearly interpolating between discrete count measurements every "Δt".

In this context, travel times from cumulative curves can be averaged in several ways. The fact that different averaging procedures lead to conceptually different travel time estimations is not taken by the related literature [Nam and Drew, 1996; Oh et al., 2003; van Arem et al., 1997], and plays an important role in the assessment of the method.

For instance, it should be clear that the area enclosed between "$N(x_u,t)$" and "$N(x_d,t)$" from time "t_{i-1}" to time "t_i" (where "$t_i - t_{i-1} = \Delta t$") is the total time travelled by vehicles in the (x_u, x_d) - (t_{i-1},t_i) space – time region (see Fig 11). Equivalently, in case of considering the curve "$V(x_d,t)$" instead of "$N(x_u,t)$" this area would correspond to the total delay suffered by the vehicles in the period (the concept of this equivalence is valid for the rest of the definitions of the section). This statement is true even if there is passing within the traffic stream. This procedure may be useful for determining these aggregate measurements in the region, which for instance may allow continuously computing the vehicle hours travelled (VHT). However it is not adequate in order to compute average travel times on the highway section, as not all the vehicles travel the whole section in the time period (it is even possible that none of the vehicles travel from "x_u" to "x_d" in the period if the travel time is long enough in relation to "Δt").

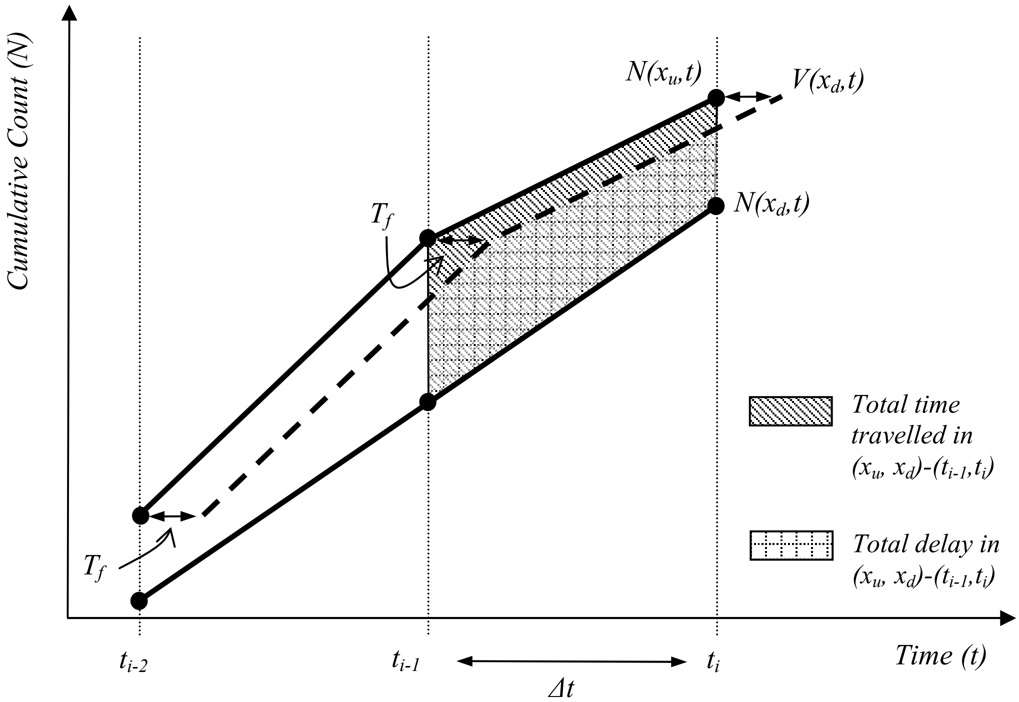
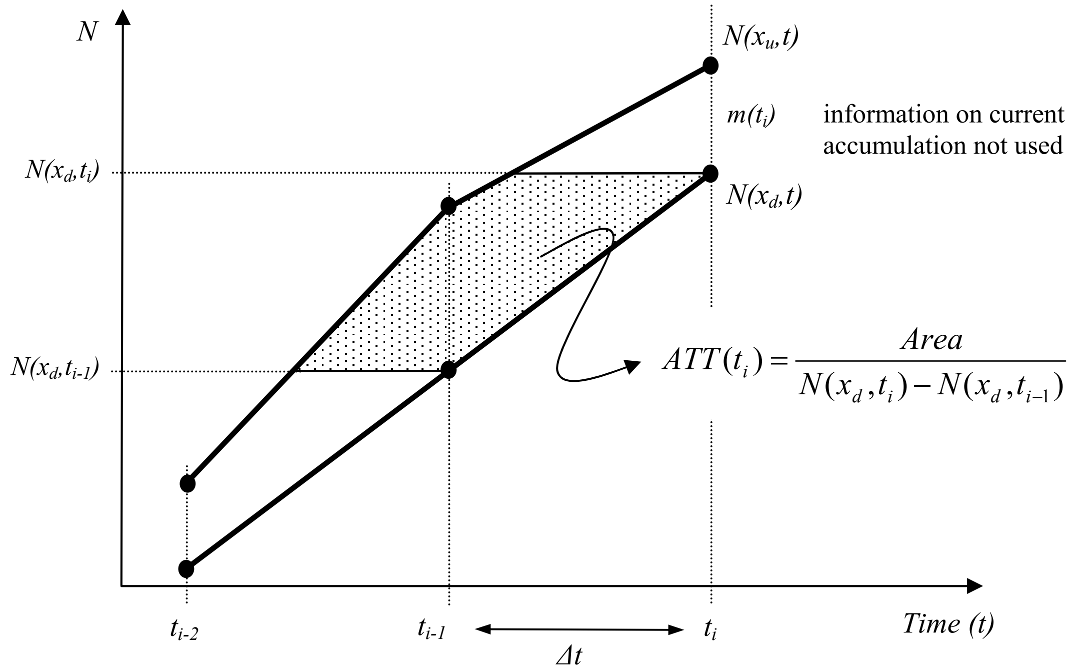


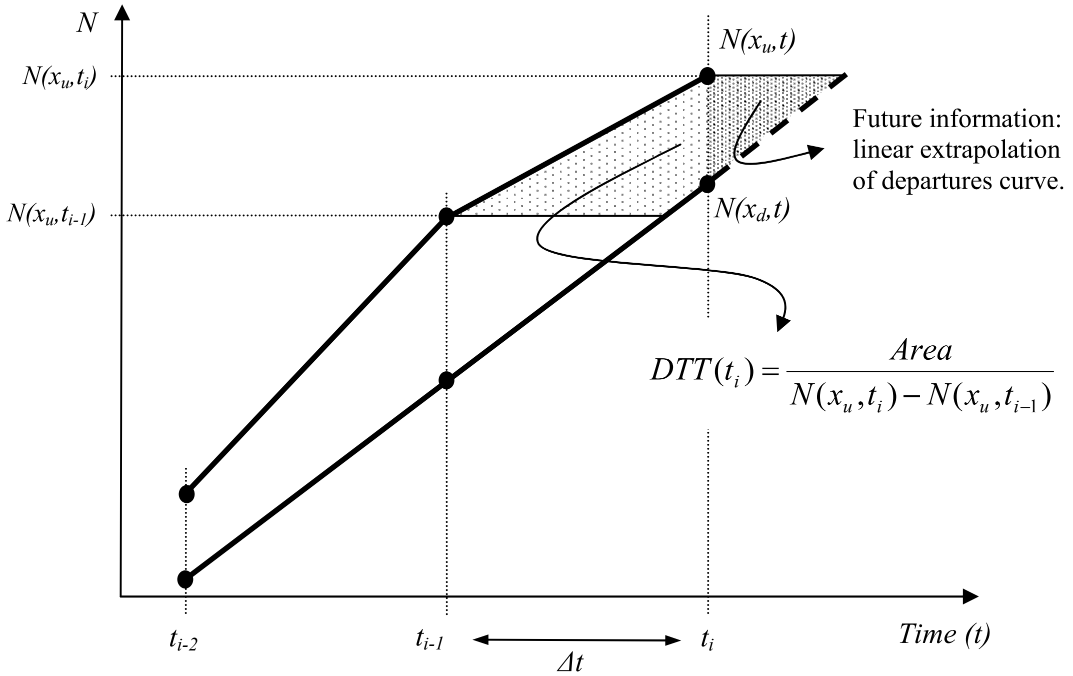
FIGURE 11 Total time travelled and total delay suffered by vehicles in the (x_u, x_d) - (t_i,t_{i+1}) space – time region.

For the objective of computing the average travel time of a group of vehicles between "x_u" and "x_d", the area enclosed between "$N(x_u,t)$" and "$N(x_d,t)$" should be limited by horizontal limits in the N-t plot corresponding to the first and last vehicle in the group considered. This area computes the total travel time of the group of vehicles, and must be divided by the number of vehicles in the group to obtain average travel times. Note that this procedure is equivalent to computing the arithmetic average of each individual vehicle's travel time in the group. These assertions are only true if there is no passing within the traffic stream.

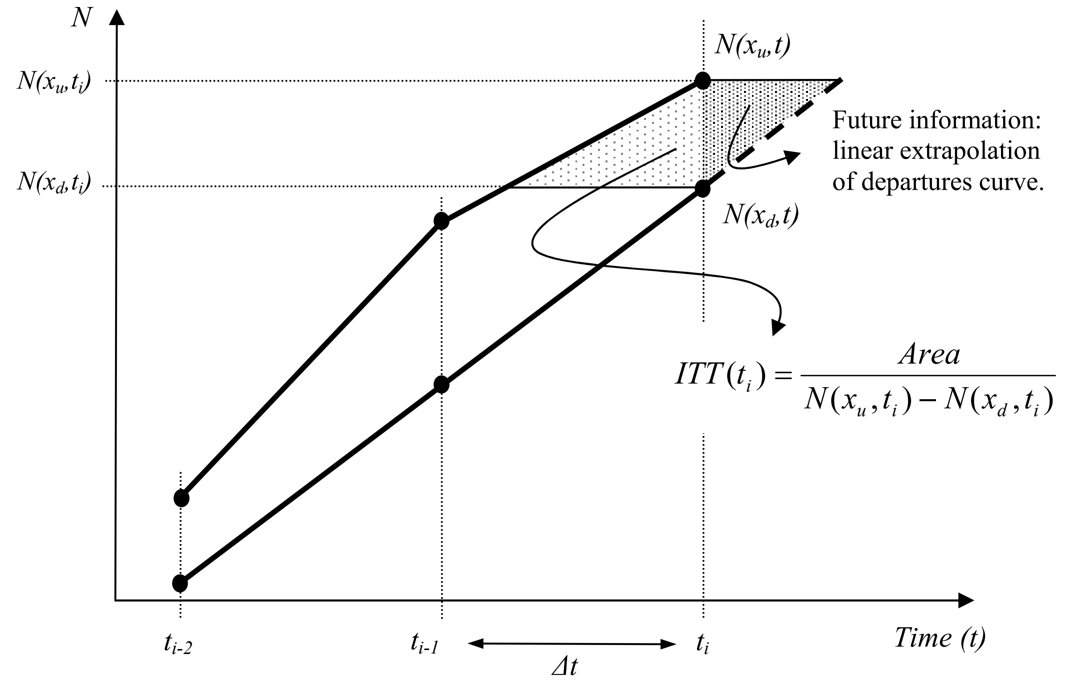
Some care must be taken in the selection of the group of vehicles. Different selections lead to conceptually different average travel times (see Fig 12). For instance if the vehicles considered are those which reach "x_d" during "Δt", arrival based travel times (ATT) would be obtained. In contrast, if the considered vehicles are those which depart from "x_u" during "Δt", the average travel times obtained would be departure based (DTT). Note that some extrapolation to future information is needed in this last case. The usual assumption is to consider that traffic conditions will remain constant in the very next future, which is translated to a linear extrapolation of the departure curve. Instantaneous travel times (ITT) would be obtained if only the vehicles contained in between "x_u" and "x_d" at time "t_i" are considered. Again, future information is needed. Other particular selections may lead to other results, like Nam and Drew (1996) where only the vehicles which completed the whole trip between "x_u" and "x_d" during "Δt" were considered. Obviously this group of vehicles will only exist in the case that individual travel times in the stretch are significantly shorter than "Δt". Assuming that "Δt" must be small, in order to track travel time variations and to provide frequent updated information, this travel time definition may only exist in short and free flowing stretches. It is evident the limited usability of this definition.



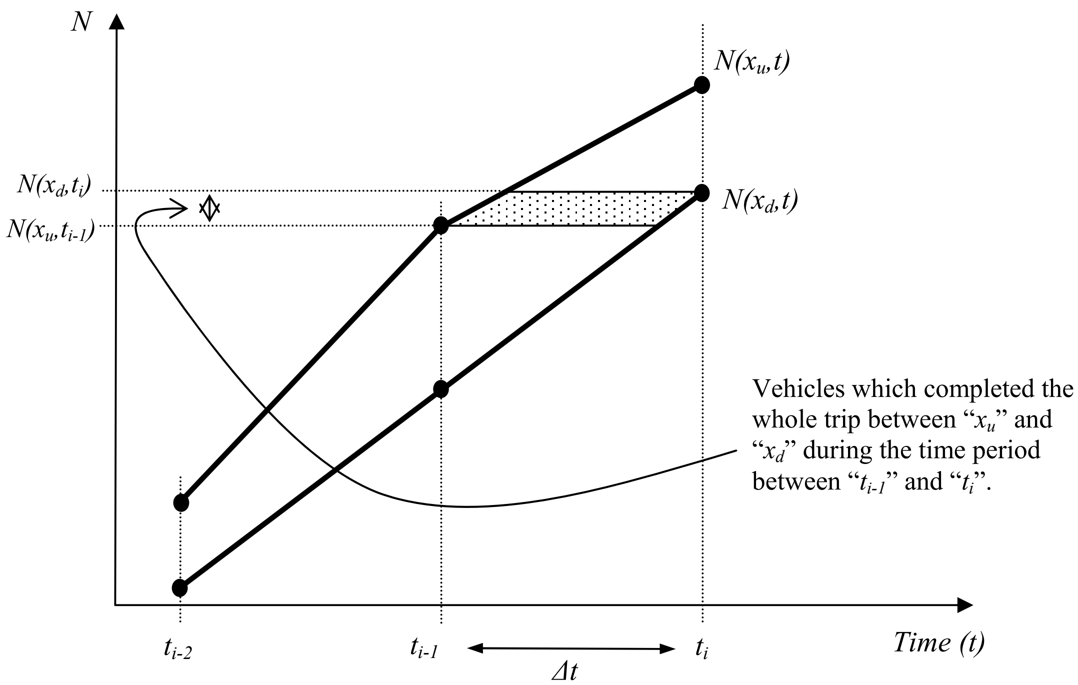
a)



b)



c)



d)

FIGURE 12 Average travel time definitions from N-curves.

2.2.2.3. Major drawbacks in using N-curves for travel time estimation

Construction of cumulative count curves at detector locations are a powerful tool for computing travel times. In practice, the unique requirement is the conservation of vehicles, which means that the count at all on/off ramps must be monitored. The method is independent of the physical characteristics of the section, which may influence the queue location but not the average travel time, and does need neither any type of calibration nor any empirical parameter. This makes the method very appealing for travel time estimation, particularly in low surveillance environments.

Nevertheless some complications appear in practical implementations of the method. These are described next.

The Effects of Passing

In case passing takes place within the traffic stream, individual vehicle's travel time cannot be obtained. This is due to the fact that in the cumulative arrivals and departures curves a particular vehicle is identified by their ordinal position in the traffic stream. In case of passing, vehicles do not maintain their order, and the vehicle downstream reidentification is not exact. This context may be physically interpreted as that the cumulative curve does not count particular vehicles but specific positions within the traffic stream [Daganzo, 1997]. Under this assumption, with every passing maneuver, vehicles change "*N*" tags as they switch positions. This eliminates the difficulty of none monotonically increasing N-curves after a passing maneuver, as position tags will never pass each other. Of course this change prevents using the N-curves for tracking individual vehicles.

Although individual vehicle's travel time cannot be accurately obtained in case of passing traffic, there are some situations where average travel times can be considered approximately true, despite the passing. Note that as a result of the passing, it is possible that some vehicles in the group considered only entered or only exited the highway section. If the total travel time of those vehicles is a very small fraction of the total travel time, the passing can be considered as insignificant, and the results approximately true. In general, this exception holds, as Muñoz and Daganzo (2002) proved that freeway traffic can be considered a FIFO system (i.e. non significant passing) as multilane behavior (i.e. significant passing) only persist for a few kilometers upstream of an off-ramp. Therefore, passing in a highway section without any junction does not imply a serious drawback to the method.

Inner Section On/Off Ramps

Another problem arises in case of an inner section on/off ramp. In this case two families of vehicles can be clearly identified: vehicles which travel along the whole stretch and vehicles which use the on/off ramp and therefore their trip on the highway stretch is uncompleted. The travel time estimation method must be capable of computing the travel time on the whole stretch. But if the on/off ramp counts are considered as standard input/outputs, only the average travel time across all vehicles will be obtained. Considering that partial trips will usually experience shorter travel times, the average travel time for the whole section will be underestimated. The bias will be bigger for bigger partial flows.

Detector Count Drift

It is widely known and demonstrated that detector counts in reality are not perfect [Nam and Drew, 1996; Oh et al., 2003; van Arem et al., 1997]. Each particular detector has its own and different trend to undercount. This small amount of "lost" vehicles does not have any important implication in the computation of average flows over short time intervals. It is when cumulative curves are constructed for long time periods that the overall amount of lost vehicles stands out as relevant. Recall that average travel times are sensitive to the differences in two cumulative count curves (i.e. the vehicle accumulation in the highway section). This accumulation is a very small fraction of the total vehicle cumulative count of either curve. This also implies some problems in the visualization scales of the N-t plots [Cassidy and Windover, 1995]. This means that small differences in the fraction of lost vehicles in both detectors will have a dramatic effect over the computed vehicle accumulations. The accumulated errors could be by far larger than the measurement objective (i.e. the vehicles accumulation), implying the obtained average travel times to be completely flawed. This issue has been partially treated in the related literature [Nam and Drew, 1996; Oh et al., 2003; van Arem et al., 1997].

N-Curves Initialization

The drift phenomena in cumulative curves advices a frequent reset of the curves in order to avoid the continuously growing bias in the vehicles accumulation. At every initialization, the initial accumulation is needed (see Fig. 13) which it is not easily measured with point detectors. A first simplistic approach could be to reset at a time when there are no vehicles in the section (i.e. null accumulation). Despite the difficulties in determining this situation, in relatively long highway sections and heavy traffic demands (when travel time information has more interest) this situation would be rare, and would not respond to the required frequent reset. An alternative could be to consider a previously known initial travel time, "tt_0", as a proxy for the initial accumulation, "m_0", given:

$$m_0 = q_0 \cdot tt_0 \tag{1}$$

Where "q_0" stands for the initial flow. "q_0" is highly insensitive to the detector drift and can be obtained as the upstream count in the previous time interval divided by "Δt". The difficulties here reside in the accurate a priori knowledge of "tt_0". This knowledge can only be gained in case of free flowing traffic. In this situation the average travel time is known, as "tt_0" can be accurately obtained by assuming that the detectors measured average speed are representative of the whole stretch. It is recommended to obtain the average speed in the stretch as the harmonic average of the average speeds measured at "x_u" and "x_d" (i.e. midpoint algorithm – see Appendix A2).

Therefore the reset of the cumulative curves can be achieved as frequently as desired, provided that at the reset instant free flowing conditions prevail in the whole stretch.

Note that in case of using the virtual arrivals curve "$V(x_d,t)$" instead of "$N(x_u,t)$", and computing delays instead of travel times, the process detailed before could be simplified (see Fig. 13). In a free flowing situation the vehicles' excess accumulation is null. In addition if it is assumed that traffic flows evolve smoothly in the short time periods considered and that free flow travel times are small in relation to the desired precision in the travel time estimation (of the order of 1 minute), then the N-curve initialization is as simple as directly construct in the same coordinate N-t axis the curves "$V(x_d,t)$" and "$N(x_d,t)$". This last common process is adequate in case the length of the highway section is shorter than 2 km.

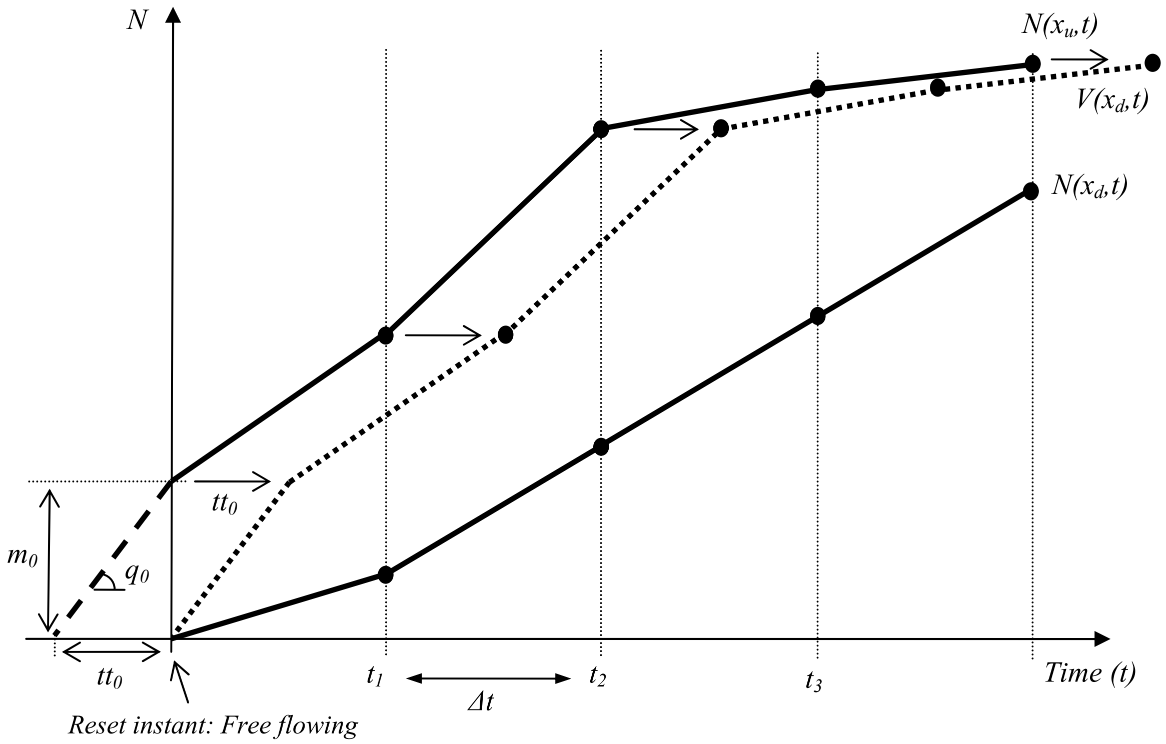


FIGURE 13 N-curves initialization.

N-Curves Linear Interpolation

The use of piecewise linear approximations of the cumulative count curves, necessary given the discrete periodicity used by loop detectors to report data (i.e. every "Δt"), adds some error to the method. Nevertheless, given that the interpolation limits are a small time aside, "Δt", and the smooth evolution of traffic in this small time periods, the errors committed can be considered as insignificant.

2.2.3. Loop Detectors and Travel Times: Summary

The main conclusion of this section is that there is only one benefit of using loop detector data to compute travel times. This is, that loop detectors are already installed out there, and the marginal cost of this new application is small. And of course this is a major benefit. Apart from this, the rest are problems, which different methods suffer more or less. Table 3 summarizes the different problems that affect the different methods. This means that loop detectors usage as travel time estimation equipment is result of reminiscence of the past and given the inertia of the already installed equipment.

Therefore, in new specific projects for systematic travel time estimation, where prevailing equipment is null, traditional loop detectors should not be considered as an option.

Surely, in the curse of time, loop detector technology will be modernized. It is a perception of the author that given the structure of traffic monitoring systems and the interests of worldwide traffic agencies, the new detectors which finally will beat traditional loops must be capable of doing (at least) exactly the same functions which old ones do (count, occupancy and speed for all vehicles). They will be installed at exactly the same spots. So that from the traffic management center point of view in the daily life nothing will be changed. But of course the new detectors will be cheap, will benefit from a long lifespan, will be less intrusive into the pavement, will require low maintenance, will communicate wireless, will not need a power supply, and will be less prone to malfunctioning. The MeMS detectors developed by UC Berkeley engineers are a good example [Hill et al. 2000]. This will be the time to incorporate to these detectors more advanced features, like reidentification capabilities.

TABLE 3 Methods of Travel Time Estimation From Loop Detectors: Benefits and Drawbacks

	Benefits	Drawbacks
Reidentification of vehicles (electromagnetic signature, characteristic length or platoon structure)	• Spatial measurement (direct travel time measurement). • Active research with promising results.	• Needs detector retrofitting (new hardware). • In case of in between junctions, and in order to avoid false reidentifications, only rare vehicles are identified. Biased measurement in free flow, but not a serious flaw in congested conditions. • Platoon structure is lost in congestion. • Clock sincronization between detectors.
Spot speed methods	• Instant travel time measurement. • Intuitive algorithm. • Directly applicable with current detectors in almost any configuration. • Intelligent smoothing helps in improving the estimations.	• For accurate measurement needs intensive (and costly) monitoring (i.e. one detector every 500m) • For accurate measurement needs the computation of space mean speeds. • Even fulfilling the previous conditions, travel time estimations will be flawed in congestion conditions.
Cumulative count curve method	• Spatial measurement (it is like a direct measurement under FIFO assumption). • Instant and trajectory based travel time measurements available. • Directly applicable with current detector technology.	• Needs closed configuration of the monitoring (all input / outputs being monitored). • Detector drift is a major problem. • Methods need to account for inner section input / output flows. • Care must be taken with the necessary frequent reset.

In addition to Table 3 contents, there is one problem which affects all methods. This is the frequent malfunctioning of loop detectors, which is translated into empty holes in the database.

Being fair, there is also one benefit of loop detector data, and this is the instantaneity. This property, which can also be obtained with the direct tracking of vehicles, but not in the AVI methods, refers to the ability of the measurement equipment of providing an instant picture of what is happening, avoiding information delays.

3. DATA FUSION AND TRAVEL TIME FORECASTING

Forecasting is not the main objective of this thesis. This must be stated here. The word "forecasting" has connotations of uncertainty, chance, fate, or even other metaphysic implications which are far beyond human knowledge. These connotations have been transferred to traffic information (have you ever wondered why traffic information bulletins are usually grouped with weather forecasts? Like if both were uncontrollable natural forces).

It is obvious that previously to face the challenging problem of forecasting, the ability of measuring the objective variable must be mastered. It is reflected in the present thesis that this is not the case if we consider highway travel times and the currently installed surveillance equipment. In general, all forecasting methods need to base their predictions on measurements. Therefore all scientific approaches to travel time measurement are necessary, even for forecasting, and play an important role in research evolution. This does not mean that research devoted to travel time forecasting is in the wrong track. It only means that in order to obtain the maximum benefits of this research and become fully applicable in most of the highway network, research on travel time measurement is equally necessary.

This thesis is devoted to travel time measurement and the usefulness of this information for real time traffic information systems. However, one must realize that a real time travel time information system also needs to predict over the very short term. This fact justifies the inclusion of the present chapter in this thesis report.

When a driver enters a highway he would desire being told how long will take his trip. This could be materialized in a futuristic example, as an individualized message from the radio in the car saying "*The travel time to your destination will be ... minutes*". In reality this is a short term forecast. As it is a forecast, and future is always uncertain, a modification of the previous message to account for this fact could be "*The expected travel time to your destination is ... minutes*". The word "expected" adds the idea of uncertainty in the information, and it means that this is the better information it can be provided, taking into account the actual traffic conditions and the typical (or recurrent) evolution in similar situations. In fact, the previous message could be further modified to quantify this uncertainty. An option could be "*There is a 90% probability for the travel time to your destination being between ... and ... minutes*". The confidence level could be avoided for marketing options, but the confidence interval could be kept as important information. One hundred percent confidence in the information cannot be guaranteed and still provide informative confidence intervals. This is due to the fact that non typical behavior may arise or non-recurrent events can happen (e.g. vehicle breakdowns, accidents…). By definition, non-recurrent events are those that cannot be anticipated, and therefore can only be taken into account once they have happened and become actual information. Despite all these uncertainties, what actually could and should (at least) be told to the driver is "*The current travel time to your destination is … minutes*". This is not his desired information, but at least it is certain (it corresponds to the last travel time measurement), and by sure it is better than unreliable predictions.

What can be concluded from the previous paragraph is:

- Last updated travel time measurement is the best information that can disseminated on real time without falling in the uncertainties of forecasting. For long trips this can be significantly different than the travel time which finally experiments the driver. Frequent update of the measurements helps in informing on the evolution.
- Forecasting is uncertain. Quantification of the uncertainties should be provided.
- Forecasting is based on typical recurrent conditions. Non recurrent events will always remain as unpredictable, until they happen and can be measured. Therefore a forecasting method must also consider actual traffic conditions (measurement) to account for non-recurrences which have already happened.

In view of this comments it seems clear that the key element to be considered in travel time forecasting is the horizon of the forecast. Two temporal horizons can clearly be distinguished: long term and short term forecasting. Long term forecasting refers to the fact that the instant of interest (horizon of the forecast) is far enough in time so that the current traffic conditions do not affect the forecasted travel time. The effects of all the non recurrent events which are currently happening will be vanished by the prediction time. In this long term forecasting, historical information of similar time periods is the basis for an accurate forecast. As an order of magnitude, it could be considered that long term horizons start from the next day, from the instant of making the prevision. Besides, short term forecasting refers to a forecasting horizon where current traffic conditions prevail. The current measured travel times, already affected by the present non-recurrences are by far more informative than historical information. This horizon has a magnitude of the next hour. Obviously this classification is neither strict nor discrete, and in between a continuous gradation of medium term forecasts, where both, current and historical information matter, can be found.

Different forecasting horizons respond to different objectives. For instance long term forecasting provide a useful trip planning tool for users and may help traffic agencies in setting in advance some operational schemes. Short term forecasting is, as it has been stated, useful for real time travel time information, where the horizon of the prevision is precisely this future travel time estimation. Wide range of medium term horizons may be useful for traffic agencies to assess the real time decision making process in setting different operational schemes on the highway.

Depending on the forecast horizon different methodologies could be applied. Each technique is suitable for a particular horizon and there's no methodology to foresee the traffic conditions in all the horizons. For instance, long term prediction is usually based on statistical methods where the main problem to solve lies in determining which similar episodes should be grouped [Chrobok et al., 2004; Danech-Pajouh, 2003; Van Iseghem, 1999] as a unique behaviors. The cluster analysis method represents a useful statistical tool to accomplish this objective as presented in Soriguera et al. (2008) and Rosas et al. (2008). The results of this analysis consist in a year calendar representing the grouping of different types of days. Note that this statistical analysis can be directly applied to a travel time database (provided it exists from a systematic measurement over some years, which is still very rare), in which case, travel time patterns (with its mean and percentiles) would be directly obtained. The alternative could be to apply this pattern characterization to the traffic demand (origin – destination matrixes) if this is the more common available information. Then, a traffic model would be needed to translate these O-D patterns into travel time patterns.

Several benefits appear if using this modeling approach. First of all, the demand is related to the travel behavior of people. Although the behavior of individuals may be irrational, on an aggregated scale the behavior of the demand is rational and with a smooth evolution. Therefore it should be easier to predict than the volatile travel times, which are a derivative variable from the traffic demands. Travel times are characterized by a constant free flow travel time for a wide range of traffic conditions with extremely sensible increases when the demand reaches the congestion threshold. In those situations little demand variations may imply significant travel time changes. Therefore, travel times must be more difficult to predict directly.

An additional benefit appear in the medium term forecast, where in addition to the historical information, non-recurrent or rare events which are actually taking place must be considered (e.g. road works, lane closures, accidents, bad weather…). Surely travel times in these situations could also be found in the database, but their statistical significance would be very low and the number of possible combinations may result prohibitive. As the demand would remain unaffected (or at least less affected), this non-recurrences are more easily taken into by modeling. A similar situation appears in case of a modification of the infrastructure (e.g. the elimination of a severe bottleneck). In this case the historical series of travel times will be lost. However the modifications in the modeling approach would be slight.

The difficulties, like in all the modeling approaches, may appear by the excess of calibration parameters, in addition to the associated modeling errors.

In relation to the short term forecasting, several methods are under research. Some of them are based in spectral analysis, autoregressive time series analysis or in Kalman filtering [Clark et al., 1993]. None of them seems to provide enough accuracy, and errors around 25% are reported in Blue et al. (1994). More recently the research interest is focused on neural network models and artificial intelligence [Dia, 2001; Dougherty and Cobbett, 1997]. Between them, data fusion schemes also appear as an alternative. This approach is treated in more detail in the next section, as it is the approach used in the Appendix A4 contribution.

3.1. DATA FUSION SCHEMES

Multiple source data fusion (DF) consists in the combined use of multidisciplinary techniques, analogous to the cognitive process in humans, with the objective of reaching a conclusion in relation to some aspect of reality which allows taking a decision. In terms of the accuracy of estimation, the combined use of data from multiple sources makes possible to achieve inferences, which will be more efficient and potentially more accurate than if they were achieved by means of a single source. Data fusion schemes are widely used, mainly in digital image recognition or in the medical diagnosis. During the last decade it has also been applied with data related to the transportation field [Hall & McMullen, 2004].

In the next section, a conceptual classification of the different data fusion techniques found in the literature is presented.

3.1.1. Behavioral Classification of Data Fusion Operators

This section aims to describe the different behaviors that a fusion operator may have. First of all, it is necessary to define the following notation:

- "x_i" is a variable which represents the credibility associated to a particular data source "i". "x_i" takes values between 0 and 1.
- "$F(x_1, \ldots, x_n)$" is the credibility resulting from the data fusion operator. It also takes values between 0 and 1.

In order to simplify the concepts, assume that two types of information are going to be fused. In this case "$x_i=(x, y)$". In this situation and according to Bloch (1996), data fusion techniques can be classified, in relation to their behavior as severe, cautious and indulgent:

1. *Severe*: a fusion operator is considered severe if performs with a conjunctive behavior. This is:

$$F(x,y) \leq \min(x,y) \tag{2}$$

2. *Cautious*: a fusion operator is considered cautious if it behaves like a compromise. This is:

$$\min(x,y) \leq F(x,y) \leq \max(x,y) \tag{3}$$

3. *Indulgent*: a fusion operator is considered indulgent if performs with a disjunctive behavior. This is:

$$F(x,y) \geq \max(x,y) \tag{4}$$

3.1.2. Contextual Classification of Data Fusion Operations

This section classifies the fusion operators in terms of their behavior with respect to the particular values of the information to be combined, and to the use of other external information available: the context.

1. *Context Independent Constant Behavior Operators (CICB)*: This class of operators it is constituted by those operators with the same behavior independently of the particular values of the information to fuse. In addition, the data fusion does not take into account any other external information (apart from the values to fuse) regarding the context of the fusion. The operator will have always with the same behavior: severe, cautious or indulgent, whichever, but always the same. The most famous techniques in this class are the Bayesian fusion or the Dempster-Shafer technique.

2. *Context Independent Variable Behavior Operators (CIVB)*: This second class of operators groups those operators which, being independent from the context of the fusion, they depend on the values of *"x"* and *"y"*. Therefore, its behavior may change depending on the values of the variables to fuse. Examples of this class are the artificial intelligence or the expert systems.

3. *Context Dependent Operators (CD)*: the behavior of this class of operators depends not only on the value of the variables to fuse, but also on the global knowledge of the context of the fusion (e.g. knowledge of the credibility of different sources in different situations). Some applications of the fuzzy sets technique are CD operators.

3.1.3. Mathematical Classification of Data Fusion Operations

In accordance to Hall & McMullen (2004), data fusion techniques can also be grouped considering the mathematical logic used to take into account the lack of credibility of data. For instance this is the main difference between Bayesian, Dempster – Shafer and fuzzy sets techniques, while artificial intelligence methods are differentiated by their learning process.

1. *Probabilistic logic*: It is the most widespread mathematical logic, with robust solid mathematical foundations given by the classic probability theory. They need an empirical construction of the probability density functions and of the conditional probabilities. These may impose severe restrictions to the method in complex problems.

2. *Evidential logic*: Mainly represented by the evidential theory of Dempster – Shafer, which allows to include the confidence given to the probability of a particular event. It is useful in those situations where the probability density functions cannot be considered as correctly measured, but only approximated. A level of credibility is given to these functions and the ignorance can be explicitly considered.

3. *Fuzzy logic*: Fuzzy logic, derived from the fuzzy sets theory firstly developed in 1965 by Lotfi Zadeh at UC Berkeley, still today is a highly controversial theory. In fuzzy sets theory, the belonging to a set property is represented by a value between 0 and 1. Equivalently, in fuzzy logic the veracity of an assertion can also vary between 0 and 1, and it is not limited to true or false as in the bivalent logics. In this sense, fuzzy logic is multivalent.

3.1.4. Bayesian Data Fusion

A little more attention is given to Bayesian data fusion schemes as it is the one used in Appendix A4. This fusion technique, based on the Bayes' Theorem of classical probability theory, belongs to the class of algorithms which use a priori knowledge of the observed variables in order to infer decisions on the identity of the objects being analyzed. The Bayesian method provides a model to compute the posteriori probability of a given context.

Analytically, the Bayesian data fusion method can be formulated as follows. If "E" is the object to evaluate and "x_1", "x_2" the information elements obtained from two sensors, from the Bayes theorem it can be stated:

$$\begin{aligned} p(E \mid x_1, x_2) &= \frac{p(E, x_1, x_2)}{p(x_1, x_2)} = \frac{p(x_2 \mid E, x_1) \cdot p(E, x_1)}{p(x_1, x_2)} = \\ &= \frac{p(x_2 \mid E, x_1) \cdot p(x_1 \mid E) \cdot p(E)}{p(x_1, x_2)} \end{aligned} \tag{5}$$

Assuming independence between the measurements of different data sources, Equation 5 can be simplified to:

$$p(E \mid x_1, x_2) = \frac{p(x_2 \mid E) \cdot p(x_1 \mid E) \cdot p(E)}{p(x_1) \cdot p(x_2)} \tag{6}$$

Generalizing the method for "n" sources of information, we obtain:

$$p(E \mid x_1, \ldots, x_n) = \frac{p(x_n \mid E, x_1, \ldots, x_{n-1}) \cdots p(x_1 \mid E) \cdot p(E)}{p(x_1, \ldots, x_n)} \tag{7}$$

Again, assuming independence between data sources:

$$p(E \mid x_1, \ldots, x_n) = p(E) \frac{\prod_{i=1}^{n} p(x_i \mid E)}{\prod_{i=1}^{n} p(x_i)} \tag{8}$$

Finally, if "$\Omega = \{E_1, \ldots, E_r\}$" is defined as the set of "r" possible states, the final decision may be reached according to the following criteria:

- *Maximum a posteriori probability rule*: The most probable state is the one with higher a posteriori probability.

$$E_k = \arg\max_{1 \le i \le r} \{p(E_i \mid x_1, \ldots, x_n)\} \tag{9}$$

- *Maximum likelihood rule*: The most probable state is the one with a higher value in the likelihood function.

$$E_k = \arg\max_{1 \le i \le r} \left\{ \prod_{j=1}^{n} p(x_j \mid E_i) \right\} \tag{10}$$

Both decision rules converge to the same decision when the a priori probabilities are uniform, "$p(E_i) = 1/r$".

Therefore, the Bayesian data fusion method is a severe technique where the fusion operator is independent on the context and on the value of the variables to fuse. The main advantage of the method is a solid mathematical background, where credibility is formulated as a probability function. Using this technique, the a priori knowledge can be expressed in stochastic terms in order to obtain the most probable state of the system.

In practice, it is needed to obtain the conditional probability functions, "$p(x_j \mid E_i)$" and the a priori probability functions, "$p(E_i)$" which model the contribution to the final knowledge of each source. It is a common assumption when there is a completely ignorance to consider uniform a priori probabilities, "$p(E_i) = 1/r$", then the a priori knowledge does not contribute to the final decision, and estimate "$p(x_j \mid E_i)$" from a statistical learning method.

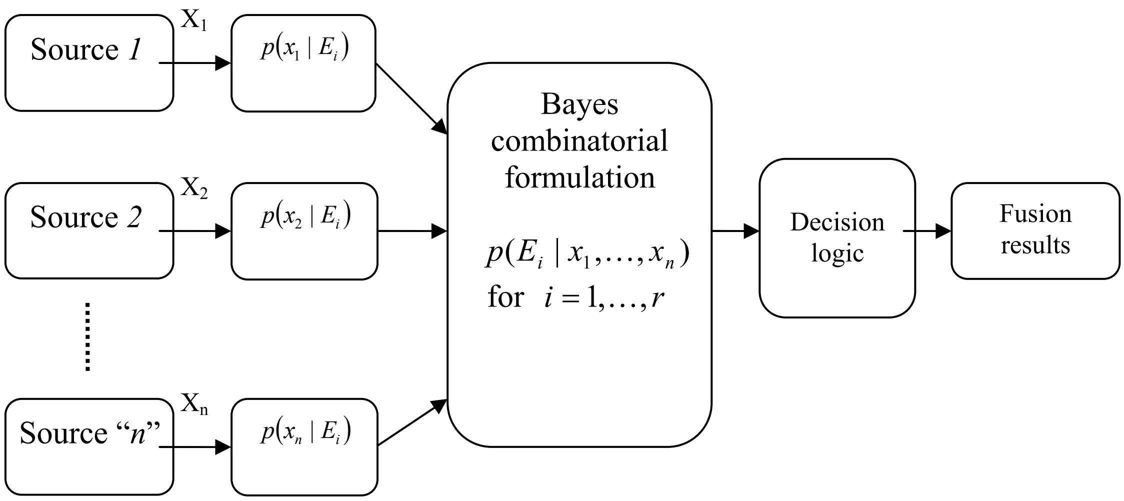


FIGURE 14 Bayesian data fusion method.

3.1.5. Main Benefits and Drawbacks of Data Fusion Schemes

Before deciding on the application of data fusion techniques it is interesting to know the overall expected benefits and the potential drawbacks in the application. Results obtained in Nahum and Pokoski (1980) suggest (see Fig. 15):

- Combining low credibility sources of information (low probability of accurate measurement of single data sources "$P_N < 0.5$") does not contribute in a better final estimation. Therefore an initial requirement to obtain some benefit from data fusion is a minimum precision of data sources.
- Combining high precision data sources ("$P_N > 0.95$") do not contribute in significant benefits. If sensors are precise, there is no point in working for further accuracy.
- Increasing the number of sensors, increases the benefits of the fusion. However the marginal benefits decrease with the increase in the number of sensors.

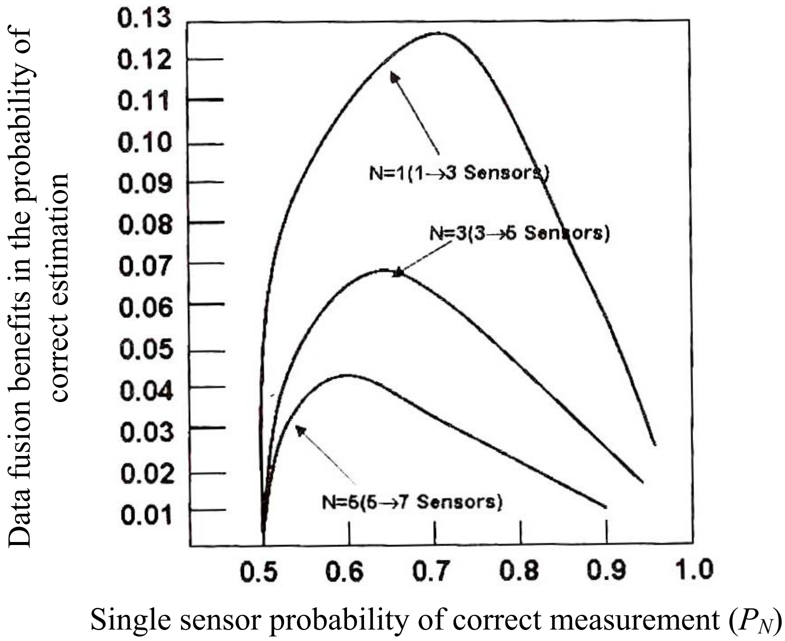


FIGURE 15 Expected benefits of data fusion in relation to the number and accuracy of sensors. Source: Nahum and Pokoski (1980).

In this context, the main advantages of the data fusion schemes are:

- Operational robustness, as one sensor may contribute in the final estimation while the others are inoperative or only partially operative. This contributes in a better temporal coverage of the estimations.
- Increase of the spatial coverage of the measurements, as different sensors can be installed in different spots with different spatial coverage.
- Reliability increase, as the veracity of the information is contrasted by redundant data sources.

The main drawbacks of data fusion may be summarized as:

- Need for a minimum number and a minimum quality of the contributing data sources. This aspect steers the results of the fusion.
- Previous knowledge of the quality of the data provided by each type of sensor. This, although not being restrictive, it may improve significantly the quality of the results.
- There is not a "perfect" data fusion operator. Each algorism has its own strengths and weaknesses.
- Common lack of training data, necessary for the statistical learning of the algorithms.
- Dynamic process. It is difficult to evaluate the results, as the efficiency of the method will improve gradually with the learning of the method.

Finally, it is interesting to stress the computational effort implied by the data fusion scheme in relation to the whole computational effort required by a data information system. According to Hall & McMullen (2004) it can be estimated in an 11%.

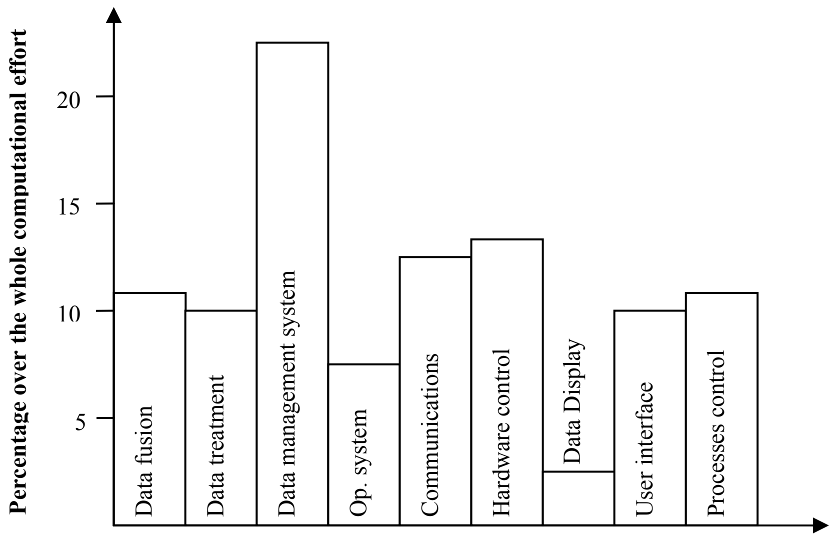


FIGURE 16 Data fusion computational effort. Source: Hall & McMullen (2004).

4. TRAVEL TIME INFORMATION DISSEMINATION

A fundamental component of a travel time information system is the dissemination of the information. Dissemination techniques can be clearly divided into two main types: pre-trip information and on-trip information. Pre-trip information allows trip planning while on-trip information enables network users to (possibly) modify the initial planning according to current traffic conditions.

To be effective, traffic information must be short, concise, quantified and specifically addressed to the receptor. Travel time information itself fulfills the three first conditions, and the dissemination technology must fulfill the last one. That is, the travel time information to be conveyed must be that of specific interest to the driver.

There are several different techniques for travel time dissemination, each one related to a particular technology. Their detailed characteristics are shown in Table 4.

4.1. INFORMATION BEFORE DEPARTURE

Pre-trip information allows the user to decide the time of travel, the mode of travel, or even cancel the trip all together. Pre-trip information reduces the risk of delivering goods late or arriving late at the destination in general.

There are many options available to disseminate pre-trip information. Traditionally, newspaper or radio has been a source of traffic information, especially in case of a special event such as a sporting event or a festival. These types of information basically warn network users of the possible delays and provide information on the extent of the disturbance on the network. Even though these types of traditional measures might be considered as too static, they may provide valuable information to the users of the network and mitigate possible unreliability impacts, if correctly targeted.

Nowadays, most of the dissemination techniques associated with pre-trip information are internet-based services while some also provide up-to-date information into mobile phones. A number of service companies offer calculation of journey times or travel information with added value. The first websites grew up in the mid 1990s. Most of them originally provided traffic information for a certain region but recently they have been extended to cover whole network. These websites initially targeted the general public but then began to offer professional solutions such as geographical location of clients.

TABLE 4 Travel Time Dissemination Techniques.

Techniques	Characteristics
Radio Broadcasts	• Traffic information bulletins • Capable of disseminating pre-trip and on-trip information • No user discrimination. Each driver must carefully listen to the whole bulletin and select his own information of interest. • Discrete information times, subject to the scheduled bulletins. • In case of short range dedicated radio signal, this last two limitations can be overcome.
TV Broadcasts	• Only pre-trip information • No user discrimination • Discrete information times
Press	• Only pre-trip information • No user discrimination • Discrete information times
Traffic Call Center	• Capable of disseminating pre-trip and on-trip information, with the limitation of on-trip telephone calls • It is a service on demand. Driver must ask for it and usually pay a price for it. This implies a limitation of access to the information.
Variable Message Signs	• Capable of disseminating pre-trip and on-trip information • Specifically addressed to the driver, as only inform the drivers who travel below them. • Continuous and very accessible information
Internet	• Only pre-trip information (on-trip using smart mobile phones) • It is a service on demand. User must log-in and ask for a specific itinerary
Car Navigator (RDS-TMC radio signal)	• Capable of disseminating pre-trip and on-trip information • Specifically addressed to the driver (GPS/GSM/UMTS) • Continuous and very accessible information
Cellular Phone (text service)	• Equivalent to a call center with the improvement that you can subscribe to a particular corridor and receive information without asking every time for it.
Information Points	• Capable of disseminating all types of information and discriminate between users. However its accessibility is very low because the driver must stop at the service area to obtain the information.

The *continuous traffic map*, for example is a Catalan website directly managed by the Catalan traffic authority (Servei Català del Trànsit). It provides traffic information on a schematic map of the Catalan road network. Several zoom levels are available to improve the visualization. The traffic data collected in real time by roadside equipment (mainly loop detectors and cameras) is represented on the map. Specifically, the traffic density is represented in a three level color scale, the images on the traffic cameras are available, and also the information displayed on the variable Message signs. Icons pinpoint incidents, road works or bad weather conditions are also reported.

4.2. INFORMATION EN ROUTE

Using on-trip information may mitigate undesired impacts in a cost-effective way. Depending on the information, network users may decide to change their route, if alternative is available, still arriving on time at their destination. Users may also reduce the impact of arriving late by rescheduling their deliveries or planned activities and hence reduce the ripple or snowball effect of them being late. Even in the case there is no possibility to react, just the information of being late reduces the stress related to not knowing how long the possible delay may last.

Electronic information signs are now a familiar sight across the world on motorways and trunk road network. These signs are the main technology to provide on-trip information. The warn drivers of emergencies, incidents and road management. They are aimed at improving safety and minimizing the impact of congestion. Variable Message Signs (VMS) is a term often used to describe these signs. The main purpose of VMS is to communicate information and advice to drivers about emergencies, incidents and network management, aimed at improving safety and minimizing the impact of congestion. Messages displayed on VMS are often limited to those that help drivers complete their journey safely and efficiently. There are a number of types of VMS in use around the world and they provide the capability to display a wide range of warnings, messages and other traffic information.

The telephone is another way of transmitting on-trip travel time information to the driver. It should be one easy-to-remember number regardless of the traveler's location. The U.S. Department of Transportation (USDOT) petitioned the Federal Communications Commission to designate a nationwide three-digit telephone number for traveller information in 1999. This petition was formally supported by 17 State DOTs, 32 transit operators, and 23 Metropolitan Planning Organizations and local agencies. On July 21, 2000 the Federal Communications Commission designated "511" as the single traffic information telephone number to be made available to states and local jurisdictions across the country. An interesting point here is that the number is national but information is local.

Dynamic vehicle guidance and navigation services which include real time traffic data is at an early stage of development. However the integration of a status quo within the transport systems is essential to provide reliable travel time information. Further development is necessary in order to tap the full potential. This especially includes the cooperation of different players in the transport system, such as road administrations on various levels as well as public and private transport service providers. A major step would be the intermodal integration of guidance and navigation applications. Today capable techniques are mostly available; the cooperation between the different administrative levels needs to be improved. This would lead to integrated traffic information, which would enable also commuters – like combined traffic today – to choose the best available multimodal connection for the actual trip. Closely related car integrated systems for driver assistance will help to minimize the probability of incidents which influence reliability negatively.

Most of the applications currently are provided on commercial basis. Many navigator models, for example, already provide real-time information on incidents, weather, and traffic to mobile phones and car navigators. They calculate estimated travel times and take into account incidents in order to improve the estimated travel time. However, they are often limited by their capability to take into account changes in the traffic conditions due to new information in real-time. Most portable GPS devices today offer only a single route choice to the destination. This has one major drawback. In case of an incident, all drivers will follow the same advice given by the navigator. This will result congestion and delay on the new route. Some applications have recently emerged which provide alternative route options.

4.3. COMPARISON OF DIFFERENT TRAFFIC INFORMATION DISSEMINATION TECHNOLOGIES

Taking into account these factors, multicriteria analysis of different traffic information dissemination technologies has been performed (Table 5).

TABLE 5 Multicriteria Analysis of Traffic Information Dissemination Technologies

Mark	Technology	Description	Pre-Trip	On-Trip	Specifically addressed	On demand	User Friendly
10	Car Navigator	On vehicle device RDS-TMC	√	√	√	X	√
10	VMS	Variable Message Signs	√	√	√	X	√
8	Radio broadcasts	Radio bulletin	√	√	X	X	√
7,5	Cellular Phone	Text services	√	√	√	√	=
7	Phone	Traffic Call Center	√	√	√	√	X
2,5	TV Broadcasts	TV bulletin	√	X	X	X	√
2,5	Press	Conflictive days announcements	√	X	X	X	√
2,5	Internet	Online services	√	X	√	√	√
2	Information Points	Service Area information	√	√	√	√	XX

Note: √ - Good, = - Medium, X – Bad, XX – Very bad.

From table 4 results it can be concluded that car navigators and VMS are the technologies with higher dissemination potentialities. These results are in accordance to the current practices of Spanish operators who are installing VMS in most of the primary network. Moreover results also agree with user perceptions, as car navigation devices are currently bestselling car items.

5. VALUE OF TRAVEL TIME INFORMATION

Funding information infrastructure is still a challenge. Traffic information is often organized by the state for national roads, by concessionaires on toll motorways, and departments and towns for their local road networks. Information equipment is generally financed by the various authorities responsible for their own networks.

Network users, for their part, have grown accustomed to considering that information should be provided free of charge. They see no reason to pay for access to information once they have paid for their car with all its accessories and paid their taxes, etc. In spite of this unwillingness to pay, there is abundant evidence that travelers place a high value on travel time information.

A study of motorists' preferences by Harder et al. (2005) found that travelers would be willing to pay up to $1.00 per trip for convenient and accurate travel-time predictions, such as when traffic is delayed which alternative routes would be faster. An interesting question that arises from this result is when travel time information has a great added value that will make drivers willing to pay for it?

The answer to this question should be when travel time information allows a benefit greater than their cost. A quantitative response to this question will not be provided here, and stands out as an interesting issue for further research. However it is evident that the value of travel time increases when the provided information is really informative.

This means that in case a freeway corridor is always free flowing, the free flow travel time information will be almost meaningless and without any value. Then, for travel time information being valuable it needs to vary (travel time variability)? Travel time will be informative when it varies, that is in congestion episodes? Not always. Note that if exactly the same congestion exists every day at exactly the same time, the travel time information will also tend to be non informative (at least for commuters). Note here that a first distinction appears: Travel time information may be valuable for a sporadic user, and meaningless for a commuter. It all depends on their baseline level of information, obtained from the experience.

In reality, what affects the value of travel time information is its ability to inform of unexpected travel times (larger, but also shorter). And here, the word unexpected matters. The value of travel time information is related to reducing the uncertainties. This concept is known as travel time reliability.

This unexpected behavior depends on two components: the baseline level of information of the user and the traffic characteristics of the highway. This means that in some infrastructures travel time information can be valuable while in others always meaningless for almost all drivers. It also means that the value of travel time information will not be the same for all drivers (also if the socioeconomic differences between drivers, which affect the willingness to pay, are not taken into account). These issues are analyzed in the present chapter.

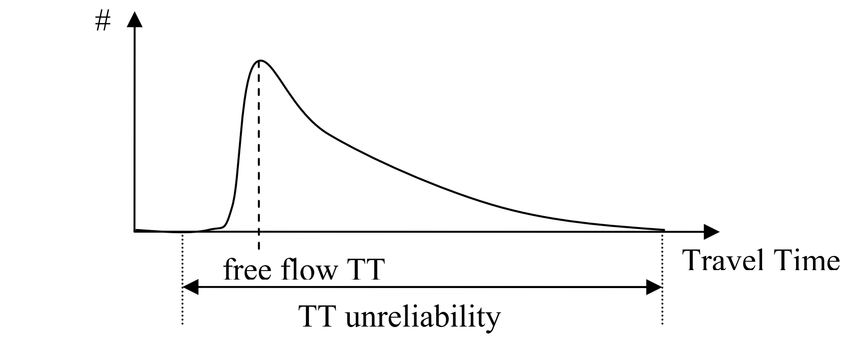
The drivers' aversion to road travel time unreliability results from its high "costs". The costs of unreliability are due to two main reasons: arriving too late and arriving two early. Both situations imply an extension of the waiting time (on route or at destination), with the aggravating circumstance of loosing meetings, connections… the undesirable snowballing effect of arriving too late. To prevent the later, drivers allow extra time (buffer time) for the journey, increasing the probability of arriving too early. It should be clear that as travel time unreliability increase, so do the waiting times. Even on the unlikely situation of being on time, in unreliable road conditions, the anxiety and stress caused by the uncertainty in the decision-making about departure time and route choice imply an additional "cost" for the driver. The costs of unreliability seem to be clear.

The value of travel time information is directly related to its ability of reducing unreliability and its associated costs. As it will be explained next, travel time information by itself can mitigate the unreliability and its consequences. The information does not stop an incident happening but rather reduces the costs that arise from the incident.

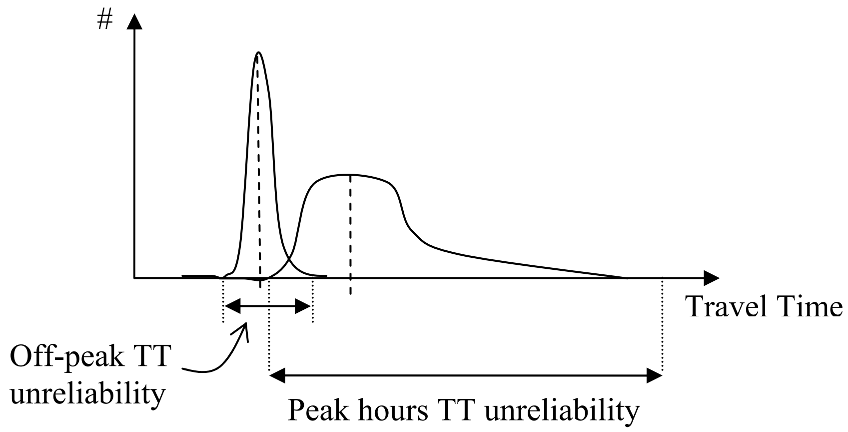
5.1. TRAVEL TIME: VARIABILITY, RELIABILITY AND VALUE OF INFORMATION

On one hand, travel time reliability can be defined as the lack of unexpected delays in a road stretch. Then an itinerary could be considered as reliable, in terms of travel time, when the actual travel time of a particular vehicle is close to its expected travel time. Take into account that the expected travel time may include the expected recurrent delays. On the other hand, travel time variability can be defined as the variation in travel time on the same trip traveled in different times of the day or in a different day in the week.

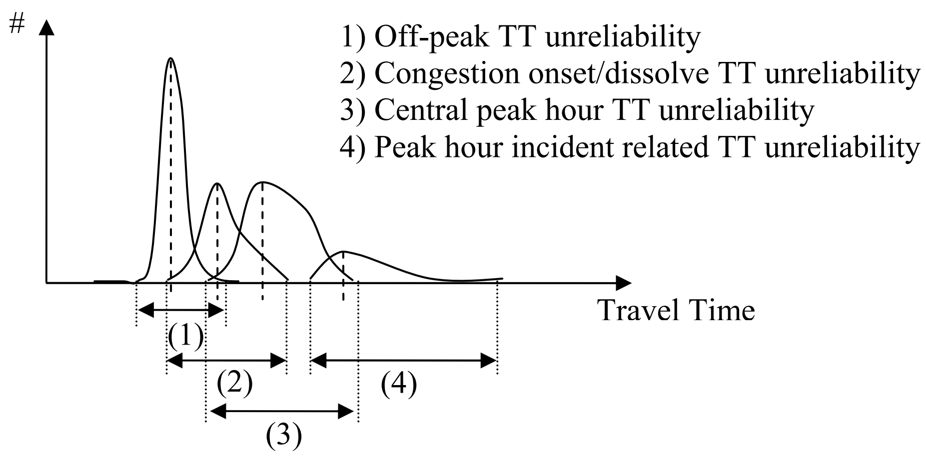
Note that, with these definitions, travel time reliability depends on the driver's expected travel time. This expectation varies with the driver's information, which could be result of experience gained from past trips or directly provided by the road operator.
In the case of a sporadic driver with no travel time information from the operator, it is probable that his knowledge is limited to the free flow travel time (from the expected average speed for the type of road and the distance). Note that in this situation (see Fig. 17a) the travel time unreliability is very high if traveling in a heavy traffic freeway stretch with its associated travel time variability. This results from a very wide travel time frequency distribution with a high range of possible travel times.



a) Information limited to free flow travel



b) Different travel time expectations for peak



c) Detailed travel time expectations for the congested situation

FIGURE 17 Variation of Travel time (TT) unreliability in relation to the drivers information in a multilane freeway.

If the knowledge of the driver improves, due to his acquired experience on the corridor or due to travel time information provided by the operator, and he knows if traveling in peak or non-peak hour and the associated expectations for the travel time (e.g. daily commuter with knowledge of recurrent delays – expected delays at same time of the day in some type of days), see Fig. 17b, the travel time unreliability is reduced due to the reduction of possible travel times in each traffic condition (congested or not). Finally, if the driver has a very good knowledge of the traffic conditions in his trip (very accurate travel time information on the freeway), different expectations of travel time within the congested period can also be predicted (Fig. 17c). Then, travel time unreliability can be defined as the width of the frequency distributions of travel time around the driver expected average travel time and considering his a priori knowledge.

In this context the relationship between travel time variability and reliability is not direct, as a heavy peak hour freeway with high travel time variations within a day, could also be a very reliable freeway if accurate information is provided to the driver and an efficient incident management system is fully working in order to avoid non recurrent incidents. In contrast, a road stretch with less travel time variability could be very unreliable for the driver if no information is provided and frequent incidents imply serious unexpected delays.

The conclusion is clear, as higher is the information provided to the driver, closer is the expected travel time to the real travel time and so, higher is the reliability of the infrastructure. Therefore, information reduces the unreliability, but it will be when reduces unreliability costs that will become most valuable. As stated before, reducing the unreliability reduces stress, and this has a cost. Then, even not being able to modify any characteristic of the trip (i.e. the driver is trapped in the highway) if the information reduces unreliability it has a value: it reduces stress and makes rescheduling the events at destination possible, smoothing the ripple effect.

However, the value of information will be higher when making possible to reduce the other costs of unreliability (e.g. avoiding arriving late or early due to an excessive buffer time). To accomplish this objective and acquire value the travel time information has to reach the user enough in advance (pre-trip information) to be able to modify the instant of departure (to account for unexpected increases or decreases in travel times), or modify the mode or the route choice. Once on route, the instant of departure it is already decided. There only remains to act in the route or mode choice. Then, the location on the highway where on-trip travel time information is provided affects significantly its value. It must allow route choice (in advance of main junctions) or mode choice (before park and ride stations).

To finish this discussion, a note of the system wide effects of travel time information is pertinent. In a highway network where all drivers have perfect information, they will distribute over the different route options taking into account their own benefits. Considering the increase of the cost of traveling through a link with the increase of the demand (see Fig. 18), the system will reach a user equilibrium (recall the Wardrop (1952) principles on network equilibrium). This user equilibrium, which could be considered to be reached in a current day (with no non-recurrent incidents) in a metropolitan highway network where most of the drivers are commuters with good knowledge of recurrent conditions, it does not have to correspond to the system optimum. With the universal dissemination of travel time information, the user equilibrium will also be reached in non-recurrent conditions. This means that if every driver is provided the same real time information of current traffic conditions, the system will evolve to user equilibrium where any driver by itself will not be able to improve his performance by switching routes. In this case the value of the information will be again very low. This means that he value of the travel time information diminishes with the number of drivers which have the same information. In general, the most valuable information is the one that few people know. In addition it is probable that the user equilibrium reached in case of universal information in non-recurrent conditions does not represent an optimum, or even the paradox case where the equilibrium reached is even worst, from a system wide point of view, than the original situation without information. The fact that information may not only re-route drivers but also divert them to other transportation networks (possibly less sensible to the demand) should be considered as a positive aspect of the information in this case.

This opens up a new research direction which should try to establish better strategies for the dissemination of the information with the objective to improve the system wide performance. It is probable that not all the drivers have to receive the same information, and here equity issues will appear.

5.2. TRAVEL TIME UNRELIABILITY EXPECTED BEHAVIOUR IN MULTILANE FREEWAYS

Figure 17 represents the approximate qualitative behavior of travel time frequency distributions for a heavy traffic multilane freeway. Some aspects that may be interesting for the reader are the following:

- Travel time distributions over long time periods (several hours) are highly skewed with a long right tail. This is due to the inclusion of different traffic states (Fig. 17a). As the time window is reduced, the travel time distributions tend towards a normal (Fig. 17b and 17c).

- Travel time unreliability in congested situations is significantly higher than in free flow situations (Fig. 17b and 17c). This results from the random behavior of traffic demand within a particular traffic flow pattern and from the great no linearity between traffic density and travel times (see Fig. 18). Note that in the congested situations, little variations in density produce high variations in travel time, while the same variation in density in the free flow zone, does not imply a meaningful change in travel times.

- For the same reason, catastrophic delays due to incidents in the freeway only occur in the heavy traffic periods, while very lower effects are expected in light traffic conditions.

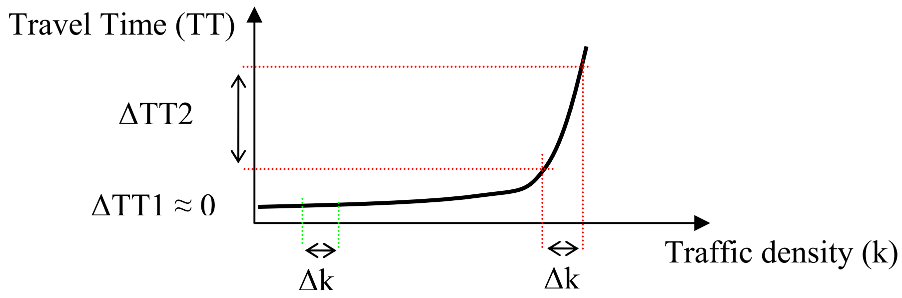


FIGURE 18 Density vs travel time diagram.

5.3. SOURCES OF TRAVEL TIME UNRELIABILITY: WHAT COULD BE DONE?

As seen (Fig. 17c) even if the recurrent delays are known and provided to drivers by the highway operator, some unreliability remains. This could be considered as the "baseline" travel time unreliability.

Baseline unreliability, the source of value for travel time information, responds to two main phenomena: the random variation on transportation facilities demand and the probability of an incident in the freeway. It is shown in Figure 18 that little variations on the demand for transportation between to similar days at the same time can produce considerable variations in travel time (in heavy traffic conditions). Additionally to this first source of unreliability it should be also taken into account the incident related unreliability.

For incident related unreliability it should be understood that "something" is happening in the freeway that implies additional delay beyond the usual random variations. These incidents can arise from two situations: an unusual increase in the demand (e.g. due to an especial event) or a reduction in the supply (i.e. capacity) of the freeway (e.g. vehicle breakdown, crash, road works, and bad weather). Note that bad weather sometimes can also imply an increase in the demand.

As stated previously, in general unreliability can be mitigated by information. In addition, each source of baseline unreliability must be faced in a different way. Table 6 present some possible strategic and operational measures to reduce the travel time unreliability in a freeway stretch. Note that each of the unreliability sources produces considerably more negative effects on travel times in heavy traffic conditions. If for some reasons (e.g. budget limitations) the mitigation measures can only be applied in reduced time windows, priority have to be given to congested periods with no doubt.

Baseline unreliability cannot be eliminated but should be limited. The question that remains for the policy makers is which should be the admissible unreliability threshold? And how bonus/malus economic incentives applied to operators could help to reach an admissible situation? But the answer is still far away. Note that a previous step is measuring unreliability, which means measure travel times (also a first mitigating measure), and as it has been stated in this thesis this is not a trivial task given the actual surveillance equipment.

TABLE 6 Sources of Travel Time Unreliability and Possible Mitigation Measures

Source of TT unreliability	Possible mitigation measures
Recurrent demand variations	Pre-trip information in travel times.
Random variation on transportation demand	Operational measures to lessen random variations (e.g. variable road pricing, ramp metering, dynamic flow control…)
Scheduled especial event	Information to road users in advance, provide information on alternative routes or alternative transportation modes, increase supply when possible by switching direction of some lanes.
Vehicle crash or breakdown	Rapid response strategy. To be quick and effective in reaching the incident point and clearing the freeway.
Road Works	Scheduling strategy of the works taking into account traffic patterns and minimum capacity affection. (i.e. off-peak road works).
Bad weather	Difficult to mitigate this baseline unreliability. Increase of reliability can be achieved by means of information by obtaining specific patterns for bad weather conditions and accurate weather forecasts.

5.4. TRAVEL TIME INFORMATION SYSTEM: LEVELS OF APPLICATION IN A ROAD NETWORK

The value of travel time information varies greatly with the variability and mainly unreliability of travel times on a particular stretch of road. The value that drivers give to travel time information is different if traveling in an unreliable stretch or in a very reliable one. In turn, this variability and reliability of travel times depends on the physical characteristics of the facility (number of lanes, slope, junctions …) and on its relation to the demand for travelling through it.

Because always exist a budget limitation in the intensive monitoring of the network, and road networks are vast (Catalonia has more than 12,000 km of roads), priorities must be given to certain locations. The level of surveillance will directly affect the accuracy of the resulting travel time information. The basic criteria in the deployment of a travel time information system on a highway network should be the value of travel time information in each stretch and therefore the unreliability of travel times. Within the high travel time variability stretches (i.e. congested stretches) the value of the information provided by the road manager to the driver will be higher in those situations less predictable by driver experience in the corridor. This, points out the fact that for example in a freeway to reach the central business district of a big city, very congested every working day but in a similar magnitude, travel time information is less valuable due to the previous knowledge gained by commuters of the conditions in the freeway. In contrast, another facility also with heavy traffic near capacity, where the breakdown is less predictable and travel time for a particular departure time ranges from free flow travel time to severe and unexpected delays caused by frequent incidents, road works, bad weather or increased demand due to special events or seasonality of the facility, implies great value for travel time information provided by the operator. This means that the criteria for selecting priority corridors respect travel time information has to consider not only travel time variations in the whole day (level of congestion) but also the frequency of congestion (i.e. variation of travel times across days for a given departure time).

Finally, the last two indicators to take into account should be the demand of the corridor (hourly traffic volume in rush hours), and the monetary cost of implementing the surveillance equipment (dependent on the already existing equipment). From a cost-benefit analysis will result that priority must be given to those corridors where less money benefits more drivers.

TABLE 7 Factors to Take into Account in a Multicriteria Analysis for Selecting Priority Corridors to Implement a Travel Time Information System

Factor	Measures	Priority to
Congestion level for a particular type of day	• Travel Time Index: $TTI = \dfrac{Max\ TT\ rush}{Free\ \ Flow\ TT}$ • Peak Delay = Max TT rush – Free Flow TT	Higher TTI Greater delays
Frequency of congestion	• 90th percentile – median for the peak hour TT distribution • # days with congestion/total # of days within each group of types of day	Higher values of these unreliability measures
Average hourly traffic volume in the rush period	• Total volume of vehicles served in the rush period / duration of the rush	Higher volumes
Surveillance implementation cost	• Depends on the selected technology and the already existing equipment	Lower costs

These concepts have been applied to the Catalan road network in Soriguera et al. (2006) taking into account the levels of existing surveillance, the frequency of congestion and the AADT (Annual Average Daily Traffic). Considering also the severity of this congestion, the rush hour durations, and the most suitable technology to measure travel time in each corridor, priorities are obtained, and can be seen in Table 8.

TABLE 8 Priority corridors to implement a travel time information system in the Catalan network (northeast of Spain).

Priorit y	Corridors	Justification
1	Toll highways, primarily those following a SW-NE axis near the coast.	• Severe congestion • Medium frequencies (high unreliability) • High traffic volumes • Existing tolling infrastructure
2	Freeways around Barcelona	• Severe daily congestion • High traffic volumes • Already existing intensive surveillance equipment
3	Seasonal corridors • Winter N-S corridors (skying) • Summer coastal corridors (beach)	• Severe sporadic congestion • low frequencies (high unreliability) • Low surveillance at the current time

6. CONCLUSIONS AND FURTHER RESEARCH

It is possible to develop an accurate real-time travel time information system on closed toll highways with the existing surveillance equipment.

This sentence summarizes the research presented in this thesis. The conclusion is significant, as it has been shown the high value drivers and traffic agencies give to travel time information in order to support their decisions. In contrast, the difficulties for traffic managers to fund the information infrastructure and to integrate new technologies as they emerge while retaining sufficient homogeneity in the network has also been shown to be challenging. The conclusions of this thesis match the travel time information desires without falling in the usual requirement of more and more data.

In this last section of the thesis report the overall conclusions of the research are presented. As the thesis is structured as a compendium of papers, detailed conclusions of each part can be found in the corresponding appendix.

A methodology is proposed in the thesis which makes use of the available traffic data on closed toll highways to provide accurate travel time information in real time to the drivers entering the highway. Measuring is not enough to achieve this objective and very short term forecasting is necessary. The method uses data obtained from toll ticket data and from a non-intensive loop detector surveillance system (approximately one loop every 5 km), and makes the most of a combined use of the data, using a two level data fusion process. The different accuracy of different data sources, their different temporal alignment and their different spatial coverage allows inferring a short term forecast of travel time, which improves the original travel time estimations from a single data source. The proposed method uses toll ticket and loop detector data, but it is not technologically captive. It can be used with any two sources of travel time data, provided that one of them supplies direct measurements (e.g. the innovative Bluetooth signature matching, or the cell phone tracking) while the other indirect estimations.

The results of the data fusion process improve with the accuracy of the single source measurements. This issue is addressed in the thesis. Travel time estimation methods from loop detector data are analyzed. It is found that the current research trend, based in looking for new mathematical speed interpolation methods between point measurements in order to solve the problem of punctual measurements while requiring spatial results (i.e. travel times), it is not adequate if it is blind to traffic dynamics. The present thesis demonstrates conceptually and with an accurate empirical comparison that travel time estimation methods based on mathematical speed interpolations between measurement points, which do not consider traffic dynamics and the nature of queue evolution, do not contribute in an intrinsically better estimation, independently of the complexity of the interpolation method. Incorporating the traffic dynamics of the theory of kinematic waves, it could improve the performance of these methods. This is currently an author's issue of active research. Two improvements are proposed in the thesis. First, a method to obtain space mean speed directly from common aggregations of loop detector data, and with no additional information. Normal distribution of speeds over small space-time regions is assumed. The quantitative prove of this statement, which is only conceptually discussed in the thesis, is the objective of other of the author's current research directions, already with promising results. This solves a recurrent problem most traffic agencies face when trying to estimate travel times from loop detector data, and avoids the current practice of using the overestimated space mean speeds. Second, an intelligent smoothing process for the noisy loop detector speed measurements, reducing the fluctuations of short time interval aggregations while maintaining the immediacy of the measurements. Current smoothing practices imply delay in the information. This simple contribution could be seen as a simplistic first approach to include traffic dynamics in the estimation. These improvements, directly applicable with the existing loop detector hardware, contribute significantly in a better travel time estimation, as it is proved in the thesis.

A side comment is relevant here. Simple modifications to the standard loop detector data treatment process would suffice in order to obtain space mean speeds. In the short or medium term however, this is an impractical task. Most of the traffic management centers stay with the standard process, because that is what they are used to, that is what the archaic detector controller software is programmed for, and because it is easier to keep doing the same than to change, especially if there is not a pressing reason to do so. But this can change one day. In addition, as a legacy of a time when communication bandwidth was a technical and cost constraint, loop data treatment standards are still designed to reduce the amount of data being transmitted, and individual activations are lost. Nowadays, the availability of developed communications technology makes it possible to process and transmit individual traffic detections to the traffic management center. This would allow using the enhanced computing capabilities of the traffic management center to solve the problem, instead of the archaic roadside controllers.

Other technological problems remain, like the frequent failures of loop detectors or their inability to correctly measure over instable stop and go traffic. These issues need to be considered in the future development of the new detectors which will finally beat traditional loops. In addition they will need to be cheap, will need to benefit from a long lifespan, will need to be less intrusive into the pavement, will need to require low maintenance, will need to communicate wireless and will not need a power supply. The MeMS detectors, developed by UC Berkeley engineers, are on the track. This will be the time to incorporate to these detectors more advanced features, like reidentification capabilities. This last issue, it is surely a shared responsibility with the automotive industry, which will need to equip vehicles with an electronic wireless readable tag as an standard equipment.

Travel time estimation from the reidentification of toll tickets in a closed toll highway is also addressed in the thesis. A method is proposed capable of estimating single section travel times (i.e. time required to travel between two consecutive junctions on the main trunk of the highway) and also the exit time at each junction (i.e. the time required to travel along the exit link plus the time required to pay the fee at the toll gate). Combining both estimations it is possible to calculate all the required itinerary travel times, even those with very few observations where direct measurement would be problematic, and avoiding the information delay for real time application. The knowledge of the exit time, allows obtaining the level of service of each toll plaza at every junction, making possible to modify the number of active toll booths in accordance. An extension of the method to open toll highway schemes is currently under research. The key issue here is the optimal location of additional control points (i.e. in addition to main trunk toll plazas), to achieve the most cost – effective surveillance scheme.

The proposed method is only one of the possible applications of the enormously rich database provided by toll ticket data in closed toll highways. A closed toll highway is a privileged infrastructure (in terms of data), where the origin, destination, type of vehicle and entrance/exit times are measured in real time and for all vehicles. This is inconceivable in any other road environment, and will fulfill the desires of the most exigent highway engineer. Tolling is the data reason to exist, and all further use of the data will contribute in net benefits. The potential of the data is huge, for researchers who would like to use the infrastructure as a highway lab to test and evaluate their models, and even for the operators which may apply most of the advances in traffic engineering with privileged data inputs, which are the main drawback of these techniques in most of the applications.

Making use of this privileged situation, all the methods proposed in the thesis have been empirically tested with data obtained in the AP-7 highway. Promising results have been obtained, which are referenced in the corresponding appendixes. Now that the conceptual developments have been exposed, it is the turn of highway operators and administrations to put them into practice, so that highway users can benefit of real time travel time estimations with a low-cost scheme. This could also be the seed for turning this test site into a real highway lab and detailed traffic observatory (similar or even better than the Berkeley Highway Lab[1], a current outstanding example), encouraging further research in information and operations on highway environments.

Other research directions should not be shelved. The highway system efficiency is only one leg of the optimal transportation system. Further research is also deserved in integrated corridor management policies, not only accounting for vehicular traffic, but also for other transportation modes. The optimal supply share of the corridor infrastructures between modes should be a target. Besides, future research should not only consider vehicle to infrastructure communications (as presented in the present thesis), but also exploit the huge potential of the futuristic vehicle to vehicle information systems, which in the next future would, by sure, become a reality.

[1] The Berkeley Highway Lab (online source http://bhl.calccit.org:9006/bhl/) Accessed October 16th 2010.

REFERENCES

Abdulhai, B. and S.M. Tabib. (2003) Spatio-temporal inductance-pattern recognition for vehicle reidentification. *Transportation Research Part C* **11**(3-4), 223-239.

Barceló, J., L. Montero and L. Marquès. (2010) Travel time forecasting and dynamic OD estimation in freeways based on Bluetooth traffic monitoring. *Proc. of the 89th Annual Meeting of the Transportation Research Board*. Washington D.C.

Berechman, J. (2003) Transportation–economic aspects of Roman highway development: the case of Via Appia. *Transportation Research Part A* **37**(5), 453-478.

Bloch, I. (1996) Information combination operators for data fusion: comparative review with Classification. *IEEE Transactions on Systems, Man and Cybernetics–Part A* **26**(1).

Blue, V., G. List and M. Embrechts. (1994) Neural net freeway travel time estimation. *Proceedings of Intelligent engineering Systems through Artificial Neural Networks* **4**, 11135-11140.

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.

Cassidy, M. J. and J. R. Windover. (1995) Methodology for assessing dynamics of freeway traffic flow. *Transportation Research Record: Journal of the Transportation Research Board* **1484**, 73-79.

Chen, C., J. Kwon, A. Skabardonis and P. Varaiya. (2003) Detecting errors and imputing missing data for single loop surveillance systems. *Transportation Research Record: Journal of the Transportation Research Board* **1855**, 160-167.

Chrobok, R., O. Kaumann, J. Wahle and M. Schreckenberg. (2004) Different methods of traffic forecast based on real data. *European Journal of Operational Research* **155**, 558-568.

Clark, S. D., M. S. Dougherty and H. R. Kirby. (1993) The use of neural network and time series models for short term forecasting: A comparative study. *Proceedings of PTRC*. Manchester.

Coifman, B. and E. Ergueta. (2003) Improved vehicle reidentification and travel time measurement on congested freeways. *ASCE Journal of Transportation Engineering* **129**(5), 475-483.

Coifman, B. and S. Krishnamurthya. (2007) Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C* **15**(3), 135-153.

Coifman, B., and M. Cassidy. (2002) Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research Part A* **36**(10), 899-917.

Daganzo, C. F. (1983) Derivation of delays based on input-output analysis. *Transportation Research Part A* **17**(5), 341-342.

Daganzo, C. F. (1997) *Fundamentals of Transportation and Traffic Operations*. Pergamon. Elsevier Science Ltd.

Daganzo, C. F. (2002) A behavioral theory of multi-lane traffic flow. *Transportation Research Part B* **36**(2), 131-158.

Daganzo, C.F. (1995) Requiem for second-order fluid approximations of traffic flow. *Transportation Research Part B* **29**(4), 277-286.

Dailey, D.J. (1993) Travel time estimation using cross-correlation techniques. *Transportation Research Part B* **27**(2), 97-107.

Danech-Pajouh, M. (2003) Les modèles de prévision du dispositif Bison futé et leur évolution. *Recherce Transports Sécurité* **78**, 1-20. In French.

Dia, H. (2001) An objected-oriented neural network approach to short-term traffic forecasting. *European Journal of Operational Research* **131**, 253-261.

Dougherty, M. S. and M. R. Cobbett. (1997) Short-term inter-urban traffic forecasts using neural networks. *International Journal of Forecasting* **13**, 21-31.

Federal Highway Administration. (2010) *UCR – Urban congestion report*. U.S. Department of Transportation.

Hall, D., S. A. and H. McMullen (2004), *Mathematical techniques in multisensor data fusion*. Archtech House Publishers.

Harder, K. A., J. R. Bloomfield, D. M. Levinson and L. Zhang. (2005) *Route preferences and the value of travel-time information for motorists driving real-world routes*. University of Minnesota Department of Civil Engineering.

Herrera, J. C., D. B. Work, R. Herring, X. Ban, Q. Jacobson and A. M. Bayen. (2010) Evaluation of traffic data obtained via GPS-enabled mobile phones: The Mobile Century field experiment. *Transportation Research Part C* **18**(4), 568-583.

Highway Research Board. (1950) *Highway capacity manual: practical applications of research*. National Research Council, Washington D.C.

Hill, J., R. Szewczyk, A. Woo, S. Hollar, D. Culler and K. Pister. (2000) System architecture directions for networked sensors. *Proceedings of ASPLOS 2000*, 93-104.

Huang, T. and S. Russell. (1997) Object identification in a Bayesian context. *Proc. of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97)*, Nagoya, Japan. Morgan Kaufmann.

Kuhne, R. and S. Immes. (1993) Freeway control systems using section-related traffic variable detection. *Proc. of Pacific Rim TransTech Conference* **1**, 56-62.

Kwon, T. M. (2006) Blind deconvolution processing of loop inductance signals for vehicle reidentification. *Proc. of the 85th Annual Meeting of the Transportation Research Board*.

Lucas, D.E., P.B. Mirchandani and N. Verma. (2004) Online travel time estimation without vehicle identification. *Transportation Research Record: Journal of the Transportation Research Board* **1867**, 193-201.

MacCarley, A. C. (2001) *Video-based vehicle signature analysis and tracking system phase 2: algorithm development and preliminary testing*. California PATH Working Paper, UCB-ITS-PWP-2001-10.

Makigami, Y., G.F. Newell and R. Rothery. (1971) Three-dimensional representations of traffic flow. *Transportation Science* **5**, 302-313.

May, A., B. Coifman, R. Cayford and G. Merrit. (2004) *Automatic diagnostics of loop detectors and the data collection system in the Berkeley Highway Lab*. California PATH Research Report, UCB-ITS-PRR-2004-13.

Muñoz, J. C. and C. F. Daganzo. (2002) The bottleneck mechanism of a freeway diverge. *Transportation Research Part A* **36**(6), 483-505.

Nahum, P.J. and J. L. Pokoski. (1980) NCTR Plus Sensor Fusion Equals IFFN. *IEEE Transactions on Aerospace Electronic Systems* **16**(3), 320-337.

Nam, D. H., and D.R. Drew. (1996) Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. *ASCE Journal of Transportation Engineering* **122**(3), 185-191.

Newell, G. F. (1982) *Applications of queuing theory*, 2nd ed. Chapman and Hall, London.

Newell, G. F. (1993) A simplified theory of kinematic waves in highway traffic, I general theory, II queuing at freeway bottlenecks, III multi-destination flows. *Transportation Research Part B* **27**(1), 281-313.

Nishiuchi, H., K. Nakamura, S. Bajwa, E. Chung and M. Kuwahara. (2006) Evaluation of travel time and OD variation on the Tokyo Metropolitan Expressway using ETC data. *Research into Practice: 22nd ARRB Conference Proceedings Information*, Australian Road Research Board.

NYSI&IS. (1970) *Automatic license plate scanning (ALPS) system – Final report*. New York State Identification and Intelligence System, Albany, NY.

OECD/JTRC (2010) *Improving reliability on surface transport networks*. OECD Publishing, Paris.

Oh, J.S., R. Jayakrishnan and W. Recker. (2003) Section travel time estimation from point detection data. *Proceedings of the 82nd Transportation Research Board Annual Meeting*. Washington D.C.

Palen, J., (1997). The need for surveillance in intelligent transportation systems. *Intellimotion* **6**(1), 1-3.

Petty, K.F., P. Bickel, M. Ostland, J. Rice, F. Schoenberg, J. Jiang and Y. Rotov. (1998) Accurate estimation of travel time from single loop detectors. *Transportation Research Part A* **32**(1), 1-17.

Robusté, F., C. Vergara, L. Thorson and M. Estrada. (2003) Nuevas tecnologías en la gestión de autopistas: El peaje y los sistemas inteligentes de transporte. *Revista de economía industrial* **353**, 33-46. In Spanish.

Rosas, D., F. Soriguera and E. Molins. (2008) Construcción de patrones de tiempo de viaje por carretera, *Proceedings of the VIII Congreso de Ingeniería del Transporte,* La Coruña. In Spanish.

Sherali, H. D., J. Desai and H. Rakha. (2006) A discrete optimization approach for locating Automatic Vehicle Identification readers for the provision of roadway travel times. *Transportation Research Part B* **40**(10), 857-871.

Soriguera, F. and F. Robusté. (2011a) Requiem for Freeway Travel Time Estimation Methods Based on Blind Speed Interpolations between Point Measurements. *Accepted for publication in IEEE Transactions on Intelligent Transportation Systems.*

Soriguera, F. and F. Robusté. (2011b) Estimation of Traffic Stream Space-Mean Speed from Time Aggregations of Double Loop Detector Data. *Transportation Research Part C* **19**(1), 115-129.

Soriguera, F., D. Rosas and F. Robusté. (2010) Travel Time Measurement in Closed Toll Highways. *Transportation Research Part B* **44**(10), 1242–1267.

Soriguera, F. and F. Robusté. (2009) Highway Travel Time Accurate Measurement and Short-Term Prediction Using Multiple Data Sources. *Transportmetrica iFirst* 1-25.

Soriguera, F., E. Molins and E. Alberich. (2008) Metodología de elaboración de patrones de intensidad de tráfico, *Proceedings of the VIII Congreso de Ingeniería del Transporte*, La Coruña. In Spanish.

Soriguera, F., L. Thorson, J. A. Grau and F. Robusté. (2006) Bases de un sistema de medición y previsión del tiempo de viaje por carretera en Cataluña. *Proceedings of the VII Congreso de Ingeniería del Transporte*, Ciudad Real. In Spanish.

Turner, S. M., W. L. Eisele, R. J. Benz and D. J. Holdener. (1998) *Travel time data collection handbook*. Research Report FHWA-PL-98-035. Federal Highway Administration, Office of Highway Information Management, Washington, D. C.

van Arem, B., M.J.M. van der Vlist, M.R. Muste and S.A. Smulders. (1997) Travel time estimation in the GERIDIEN project. *International Journal of Forecasting* **13**, 73-85.

Van Iseghem, S. (1999) Forecasting traffic one or two days in advance. An intermodal approach. *Recherce Transports Sécurité* **65**, 79-97.

Vanderbilt, T. (2008) *Traffic: why we drive the way we do (and what it says about us)*. Alfred. A. Knopf publisher, New York, Toronto.

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

Yim, Y. and R. Cayford. (2006) Field operational test using anonymous cell phone tracking for generating traffic information. *Proc. of the 85th Annual Meeting of the Transportation Research Board*.

APPENDIX A1

Estimation of Traffic Stream Space-Mean Speed from Time Aggregations of Double Loop Detector Data

ABSTRACT

In one of the very first papers on traffic flow theory back in 1952, Wardrop presented the difference between the space mean speed (SMS) and the time mean speed (TMS) of a group of traveling vehicles, and derived a relationship suitable for estimating TMS, given SMS and the speed variance over SMS. As time goes by, traffic practitioners have tended towards computing TMS instead of SMS, mainly when using double loop detectors, and nowadays this is the usual practice in traffic management centers. Therefore, the useful relationship between TMS and SMS should go the other way around in relation to Wardrop's. Recently, the complementary relationship, suitable for estimating SMS from TMS and the speed variance over TMS, has been proved. However this is not enough, as speed variance is usually not available.

The present paper develops a probabilistic method to estimate SMS from TMS without the previous knowledge of speed variance and only using the usual time aggregations of double loop detector data. The main assumption of the method – the normality of vehicle speed distribution – is discussed and a formulation to obtain the expected error of the estimation is derived.

The results obtained with test data from the AP-7 highway, near Barcelona in Spain, show that the developed methodology is able to estimate SMS with an average relative error as low as 0.5%.

Keywords: Space mean speed, time mean speed, traffic fundamental variables, speed variance, traffic flow theory.

1. INTRODUCTION

On the one hand, one can define an average of the speeds of the "*m*" vehicles passing a fixed location "x_1" over some observation period "*T*" (i.e., an average across time or time mean speed – TMS). On the other hand, an average at a particular instant "t_1", of the speeds of the "*n*" vehicles contained on a road segment of length "*L*" (i.e., an average across space or space mean speed – SMS) can also be defined. There is no reason for these two averages to be the same.

Think of a closed loop track where two classes of vehicles, fast and slow (e.g., half and half), are being driven without interactions. Realize that the fraction of fast vehicles seen by a stationary observer is greater than the fraction of these same fast vehicles seen in an aerial photograph of the track (e.g., 50% in this example). This is obvious considering that the fast vehicles will pass the stationary observer's location more often than the slow ones will. This is an intuitive way to see that TMS will always be greater than SMS, unless all the vehicles travel at the same speed. In this last situation, both means are obviously equal. From the previous example, it should come as no surprise that TMS is a flow weighted average of speed, while SMS is a density weighted average of speed. These concepts are explored in detail in the traffic operations textbook by Daganzo (1997).

In practice, the most common technology for measuring vehicle speeds is the use of inductive loop detectors, massively installed on highways worldwide, with increased use since the early 1990s and on metropolitan highways, where densities can reach one detector every 500m. The purpose of these detectors is to monitor congestion and to provide information for traffic control operations. Loop detectors are presence-type detectors, able to detect the presence of a vehicle above them. Applying some basic operations to these raw presence measurements (see Section 3 for details) the individual speeds of every vehicle can be obtained. This results in a vast amount of data on heavy traffic highways with high density of detectors. Nowadays, the availability of developed communications technology makes it possible, and sometimes even economically feasible, to process and transmit individual traffic detections to the traffic management center (TMC) in charge of controlling and monitoring traffic for a whole metropolitan area. However, as a legacy of a time when communication bandwidth was a technical and cost constraint, loop data treatment standards are still designed to reduce the amount of data being transmitted. Individual measurements are aggregated or averaged across time (typically every 30sec, 1min or 3 min) at the detector site by the roadside detector controller. The common averaging operations (in particular the Spanish loop data treatment standards) result in time averages of loop detector data, specifically time mean speeds, to be sent to the TMC. Although simple modifications to this process would suffice in order to obtain space mean speeds most of the TMCs stay with the standard process, because that is what they are used to, that is what the archaic detector controller software is programmed for, and because it is easier to keep doing the same than to change, especially if there is not a pressing reason to do so. Despite this, the increasing need for accurate traffic data, in particular space mean speeds, and the technical and economical development of communications, sets a trend to transmit individual traffic detections to the TMC. In this situation, already implemented in some agencies in the USA, both speed means will be directly available. However, this is not the common case worldwide, yet.

Meanwhile, the usual structure of traffic data gathering only around temporal averages has some drawbacks, particularly in relation to traffic stream models. Note that the traffic fundamental equation which establishes a relationship between flow "*q*", density "*k*" and average speed "$\bar{v}$" (i.e., $q = k \cdot \bar{v}$) only holds when variables are accurately measured and average speed is defined as a space mean speed [Cassidy and Coifman, 1997]. This has raised some confusion in the literature, even disputing the empirical fulfillment of the fundamental equation. The whole problem lies in having defined and measured the traffic variables in some other ways (e.g., using the time mean speed). In addition, well-defined bivariate relations exist among traffic variables, such as flow and occupancy, when traffic conditions are approximately stationary [Cassidy, 1998]. To translate this relation to the traffic fundamental diagram (relating flow to density), it is necessary that the ratio of occupancy to density is equal to the average effective vehicle length "$\bar{g}$" (i.e., $\bar{g} = occupancy/k$), where again this average must be a space mean in order for the equation to hold [Daganzo, 1997; Cassidy, 1998]. See Section 3 for details. Therefore, space means are necessary within traffic stream models and thus critical for an accurate modeling of traffic stream behavior.

Not only traffic flow theory relies on space mean speed, but also simple practical applications. For instance, the average travel time on a road segment can be computed as the ratio between the length of the target road segment and the average speed of the vehicles, where this average must be the space mean speed for the computation to be accurate. This statement is also proved in the next section. To sum up, there is a clear need for space mean speeds either for modeling or practical purposes, and this average speed is not available from the common operations of the TMCs.

The present paper develops a methodology to accurately estimate the space mean speed of a traffic stream from time mean speed only using the common available time aggregations of loop detector data already available at TMCs, without any kind of modification to the detector roadside controller standard computations. The paper is organized as follows: firstly, in Sections 2 and 3, the traffic fundamental variables are defined and common methods to obtain them from loop detectors are presented. Next, in Sections 4 and 5, the proposed method is developed and a formulation for the confidence intervals for the results is obtained. Then, Section 6 presents the test data obtained from a double loop detector on a Spanish highway and Section 7 illustrates some results from the application of the method. Section 8 discusses the assumptions of the method and finally, general conclusions and issues for further research are discussed.

2. BACKGROUND

As stated before, traffic variables are usually measured at a particular spot (i.e., the loop detector site) and over a period of time. Time averages are clear, but how can space means be calculated from this punctual surveillance infrastructure, if spatial measurements over a length of the road are required? The answer is that they cannot be calculated using the simple definition of the space mean. To solve this problem, Edie (1965), proposed a family of definitions for "*q*", "*k*" and "$\bar{v}$" valid for any region "*A*" in the space (x) - time (t) plane. These are:

$$q = \frac{\sum_i x_i}{|A|} \tag{1}$$

$$k = \frac{\sum_i t_i}{|A|} \tag{2}$$

$$\bar{v} = \frac{\sum_i x_i}{\sum_i t_i} \tag{3}$$

Where:

"x_i" is the distance traveled by the ith vehicle in the region "*A*".

"t_i" is the time spent by the ith vehicle in the region "*A*".

"$|A|$" is the Euclidean area of region "*A*", having dimensions of (length)·(time), where the length dimension corresponds to the physical length (e.g. km) of the road segment, independently of the number of lanes.

The usual definitions of traffic variables could be considered as special cases of Edie's generalized definitions. For example, the definition of SMS as the arithmetic average of the "*n*" speed measurements taken at a particular instant "t_1" of the vehicles contained on a road segment of length "*L*" has the interpretation as the special case in which "*A*" is chosen as a narrow rectangular strip in the *x-t* plane with a length equal to "*L*" in space and differential temporal width "*dt*". Only area wide traffic sensors like the out-dated aerial photographs or the emerging video imaging or GPS tracking are able to measure in such *x-t* regions.

The generalized definition of average speed proposed by Edie is neither a time mean speed nor a space mean speed (it is only equal to SMS in the particular case exemplified above, or in case of stationary traffic where Equations 1, 2 and 3 do not depend on the measurement region "*A*", and therefore the generalized definitions of traffic characteristics for all space-time measurement regions are equivalent). Nevertheless, the generalized definition of average speed is the one that relates flow and density in any region of the *x-t* plane. Therefore Edie's definitions are appropriate definitions to treat traffic variables. Note that the traffic fundamental equation is true by definition when the variables are defined in the manner described by Edie (Equations 1, 2 and 3). In an abuse of notation, in the rest of the paper, the generalized definition of average speed (Equation 3) computed over a narrow rectangular strip in the *x-t* plane with a differential spatial width "*dx*" and a time length equal to "*T*" (this is the measurement region of a loop detector on a highway), will be named space-mean speed, "$\overline{v_s}$" (or SMS indistinctly), although the measurements do not have the spatial implications of the original space mean speed definition, unless traffic is stationary. This is:

$$\overline{v_s} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} t_i} = \frac{n \cdot dx}{\sum_{i=1}^{n} \frac{dx}{v_i}} = \frac{1}{1/n \cdot \sum_{i=1}^{n} \frac{1}{v_i}} \tag{4}$$

In contrast to the time mean speed "$\overline{v_t}$" (or TMS) in the same *x-t* region:

$$\overline{v_t} = \frac{\sum_{i=1}^{n} v_i}{n} \tag{5}$$

Note that space mean speed at a loop detector site (recall that from now on this refers to the generalized definition of speed on a detector measurement region) is the harmonic mean of the speeds of vehicles passing the detector spot during a time interval, while the time mean speed is the arithmetic mean of these individual speeds. As stated before, the space mean speed is the one that relates the average travel time over a highway section, "$\overline{TT}$", with the length of this highway section. If each vehicle speed is assumed to remain constant within the section:

$$\overline{TT} = \frac{\sum_{i=1}^{n} TT_i}{n} = \frac{\sum_{i=1}^{n} \frac{L}{v_i}}{n} = L \cdot 1/n \cdot \sum_{i=1}^{n} \frac{1}{v_i} = \frac{L}{\overline{v_s}} \tag{6}$$

Both speed averages are related and their relationships can be found in the literature. The first and most famous relationship, was derived by Wardrop (1952), and allows the estimation of "$\overline{v_t}$" from "$\overline{v_s}$" and the individual speeds variance over "$\overline{v_s}$".Wardrop's relationship is expressed as follows:

$$\overline{v_t} = \overline{v_s} + \frac{\sigma_s^2}{\overline{v_s}} \tag{7}$$

The usefulness of Wardrop's equation is limited, because the required relationship usually goes all the way around (i.e., it is needed to obtain "$\overline{v_s}$" from "$\overline{v_t}$"), which is not possible from Wardrop's relationship. Recently, Rakha and Zhang (2005) proved the complementary relationship, previously published in Khisty and Lall, (2003):

$$\overline{v_s} = \overline{v_t} - \frac{\sigma_t^2}{\overline{v_t}} \tag{8}$$

Where "σ_t^2" is the speed variance over "$\overline{v_t}$". Recall that Equations 4, 6, 7 and 8 valid for all traffic situations over the space-time measurement region of a loop detector, will only correspond to the "true" space mean speed measured over a long highway section and at a particular time instant in case of stationary traffic.

Equation 8 is still not applicable, as the speed variance is usually not obtained from loop detector time aggregations. One possible solution for this drawback is proposed in Garber and Hoel (2002), where a linear relationship between "$\overline{v_t}$" from "$\overline{v_s}$" is claimed (e.g., $\overline{v_t} = 0.966 \cdot \overline{v_s} + 3.541$). These types of equations, which result from linear regression of a particular database, do not reflect the nature of the differences between speed averages, and are, obviously, site specific [Rakha and Zhang, 2005].

The developed methodology estimates the space mean speed of a traffic stream from the measured time mean speed without previous knowledge of speed variance and avoiding having to resort to naïve site specific regressions. The proposed method uses the time mean speed and stratified vehicle counts in different speed thresholds to estimate the speed variance over the arithmetic mean of individual speeds "$\overline{v_t}$". The main assumption of the method is the normality of the speed probability distribution. This assumption is also discussed in the paper. Finally, Equation 8 is used to obtain the "$\overline{v_s}$" estimation. A formulation to obtain a confidence interval for "$\overline{v_s}$" estimation is also developed.

3. COMMON AVAILABLE MEASUREMENTS FROM INDUCTIVE LOOP DETECTORS

An inductive loop detector consists of a wire loop installed under the pavement of a particular lane of a highway that detects the presence of a metallic object (e.g., a vehicle) above it by monitoring the change in the electromagnetic properties of the loop. When a vehicle enters the detection zone, the sensor is activated (e.g., signal = 1) and remains so until the vehicle leaves the detection zone (e.g., signal = 0). See Figure 1.

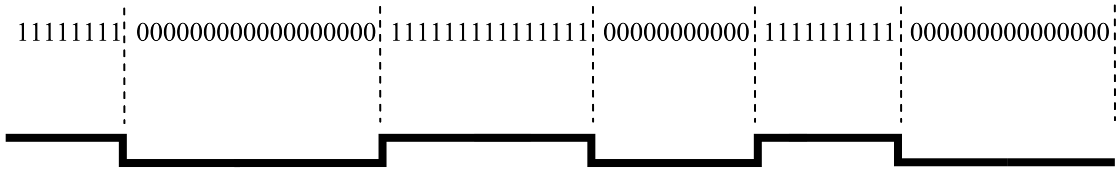
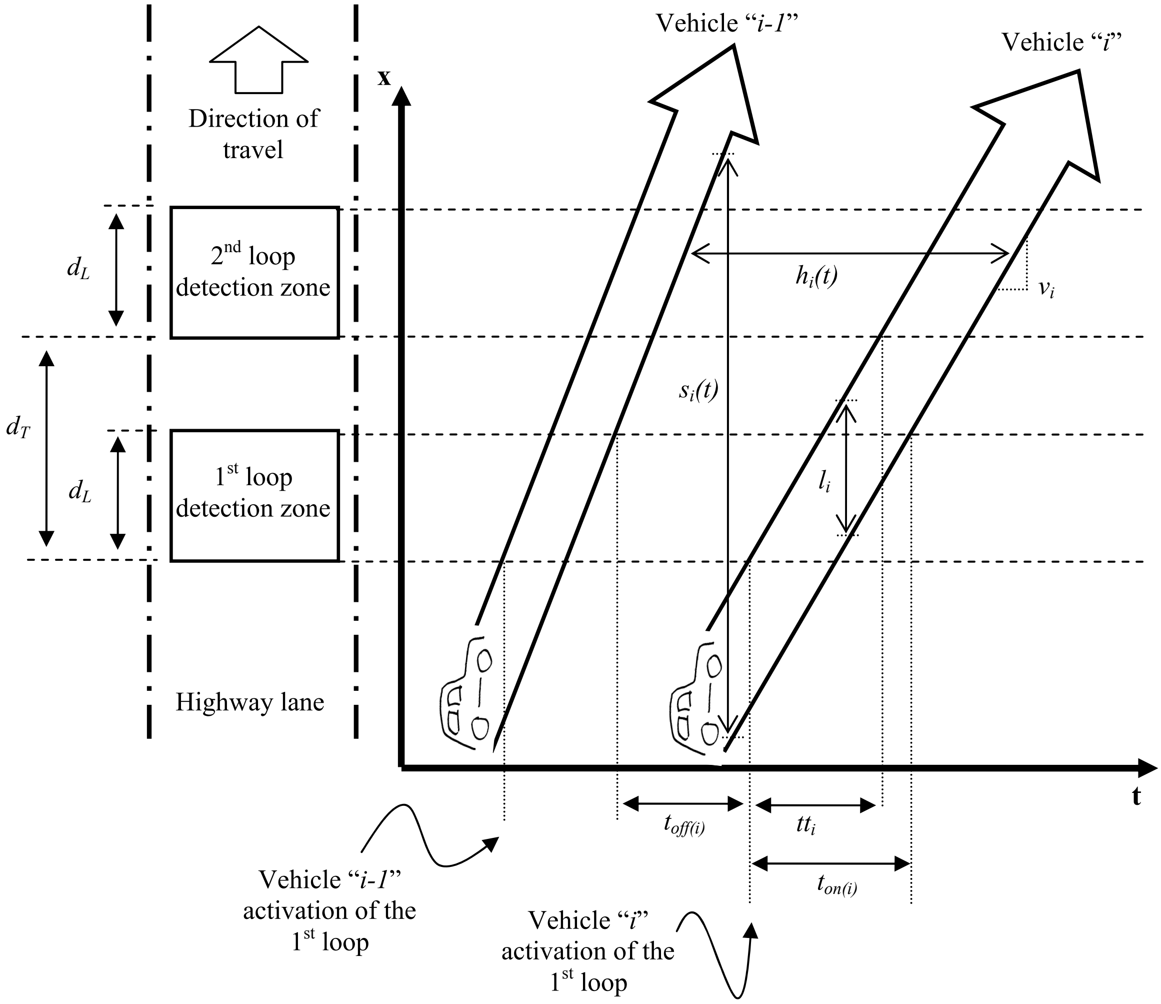


FIGURE 1 Output signals from a presence-type detector.
Source: adapted from May (1990)

Nowadays, common installations consist of two closely spaced wire loops, namely a double-loop detector, dual-loop detector or speed trap. Double loop detectors are able to accurately measure speeds in relation to the vague estimations of single loops that will be described next.



Legend:

d_L	Length of the detection zone of the loop; typically = 2m
d_T	Distance between equivalent points of 1st and 2nd loops; typically = 3.5m
v_i	Speed of vehicle "*i*"
l_i	Length of vehicle "*i*"
$h_i(t)$	Headway between vehicle "*i-1*" and vehicle "*i*", as a function of time
$s_i(t)$	Spacing between vehicle "*i-1*" and vehicle "*i*" , as a function of time
tt_i	Time between the activations of the 1st and the 2nd loop in the passage of vehicle "*i*"
$t_{off(i)}$	Off time of the 1st loop (i.e. time the 1st loop has remained off since the passage of the last vehicle, "*i-1*")
$t_{on(i)}$	On time of the 1st loop (i.e. time the 1st loop has remained on in the passage of vehicle "*i*")

FIGURE 2 Trajectory diagram of two vehicles passing over a double loop detector.

The length of the detection zone "d_L" (see Figure 2) corresponds to the "electrical" length of the loop which may not be the same as the physical length due to the fringing fields. This difference is usually ignored in practice, introducing a bias in all the measurements involving "d_L". The use of "d_L" is more relevant in single loop configurations. However, when double loops are available, the distance between equivalent points of 1st and 2nd loops "d_T" is of greater use (unless in case of estimating the vehicle length). Note that in this case the "electrical" distance corresponding to "d_T" is equivalent to the physical distance, provided that both loops are identical and they are equally and carefully installed. In any case a best practice rule is to calibrate the "electrical distances "d_L" and "d_T" regularly.

From the passage of a vehicle over a double loop detector, four basic measurements can be obtained (see Figure 2):

1. Instant of activation of the first loop (i.e., a vehicle has entered the detection zone).
2. Time between the activations of the first and the second loop, “tti”, in the passage of vehicle “i”.
3. Off time of the first loop (i.e., time the first loop has remained off since the passage of the last vehicle, “i-1”), “toff(i)”.
4. On time of the first loop (i.e., time the first loop has remained on), “ton(i)”, in the passage of vehicle “i”.

Note that 1, 3 and 4 could equally be defined in relation to the second loop.

From these four basic measurements all the microscopic traffic variables can be easily obtained:

$$v_i = \frac{d_T}{tt_i} \tag{9}$$

$$h_i = t_{off(i)} + t_{on(i)} \tag{10}$$

$$s_i = h_i \cdot v_{i-1} \tag{11}$$

$$l_i = v_i \cdot t_{on(i)} - d_L \tag{12}$$

Averaging these measurements in a way consistent with Edie’s generalized definitions, over a time period “*T*” in which the detector has had “*n*” activations, the macroscopic characteristics of the traffic stream can be accurately obtained as:

$$q = \frac{\sum_{i=1}^{n} x_i}{|A|} = \frac{n \cdot d_L}{T \cdot d_L} = \frac{n}{T} \tag{13}$$

$$\overline{v_s} = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{v_i}} = \frac{1}{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{tt_i}{d_T}} = \frac{n \cdot d_T}{\sum_{i=1}^{n} tt_i} \tag{14}$$

$$k = \frac{\sum_{i=1}^{n} t_i}{|A|} = \frac{\sum_{i=1}^{n} \frac{d_L}{v_i}}{T \cdot d_L} = \frac{\sum_{i=1}^{n} \frac{tt_i}{d_T}}{T} = \frac{\sum_{i=1}^{n} tt_i}{T \cdot d_T} \tag{15}$$

Obviously, the traffic fundamental equation (i.e. $q = k \cdot \overline{v}$), holds.

As stated before, these calculations (Equations 13-15) are not a common practice at detector roadside controllers, although exceptions exist. Usually, and in particular if one considers the Spanish standards in loop data treatment, the data traditionally sent to the TMC every aggregation period, “*T*”, results from the following variables:

- "n", the traffic count of the detector during "T"
- "$\overline{v_t}$", the time mean speed

$$\overline{v_t} = \frac{\sum_{i=1}^{n} v_i}{n} = \frac{\sum_{i=1}^{n} \frac{d_T}{tt_i}}{n} = \frac{d_T \cdot \sum_{i=1}^{n} \frac{1}{tt_i}}{n} \qquad (16)$$

- "occ", the occupancy of the 1st loop, defined as the time the 1st loop has remained on over the whole aggregation period:

$$occ = \frac{\sum_{i=1}^{n} t_{on(i)}}{T} \qquad (17)$$

- "$\overline{l_t}$", the average across time of vehicle lengths:

$$\overline{l_t} = \frac{\sum_{i=1}^{n} l_i}{n} = \frac{\sum_{i=1}^{n} (v_i \cdot t_{on(i)} - d_L)}{n} = \frac{\sum_{i=1}^{n} \frac{d_T}{tt_i} \cdot t_{on(i)}}{n} - d_L = \frac{d_T \cdot \sum_{i=1}^{n} \frac{t_{on(i)}}{tt_i}}{n} - d_L \qquad (18)$$

- "n_{v*}", the count of vehicles travelling at a speed lower than a speed threshold "$v*$". This is simply a stratification of the traffic count "n" during "T" in two groups considering the individual vehicle speed measured at the detector site. "n_{v*}" can be expressed in terms of "tt_i" as the aggregate count of vehicles during "T" that fulfill "$tt_i \geq d_T / v*$" (usually two speed thresholds are considered $v_1 = 50$ km/h and $v_2 = 100$ km/h, resulting "n_{v1}" and "n_{v2}")

- "n_{L*}", the count of vehicles whose length is shorter than a length threshold "$L*$". Again, this is simply a stratification of the traffic count "n" during "T" in two groups considering the individual vehicle length measured at the detector site. Analogously "n_{L*}" can be expressed in terms of "tt_i" and "$t_{on(i)}$" as the aggregate count of vehicles during "T" that fulfill "$t_{on(i)} / tt_i \leq (L*+d_L) / d_T$" (usually two length thresholds are considered $L_1 = 5$ m and $L_2 = 10$ m, resulting "n_{L1}" and "n_{L2}" which are commonly used to estimate the traffic composition between cars and long vehicles)

It is important to realize that from these calculations performed at the detector controller and sent to the TMC it is not possible to obtain directly either the space mean speed "$\overline{v_s}$" or the density "k" defined in Equations 14 and 15, resulting in the drawbacks stated in the introduction to the paper. In particular, note that "k", which could be obtained as the occupancy divided by the average effective length "$\overline{g}$" (i.e., average length of the vehicle plus the length of the detection zone), cannot be obtained from the variables sent to the TMC:

$$k=\frac{\sum_{i=1}^{n}t_i}{|A|}=\frac{\sum_{i=1}^{n}\left(t_{on(i)}-\frac{l_i}{v_i}\right)}{T\cdot d_L}=\frac{Occ}{d_L}-\frac{\sum_{i=1}^{n}l_i\cdot p_i}{T\cdot d_L}=\frac{Occ}{d_L}-\frac{{}^{1}\!/_{n}\cdot\sum_{i=1}^{n}l_i\cdot p_i}{{}^{1}\!/_{q}\cdot d_L}=$$

$$=\frac{Occ}{d_L}-\frac{{}^{1}\!/_{n}\cdot k\cdot\overline{v_s}\cdot\sum_{i=1}^{n}l_i\cdot p_i}{d_L}=\frac{Occ}{d_L}-\frac{{}^{1}\!/_{n}\cdot k\cdot\frac{n}{\sum_{i=1}^{n}p_i}\cdot\sum_{i=1}^{n}l_i\cdot p_i}{d_L}= \tag{19}$$

$$=\frac{Occ}{d_L}-\frac{k\cdot\left[\frac{\sum_{i=1}^{n}l_i\cdot p_i}{\sum_{i=1}^{n}p_i}\right]\cdot}{d_L}=\frac{Occ}{d_L}-\frac{k\cdot\overline{l_s}\cdot}{d_L}$$

Where "p_i" is the pace of vehicle "i", obtained as the inverse of its speed. Then:

$$k\left[1+\frac{\overline{l_s}\cdot}{d_L}\right]=\frac{Occ}{d_L} \tag{20}$$

$$k=\frac{occ}{d_L+\overline{l_s}}=\frac{occ}{\overline{g}} \tag{21}$$

Note that "$\overline{l_s}$", the average vehicle length weighted by the paces is different in relation to "$\overline{l_t}$" commonly obtained from loop detector time aggregations of data. They could be considered as equal if it is assumed that vehicle velocity and length are uncorrelated.

From all these calculations, it can be concluded that three time aggregations of data at the detector controller would be enough to compute all the traffic fundamental variables in a number of different ways, once sent to the TMC. These basic time aggregations should be "n", "$\sum_{i=1}^{n}tt_i$" and "$\sum_{i=1}^{n}t_{on(i)}$". Among these, "$\sum_{i=1}^{n}tt_i$" is the only aggregation that is not commonly computed. Instead, "$\sum_{i=1}^{n}\frac{1}{tt_i}$" is available.

Although these simple modifications at the roadside detector controller and in the TMC computation procedures would solve all the problems of accurate measurement of traffic variables, it is in practice a hard task to modify these standards. Therefore, traffic researchers have to look for methodologies to estimate, for instance space mean speed, from the available time aggregations of loop detector data. This is what the present paper aims to. Take into account that the proposed methodology relies heavily in the availability of "n_{v*}" at the TMCs. While this is fulfilled in all Spanish freeway traffic management centers, and it is quite common in Europe, it is certainly not a standard in the USA. If "n_{v*}" is not reported to the TMC in the normal functioning of the system, the proposed method can't be applied. Modifications to the roadside controller in order to obtain "n_{v*}" are on the wrong track, as in this case it would be simpler to work for applying Equation 4 or 14 directly.

The proposed methodology presented in the next section could also be useful for application in the case of single loop detectors. There is only one difference in relation to using the double detector, although this is a big difference, and that is the impossibility of measuring "tt_i". Therefore, vehicle individual speeds cannot be calculated, which implies that neither spacing nor vehicle length can be obtained. The traditional practice to overcome this limitation is based on the assumption of a constant average vehicle length, which should be site specific, depending on the composition of traffic and which lane of the highway is being considered. Several studies reveal that this assumption provides speed estimates that are much too inaccurate to be used for real time management and traveler information systems, and present research efforts to improve speed estimates from single loop detectors [Dailey, 1999; Coifman, 2001; Coifman et al., 2003; Hellinga, 2002; Lin et al., 2004; Wang and Nihan, 2000, 2003], achieving promising results, although very few references take into account the practical difficulties in modification of the loop detector controller procedures. Taking this into account, the vehicle "*i*" individual speed could be obtained from Equation 12 as:

$$v_i = \frac{\bar{l} + d_L}{t_{on(i)}} \tag{22}$$

Where "$\bar{l}$" is the assumed average spatial vehicle length. If the average speed is computed as the arithmetic average of these speeds, the problems in obtaining the space mean speed remains and the proposed methodology holds, considering that the quality of the results will be clouded by the original inaccuracy of speed estimates. However, the common practice when using single loop detectors is to compute average speed from time aggregation of "$t_{on(i)}$", as:

$$\bar{v}_S = \frac{n \cdot \left(\bar{l} + d_L\right)}{\sum_{i=1}^{n} t_{on(i)}} = \frac{q \cdot \left(\bar{l} + d_L\right)}{occ} \tag{23}$$

Note that Equation 23 is equivalent to Equation 21. Therefore, the space mean speed is directly obtained from these calculations, and there is no need to use the proposed method.

4. A METHOD FOR ESTIMATING SPACE MEAN SPEEDS FROM TIME AGGREGATIONS OF LOOP DETECTOR DATA

Consider a particular highway lane, and assume that the vehicle speed distribution over the time aggregation period "*T*" follows a normal distribution (this assumption will be discussed in Section 8). Then, the probability of a vehicle traveling at a speed lower than "*v**" (a particular speed threshold) could be estimated as:

$$\Pr(V \leq v^*) \approx \frac{n_{v^*}}{n} = F(Z_{(v^*)}) \qquad (24)$$

Where "*F(z)*" is the cumulative distribution function of a standard normal distribution (i.e., expectation equal to zero and standard deviation equal to 1). "$Z_{(v^*)}$" is the standardized speed threshold which can be expressed as:

$$Z_{(v^*)} = \frac{v^* - \overline{v_t}}{\sigma_t} \qquad (25)$$

Approximating the expectation of the speed random variable, "*V*", by its unbiased estimator "$\overline{v_t}$" (i.e., the arithmetic mean of individual speeds or time mean speed). Recall that "σ_t" is the standard deviation of vehicle speeds over the arithmetic mean, "$\overline{v_t}$". Then:

$$F^{-1}\left[\frac{n_{v^*}}{n}\right] = \frac{v^* - \overline{v_t}}{\sigma_t} \qquad (26)$$

Where"$F^{-1}(z)$" is the inverse cumulative distribution function of a standard normal probability distribution. Therefore the unique unknown in Equation 26 is "σ_t":

$$\sigma_t = \frac{v^* - \overline{v_t}}{F^{-1}\left[\frac{n_{v^*}}{n}\right]} \qquad (27)$$

Finally, applying Equation 8, the space mean speed of the traffic stream, "$\overline{v_s}$" can be obtained.

Note that the proposed method can be applied for any speed threshold, "*v**", given that "$v^* \neq \overline{v_t}$", "$n_{v^*} \neq 0$" and "$n_{v^*} \neq n$", although very unreliable estimates would be obtained if these inequalities hold only by little (and consequently poor confidence intervals for the estimation would be obtained, as it will be seen next). In addition, using the observed cumulative frequency to replace the theoretical probability in Equation 24 would be problematic in case of very low "*n*", leading also to an unreliable estimation. This situation, which would arise frequently during non-peak times in case of using a short aggregation period "*T*" (e.g. the 20 or 30 seconds quite common in North America), must be avoided by using a higher low threshold for the aggregation period, which can always result from the addition of shorter periods. All these sources of error are analyzed in more detail in Section 7.

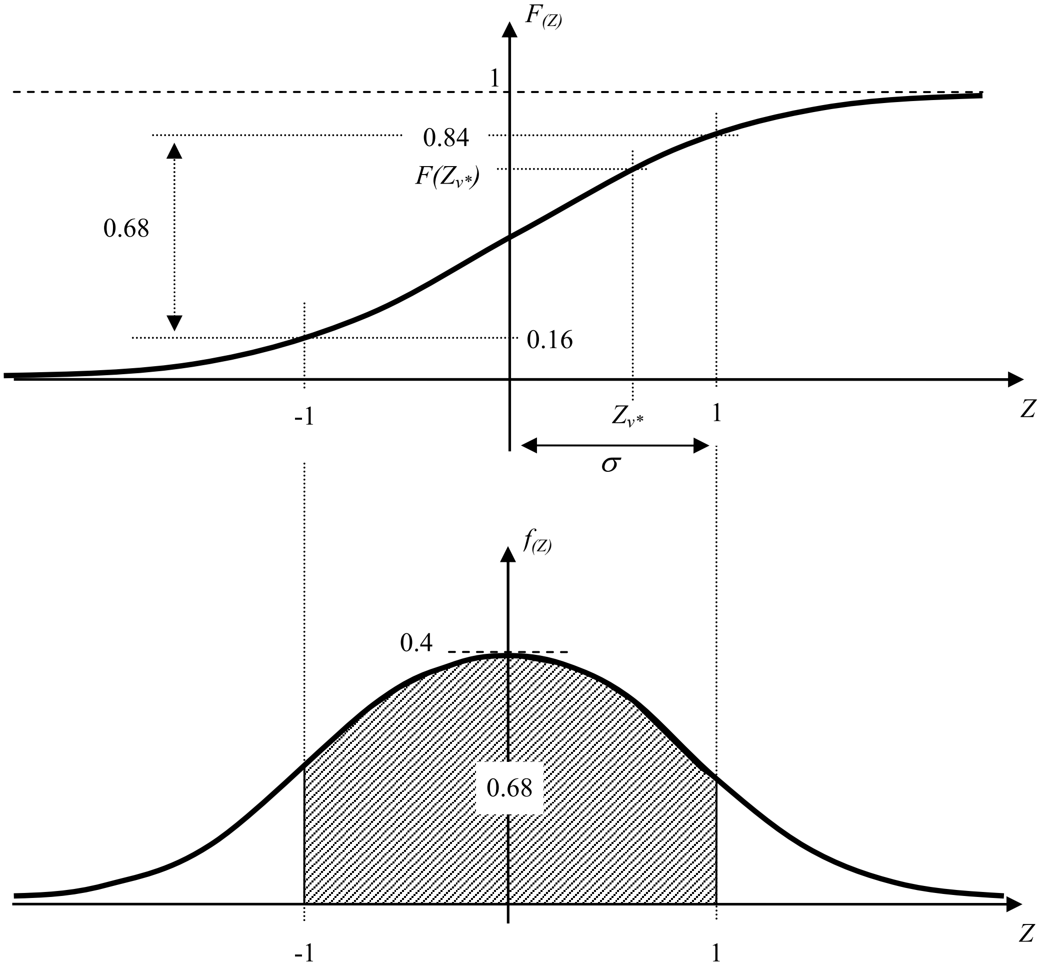


FIGURE 3 Plot of the cumulative distribution function, "*F(z)*", and of the probability density function, "*f(z)*", of a standard normal probability distribution.

5. CONFIDENCE INTERVALS FOR THE ESTIMATED SPACE MEAN SPEEDS

This section provides an analytical formulation for calculating confidence intervals for the space mean speed estimation. The confidence interval delimits the error of the estimation within the limits of the interval, considering a particular confidence level. This allows the user to set a maximum acceptable error so that there is a low probability (i.e., the complementary of the confidence level on the interval) of exceeding it.

To formulate this confidence interval, imagine an observer located at the loop detector site whose objective is to classify the passing vehicles into two categories: those whose speed is lower than a particular threshold "$v*$" (that will sum n_{v*} after the passage of "n" vehicles in the time period "T") and the rest. Each classification can be seen as a Bernoulli trial (i.e., an experiment whose outcome is random and can be either of two possible outcomes, "success = 1" and "failure = 0"). In the passing of every vehicle, the random variable "X" which defines the result of the classification can be expressed as:

$$X = \begin{cases} 1 & if \quad v \le v^* \\ 0 & if \quad v > v^* \end{cases} \tag{28}$$

If "p" is the probability of obtaining "X=1" and "$1-p$" the probability of obtaining "X=0" (the probability distribution of "X"), then the expected value of "X" and its variance are given by:

$$E[X] = 1 \cdot p + 0 \cdot (1-p) = p \tag{29}$$

$$Var[X] = E[X^2] - (E[X])^2 = p - p^2 = p(1-p) \tag{30}$$

If one considers the speeds of the passing vehicles as independent observations of the random variable "V", the classification of the "n" vehicles that pass the loop detector over a period of time "T", results in a Bernoulli process (i.e., repeated independent but identical Bernoulli trials), with a "p" probability of success. In this context an unbiased estimator for "p" (in fact the maximum-likelihood estimate) is given by:

$$\hat{p} = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{n_{v*}}{n} \tag{31}$$

Therefore, in order to obtain a confidence interval for "$\frac{n_{v*}}{n}$" it is only necessary to calculate the standard deviation of the estimator "$\hat{p}$".

$$Var[\bar{x}] = Var\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = \frac{n \cdot Var[x_i]}{n^2} = \frac{p(1-p)}{n} \tag{32}$$

And to obtain the standard deviation, simply:

$$\sigma_{\bar{x}} = \sqrt{\frac{p(1-p)}{n}} \tag{33}$$

Given that $(X - \bar{x})/\sigma_{\bar{x}}$ has approximately a standard normal distribution when "X" is binomial and "n" is large, for a confidence level of 68%, the resulting confidence interval has its limits one standard deviation apart from the expected value (see Figure 3). This means that there will be a probability of 0.68 for the true expected value of "$\frac{n_{v*}}{n}$" to be contained in the confidence interval defined by:

$$\left\{ \frac{n_{v*}}{n} - \sqrt{\frac{\frac{n_{v*}}{n}\cdot\left(1-\frac{n_{v*}}{n}\right)}{n}} \quad , \quad \frac{n_{V*}}{n} + \sqrt{\frac{\frac{n_{v*}}{n}\cdot\left(1-\frac{n_{v*}}{n}\right)}{n}} \right\} \tag{34}$$

Note that in Equation 34, "*p*" has been substituted with its expected value.

Analogously, the true value of "*p*" can be expressed as the estimation of the expected value plus an error:

$$p = \hat{p} \pm \varepsilon_p \tag{35}$$

$$\varepsilon_p \in \left[-\sqrt{\frac{p(1-p)}{n}}, \sqrt{\frac{p(1-p)}{n}} \right] \tag{36}$$

And there is a probability of 0.68 that Equation 36 holds. Doubling the limits of the confidence interval defined in Equation 36 (i.e., two standard deviations), the confidence in Equation 36 increases approximately to 0.95.

It is equally simple to solve Equation 36 for "*n*", and therefore obtain an equation to estimate the number of vehicle passages "*n*" required to estimate "*p*" for a particular error level.

The maximum errors for "*p*" (i.e. the confidence interval limits) estimated in Equation 36 must be propagated using Equations 25-27 to finally obtain a confidence interval for "$\overline{v_s}$". This process is formulated in the following equations:

$$\begin{aligned} \varepsilon_{Z(1)} &= F^{-1}(p+\varepsilon_p) - F^{-1}(p) \\ \varepsilon_{Z(2)} &= F^{-1}(p-\varepsilon_p) - F^{-1}(p) \end{aligned} \tag{37}$$

$$\begin{aligned} \varepsilon_{\sigma_t(1)} &= -\frac{(v^* - \overline{v_t})\cdot\varepsilon_{Z(1)}}{Z\cdot(Z+\varepsilon_{Z(1)})} \\ \varepsilon_{\sigma_t(2)} &= -\frac{(v^* - \overline{v_t})\cdot\varepsilon_{Z(2)}}{Z\cdot(Z+\varepsilon_{Z(2)})} \end{aligned} \tag{38}$$

Note that the confidence interval for the estimated standard deviation of the speed is not symmetrical around the expected value. Finally, the confidence interval for the estimated "$\overline{v_s}$" is expressed as:

$$\begin{aligned} \varepsilon_{\overline{v_t}(1)} &= -\frac{\left(\varepsilon^2_{\sigma_t(1)} + 2\cdot\sigma_t\cdot\varepsilon_{\sigma_t(1)}\right)}{\overline{v_t}} \\ \varepsilon_{\overline{v_t}(2)} &= -\frac{\left(\varepsilon^2_{\sigma_t(2)} + 2\cdot\sigma_t\cdot\varepsilon_{\sigma_t(2)}\right)}{\overline{v_t}} \end{aligned} \tag{39}$$

6. THE DATA

The data extracted from the double loop detector shown in Figure 4 is used in the following section to prove the accuracy of the proposed methodology. The database used is very rich, in the sense that it includes individual vehicle actuations (i.e., the instants of actuation of each loop and the measured on and off times for the passage of every vehicle). This allows calculating the space mean speed of the vehicles on a per lane basis, as the harmonic mean of individual vehicle speeds, which will be the ground truth to which the results of the method will be compared. The space mean speed is not available in the regular functioning of the detector controller, as data is aggregated and time averaged every 3 minutes to be sent to the TMC. The computational procedures performed in this process are the common ones described in Section 3, so that only time mean speeds are available in the regular operation. The variables used as inputs for the method considering the time aggregation period "T"= 3 min are: "n", "$\overline{v_t}$", "n_{v1}" and "n_{v2}", where the speed thresholds "v_1" and "v_2" are set to 50 km/h and 100 km/h respectively. This data was taken from the three lanes of the AP-7 highway at the detector site and over a whole day.

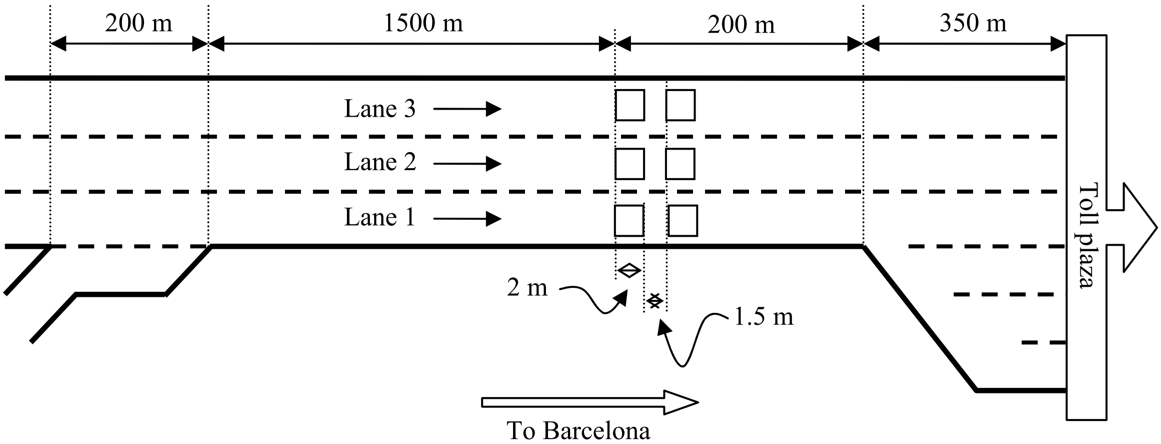


FIGURE 4 Data collection site, AP-7 highway, near Barcelona, Spain.

The test data, plotted in Figure 5, was obtained throughout a sunny Sunday on September 7th, 2008. On that day, congestion grew during evening and night. This congestion resulted from a high demand on the southbound direction of the AP-7 highway, caused by drivers wanting to return to Barcelona after spending the day or the weekend on the coast, as the AP-7 highway goes along the whole Spanish Mediterranean Coast.

The raw database was made up of 58,003 vehicle actuations for the three lanes and 24 hours of the day. 6% of these were vehicles longer than 5 meters (i.e. vans, trucks, buses …). The database can be decomposed in a per lane basis, resulting 17,496 vehicle actuations in the rightmost lane (i.e. lane 1) with a 14% of long vehicles, 22,575 vehicles in the central lane (i.e. lane 2) with a 5% of long vehicles and 17,932 vehicles in the leftmost lane (i.e. lane 3), with an insignificant number of long vehicles. The data plotted in Figure 5 results from the 3 minute aggregation of these raw data (i.e. 480 time periods). The vehicle counts in each period range from 10 to 309 vehicles considering the three lanes as a whole. In a per lane basis, 3 minute counts range from 3 to 109 for lane 1, 4 to 106 for lane 2 and 0 to 123 for lane 3.

The possibility of computing both speed means, "$\overline{v_s}$" and "$\overline{v_t}$", as the harmonic and arithmetic means, respectively, of individual speeds over the aggregation period of 3 minutes of duration, allows gaining some empirical evidence about the differences between these two average speeds.

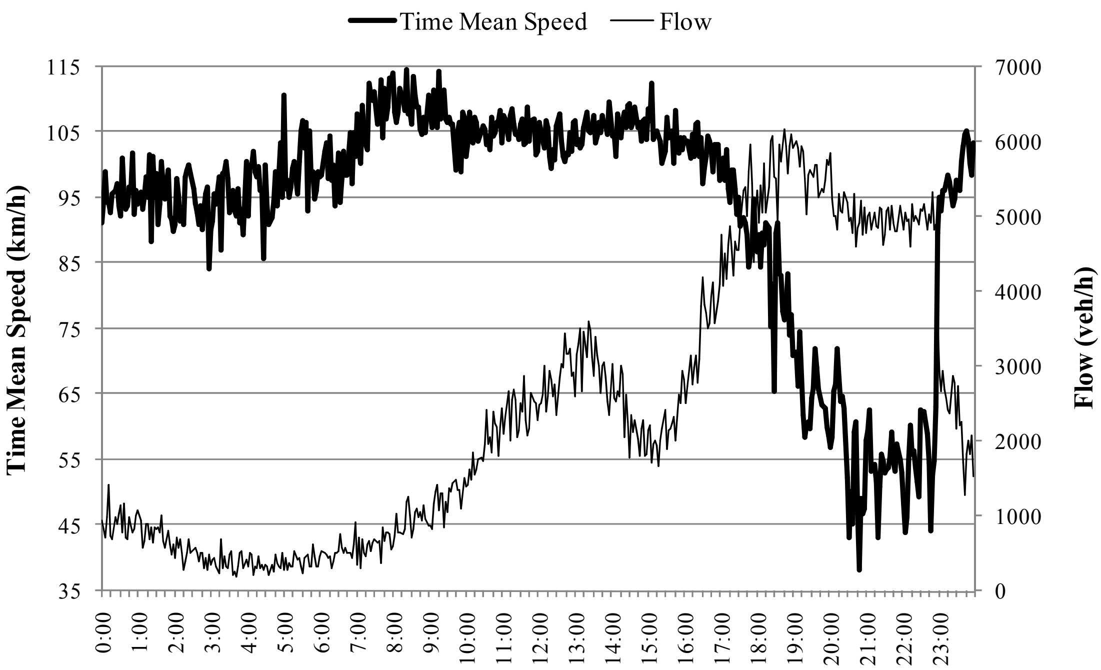


FIGURE 5 Flow and time mean speed over the three lanes and 3 minute periods at the detector site.

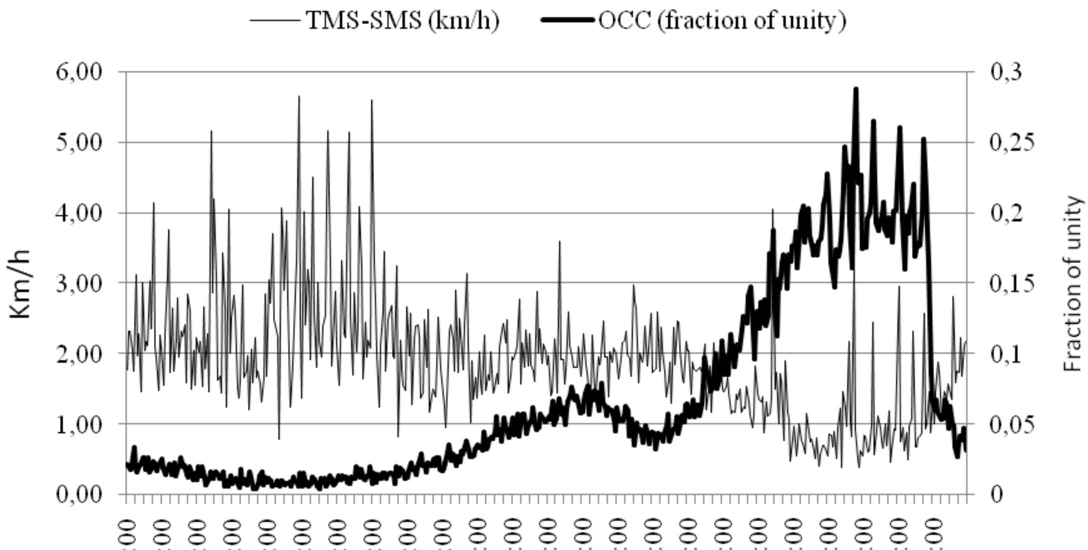


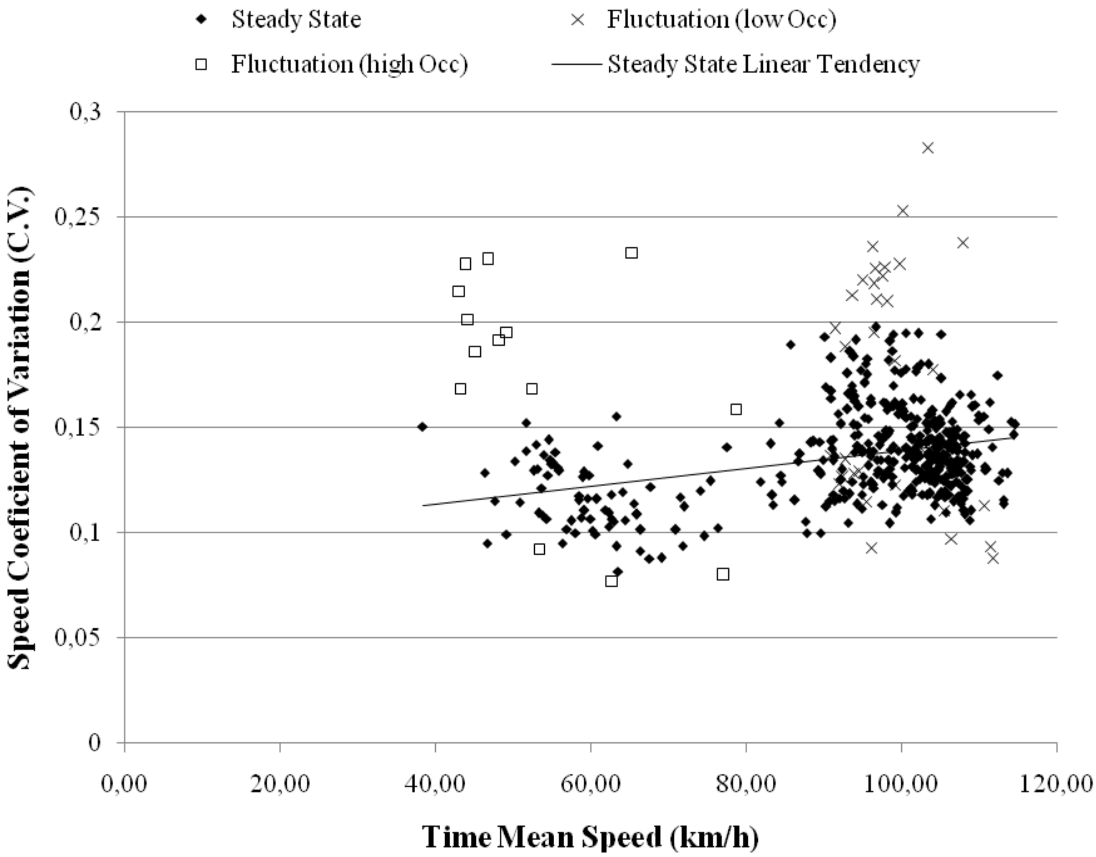
FIGURE 6 Difference between time mean speed and space mean speed over 3 minute periods considering the whole highway section.

It is interesting to note (from Equations 7 and 8) that the differences between "$\overline{v_s}$" and "$\overline{v_t}$", are directly related to the vehicle speed coefficient of variation "*C.V.*" defined as the standard deviation of the speed sample divided by the sample mean speed:

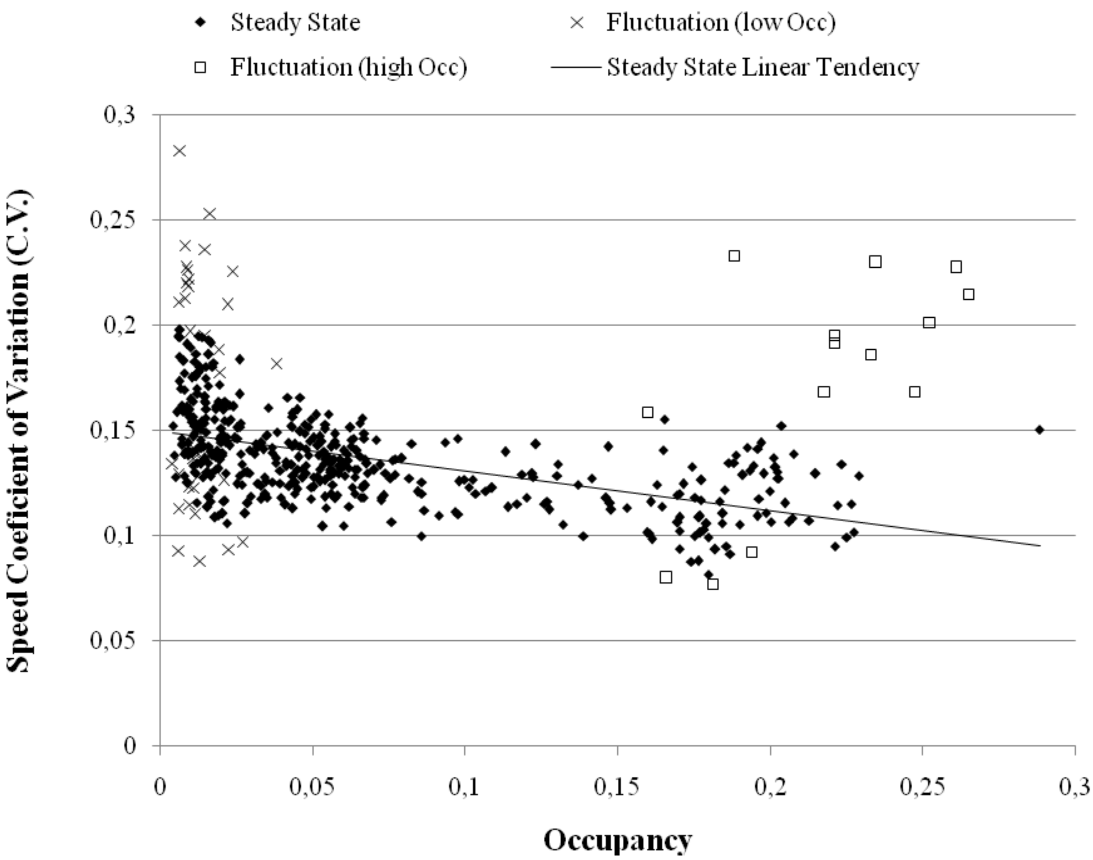
$$\overline{v_t} - \overline{v_s} = \frac{\sigma_t^2}{\overline{v_t}} = \frac{\sigma_s^2}{\overline{v_s}} = CV \cdot \sigma = CV^2 \cdot \overline{v} \qquad (40)$$

Larger differences should be obtained as the speed coefficient of variation becomes larger, and for a given "*C.V.*", greater absolute differences would occur when the mean speed is high. From Equation 40 it can be derived that relative differences are equal to the square of the speed coefficient of variation.

The speed coefficient of variation is not a constant parameter but a variable depending on the characteristics of the traffic stream. It should not come as a surprise that the speed variance is larger in a situation with no interactions between vehicles and no vehicle performance limitations (i.e., low traffic densities), and smaller when a high density of traffic stream forces all the vehicles to travel at the same speed. Therefore, as a monotonically decreasing relation exists between speed and density (see Figures 8 and 9), the speed variance should decrease as mean speed does. The relative reduction of the speed standard deviation in relation to the reduction of the mean speed results in the behavior of the speed coefficient of variation.



a)



b)
FIGURE 7 Scatter plot of speed coefficient of variation against time mean speed (a) and against occupancy (b) over 3 minute periods considering the whole highway section.

There is empirical evidence (Figure 7a) that speed C.V. decreases given a reduction of the mean speed resulting from a density increase in the highway. From this logic and considering Equation 40, it is evident that differences between time mean speed and space mean speed are reduced when the traffic mean speed decreases. However, it can be seen from the data (Figures 5, 6 and 7) that big differences can also be obtained when the traffic mean speed is low, and these would be the maximum relative differences in relation to the space mean speed. This fact, reported in the literature [Heidemann, 1986; May, 1990; Rakha and Zhang, 2005], adds some confusion to the matter. On the one hand the speed coefficient of variation should decrease with mean speed and on the other hand maximum empirical values are obtained for low speeds. The evidence of this contradiction resides in empirical measurements over transitional time periods. The theoretically derived behavior of the speed CV is completely true when considering a time interval contained in a unique stationary traffic state where stop&go traffic does not arise. Traffic states characterized by low mean speed and high density are likely to suffer stop&go situations. Therefore, it is also likely that the time period for calculation includes some transitions between stopped traffic and moving traffic, resulting in an increase of the speed coefficient of variation within this period. This does not rule out that if the next time period does not include any traffic breakdown, the resulting speed CV will be at a minimum. Figures 6 and 7b illustrate this behavior, as they show that sharp increases in mean speed differences obtained between 6:00 pm and 11:00 pm (congested period) match exactly with sharp increases in loop detector occupation, indicating that a traffic breakdown has occurred within the time period and vehicles have stopped for a while above the detector. So the expected behavior of the speed CV at low speed and high density traffic states is in its lower range and suffers random increases to higher values. The magnitude of these fluctuations depends on the duration of the calculation period, being maximum for durations similar to the time of transition from stopped to moving traffic or vice versa. This supports the relative differences between mean speeds of 30% reported in Rakha and Zhang (2005) for 30 second time periods. Numerical values for the 3 minute period considered in the present paper are shown in Table 1.

It is also noticeable from Figures 5, 6 and 7 that big fluctuations of the speed CV, and accordingly of the differences between space mean speed and time mean speed (Equation 40), occur when traffic is free-flowing, mean speed is high but flow is very low (e.g., from 0:00 am to 8:00 am, resulting in a low occupancy). These fluctuations, which can be either an increase or a decrease in the speed CV, must not be attributed to a real change in the mean behavior of the drivers on the highway, but a statistical error in the estimation of population dispersion due to a very small sample size (i.e., small count "*n*" in the three minute period).

TABLE 1 Numerical Differences Between Space Mean Speed and Time Mean Speed and the Corresponding Speed Coefficient of Variation for an Aggregation Period of 3 Minutes

		Occupancy < 0.025				0.025<Occupancy< 0.15				Occupancy > 0.15			
		Sec.	**L1**	**L2**	**L3**	**Sec.**	**L1**	**L2**	**L3**	**Sec.**	**L1**	**L2**	**L3**
Speed $C.V.$	**Med**	0,15	0,14	0,13	0,11	0,13	0,12	0,11	0,10	0,12	0,10	0,10	0,11
	Max	0,28	0,24	0,35	0,31	0,18	0,21	0,18	0,15	0,23	0,19	0,21	0,30
	Min	0,09	0,04	0,04	0,01	0,10	0,07	0,07	0,07	0,08	0,07	0,06	0,06
$TMS-SMS$ (km/h)	**Med**	2,24	0,57	1,70	0,71	1,81	1,24	1,08	1,12	0,84	0,56	0,61	0,70
	Max	5,67	0,44	7,69	7,59	3,59	3,68	2,84	2,32	4,05	1,69	2,69	6,46
	Min	0,79	0,15	0,14	0,00	0,89	0,47	0,32	0,48	0,38	0,22	0,22	0,22
$(TMS-SMS)/SMS$ (fract. unity)	**Med**	0,02	0,02	0,02	0,01	0,02	0,01	0,01	0,01	0,01	0,01	0,01	0,01
	Max	0,06	0,08	0,08	0,07	0,04	0,05	0,03	0,02	0,07	0,03	0,05	0,11
	Min	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,00	0,01	0,00	0,00	0,00

Note: Sec. = whole highway section; L1 = Lane 1; L2 = Lane 2; L3 = Lane 3.

From Table 1 it can be seen that for the considered 3 minute interval of aggregation and for the highway section as a whole, the speed $C.V.$ ranges from 0.08 to 0.28, but the values above 0.15 result from different types of fluctuation. It is also interesting to note the absence of fluctuations in the central range of occupancies, due to the considerable amount of free-flowing traffic. When analyzing the results on a per lane basis, the speed $C.V.$ is significantly lower in relation to the whole section. This is due to the fact that each driver selects a particular lane in accordance with his speed preferences, and therefore, variability is reduced. As an exception, lane 1 variability is higher due to the coexistence of cars and trucks.

7. ILLUSTRATING SOME RESULTS

The presented method for estimating the space mean speed from time aggregations of loop detector data was tested with the data presented in the previous section. Two estimations could be obtained for each time interval of aggregation, one from "n_{v1}" and the other from "n_{v2}". However, this almost never happens (i.e., in the lane where it happened more often, it only happened in 3 time periods out of the 480 of the whole day), due to the fact that it is very rare for the required conditions to obtain the estimation, "$n_{v*} \neq 0$" and "$n_{v*} \neq n$", are fulfilled for both speed thresholds, "v_1" and "v_2", simultaneously. In those cases where two estimations are obtained, the one with a smaller error (see Equation 39) is selected. Therefore, the usual result of the method is only one estimate, or none, as it is not rare that for both speed thresholds the conditions are not fulfilled. For instance, from Figure 8 it can be seen that it is likely that no estimation is obtained when time mean speed is around 70 km/h, as it is probable that "$n_{v1} = 0$" and "$n_{v2} = n$". The percentage of estimation can be seen in Table 2.

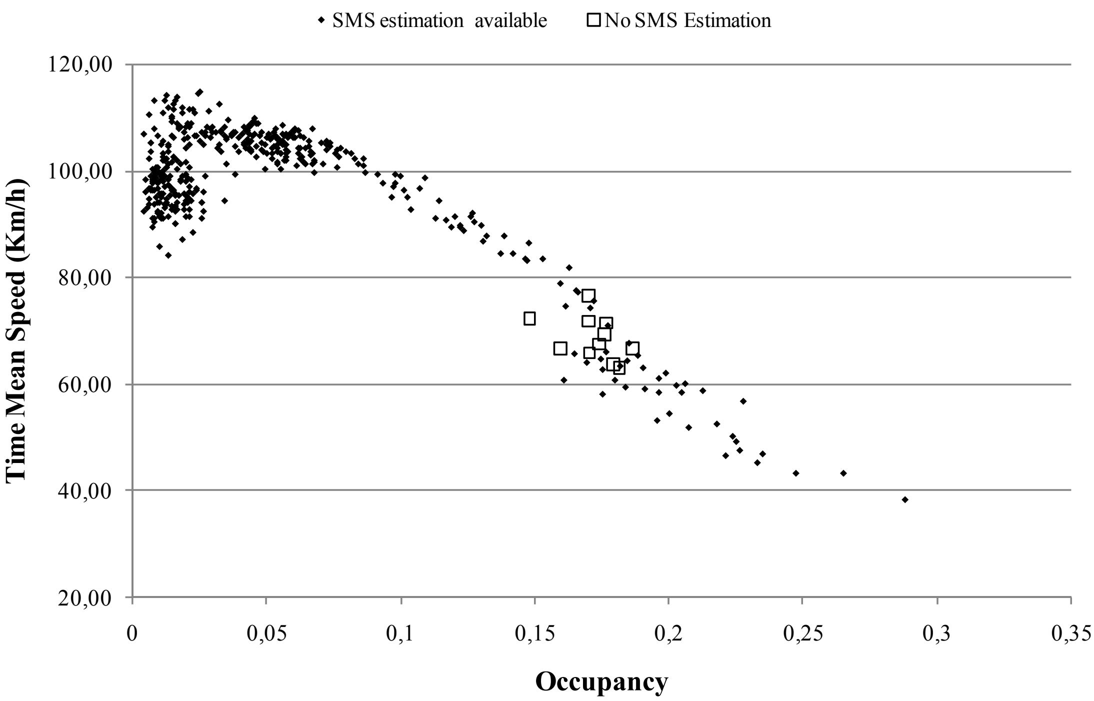
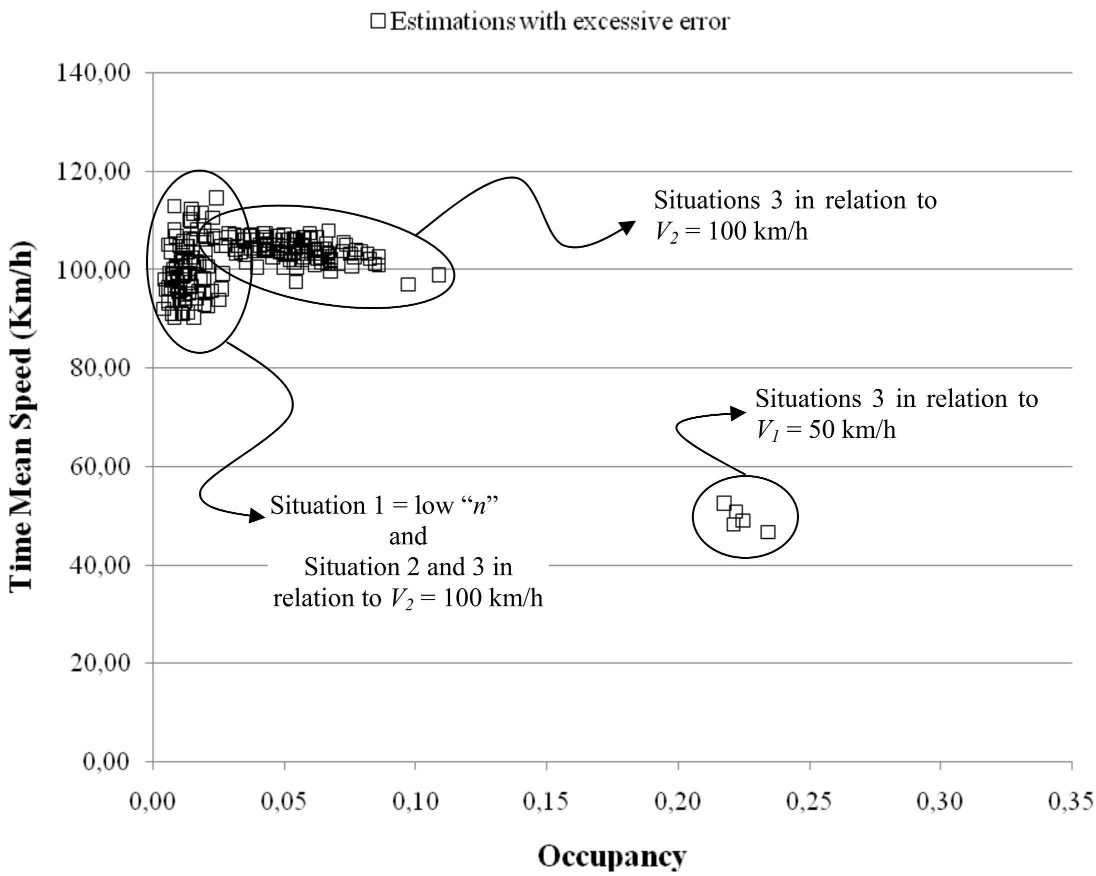


FIGURE 8 Available SMS estimations on a scatter plot of TMS against occupancy over 3 minute periods considering the whole highway section.

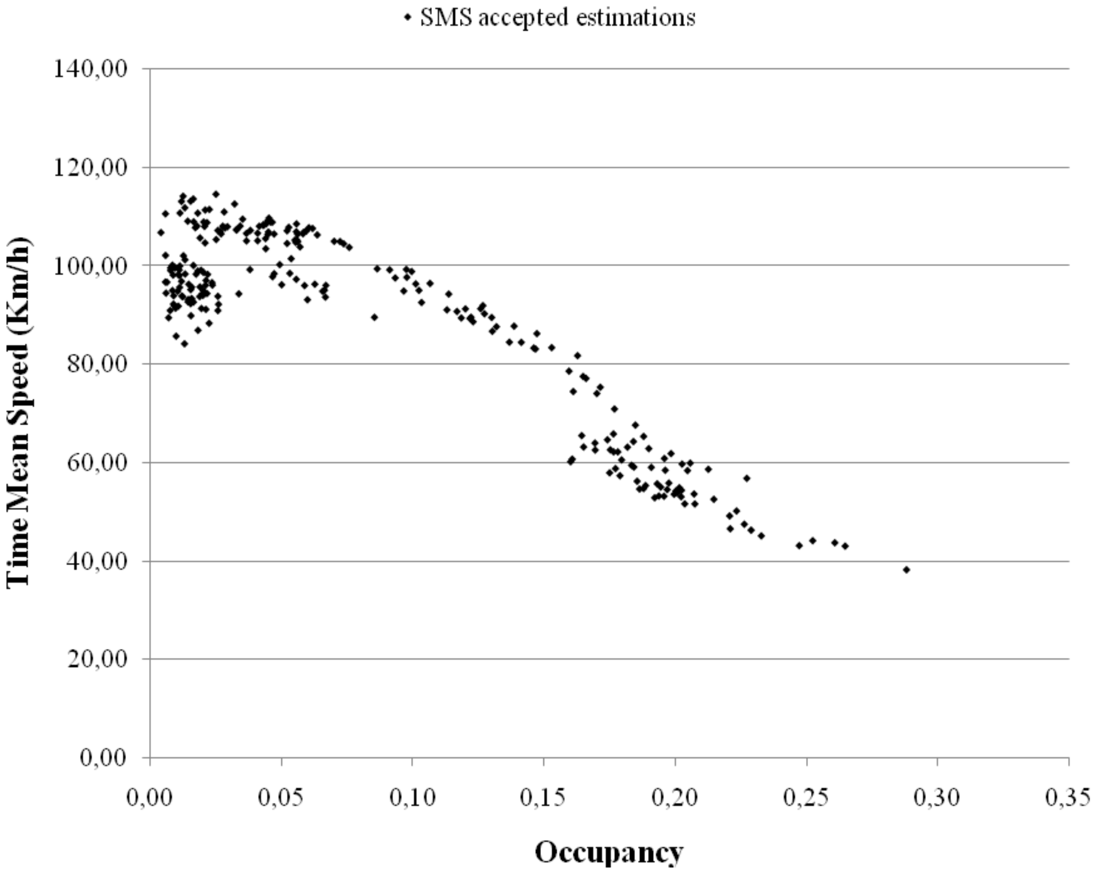
In addition, space mean speed estimation is only considered valid if the maximum error that can be made with a probability of 0.68 is lower than a particular threshold. This threshold is set to 3.5 km/h, taking into account that this is approximately the maximum difference between time mean speed and space mean speed without considering variance fluctuations resulting from measurement errors (see Figure 6 and Table 1). It would be fruitless to estimate "$\overline{v_s}$"from "$\overline{v_t}$", with an error larger than the original one, resulting from simply considering both speed means as equal, which is the common practice. The percentage of accepted estimation can be seen in Table 2.

There are three situations where the maximum probable error could be high, and thus three limitations for the applicability of the method. These situations are identified in Figure 9:

1. Low counts "*n*" on the three minute period result in high probable errors on the estimation of "*p*", as the observed cumulative frequency used to replace the theoretical probability (Equation 24) is problematic and the confidence interval using the normal approximation (Equation 36) would be very poor.
2. When "n_{v*}" is very small or approaches "*n*", as it would result with "$F(Z_{(v*)})$" close to zero or one. In this situation, small errors in the estimation of "*p*" turn into great errors in the estimation of "$Z_{(v*)}$" (Equation 37). This situation is likely to happen when an estimation exists and "*v**" is more than two standard deviations apart from "$\overline{v_t}$"
3. When "$\overline{v_t}$"approaches "*v**", as it would result with "$Z_{(v*)}$" close to zero. This situation results in high probable errors in the estimation of "σ_t" (Equation 38).



a)



b)
FIGURE 9 Scatter plot of time mean speed against occupancy over 3 minute periods considering the whole highway section, a) Ruled out estimations due to possible excessive error, b) Accepted estimations.

In light of previous remarks, there are time intervals without an accepted estimation of "$\overline{v_s}$". In these intervals, it is assumed that the speed variance remained constant since the last accepted estimation. With this assumption, "$\overline{v_s}$" is easily obtained using Equation 8.

Note from Figure 7 and Table 2 that the number of discarded estimations is significant in all situations. Some of the discarded periods correspond to night hours with very low "*n*" on the three minute period (situation 1). From this it can be derived the importance of the duration of the aggregation time interval in the method. For a given traffic demand, as longer are the intervals, higher counts will be obtained and thus greater accuracy in the estimation. The selection of the aggregation period (whose lower limit is the baseline aggregation period at the roadside detector controller) must respond to a trade-off between the average sample size and the accurate tracking of speed evolution for a given demand pattern. In addition, the assumption of normality of the speed distribution within the time period must be acceptable. This last requirement is analyzed in the next section. Then for periods with very low flows (e.g., night hours) it could be advisable to expand the aggregation interval in order to obtain a larger number of estimations (e.g. 15 minutes could be adequate). This would not imply any shortcoming to the tracking of speed evolution as for these free flowing and low flow periods speed is likely to remain fairly stable around free flow speed. For higher demand periods, a 3 minute aggregation interval is considered adequate to fulfill the previous mentioned trade-off. Take into account that longer intervals may smooth rapid evolving traffic conditions, while shorter intervals (many traffic agencies, particularly in the USA, use baseline time intervals which are 20 or 30 seconds long) result in small traffic counts "*n*" (maximum of 10 to 15 vehicles during peak hours, and can be zero in many night intervals) which would result in almost all estimations being discarded due to small statistical significance. Therefore, in order to apply the proposed method to this short baseline time intervals, they should be aggregated until the three minute period is achieved. In case higher granularity in the space mean speed estimation is desired to match the detector baseline time intervals, Equation 8 could be applied every 20 or 30 seconds, provided that the speed variance is obtained from a robust three minute estimation and assuming that it will remain constant until next acceptable estimation.

Finally, results of the proposed methodology are presented in Table 2, and are compared with the common practice of considering SMS and TMS as equal.

Table 2 proves the accuracy of the proposed method for estimating "$\overline{v_s}$". The mean error of the estimation is below 1 Km/h in all cases, which represents a relative error below 1%. Benefits in relation to considering "$\overline{v_s}$" equal to "$\overline{v_t}$" are significant in all situations. Although positive, results are poorer when considering maximum errors. This results from the fact that maximum errors are obtained in intervals when the normality assumption is less appropriate and the method is not capable of producing an accurate estimation.

Analyzing the results for the whole day (i.e. accepted estimations plus estimations maintaining the variance) it can be seen that the mean error increases slightly in relation to considering only accepted estimations. This supports the assumption that in most cases the speed variance stays almost constant for short time periods. Therefore, the existence of some periods where the size of, or the speeds in, the sample are not adequate does no limit the applicability of the method. As before, maximum errors do not behave in such a good way, and their reduction using the proposed method is less significant.

Also notice from Table 2 that per lane results are slightly better than considering the highway section as a whole. This is due to the fact that the assumption of normality of speed distribution is more acceptable on a per lane basis. This aspect is discussed in the next section. Therefore the recommended process for estimating the space mean speed for a whole highway section would be to apply the proposed method on a per lane basis and average the result over the whole section, taking into account:

$$\overline{v}_s^{(\mathrm{sec}\,tion)} = \frac{1}{\left(\dfrac{1}{\sum_i n^{(i)}}\right) \cdot \sum_i \left(n^{(i)} \Big/ \overline{v}_s^{(i)} \right)} \qquad (41)$$

Where "*i*" refers to the number of lanes.

TABLE 2 Results of the SMS Estimation from 3 Minute Time Aggregations of Loop Detector Data

			Whole section		Lane 1		Lane 2		Lane 3	
			Accepted	All day	Accepted	All day	Accepted	All day	Accepted	All day
No estimation			2 %		9 %		8 %		25 %	
Excessive error estimations			44 %		29 %		37 %		16 %	
Absolute error in relation to SMS	Estimated SMS	Mean (km/h)	0.65	0.79	0.46	0.66	0.48	0.71	0.32	0.63
		Max (km/h)	5.17	5.17	3.62	6.11	5.14	7.54	2.84	7.59
	TMS	Mean (km/h)	1.67	1.89	1.18	1.31	1.15	1.36	0.92	1.06
		Max (km/h)	5.17	5.67	3.68	6.44	6.08	7.69	3.94	7.59
Relative error in relation to SMS	Estimated SMS	Mean (%)	0.74	0.87	0.53	0.76	0.52	0.76	0.36	0.68
		Max (%)	5.44	5.44	3.74	7.41	5.10	7.76	5.08	9.33
	TMS	Mean (%)	1.95	2.07	1.41	1.53	1.28	1.46	1.05	1.16
		Max (%)	7.24	7.37	4.72	7.82	6.10	7.92	11.30	11.40

Note: Accepted = only considering accepted estimations; All day = Maintaining the variance for those intervals without estimation.

8. SOME REMARKS ABOUT NORMALITY OF THE SPEED DISTRIBUTION

The main assumption in the proposed method is the normality of speed distribution over the aggregation period. In traffic engineering, normal and log-normal distributions, as well as gamma distributions, have been traditionally used to model vehicular speed [Haight, 1963; Gerlough and Huber, 1975]. The latter two types of distribution, log-normal and gamma, are sometimes selected for two practical reasons. Firstly, in order to avoid the theoretical difficulty of negative speeds given by the left tails of normal distributions. Secondly, due to the difficulty of drawing statistical inference about vehicular speed which is analytically intractable under the normality assumption for individual speed measurements [Li, 2009]. Composite distributions have also been proposed [May, 1990]. However, it seems clear that this type of bimodal distribution arises from the mix of different populations: for instance, driver types, vehicle types or lane usage, whose speeds follow one of the previous distributions, each one with different parameters.

It seems reasonable to postulate that the speed distribution follows a normal distribution if considering a single lane where there is only one type of vehicles traveling (i.e., cars or trucks) and over a time period where traffic remains time stationary. Note that given these conditions all the different subpopulations have been set apart. It can be assumed that statistically different populations of drivers (aggressive or calm) are also set apart by their lane selection behavior. In the case of dense traffic, approaching the desired speed for fast vehicles is almost impossible because of interactions with other drivers in the same lane. This, results in a uniform behavior of all vehicles, thus the type of vehicle limitation can be dismissed in this case. The skew that may appear in the speed distribution in the case of time varying traffic conditions (i.e., congestion onset or congestion resolution within the time aggregation interval) is eliminated by the stationarity requirement.

Therefore, the speed distribution normality assumption is more prone to hold in the following situations:

- For low densities, on lane 2 and lane 3, on a per lane basis. Lane 1 is more prone to a composite distribution in this case due to the coexistence of cars and trucks.
- For moderately high densities, every lane on a per lane basis. But not the highway section as a whole, as each lane could have a different mean speed.
- For extremely high densities every lane on its own, and the highway section as a whole.

The obtained results confirm these assumptions, as it is under these situations where best accuracy is obtained.

Although for most of the three minute periods in congested traffic speed distribution normality assumption can be assumed to hold, in the case of rapidly changing traffic conditions (i.e., the passage of a shock wave) or in the case of instable traffic behavior (i.e. stop&go traffic) the normality assumption could be broken due to the loss of time stationarity in the three minute period. Is in these periods when the proposed algorithm will be less accurate.

The normality assumption can also be assumed to hold in the longer aggregation periods (e.g., 15 minutes) which may be used in case of very low flows, because in these traffic conditions the time stationarity of traffic is likely to span for even longer intervals.

In free flowing situations, the speed gradient between lanes will be more marked in case the loops are located close to an on/off ramp, as the rightmost lane will be affected by vehicle acceleration or deceleration. In this lane, the speed variance is likely to increase due to different aggressiveness of drivers when facing the entrance/exit to the freeway. In spite of this, the proposed method would be equally valid in a per lane basis.

A posted speed limit that falls within the interquartile range of the free speed distribution (the usual situation) would bias the speed distribution towards the speed limit (if strict enforcement is applied). Other speed limits would have no effect (e.g., in the case of speed limits above the free speed range) or would create a very narrow speed distribution around the speed limit (e.g., in the case of speed limits below the range of free speeds). These last situations are not a common practice. The posted speed limit in the test location is 120 km/h. This limitation does not have any effect on lane 1 and 2 speed distributions, as it is above the free speed range for these lanes. However, lane 3 speed distribution would be somehow truncated for speeds higher than the speed limit. This does not imply any drawback to the proposed methodology, as it is only necessary that the accumulated frequency of observations above and below the speed thresholds "v_1" and "v_2", set to 50 km/h and 100 km/h respectively, are consistent with the normal assumption. The transfers of observations from highest speeds to speeds around 120 km/h caused by the posted speed limit do not jeopardize the accuracy of the results.

9. CONCLUSIONS AND FURTHER RESEARCH

There is a need for traffic stream space mean speed data in order to monitor traffic and accurately model its evolution. TMCs' standardized computation procedures are not suitable for obtaining this average speed, as databases are structured around time means. Although small modifications of the standard process would suffice, this turns out to be an impractical task. Researchers have proposed relationships between time mean speeds and space mean speeds, but they rely on the knowledge of the speed standard deviation, which is usually not known. A method has been proposed in the present paper to obtain space mean speeds directly from common time aggregations of loop detector data. Particularly, the method relies only on the knowledge of time mean speed and the count stratification over a speed threshold in the time period considered. The main assumption of the methodology is the normality of the speed distribution, which is postulated to hold under certain conditions. Further research would be necessary to prove these assumptions. However, the accuracy of the results of applying the method to the data obtained from a double loop detector on a Spanish highway seems to indirectly prove the assumptions. The method is capable of estimating the space mean speed over 3 minute periods on every lane of a highway and in different traffic conditions, with a mean relative error of approximately 0.5%. Although such accuracy would not be actually necessary in some practical traffic engineering applications (e.g. congestion and incident detection or freeway performance evaluation) where coarse speed estimation would suffice, in other applications an accurate space mean speed estimation makes a big difference. For instance, freeway travel time estimations from spot speed measurements, construction of traffic diagrams, modeling traffic state evolution, accurate determination of traffic state as an input for operational policies… all of them highly sensitive to the speed estimation.

REFERENCES

Cassidy, M.J. (1998) Bivariate relations in nearly stationary highway traffic. *Transportation research Part B* **32**(1), 49-59.

Cassidy, M.J. and B. Coifman. (1997) Relation among average speed, flow and density and the analogous relation between density and occupancy. *Transportation Research Record: Journal of the Transportation Research Board* **1591**, 1-6.

Coifman, B. (2001) Improved velocity estimation using single loop detectors. *Transportation Research Part A* **35**(10), 863-880.

Coifman, B., S. Dhoorjaty and Z. Lee. (2003) Estimating the median velocity instead of mean velocity at single loop detectors. *Transportation Research Part C* **11**(3-4), 211-222.

Daganzo, C.F. (1997) *Fundamentals of transportation and traffic operations*. Elsevier, New York.

Dailey D.J. (1999) A statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B* **33**(5), 313-322.

Edie, L.C. (1965) Discussion of traffic stream measurements and definitions. *Proc. 2nd International Symposium on the Theory of Traffic Flow*, J. Almond, ed., OECD, Paris, 139-154.

Garber, N. and L. Hoel. (2002) *Traffic and highway engineering*. 3rd edition, Brooks/Cole.

Gerlough, D.L. and M.J. Huber. (1975) *Traffic flow theory: A Monograph*. Transportation Research Board, Washington, D.C.

Haight, F.A. (1963) *Mathematical theories of traffic flow*. Academic Press, London.

Heidemann, D. (1986) An analytic calculation method for speed distribution in dependence on traffic density. *Z. Für Verkehrswiss* (Traffic Engineering), 57-68.

Hellinga, B. (2002) Improving freeway speed estimates from single loop detectors. *Journal of Transportation Engineering* **128**(1), 58-67.

Khisty C.J. and B.K. Lall. (2003) *Transportation engineering: an introduction*. Prentice Hall.

Li, B. (2009) On the recursive estimation of vehicular speed using data from a single inductance loop detector: A Bayesian approach. *Transportation Research Part B* **43**(4), 391-402.

Lin, W., J. Dahlgren and H. Huo. (2004) Enhancement of vehicle speed estimation with single loop detectors. *Transportation Research Record: Journal of the Transportation Research Board* **1870**, 147-152.

May, A. D. (1990) *Traffic flow fundamentals*. Prentice Hall.

Rakha H. and W. Zhang. (2005) Estimating traffic stream space-mean speed and reliability from dual and single loop detectors. *Transportation Research Record: Journal of the Transportation Research Board* **1925**, 38-47.

Wang, Y. and L.H. Nihan. (2000) Freeway traffic speed estimation with single-loop outputs. *Transportation Research Record: Journal of the Transportation Research Board* **1727**, 120-126.

Wang, Y. and L.H. Nihan. (2003) Can single-loop detectors do the work of dual-loop detectors? *Journal of Transportation Engineering* **129**(2), 169-176.

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

APPENDIX A2

Requiem for Freeway Travel Time Estimation Methods Based on Blind Speed Interpolations between Point Measurements

ABSTRACT

The accuracy of real-time travel time information disseminated on metropolitan freeways is one of the key issues in the development of advanced traveler information systems. Although very accurate estimations could be obtained if suitable and intensive monitoring systems were available, travel time estimations must usually rely on data obtained from the preexisting surveillance equipment installed on freeways: loop detectors. Travel time estimation from loop measurements has attracted extensive research in the last decade, resulting in numerous methodologies. Among these, the ones that rely on spot speed measurements at detector sites in order to obtain the travel time estimation on the target stretch are the most intuitive. The key issue concerning these methods is the spatial generalization of point measurements over a freeway link. Multiple approaches can be found in the literature, ranging from the simplest, and mostly implemented in practice, constant speed approach, to recent and more complex mathematical interpolations.

The present paper shows that all speed interpolation methods that omit traffic dynamics and queue evolution do not contribute to better travel time estimations. All methods are inaccurate in congested and transition conditions, and the claimed relative benefits using various speed interpolation methods result from context specific experiments. Therefore, these methods should be used carefully, and not taken as perfect. Lacking a better approach, it is recommended to avoid overcomplicated mathematical interpolations and focus the efforts on intelligent smoothing of the noisy loop detector data, reducing the fluctuations of short time interval aggregations while maintaining the immediacy of the measurements.

Keywords: Travel time estimation, loop detector data, speed trajectory interpolation.

1. INTRODUCTION AND BACKGROUND

Information on the expected travel time along a congested freeway corridor is perhaps the most valuable traffic information for commuters in order for them to improve the quality and efficiency of their trips [Palen, 1997]. Pre-trip information may allow drivers the selection of time and route or even a mode shift. On-trip information is valuable for rerouting or deciding to accept park& ride options. In both situations, travel time information contributes to congestion mitigation. Even in the case where no travel time improvement was possible, travel time information would still improve the quality of the journey by reducing the uncertainty and consequently the stress of the driver. It must be highlighted that the accuracy of the disseminated information is crucial, as providing inaccurate travel time estimations can be detrimental.

Not only drivers benefit from travel time information, but also highway administrations, as travel time is an essential quantitative variable for evaluating the performance of transportation networks, or the operational efficiency of traffic management strategies, and can also be used as a robust and deterministic indicator of an incident. Travel time will become a basic input for the new real-time Advanced Traffic Management Systems (ATMS).

The abbreviation ATIS (Advanced Traveler Information Systems), groups all the technological elements necessary in order to develop a travel time information system, from the measurement of the source data to the dissemination of final information. While the deployment of technological equipment involved in the dissemination of variable traffic information moves toward a positive end (e.g. on-board traffic information devices are best-sellers, Variable Message Signs –VMS – are being widely installed on metropolitan freeways, traffic information web sites are becoming more popular, …) [OECD/JTRC, 2010], the measurement of the source travel time data is more problematic. In order to directly measure the travel time of vehicles on a freeway section, area wide monitoring is required. This means that it is necessary to record the position of the vehicle every few seconds (i.e. vehicle tracking), which could be achieved using traditional probe cars, or in its new concept by tracking GPS equipped vehicle fleets. In order to obtain a continuous flow of travel time measurements, a high percentage of equipped vehicles is necessary. Presently, this limits the practical application of this measurement technology to particular and delimited travel time experiments [Turner et al., 1998], although great hope exists for vehicle tracking once some type of sensor becomes extremely popular, for example GPS equipped cell phones [Claudel et al., 2009].

The alternative for the direct measurement of freeway travel times is not measuring the detailed trajectory of the vehicle, but only identifying the times the vehicle enters and exits the target section. Although this simplification has some implications on the nature of the measurements (see Section 2 for details), it allows a direct measurement of travel times by only identifying the vehicle at two control points. Classical AVI (Automated Vehicle Identification) technologies, such as video recognition of license plates or automatic reading of toll tags (in case of a turnpike highway equipped with an Electronic Collection System – ETC – device) [Soriguera et al., 2010] are already being used. Furthermore, promising innovative schemes like the identification of the bluetooth signature of a particular vehicle must also be considered.

In spite of this technological feasibility and accuracy of directly measuring travel times using area wide surveillance equipment, most of the Traffic Management Centers (TMCs) around the world currently rely on traditional inductive loop detectors to monitor traffic. TMCs have faced up to the new traffic operations data requirements by increasing the density of detector sites on most metropolitan freeways, up to 1 detector every 0.5 km. Therefore, any travel time estimation system hoping to be implemented on a large scale must rely on loop detector point measurements. This situation has not gone unnoticed by the transportation research community and has captured a large research effort during the last decades, as is reflected by the vast number of references that can be found in literature in relation to travel time estimation from loop point measurements.

Conceptually, two research directions could be distinguished. The first approach consists of using loop technology to identify particular vehicles, or groups of vehicles, at consecutive detector sites. Once the vehicles have been reidentified, direct travel time measurements can be obtained. The differences among several methods that fall under this category are in regard to the identification procedure. In Coifman and Cassidy (2002), Coifman and Erguera (2003) and Coifman and Krishnamurthya (2007), the length of vehicles is used to reidentify previously measured patterns. Lucas et al. (2004) reidentifies the platoon based structure of traffic, while Abdulhai and Tabib (2003) uses the particular electromagnetic signature that a vehicle produces when travelling over a loop detector. In practice, none of these methods can be applied using the standard time averaged data that is nowadays sent to the TMC from the detector roadside controller. In Dailey (1993) a statistical method is proposed to match the vehicle count fluctuations around the mean at adjacent detectors. The problem in this case (like in all platoon based identification methods) is that the platoon structure of traffic, which causes the strong correlation in vehicle count fluctuations, is lost in dense traffic or in the case of in-between junctions. Therefore, the aforementioned method is only valid for light free flowing traffic when travel time information is less valuable, as it is already known in advance. In Petty et al. (1998) this drawback is confronted by proposing to search for correlations in a wider and dynamic time window, and obtaining a probability density function of travel times. The authors claim good estimates, even for congested conditions, provided that data time aggregations are on the order of one second. Although they reported promising results, this last condition is not realistic in common present day practices. The last methods which could be grouped in this category are those that use relative differences in cumulative counts at consecutive detectors to estimate travel times [Nam and Drew, 1996; Oh et al., 2003]. Note that, in fact, this represents a vehicle reidentification under the first in first out assumption. The loop detector count drift is a major drawback in these methods.

The second, but most intuitive and simple category of methods to obtain travel time estimates from loop point measurements, uses the spot speed measurement at the detector site to generalize the speed over the target section and obtain the travel times. As a result of this apparent simplicity and straightforward application using common aggregations of loop detector data, spot speed based travel time estimation methods have been implemented worldwide by most traffic agencies. Problems with these types of methods arise mainly due to two factors: accuracy of spot speed measurements and spatial generalization of point measurements. The lack of accuracy of single loop speed estimations, usually assuming constant vehicle length, is widely accepted and has been vastly analyzed for a long time. Several methods have been proposed to enhance single loop performance in terms of speed measurements [Petty et al., 1998; Mikhalkin et al., 1972; Pushkar et al., 1994; Dailey, 1999; Wang and Nihan, 2000; 2003; Coifman, 2001; Hellinga, 2002; Coifman et al., 2003; Lin et al., 2004], however, to solve this problem most highway administrations have chosen to install double loop detectors capable of an accurate speed measurement, at least on their metropolitan freeways where intensive traffic monitoring is crucial.

The spatial generalization of point measurements, necessary in this type of travel time estimation methods, is the second factor that introduces error. The common practice is to assign a particular loop detector to a freeway portion, and assume that the speed remains constant during the whole section and during the whole time aggregation period. Even in the case of a unique stationary traffic state prevailing for the whole freeway section, travel time estimates would be flawed due to the fact that loop detector controllers usually compute and send to the TMC, time mean speeds (e.g. this is the case for the Spanish standards in loop detector data treatment), while the variable that relates average travel time with section length is the space mean speed. As local time mean speed structurally overestimates the space mean because faster observations are overrepresented, average travel times computed in this hypothetical stationary situation will be slightly underestimated (approximately 2% on average) [Soriguera and Robusté, 2011; Li et al., 2006]. This drawback could easily be solved by computing the space mean speed at the detector site (i.e. the harmonic mean of individual speeds) instead of the time mean speed (i.e. the arithmetic mean). A more problematic situation, which can be seen as an extreme of this last shortcoming, arises in the case of congested unstable behavior of traffic, when travel time estimation errors using point speed measurements and short aggregation periods can reach 30% [Rakha and Zhang, 2005] even though traffic conditions can be in average the same on the whole freeway link. This is due to the fact that, at the detector site and for the short updating time intervals (< 5 minutes), the stop&go instabilities can result in great errors in measuring a representative speed average for the whole link, because it is possible that the detector only measures average speed over one of the traffic instabilities. Smoothing data over longer time periods or wider measurement regions in space would average together many unstable, non-stationary traffic states, to hopefully converge to an unbiased global average traffic state. In addition in these stop&go situations, the measured average speeds only reflect the “go” part of the traffic and do not account for the time vehicles are completely stopped. Generally, the “stop” periods are small compared to the travel times and therefore this effect has not a significant contribution. However, in case of very congested traffic states, this last assumption cannot be accepted, and travel time will usually be underestimated, but not always, as it is also possible that the detector is measuring a very low speed instability. These flaws resulting from unstable non-stationary traffic states are not solved by computing the point estimation of space mean speed, as the spatial behavior of traffic instabilities wouldn't be captured either. Therefore, nothing can be said about the effects of using time mean speed instead of space mean speed in these situations. The paradox is that by using the wrong mean speed, improved travel time estimates could happen. However, this has to be considered as a positive accident, and not as a constant rule.

Being aware of these limitations, one must realize that considering common loop detector spacing (i.e. 0.5 km at best), one can see that they are almost negligible in relation to the errors that would arise in travel time estimation in the case of a dramatic traffic state transition on the freeway section (e.g. change from free flowing traffic to queued traffic within the section). Only the stop&go drawback may imply an exception to this last assertion. Take into account that when the spatial stationarity condition is broken, the point measurements assumed for the whole section would be totally unrealistic, and travel time errors can be huge. These will be largest when most of the segment is queued and the detector is unqueued (or viceversa).

This evidence leads to the obvious and widely demonstrated fact [Li et al., 2006; Kothuri et al., 2007; 2008] that travel time estimation methods based on spot speed measurements perform well in free flow conditions (this implies that there isn't any change of a traffic state within the freeway section), while the accuracy of the estimation in congested or transition conditions are dubious (there exists the possibility of a traffic state change within the freeway section and stop&go instabilities). It is also evident that the magnitude of these errors depends highly on the length of the link. For long links (i.e. low surveillance equipment density or increased detector spacing due to the temporary malfunctioning of a particular detector, a situation that arises too often) errors could be enormous, as the erroneous speed spatial generalization would be considered for this long highway section. Moreover, there is more probability of state transition within the link. Therefore, as the length of the links is defined by the surveillance density, there is a clear relation between the distance between loop detectors and the travel time estimation errors.

In order to solve this problem, several authors [Cortés et al., 2002; van Lint and van der Zijpp, 2003; Sun et al., 2008; Coifman, 2002; Treiber and Helbing, 2002] propose new speed interpolation models, different than the constant assumption, to better describe the spatial speed variations between point measurements, especially in congested conditions. Apart from a pair of remarkable exceptions and possible alternatives in [Coifman, 2002; Treiber and Helbing, 2002], where classical continuum traffic flow theory is used to generalize speed point measurements over the freeway section, the other new models, which are described in detail in Section 4, are tending to be mathematical exercises of interpolation, blind to traffic stream dynamics. These methods basically smooth the constant interpolation over time or space, but do not address the fact that any feature (i.e. end of a queue location) finer than the detector spacing will go unobserved.

The present paper aims to demonstrate that there is no reason to expect that a speed interpolation method which does not consider traffic dynamics and queue evolution, performs better in freeway sections that are partially congested. All these methods are inaccurate in congested and transition traffic states and the claimed benefits usually result from context and site specific validations, which sometimes can lead to counterintuitive conclusions [Li et al. 2006]. Lacking a better approach (note that the methods proposed in [Coifman, 2002; Treiber and Helbing, 2002] do not contribute in a better approach in partially congested sections, and therefore the problem remains unsolved), the simplest interpolations are recommended and an “intelligent” smoothing process is proposed in order to smooth the typical speed fluctuations of vehicle mean speed over short time intervals and traffic instabilities, while preserving the immediacy characteristic of loop measurements. Surprisingly, not much research effort has been focused on this last issue.

The remainder of this paper is organized into several sections. Travel time definitions are presented next, aiming to create a conceptual framework useful in comparing travel times obtained from different methodologies. Then, in Section 3, a trajectory reconstruction algorithm, necessary to compare direct and indirect travel time measurements is presented. Section 4 is devoted to the review of the proposed speed interpolation methods, followed by sections 5 and 6 where they are evaluated using AVI travel time data obtained on a metropolitan freeway near Barcelona, Spain. In the last part of the paper, Section 7, the proposed smoothing process useful for online travel time estimation from speed point measurements is described. Results of its application to the same set of data are also presented in Section 7. Finally, some conclusions and directions for further research are outlined.

2. TRAVEL TIME DEFINITIONS

All the studies dealing with travel time estimation referenced in the previous section compare the results of their proposed methods to some ground truth travel time data in order to evaluate the accuracy of the method. In fact, some of these studies are only devoted to that comparison [Li et al., 2006; Kothuri et al., 2007; 2008]. The nature of the ground truth travel time data used in each study is varied. For instance [Kothuri et al., 2007] uses data obtained from probe vehicle runs, and [Kothuri et al., 2008] adds data from the bus trajectories obtained from a GPS equipped bus fleet. In addition to the probe vehicle runs, [Sun et al., 2008] also considers travel time data obtained from video camera vehicle recognition. [Li et al., 2006] also uses data obtained from vehicle reidentification at control points, in this case by means of toll tags and number plate matching, while [Coifman, 2002] uses the length of vehicles to reidentify them from double loop detector measurements. In all cases, ground truth travel time data are obtained by directly measuring travel times, whether tracking the vehicle or identifying it at two successive control points. The ways in which these ground truth travel time data are obtained have some implications in the comparison procedure. Coarse comparisons can lead to the counterintuitive results found in literature because what is being compared are apples and oranges.

In the absence of these directly measured travel time data, the alternative can be simulated data using traffic microsimulators [Cortés et al., 2002; van Lint and van der Zijpp, 2003]. The same care must be taken with the simulated data, and in addition, it must take into account that the simulation is a simplification of the real traffic dynamics, and may have not been considering all the complexities of real traffic, resulting in predictable evolution of travel times. This leads to an artificial improvement of travel time estimation methods when ground truth data is obtained from simulation [Li et al. 2006].

The present section aims to rigorously present travel time definitions in order to fully understand the nature of each type of measurement. The first step is to differentiate between link (or section) travel time in relation to corridor (or itinerary) travel times. A link is the shortest freeway section where travel time can be estimated, while the corridor refers to the target itinerary whose travel time information is useful to the driver. The common practice (and the case in the present paper) is to define links limited by a pair of detector sites (this represents only some hundreds of meters in metropolitan freeways), while itineraries are defined, for instance, between freeway junctions. Therefore, an itinerary is usually composed of several links.

2.1. Link Travel Time Definitions

Consider the highway link of length "*Δx*" and the time interval of data aggregation "*Δt*" shown in Figure 1. In this configuration, the true average travel time over the space-time region "*A=(Δx, Δt)*" can be expressed as:

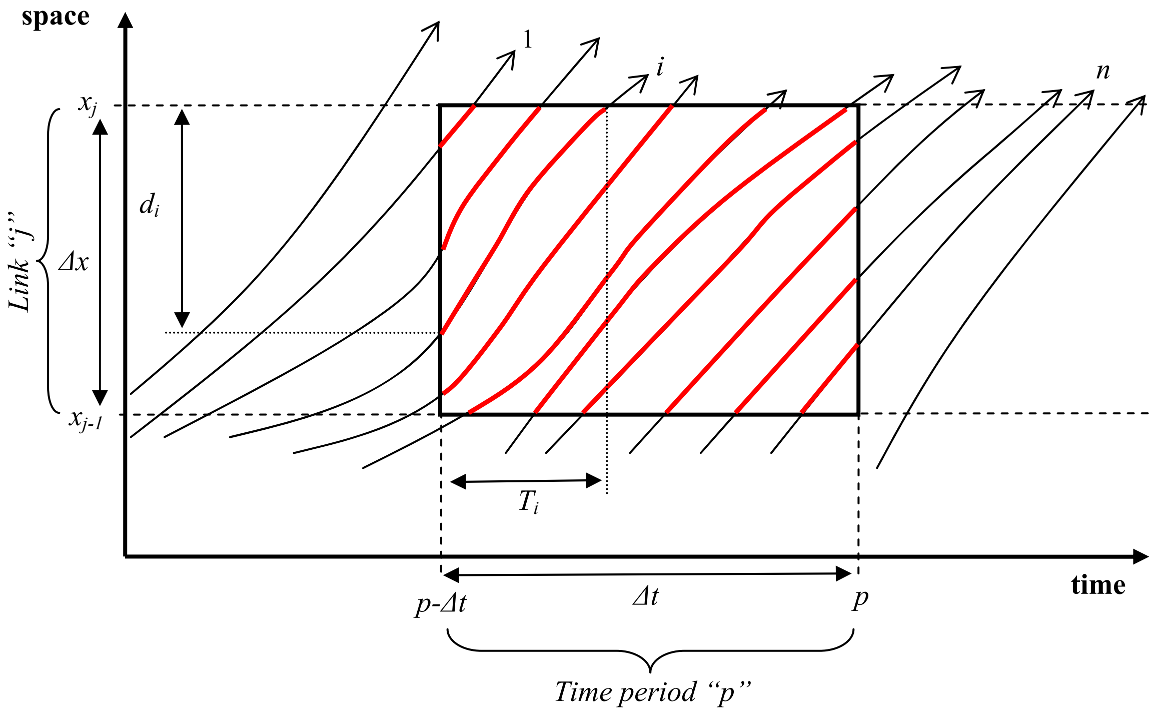
$$T_T(A) = \frac{\Delta x}{v(A)} = \Delta x \cdot \sum_{i=1}^{n} \frac{d_i}{T_i} \qquad (1)$$

Where:

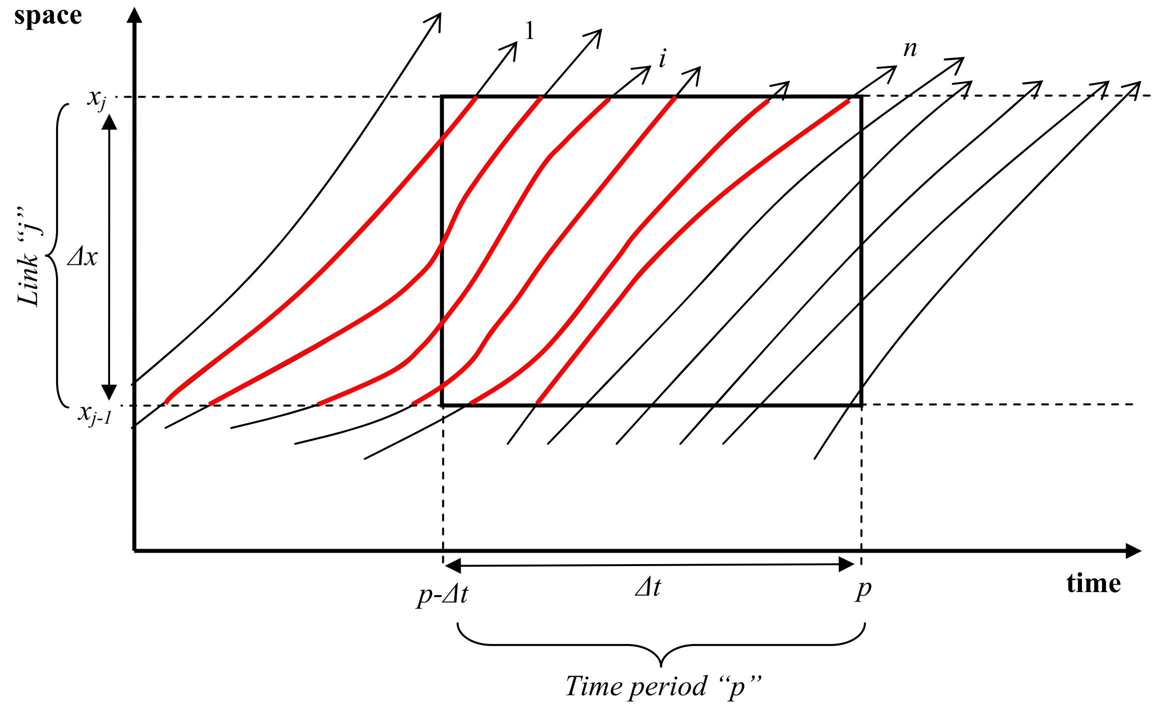
"*v(A)*" is the generalized average speed definition in the region "*A*", first proposed by Edie (1965).

"d_i" is the distance traveled by the i^{th} vehicle in the region "*A*".

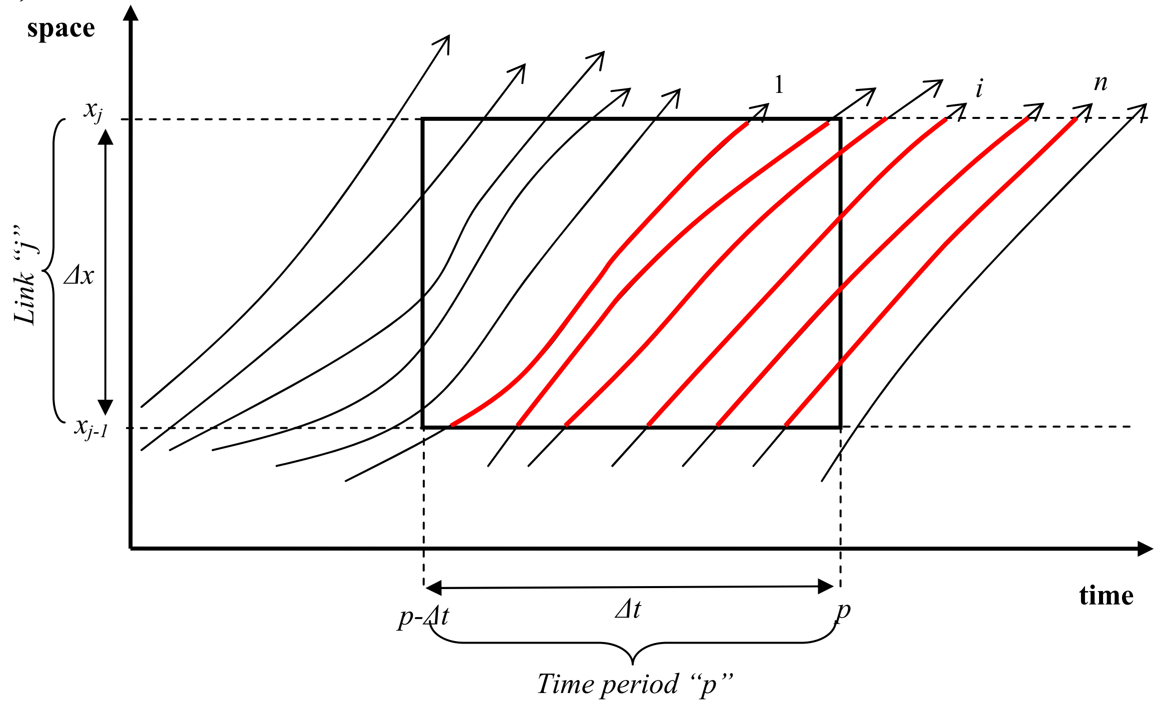
"T_i" is the time spent by the i^{th} vehicle in the region "*A*".



a)



b)



c)

FIGURE 1 Link travel time definitions in a trajectories diagram. a) True average travel time. b) Arrival based average travel time. c) Departure based average travel time.

Note: In red, parts of the vehicles' trajectories considered in the average travel time definitions.

Note that two control points at "x_{j-1}" and "x_j" where vehicles are identified are not enough to obtain this true average travel time, as the position of the vehicles travelling within the section at time instants "p-Δt" and "p" could not be obtained. It is possible that the only way to directly measure this travel time is by continuously tracking all the vehicles (or a representative sample of them).

The true average travel time "$T_T(A)$" should not be confused with the arrival based average travel time, "$T_A(A)$", defined as the average travel time in the trip along the whole link "j" of those vehicles that reach "x_j" in the time period "p" (see Figure 1b). This type of ground-truth travel time is obtained from all the direct measurements based on the reidentification of vehicles (number plates, toll tags, bluethooth devices, electromagnetic signatures, platoons, cumulative counts …). As this is the most common directly measured type of travel time, it is sometimes named MTT (measured travel times).

Following the same logic, a third average travel time can be defined. The departure based average travel time, "$T_D(A)$", is defined as the average travel time on a trip along the whole link of those vehicles that depart from "x_{j-1}" in the time period "p" (see Figure 1c).

On the one hand, "$T_A(A)$" considers the last completed trajectories on the highway link, and this may involve considering relatively old information of the traffic conditions on the first part of the link (some of the information was obtained more than one travel time before). On the other hand, "$T_T(A)$" uses the most recent information obtained in the whole link (sometimes these types of travel times are named ITT – instantaneous travel time). However, it is possible that any vehicle has followed a trajectory from which this true average travel time results. Finally, "$T_D(A)$" needs future information in relation to the instant of calculation. Therefore it is not possible to compute "$T_D(A)$" in real time operation. However, there is no problem in obtaining this future estimation in an off-line basis, when a complete database is available, including future information in relation to the instant of calculation. Note that "$T_D(A)$" would be approximately equal to "$T_A(A')$" where "A'" corresponds to the space-time region "A" moved forward one travel time unit in the time axis.

It is also possible, but not so easy, to obtain "$T_A(A)$" and "$T_D(A)$" from "$T_T(A)$". It is only necessary to compute the position of a virtual vehicle within the link as a function of time and considering the average speeds resulting from "$T_T(A)$" at different time intervals. This process, known as trajectory reconstruction, is detailed in Section 3. Note, that in order to obtain "$T_D(A)$", future "$T_T(A)$" will be needed.

As stated before, the differences between these average travel time definitions lie in the vehicle trajectories considered in the average calculation, "$T_T(A)$" being the only definition that considers all and only all the trajectories contained in "A", while "$T_A(A)$" or "$T_D(A)$" consider trajectories measured outside the time edges of "A", before or after respectively. The magnitudes of these differences depend on the relative difference between "Δt" and travel times. The longer the travel times in relation to the updating time interval are, the greater the difference will be in the group of vehicles considered in each average travel time definition (see Figure 2). "Δt" is a parameter to be set for the travel time information system, with a lower bound equal to the updating interval of the source data (e.g. time interval of aggregation of loop detector data). "Δt" should not be much longer than this lower bound in order not to smooth out travel time significant variations and maintain an adequate updating frequency (i.e. "Δt" should not go above 5 minutes). Therefore, as "Δt" must be kept small, differences between average travel time definitions depend mainly on travel times, which in turn, depend on the length of the highway link, "Δx", and on the traffic conditions.

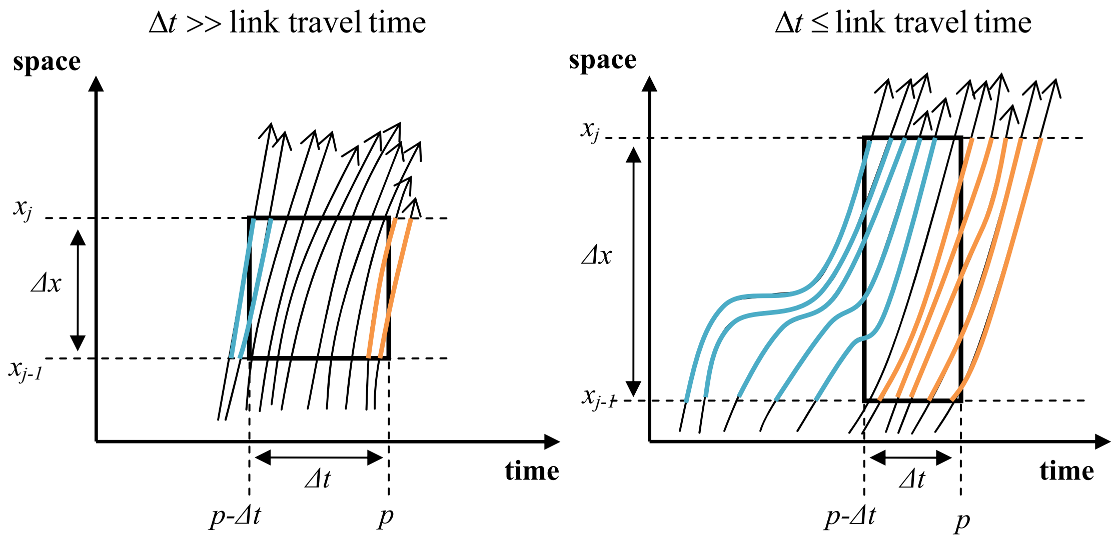


FIGURE 2 Different trajectories considered in the link travel time definitions.
Note: In blue, trajectories considered in the arrival based average travel times but not in the departure based average travel times. In orange, the opposite situation, both types of trajectories are only partially considered in the true average travel time. In black, shared trajectories.

In situations where link travel times are significantly longer than "Δt", as would happen in the case of long highway links or in the case of congested traffic conditions, the trajectories considered in several average travel time calculations will belong to different groups of vehicles (see Figure 2, right). The case may even arise where none of the vehicle trajectories are shared between different definitions. This would not have any effect on the average travel time in the case of stationary traffic, as the trajectories of the different groups of vehicles would be very similar. However, if a traffic transition occurs in the space-time regions considered in one definition but not in the others, this could result in significant differences between computed average travel times. This is the situation when the definition of average travel time plays an important role. On the contrary, in situations where link travel times are significantly shorter than "Δt" (i.e. short highway links due to high surveillance density and free flowing traffic conditions), the vehicle trajectories considered in one definition but not in the others would be very limited in relation to the total amount of shared trajectories (see Figure 2, left). Therefore, the probability and the relative weight of traffic transitions in this reduced space-time region is very low. This results in differences among definitions as being almost negligible in this case.

As commented before, it is rather difficult in practice to obtain "$T_T(A)$". However, to obtain "$T_A(A)$", only two vehicle identification points are necessary. As a result of this, there seems to be an interesting possibility of obtaining an approximation to "$T_T(A)$" by using the measurements that configure "$T_A(A)$". This approximation is as simple as only considering the trajectories which have arrived at the downstream control point during "p" time period (i.e. they belong to "$T_A(A)$" group of trajectories) and have departed from the upstream control point also during "p" (i.e. trajectories fully contained in "A" or equally, the shared trajectories between "$T_D(A)$" and "$T_A(A)$", see Figure 2). This approximation would be better as the number of shared trajectories increase. The process would converge to a perfect estimation when all trajectories are shared (i.e. "$T_T(A)$", "$T_A(A)$" and "$T_D(A)$" are equal). On the contrary, in some situations the approximation cannot be applied due to the inexistence of shared trajectories. This would result in the possibility of "$T_A(A)$" being a bad approximation to "$T_T(A)$".

Finally, note that in a real-time information dissemination scheme, "$T_T(A)$" and "$T_A(A)$" would be available for drivers entering the section at time period "$p+1$". However, neither the true average travel time, nor the arrival based average travel time at time interval "p" is the information that these drivers wish to obtain. They want to know their expected travel time, and therefore a departure based travel time at time interval "$p+1$" (sometimes this travel time is known as a PTT – predicted travel time). Therefore, the desired forecasting capabilities of measured travel time must not only span a time horizon equal to the travel time (i.e. in order to obtain the departure based average travel time at time period "p"), but an extended horizon equal to the travel time plus "Δt". This leads to the apparent paradox that while for a longer "Δt"s the true average travel time measurement is a better approach to the departure based average travel time at time period "p" (because more trajectories will be shared), the error made with the naïve assumption of considering "$T_T(A)$" as a proxy for the departure based travel time at time interval "$p+1$" (the implicit assumption here is that traffic conditions on the corridor remain constant from the measuring instant until the forecasting horizon) usually increases as "Δt" does (because of the extension of the forecasting horizon). Therefore, in some contexts (i.e. transitions, when the implicit assumption does not hold) the averaging of traffic conditions within long "Δt"s could lead to huge variations between adjacent time intervals. This is another reason for "Δt" being kept short.

2.2. Corridor Travel Time

The presented link average travel time definitions are also valid in a corridor context. The main difference in this case is that while the "Δt" remains the same as in the link basis, the increased length of the corridor in relation to the several links from which it is constituted of individual length equal to "Δx_j" results in larger travel times. This implies "$T_T(C)$", "$T_A(C)$" and "$T_D(C)$", where "C" stands for a corridor space – time region "$C = \left(\sum_{j \in corridor} \Delta x_j, \Delta t\right)$" being significantly different in non stationary traffic conditions.

Another issue to consider is how the corridor travel times could be obtained from composing link travel times. In can be easily deduced that corridor "$T_T(C)$" is obtained by simply adding up the links "$T_T(A)$" from the time period of calculation. On the contrary, corridor "$T_A(C)$" and corridor "$T_D(C)$" are not obtained from this simple addition, as it is needed to consider the vehicle trajectory in space and time.

As a conclusion to this definitions section, take into account that the common practice in the real time implementation of travel time systems based on speed point measurements is to estimate link "$T_T(A)$" by means of the available loop detectors, which are added up to obtain the corridor travel time, "$T_T(C)$", to be disseminated in real time. These true average corridor travel times are considered to be the best measurable estimation for the desired departure based average travel time at time interval "$p+1$", if one wants to avoid the uncertainties of forecasting, and assumes traffic conditions will remain constant. In particular, better than the "delayed" information from arrival based average travel times is "$T_A(C)$" (which could be directly measured in the corridor), provided that "$T_T(A)$" estimations of speed are accurate. Otherwise, this assertion could not be true. Also note that differences between true and predicted travel times depend on the corridor length and on the aggregation period "Δt", which constitute the horizon of the prediction. Therefore, in order to keep differences low and improve the accuracy of the "forecast", an advisable dissemination strategy is to keep corridor lengths as short as possible, while maintaining the interest of the driver on the disseminated information, and frequent updating, so that the time horizon of the prediction is as short as possible.

In the case of off-line travel time assessment, there is no need of trying to infer future travel times, for example by considering the latest information on the corridor as "$T_T(C)$" does. However, it is advisable to assess the real travel time that drivers actually experimented; this means reconstructing their trajectories in order to obtain "$T_A(C)$" from original "$T_T(C)$". These measurements are different in nature, and although they may be pretty similar in a link context, they will be significantly different on a corridor basis and non stationary traffic conditions. The results would be analogous to those obtained from direct measurement from an AVI device, provided that the original true link average travel times were accurate. Therefore, the same process of time and space alignment must be undertaken in a case of comparisons between "$T_T(C)$" obtained from loop detectors and "$T_A(C)$" obtained from AVI direct measurements. This process is described in detail in the next section.

3. TRAJECTORY RECONSTRUCTION PROCESS

This section aims to present the simple process necessary in order to reconstruct a vehicle trajectory from a speeds field in a discretized space-time plane. In other words, if space mean speeds are available within each link as a function of "x" (i.e. the position of the virtual vehicle within the link), and this function "$v(x)$", it is assumed that it will remain constant within each time interval "Δt" (see Figure 3). It is possible to reconstruct the trajectories that would result from the arrival based or departure based average link and corridor travel times. It is also possible to obtain the link and corridor average true travel time.

Therefore, the following process is necessary in order to convert true (or sometimes called "instant") travel times into trajectory based travel times. This process is analogous for the arrival based (backward reconstruction) or departure based (forward reconstruction) travel time averages. Taking into account that departure based travel times require future true information, only backwards reconstruction will be described in detail, but the analogous process can be easily derived [van Lint and van der Zijpp, 2003].

As van Lint and van der Zijpp (2003) describe, to reconstruct the trajectory of a virtual vehicle (i.e. to obtain the function "$x(t)$") within a cell (j, q) of the speeds field it is only necessary to solve the following differential equation:

$$\frac{\partial x}{\partial t} = v_{j,q}(x) \qquad (2)$$

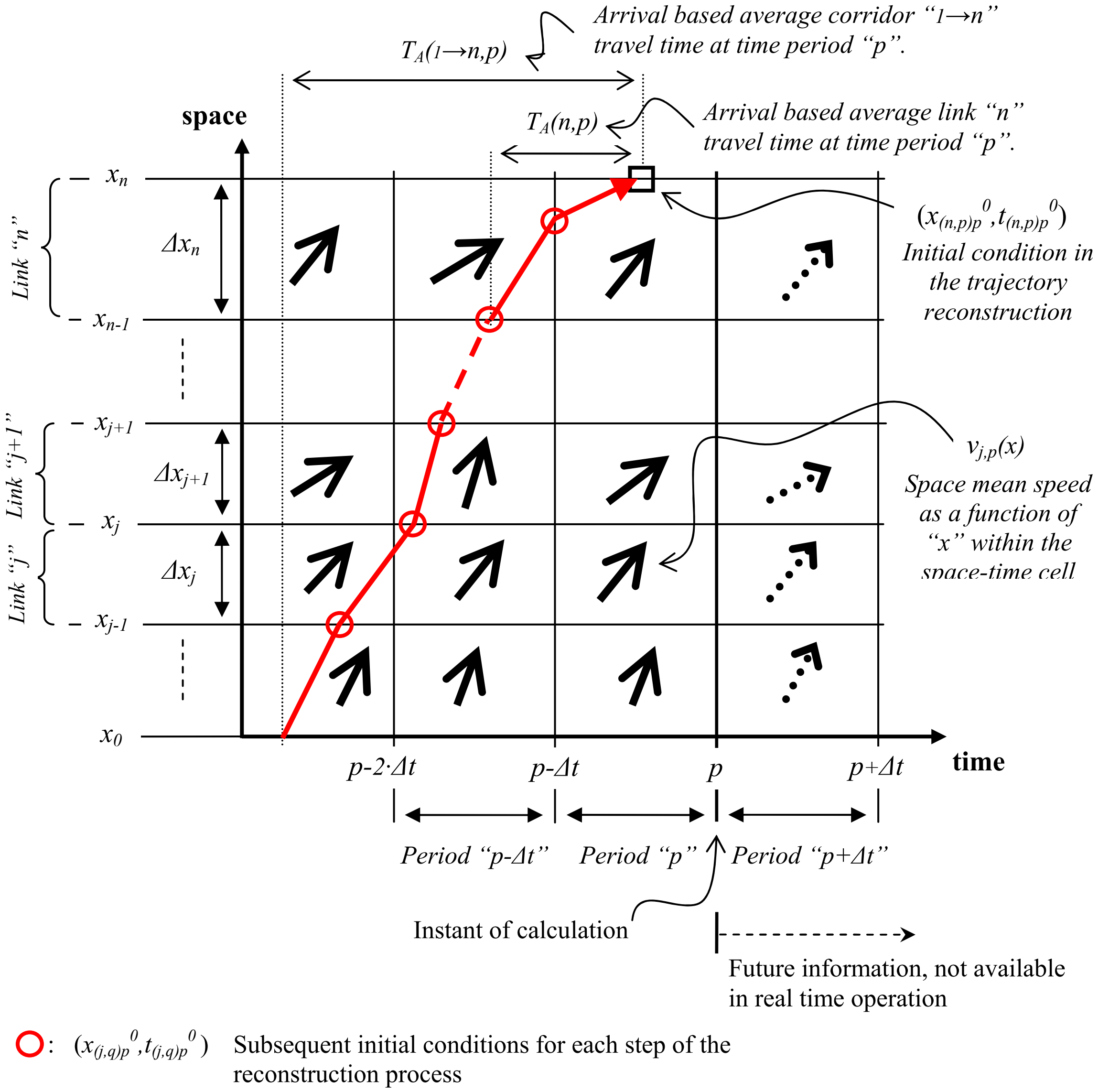


FIGURE 3 Speeds field in a space-time discretization.

Given "$v_{j,q}(x)$", which do not depend on time within a particular cell, and an initial condition "$x(t_{j,q}^0)=x_{j,q}^0$", which in this backward trajectory reconstruction corresponds to the cell exit point of the trajectory. The obtained solution will be valid for that particular cell. In order to obtain the whole trajectory along the link or the corridor to compute the arrival based travel time in time interval "p", it is only necessary to apply the aforementioned process iteratively from the last cell which crosses the trajectory, (n, p) (see Figure 3), until the start of the corridor is reached. At each step the "$v_{j,q}(x)$" (which is a function of the cell and of the time interval) and the initial condition must be updated. Note that the initial condition for subsequent cells corresponds to the entrance point in the space-time diagram of the trajectory to the previously calculated cell (see Figure 3). Then, the only initial condition needed to be set is the first one, corresponding to the time instant of calculation of the average corridor travel time. It seems adequate to consider this first initial condition, when computing the arrival based average corridor "$1\rightarrow n$" travel time at time period "p", as the midpoint of the time interval:

$$x(p-\Delta t/2)=x_n \tag{3}$$

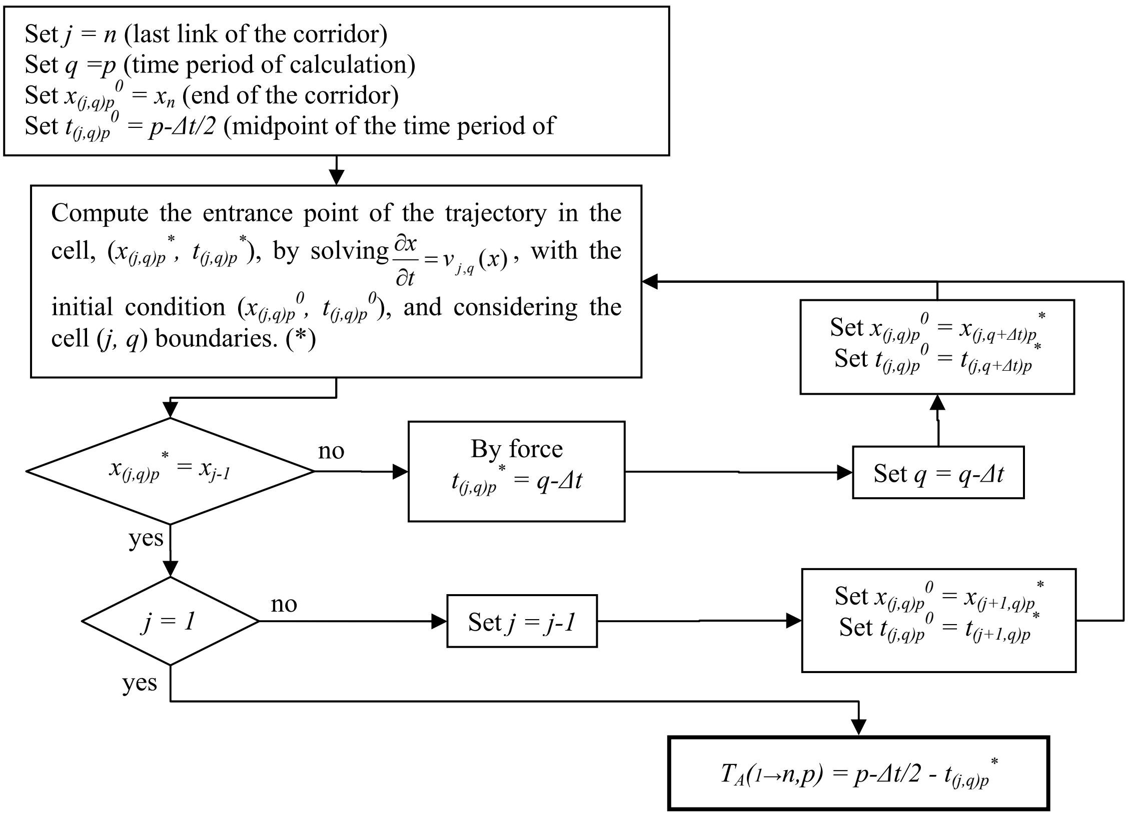
The remaining initial conditions, as each cell is confined by space and time bounds, will be defined by the instant the virtual vehicle crosses a link border, or the position within the link where the vehicle undergoes a change of time period.

The whole process for time interval "*p*" is detailed in a flowchart in Figure 4.

The computation of the true average link travel time for link "*j*" and time period "*p*", "$T_T(j,p)$", is simpler as it is only needed to solve Equation 2 without considering the time boundary of the cell. On each link, the initial condition could be "$x(p) = x_j$". Once the equation has been solved and the trajectory function "*x(t)*" is obtained, "$T_T(j,p)$" is calculated by imposing "$x(p - T_T(j,p)) = x_{j-1}$".

Finally, the true average corridor "*1→n*" travel time for the time period "*p*" is obtained as:

$$T_T(1 \to n, p) = \sum_{j=1}^{n} T_T(j,p) \tag{4}$$



(*) This calculation is explained in detail in the next section for each type of function defining *v(x)*.

FIGURE 4 Trajectory reconstruction flow chart.

4. METHODS OF LINK TRAVEL TIME ESTIMATION FROM POINT SPEED MEASUREMENTS

In long interurban trips composed of many links, where only a few suffer from congestion, the extremely inaccurate link travel time estimation in congested or transition conditions could have little effect as traveler does not care about one link's travel time, but on the aggregate travel time along the links that configure his trip. However, in the shorter commute trips across congested metropolitan freeways, where travel time information is more valuable and accuracy crucial, the situation is the other way around. It follows that the key issue in order to accurately estimate corridor travel times in these conditions (both, arrival based average travel times useful for off-line assessment or true average travel time for real time dissemination of information) is an accurate estimation of the true link average travel time within a time period, as true link average travel time is the building block for all other travel time definitions. Therefore the link level is adequate for the analysis.

It has been stated that the main problem in this link travel time estimation from point speed measurements is the lack of knowledge regarding the speed evolution in space, "*v(x)*", between measurement spots. This results in these methods being highly inaccurate when there is a traffic state transition within the link. The magnitude of these errors is directly related to the length of the link, which is inversely proportional to the loop detector density.

The present section is devoted to presenting several proposals for the estimation of "*v(x)*" between detector sites which can be found in the literature or in practical implementations. Explicit formulations to calculate ($x_{j,q}^*$, $t_{j,q}^*$), the entrance/exit points of the trajectory in a space-time cell (j, q), by solving Equation 2 will also be derived for each method.

It is interesting to note that some authors [Cortés et al., 2002; Sun et al., 2008] try to estimate "*v(t)*" between detector sites, instead of "*v(x)*". The claimed reason for such an approach is that although assuming continuous and smooth functions represent speed in time and space, the necessary trajectory reconstruction process when using "*v(x)*" results in speed discontinuities at time interval changes. This would not happen in the "*v(t)*" approach because the obtained average link travel time estimation would be directly arrival based, with the related drawbacks in real time estimation. Therefore, the trajectory reconstruction process is not necessary in a link context, but in a corridor basis. In addition, there is no reason to suspect that "*v(x)*" or "*v(t)*" should be continuous and smooth. For instance, a sharp change in speed when a vehicle encounters a queue on a freeway can be seen as a discontinuity in this function between approximate constant speeds. These discontinuities are more intuitive in relation to "*x*" as they happen on freeway spots where sudden traffic state changes arise. The artificial speed discontinuities every "Δt" within the reconstruction process are an inherent consequence of the discretization of time domain and would be small in the case of frequent updates and accurate estimations of "*v(x)*". Using "*v(t)*" eliminates this drawback indeed, but some complexity is added as the "distance" between interpolation points is not constant, but is rather the precise average link travel time. As the speed measurements at detector spots are not continuous in time (as a consequence of the discrete time domain), there is no guarantee that the iterative process required to obtain "*v(t)*" converges.

4.1. Constant Interpolation Between Detectors

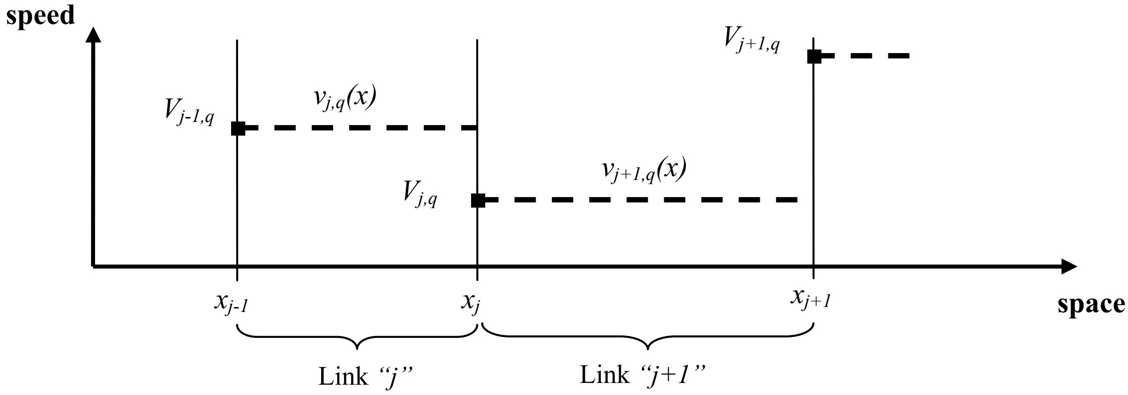
The simplest approach for the space generalization of speed between measurement points is the constant speed assumption. For its simplicity, this approach is widely used around the world [van Lint and van der Zijpp, 2003; Sun et al., 2008; Kothuri et al., 2008]. Several variants exist in relation to which speed measurement is selected to represent the whole section (see Figure 5). For instance, "$V_{j-1,q}$", the upstream speed measurement on the link at "x_{j-1}" and time interval "q" could be considered to represent "$v_{j,q}(x)$" in the whole link "j". Some case specific applications in particular links may suggest that this assumption is acceptable, but in a systematic application there is no objective reason for that. Therefore, on the same basis, the downstream measurement at "x_j" could be equally valid. Another approach could be to adopt a conservative strategy and assign the whole link the lowest of the speed measurements at "x_{j-1}" and "x_j", as in the ATIS in San Antonio, Texas [Fariello, 2002]. On the contrary, there could also be an optimistic approach in considering the largest of the measured speeds. Finally, one may also want to adopt an in-between solution [Cortés et al., 2002], and select a weighted average speed "$v_{j,q}(x)=\alpha V_{j-1,q} + (1-\alpha)\cdot V_{j,q}$" where "$\alpha \in (0,1)$".

With any of these approaches, the solution to Equation 2 which defines the virtual vehicle trajectory within a space-time cell, (j, q), is expressed as:

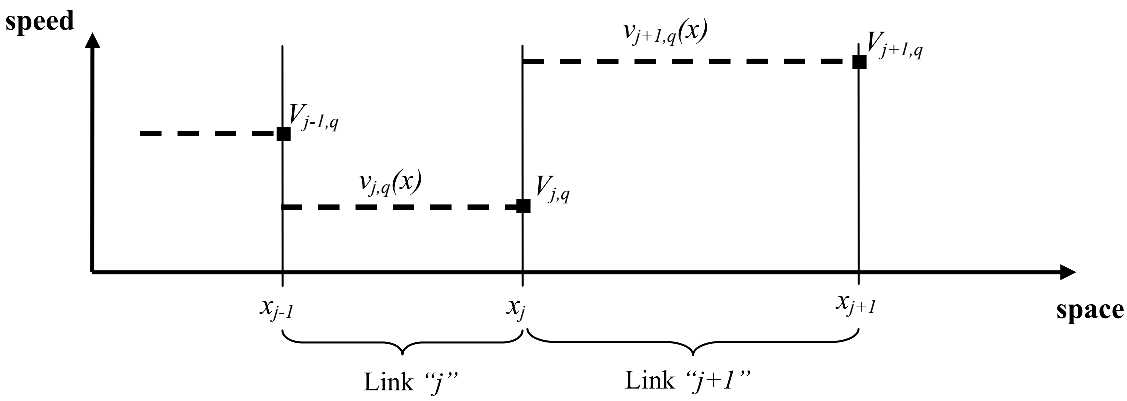
$$x_{j,q}(t) = x^0_{j,q} + \vartheta_{j,q} \cdot (t - t^0_{j,q}) \tag{5}$$

Where ($x^0_{j,q}, t^0_{j,q}$) is the cell exit point of the trajectory and "$\vartheta_{j,q}$" is the selected cell constant speed. From Equation 5, the trajectory entrance point to the cell is:

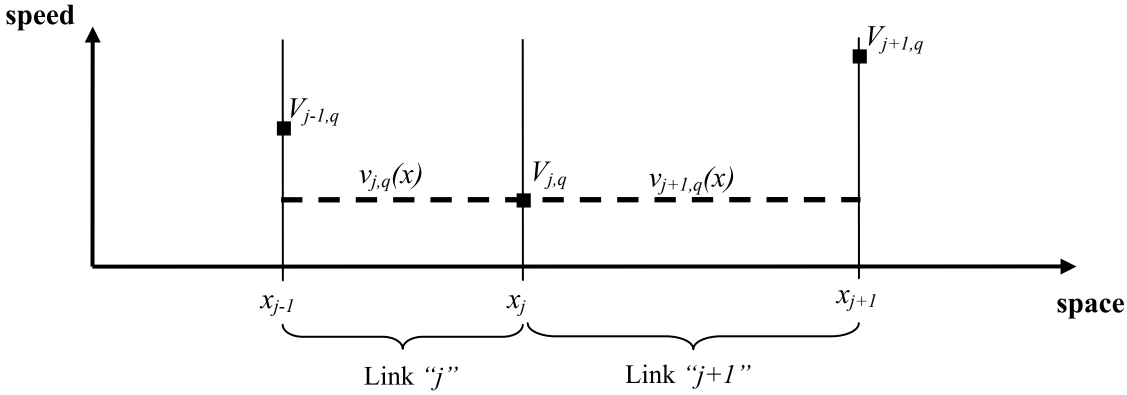
$$\left\{x^*_{j,q}, t^*_{j,q}\right\} = \begin{cases} \left\{ x_{j-1}, \ \dfrac{x_{j-1} - x^0_{j,q}}{\vartheta_{j,q}} + t^0_{j,q} \right\} & \text{if } \ x^0_{j,q} + \vartheta_{j,q} \cdot \left((q - \Delta t) - t^0_{j,q}\right) < x_{j-1} \\ \left\{ x^0_{j,q} + \vartheta_{j,q} \cdot \left((q-\Delta t) - t^0_{j,q}\right), \ q - \Delta t \right\} & \text{otherwise} \end{cases} \tag{6}$$



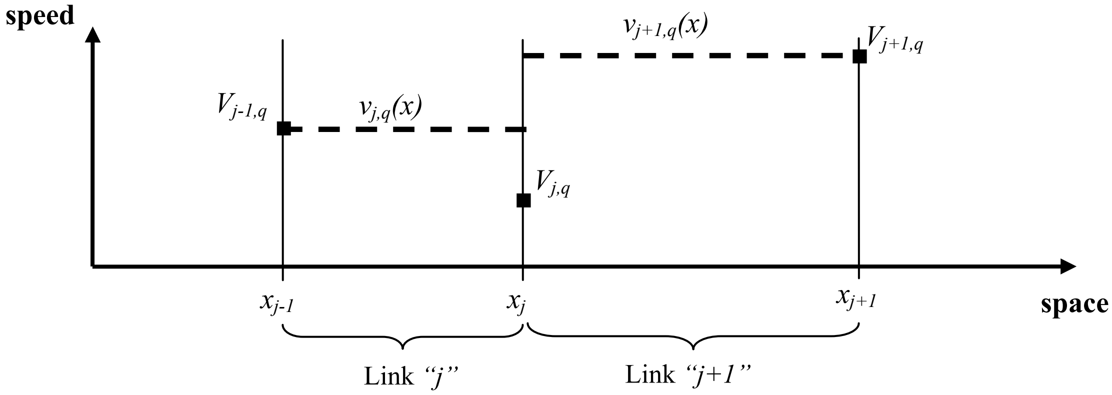
a)


speed
$v_{j+1,q}(x)$
$V_{j+1,q}$
$V_{j-1,q}$
$v_{j,q}(x)$
$V_{j,q}$
x_{j-1}
x_j
x_{j+1}
space
Link "j"
Link "j+1"

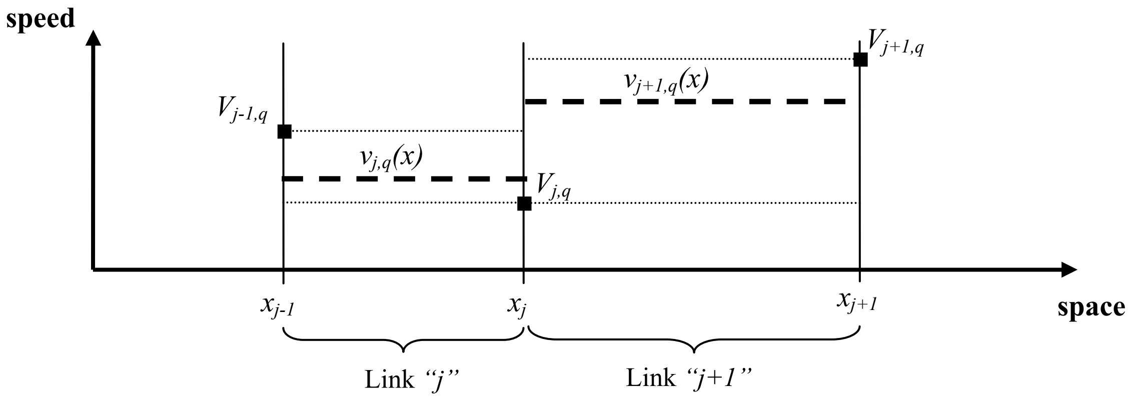
b)


speed
$V_{j+1,q}$
$V_{j-1,q}$
$v_{j,q}(x)$
$V_{j,q}$
$v_{j+1,q}(x)$
x_{j-1}
x_j
x_{j+1}
space
Link "j"
Link "j+1"

c)


speed
$v_{j+1,q}(x)$
$V_{j+1,q}$
$V_{j-1,q}$
$v_{j,q}(x)$
$V_{j,q}$
x_{j-1}
x_j
x_{j+1}
space
Link "j"
Link "j+1"

d)



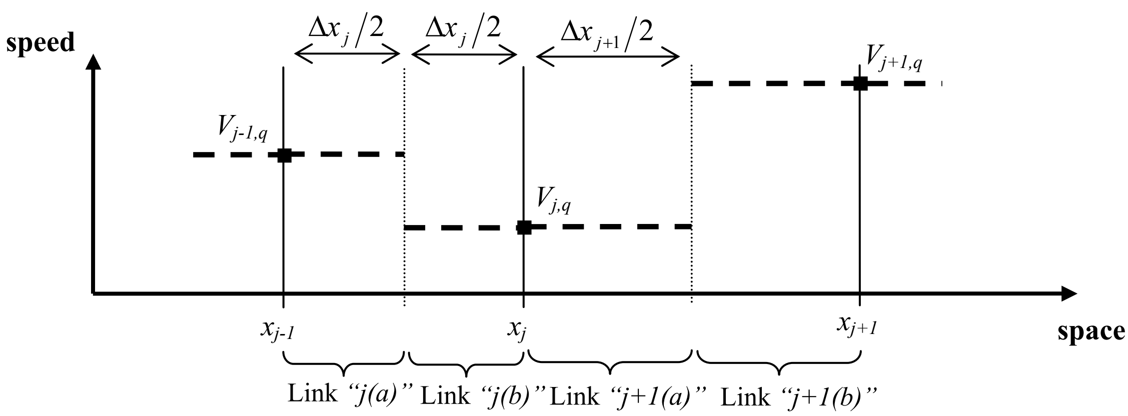
$v_{j,q}(x)=\alpha V_{j-1,q} + (1-\alpha)\cdot V_{j,q}$ with $\alpha \in (0,1)$

e)

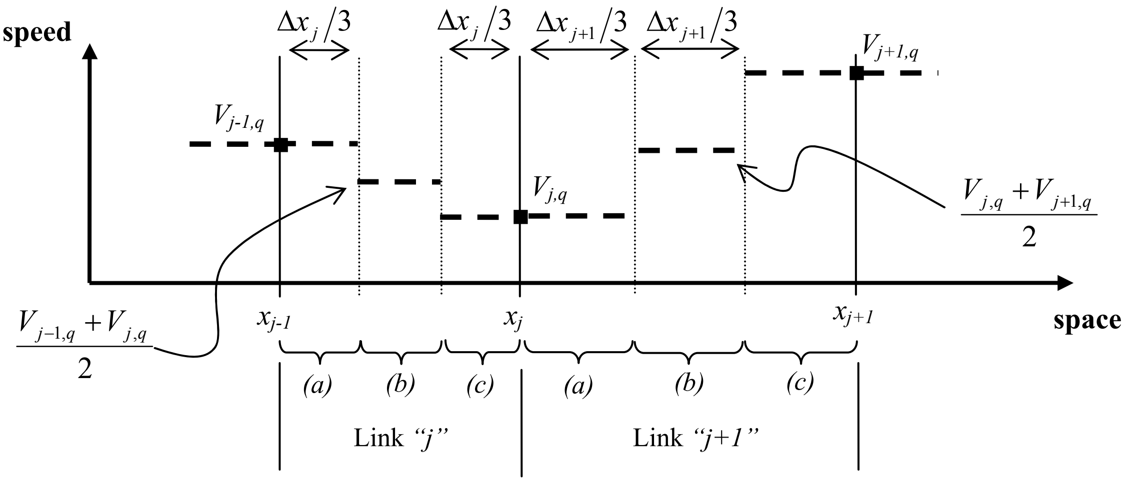
FIGURE 5 Constant speed trajectory spatial generalization. a) Upstream. b) Downstream. c)Conservative. c)Optimistic. e) Weighted average.

4.2. Piecewise Constant Interpolation Between Detectors

A simple modification of the constant speed assumption is the piecewise constant assumption between measurement points. The only difference is that the speed discontinuity is assumed to take place inside the link, and not at detector points. Therefore, the piecewise constant interpolation just redefines where a link begins and ends relatively to the detector. The details of the specific errors should change, but the net magnitude should not be much different, unless the speed discontinuity location is selected taking into account traffic dynamics within the link. Two common piecewise constant methods are the midpoint algorithm, widely used around the world [Li et al., 2006], and the thirds method (see Figure 6) used by the Minnesota DOT in the Twin Cities metropolitan area [Kwon, 2004]. In this context, Equations 5 and 6 remain valid, provided that each space-time cell is divided to maintain the constant speed assumption within the cell (see Figure 6).



a)



b)

FIGURE 6 Piecewise constant speed trajectory spatial generalization. a) Midpoint algorithm. b) Thirds method.

In some cases the piecewise constant speed trajectory within the link is simplified to a weighted average constant speed interpolation where the weighting factors are each a relative share of the link between speed measurements. Note that the resulting vehicle trajectory, "*x(t)*", reconstructed over the link or the corridor, is different, and therefore trajectory based travel times will be different depending on the method. However, true or instant travel times, which do not depend on the vehicle trajectory, would be the same, provided that the average speed is a harmonic weighed average. The following equation should be applied in order to put the case on a level with the one shown in Figure 5(e):

$$T_{T(j,q)} = \frac{\alpha \cdot \Delta x_j}{V_{j-1,q}} + \frac{(1-\alpha)\cdot \Delta x_j}{V_{j,q}} = \frac{\Delta x_j}{\dfrac{1}{\alpha \cdot \dfrac{1}{V_{j-1,q}} + (1-\alpha)\cdot \dfrac{1}{V_{j,q}}}} = \frac{\Delta x_j}{v_{j,q}} \tag{7}$$

Where $v_{j,q} = \dfrac{1}{\alpha \cdot \dfrac{1}{V_{j-1,q}} + (1-\alpha)\cdot \dfrac{1}{V_{j,q}}}$

4.3. Linear Interpolation Between Detectors

Van Lint and van der Zijpp (2003) challenge the constant speed interpolations because they result in instantaneous speed changes where vehicle trajectories would be piecewise linear. A linear speed interpolation between measurement points is proposed (see Figure 7), so that a smooth trajectory is obtained. However, no evidence is presented that drivers behave in this smooth fashion, anticipating slower or faster speed regimes, and driving experience seems to indicate that this is not sound.

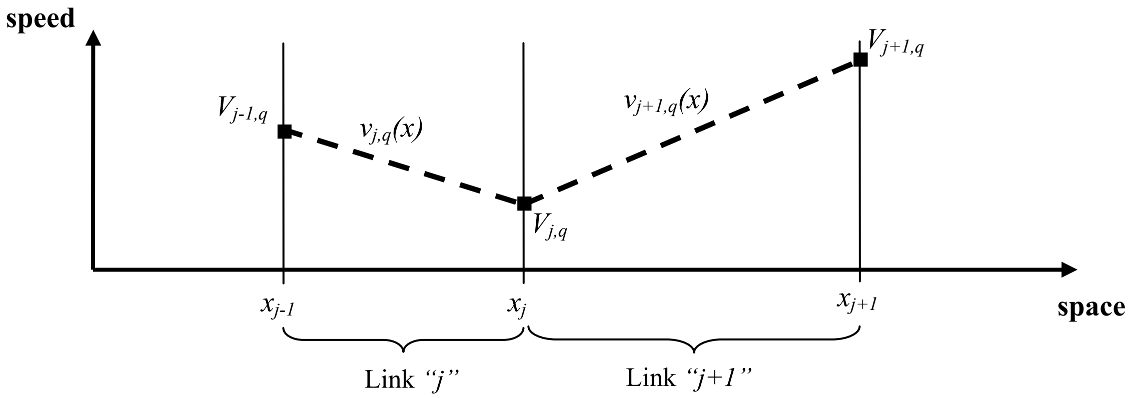


FIGURE 7 Linear speed trajectory spatial generalization.

The analytic equation for this linear speed interpolation is:

$$v_{j,q}(x) = V_{j-1,q} + \frac{(x - x_{j-1})}{(x_j - x_{j-1})} \cdot (V_{j,q} - V_{j-1,q}) \tag{8}$$

And the solution to the differential equation (Equation 2) which defines the virtual vehicle trajectory within a space-time cell (*j, q*) and an initial condition ($x^0_{j,q}, t^0_{j,q}$) is expressed as [Van Lint and van der Zijpp, 2003]:

$$\begin{aligned} x_{j,q}(t) &= x^0_{j,q} + \left(\frac{V_{j-1,q}}{\Lambda} + x^0_{j,q} - x_{j-1} \right) \cdot \left(\exp\left[\Lambda \left(t - t^0_{j,q} \right) \right] - 1 \right) \\ \Lambda &= \frac{V_{j,q} - V_{j-1,q}}{x_j - x_{j-1}} \end{aligned} \tag{9}$$

"Λ" must be significantly greater than zero to avoid numerical problems in the solving of Equation 9. Otherwise, constant speed assumption is equivalent to the linear interpolation, and solutions in Equation 6 can be used.

The trajectory entrance point to the cell depends then on the following condition:

$$x^0_{j,q} + \left(\frac{V_{j-1,q}}{\Lambda} + x^0_{j,q} - x_{j-1} \right) \cdot \left(\exp\left[\Lambda \left((q - \Delta t) - t^0_{j,q} \right) \right] - 1 \right) < x_{j-1} \tag{10}$$

Then:

$$\left\{ x^*_{j,q}, t^*_{j,q} \right\} = \begin{cases} \left\{ x_{j-1},\ t^0_{j,q} + \dfrac{1}{\Lambda} \ln \left(\dfrac{\dfrac{V_{j-1,q}}{\Lambda}}{\dfrac{V_{j-1,q}}{\Lambda} + x^0_{j,q} - x_{j-1}} \right) \right\} & \textit{if Equation } 10 \textit{ holds} \\ \left\{ x^0_{j,q} + \left(\dfrac{V_{j-1,q}}{\Lambda} + x^0_{j,q} - x_{j-1} \right) \cdot \left(\exp\left[\Lambda \left((q - \Delta t) - t^0_{j,q} \right) \right] - 1 \right),\ q - \Delta t \right\} & \textit{otherwise} \end{cases} \tag{11}$$

4.4. Quadratic Interpolation Between Detectors

Recently, a quadratic speed interpolation has been proposed by Sun et al. (2008), see Figure 8. This approach tries to mimic the drivers' behavior in relation to speed variations by allowing variable acceleration rates, as drivers may decelerate more when getting close to a congested zone or accelerate more when leaving a congested zone to become free-flow traffic. This method conceptually improves the linear interpolation, in the sense that in the linear case the drivers' behavior excessively anticipates downstream traffic conditions as a result of constant acceleration between measurement points, even before the driver notices the change in the traffic state. Note that in fact, this quadratic approach can be seen as a smoothed approximation to the piecewise constant speed interpolation. However, the problem in this quadratic interpolation is that the "sharp" changes in speed do not respond to traffic dynamics or queue evolution but only to the whims of a mathematical function.

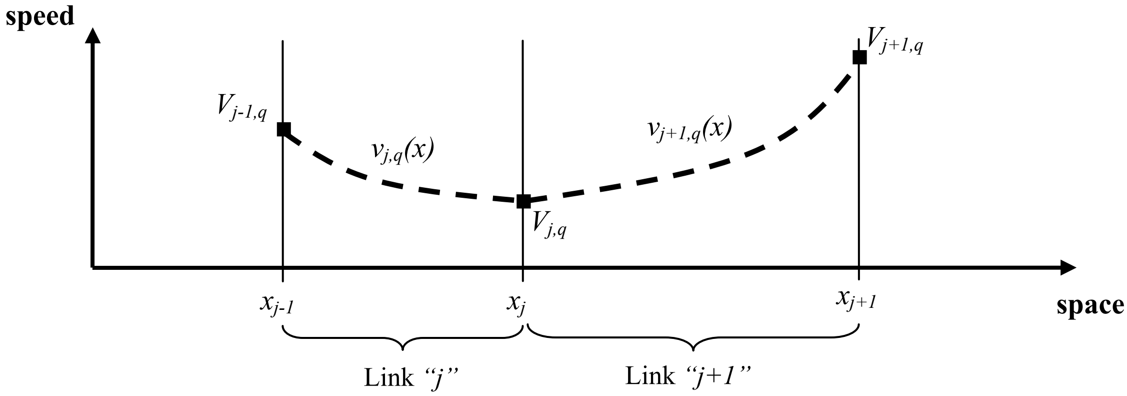


FIGURE 8 Quadratic speed trajectory spatial generalization.

The quadratic speed interpolation uses speed observations from three adjacent measurement points. Adapting the formulation presented in Sun et al. (2008) using a Lagrange quadratic interpolation polynomial, the speed trajectory as a function of "x" can be approximated as:

$$v_{j,q}(x) = V_{j-1,q} \cdot \ell_{j-1,q}(x) + V_{j,q} \cdot \ell_{j,q}(x) + V_{j+1,q} \cdot \ell_{j+1,q}(x) \tag{12}$$

Where the Lagrange basis functions are:

$$\begin{aligned}
\ell_{j-1,q}(x) &= \frac{(x - x_j)\cdot(x - x_{j+1})}{(x_{j-1} - x_j)\cdot(x_{j-1} - x_{j+1})} \\
\ell_{j,q}(x) &= \frac{(x - x_{j-1})\cdot(x - x_{j+1})}{(x_j - x_{j-1})\cdot(x_j - x_{j+1})} \\
\ell_{j+1,q}(x) &= \frac{(x - x_{j-1})\cdot(x - x_j)}{(x_{j+1} - x_{j-1})\cdot(x_{j+1} - x_j)}
\end{aligned} \tag{13}$$

Equation 12 can be rearranged as:

$$v_{j,q}(x) = a \cdot x^2 - b \cdot x + c \tag{14}$$

Where,

$$a = \frac{V_{j-1,q}}{(x_{j-1}-x_j)\cdot(x_{j-1}-x_{j+1})} + \frac{V_{j,q}}{(x_j-x_{j-1})\cdot(x_j-x_{j+1})} + \frac{V_{j+1,q}}{(x_{j+1}-x_{j-1})\cdot(x_{j+1}-x_j)}$$

$$b = \frac{V_{j-1,q}\cdot(x_j+x_{j+1})}{(x_{j-1}-x_j)\cdot(x_{j-1}-x_{j+1})} + \frac{V_{j,q}\cdot(x_{j-1}+x_{j+1})}{(x_j-x_{j-1})\cdot(x_j-x_{j+1})} + \frac{V_{j+1,q}\cdot(x_{j-1}+x_j)}{(x_{j+1}-x_{j-1})\cdot(x_{j+1}-x_j)} \quad (15)$$

$$c = \frac{V_{j-1,q}\cdot x_j\cdot x_{j+1}}{(x_{j-1}-x_j)\cdot(x_{j-1}-x_{j+1})} + \frac{V_{j,q}\cdot x_{j-1}\cdot x_{j+1}}{(x_j-x_{j-1})\cdot(x_j-x_{j+1})} + \frac{V_{j+1,q}\cdot x_{j-1}\cdot x_j}{(x_{j+1}-x_{j-1})\cdot(x_{j+1}-x_j)}$$

It can be checked that $b^2 \neq 4ac$. Then, Equation 14 leads to two solutions to differential Equation 2. Firstly, in case that $b^2 > 4ac$:

$$x_{j,q}(t) = \frac{b+\Phi+(\Phi-b)\cdot \mathrm{B}\cdot\exp(\Phi\cdot t)}{2\cdot a\cdot\left(1-\mathrm{B}\cdot\exp(\Phi\cdot t)\right)} \quad (16)$$

Where,

$$\Phi = \sqrt{b^2-4ac} \quad (17)$$

And "*B*" is the constant obtained applying the initial condition ($x^0_{j,q}, t^0_{j,q}$):

$$\mathrm{B} = \frac{2\cdot a\cdot x^0_{j,q} - b - \Phi}{(\Phi + 2\cdot a\cdot x^0_{j,q} - b)\cdot\exp(\Phi\cdot t^0_{j,q})} \quad (18)$$

In this case, the trajectory entrance point to the cell then depends on the following condition:

$$\frac{b+\Phi+(\Phi-b)\cdot\mathrm{B}\cdot\exp\left[\Phi\cdot(q-\Delta t)\right]}{2\cdot a\cdot\left(1-\mathrm{B}\cdot\exp\left[\Phi\cdot(q-\Delta t)\right]\right)} < x_{j-1} \quad (19)$$

Then:

$$\left\{x^*_{j,q}, t^*_{j,q}\right\} = \begin{cases} \left\{x_{j-1},\ \dfrac{1}{\Phi}\ln\left[\dfrac{1}{\mathrm{B}}\left(\dfrac{2\cdot a\cdot x_{j-1}-b-\Phi}{2\cdot a\cdot x_{j-1}-b+\Phi}\right)\right]\right\} & \textit{if Equation 19 holds} \\ \left\{\dfrac{b+\Phi+(\Phi-b)\cdot\mathrm{B}\cdot\exp\left[\Phi\cdot(q-\Delta t)\right]}{2\cdot a\cdot\left(1-\mathrm{B}\cdot\exp\left[\Phi\cdot(q-\Delta t)\right]\right)},\ q-\Delta t\right\} & \textit{otherwise} \end{cases} \quad (20)$$

Finally, a second solution arises if $b^2 < 4ac$:

$$x_{j,q}(t) = \frac{\Phi' \tan\left(\frac{\Phi' \cdot t}{2} + \mathrm{B}'\right) + b}{2 \cdot a} \tag{21}$$

Where,

$$\Phi' = \sqrt{4ac - b^2} \tag{22}$$

And "B'" is the constant obtained applying the initial condition ($x^0_{j,q}, t^0_{j,q}$):

$$\mathrm{B}' = \tan^{-1}\left(\frac{2 \cdot a \cdot x^0_{j,q} - b}{\Phi'}\right) - \frac{\Phi' \cdot t^0_{j,q}}{2} \tag{23}$$

The trajectory entrance point to the cell in this case depends then on the following condition:

$$\frac{\Phi' \tan\left(\frac{\Phi' \cdot (q - \Delta t)}{2} + \mathrm{B}'\right) + b}{2 \cdot a} < x_{j-1} \tag{24}$$

Then:

$$\left\{x^*_{j,q}, t^*_{j,q}\right\} = \begin{cases} \left\{x_{j-1},\ \frac{2}{\Phi'} \cdot \left[\tan^{-1}\left(\frac{2 \cdot a \cdot x_{j-1} - b}{\Phi'}\right) - \mathrm{B}'\right]\right\} & \textit{if Equation 24 holds} \\ \left\{\frac{\Phi' \tan\left(\frac{\Phi' \cdot (q - \Delta t)}{2} + \mathrm{B}'\right) + b}{2 \cdot a},\ q - \Delta t\right\} & \textit{otherwise} \end{cases} \tag{25}$$

Note that for particular values of "$V_{j-1,q}$", "$V_{j,q}$" and "$V_{j+1,q}$", the quadratic interpolation "$v_{j,q}(x)$" is not bounded by these measurements. This may result in unrealistic speeds at a particular "*x*" within the link (i.e. extremely high speeds never measured or extremely low speeds, even negative). This means that the solutions expressed in Equations 20 and 25 cannot be applied directly, and in order to obtain solutions which make sense a truncated definition of the speed evolution must be defined [Sun et al., 2008]. Therefore, Equation 12 should be rewritten as:

$$v_{j,q}(x) = \min\left[V_{\max}, \max\left(V_{\min}, V_{j-1,q} \cdot \ell_{j-1,q}(x) + V_{j,q} \cdot \ell_{j,q}(x) + V_{j+1,q} \cdot \ell_{j+1,q}(x)\right)\right] \quad (26)$$

where "V_{min}" and "V_{max}" are the speed thresholds to be set.

Given the truncated speed evolution with space within the link, the differential equation (Equation 2) is recommended to be solved numerically.

4.5. Criticism to the presented methods

The presented speed interpolation models between point measurements have been developed in order to solve the main problem of speed based freeway travel time estimation: the lack of accuracy in case of traffic state transitions within the link.

Constant and piecewise constant models imply instantaneous speed changes which in fact do not occur in real traffic. The remaining approaches seek to obtain continuous speed functions and smoother vehicle trajectories in order to avoid this drawback. For instance, the linear approach distributes the speed change in the traffic transition along the whole link.

However, it is evident from driving experience that traffic state transitions occur in specific spots of the freeway which evolve in time and space at the shockwave speed. When a driver encounters a shockwave, he adapts to the new traffic conditions in a short interval of time and space. This adaptation period depends on the acceleration/braking capabilities of the vehicle, on the driving behavior of the driver (i.e. aggressive or not) and on the perception of accident risk. Either way, it seems evident that the transition will not span for a long time-space period as the linear model assumes, which even implies the driver anticipating the perception of the traffic state change. In order to solve this problem, quadratic interpolations are proposed which imply a more rapid adaptation to speed changes.

None of these advanced methods face the key issue of the problem: where the transition occurs within the link. The proposed mathematical interpolations are blind to traffic dynamics, and hence still prone to errors, as they locate the traffic state transitions according to the whims of the mathematical functions. The improvements in travel times obtained by considering the detailed trajectory of the vehicle within the transition are negligible when compared to the benefits of accurate estimation of the location of the transition at each time period. If there is a situation where these improvements could have a significant contribution, this would be congestion dissolve episodes, where vehicles' acceleration is not so sharp, in relation to the sudden breaking to avoid collision in a congestion onset. In practice, the assumption of instantaneous speed change with the crossing of the shockwave would suffice, as it has been accepted traditionally in the context of continuum traffic flow modeling. Therefore, piecewise constant speed trajectories could be adequate.

These assertions do not imply that the presented constant or piecewise constant models perform better. They are only particular solutions for when the crossing of the shockwave coincides with the speed discontinuity location in the model (e.g. detector location, midpoint …). The piecewise constant speed interpolation method would be adequate provided that the speed change location is accurately estimated (for an online application this is equivalent to a constant weighted speed in the whole section – see Equation 7 – where the location of the speed change must be described by an appropriate and dynamic estimation of the parameter "α"). This last issue remains in practice unsolved, as one could employ queuing theory or traffic flow theory to estimate the length of the queue in between detectors, but detector counting errors rapidly accumulate and undermine the results. In specific locations, where a recurrent bottleneck location is detected, one method could be selected among others in relation to the adequacy of its assumptions. On the contrary, in uniform sections, any of the methods will result in the same average errors over sufficiently long time periods

Having said that, the contradictory results found in the literature should not be unexpected as the same method sometimes overestimates travel times and sometimes underestimates them; sometimes considering upstream speed is more accurate and sometimes it is the inverse… It all depends on the location of the traffic transition which evolves with time.

5. THE DATA

In order to provide empirical evidence of the previous statements, it is necessary to compare travel time estimates obtained from average speed data at detector sites with directly measured travel times. Although it is a difficult issue to obtain a representative number of ground truth travel time measurements within each time period "Δt", and for some authors virtually impossible if one wants to consider all the vehicles in a realistic urban freeway [Cortés et al., 2002], the AP-7 turnpike, on the north eastern stretch of the Spanish Mediterranean coast, represents a privileged test site.

The closed tolling system installed on the turnpike, whose objective is to charge every vehicle a particular toll resulting from the application of a kilometric fee to the distance travelled by the vehicle, provides collateral data for every trip on the highway, including the entry junction, the exit junction, and the entrance and exit times. This allows computing origin – destination matrices and travel times between control points on the turnpike [Soriguera et al., 2010]. In addition, the surveillance equipment installed consists of double loop detectors located approximately every 4 km.

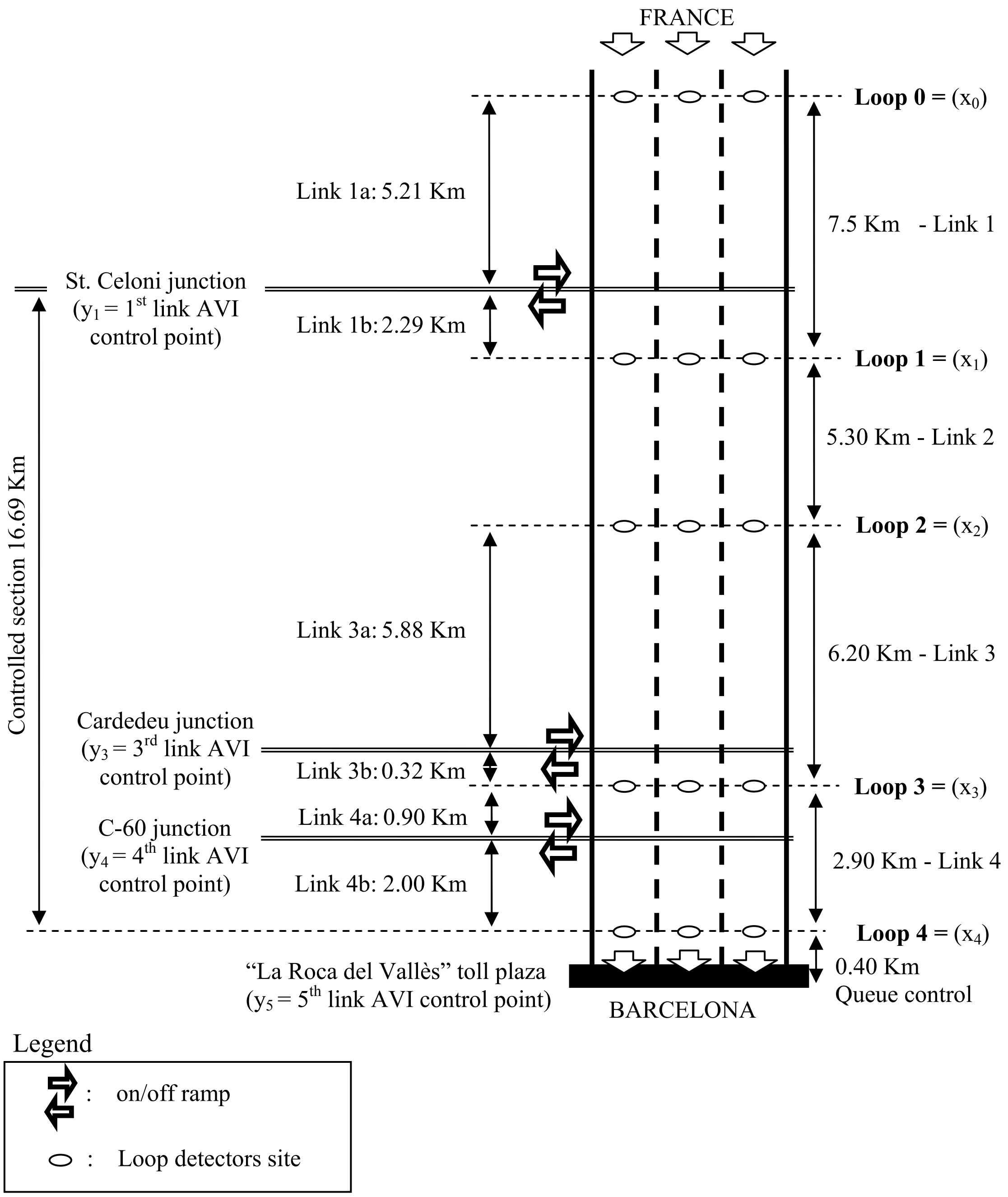


FIGURE 9 Test site layout.

This provides a perfect environment for evaluating travel time estimation methods from loop detector data. The test site, shown in Figure 9, consists of 21.9 km on the southbound direction of the AP-7 turnpike towards Barcelona, Spain. There are 5 detector sites which define 4 links in between. As discussed before, the length of these links, ranging from 2.9 to 7.5 km, is far too large in order to consider the practical application of spot speed based travel time estimation methods, as queues take long time to grow over such a long distance, and therefore travel time estimations in transition conditions would be completely flawed. However the behavior of these estimation errors is precisely the issue being analyzed here and their enlargement will be helpful in visualizing the results. Ground truth travel time data are available from control points located by each junction to the downstream exit of the turnpike at "La Roca del Vallès" where the main trunk toll plaza is located. In addition, the queue control system at this main trunk toll plaza provides measured travel times between loop 4 and the 4th AVI control point. This results in a 16.69 km long stretch where loop travel time estimations can be evaluated, as both measured and estimated travel times are available. All data are obtained as 3-minute average.

Test data were obtained on Thursday June 21st, 2007, a very conflictive day in terms of traffic in the selected stretch. Problems started around 12:39 when a strict roadblock at "La Roca del Vallès" toll plaza was set up by the highway patrol, reducing its capacity and consequently the output flow. This, in addition to the high traffic demand of the turnpike at that time, caused severe queues to grow rapidly. In view of this fact, and approximately 45 min after the setting of the roadblock, when the queue was already spanning around 5 km, the service rate of the roadblock was increased. This reduced the queue growing rate. Until 14:33 when the patrol roadblock was removed, queues had not started to dissipate. In addition, at 14:51 when queues were still dissipating, the breakdown of a heavy truck within link 4a blocked one out of the three lanes. In turn, this caused queues to start growing upstream again. Finally, when the broken down truck was removed at 16:06, the queues started to dissipate again, flowing at capacity. The high and unanticipated demand at "La Roca" toll plaza, as a consequence of the queue discharge, exceeded the capacity of the toll gates, causing small queues to grow at this location between 16:06 and 16:21. A complete sketch of traffic evolution on the test site for this particular day can be seen in Figure 10, drawn at scale to match empirical data presented in Figures 11 and 12. From Figure 11 it can also be seen how Loop 4 is impacted by the slowing for the toll plaza in free flowing conditions. Also note that the queue never reached Loop 2 location. Besides, Figure 12 shows the low impact of the truck breakdown on vehicles entering the highway at "y_3" (i.e. Cardedeu junction), although this entrance was located several hundreds of meters upstream of the induced bottleneck. This is due to the existence of an auxiliary entrance lane at "y_3" which almost allowed bypassing the breakdown location.

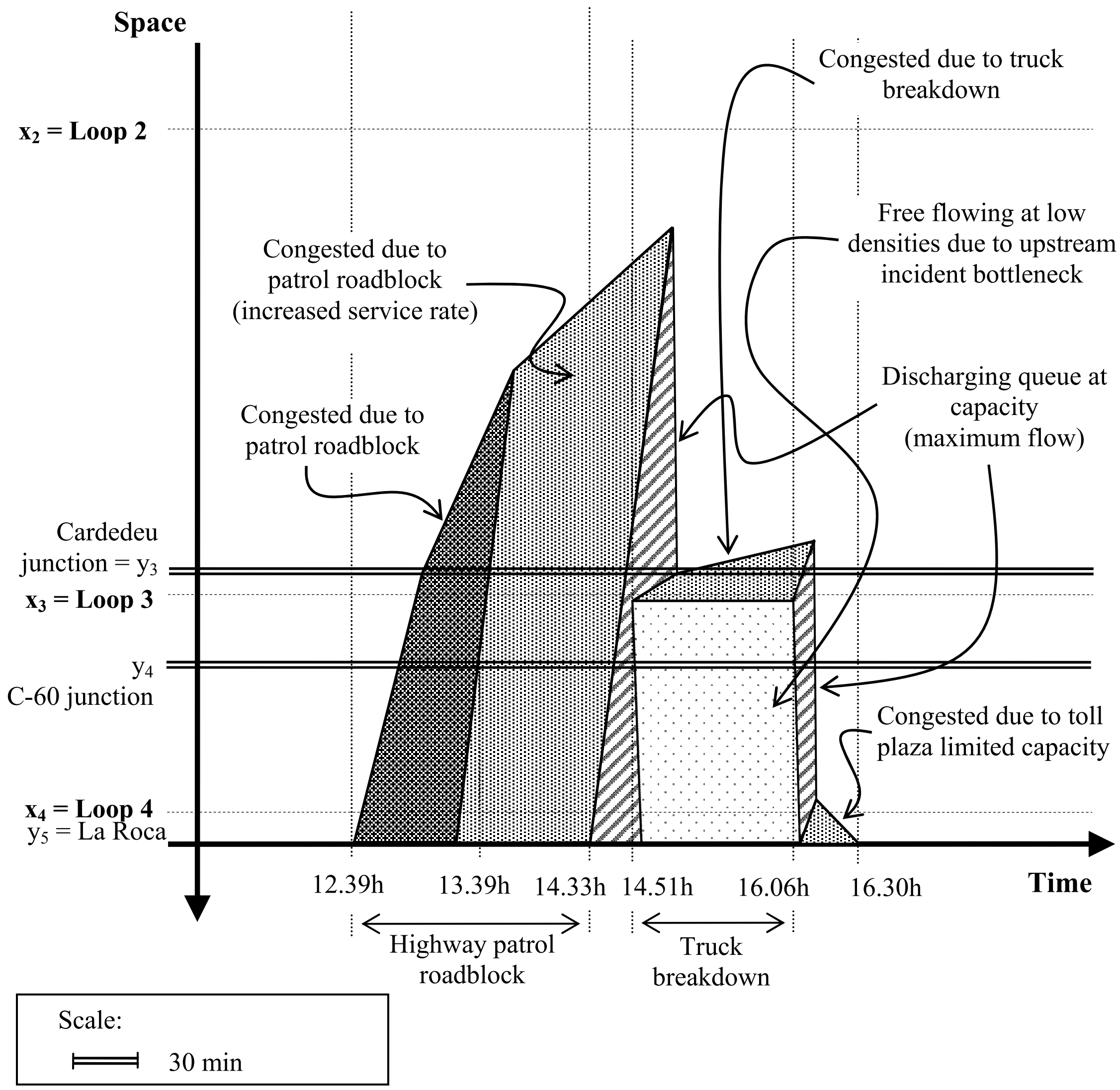


FIGURE 10 Traffic evolution on test site in the afternoon of 21st June 2007.

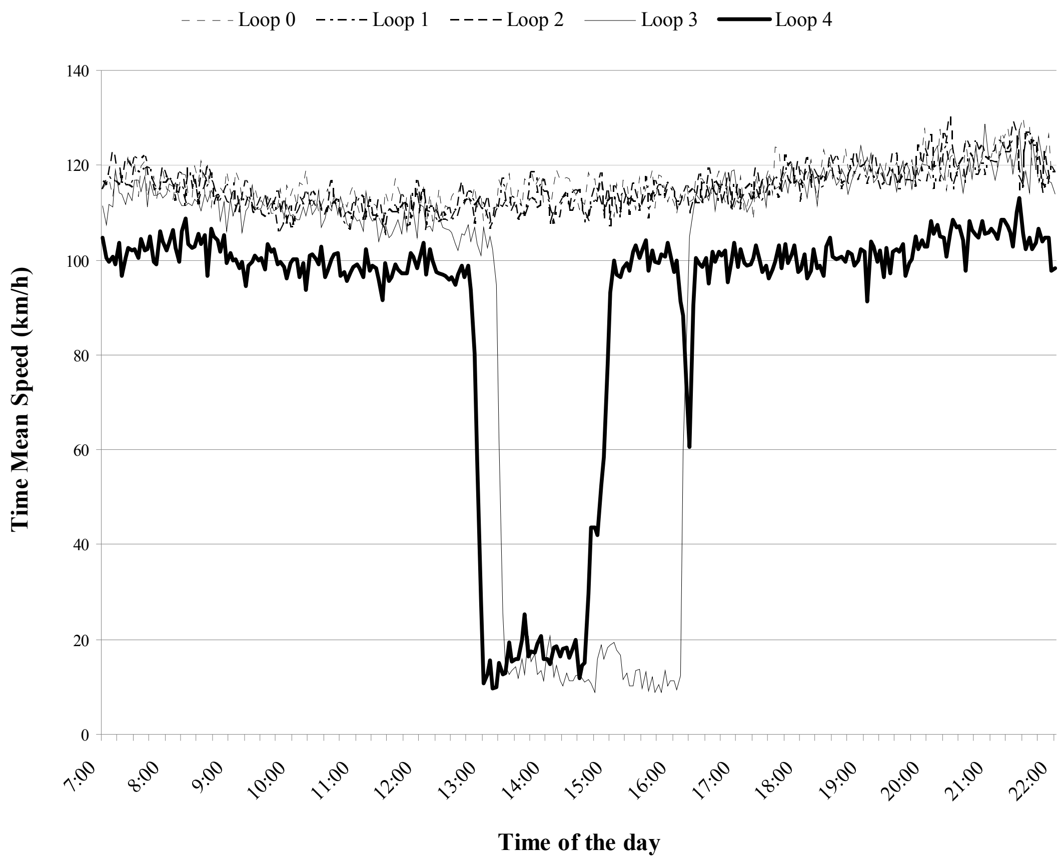


FIGURE 11 Time mean speed (3 minute average) for the whole section measured at loop detector sites.

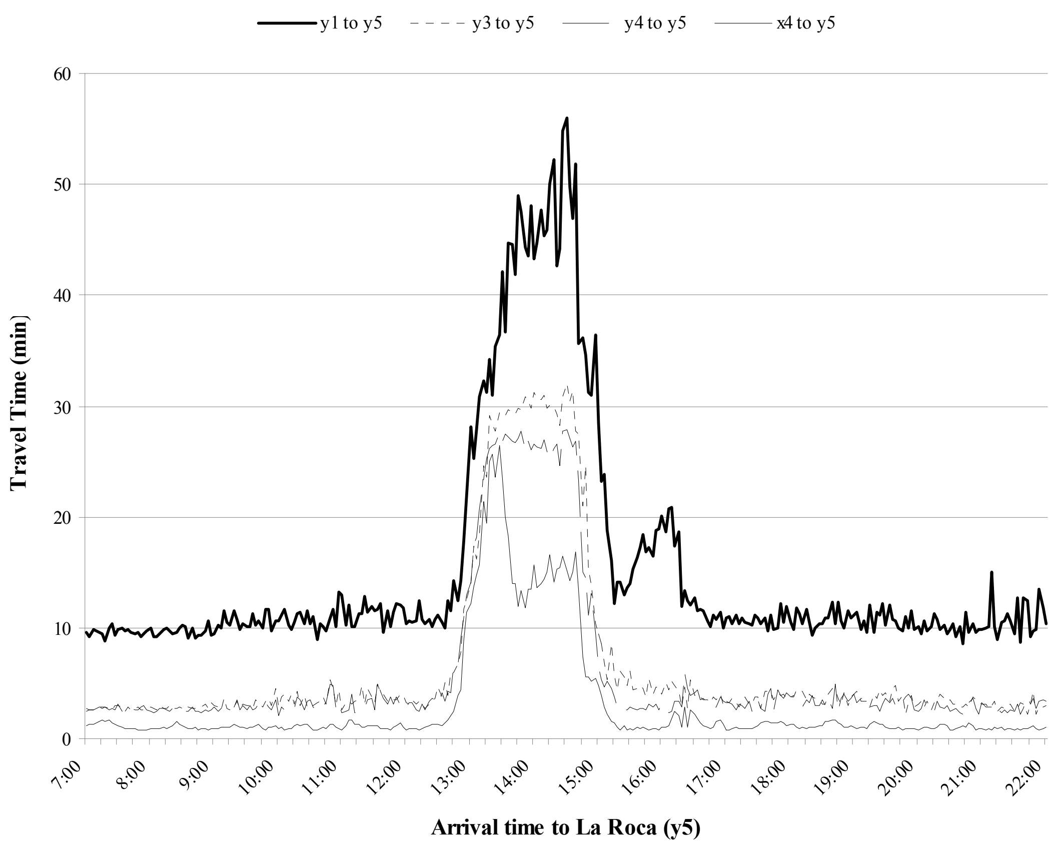
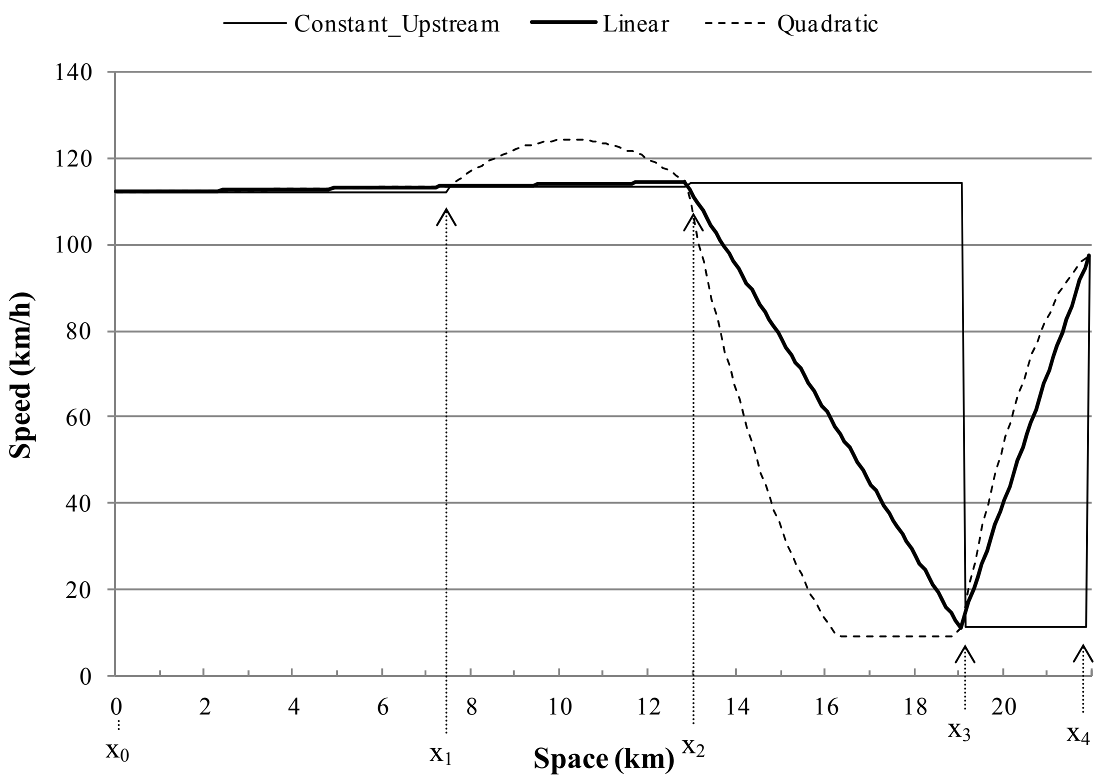


FIGURE 12 AVI measured arrival based average travel times (3 minute average).

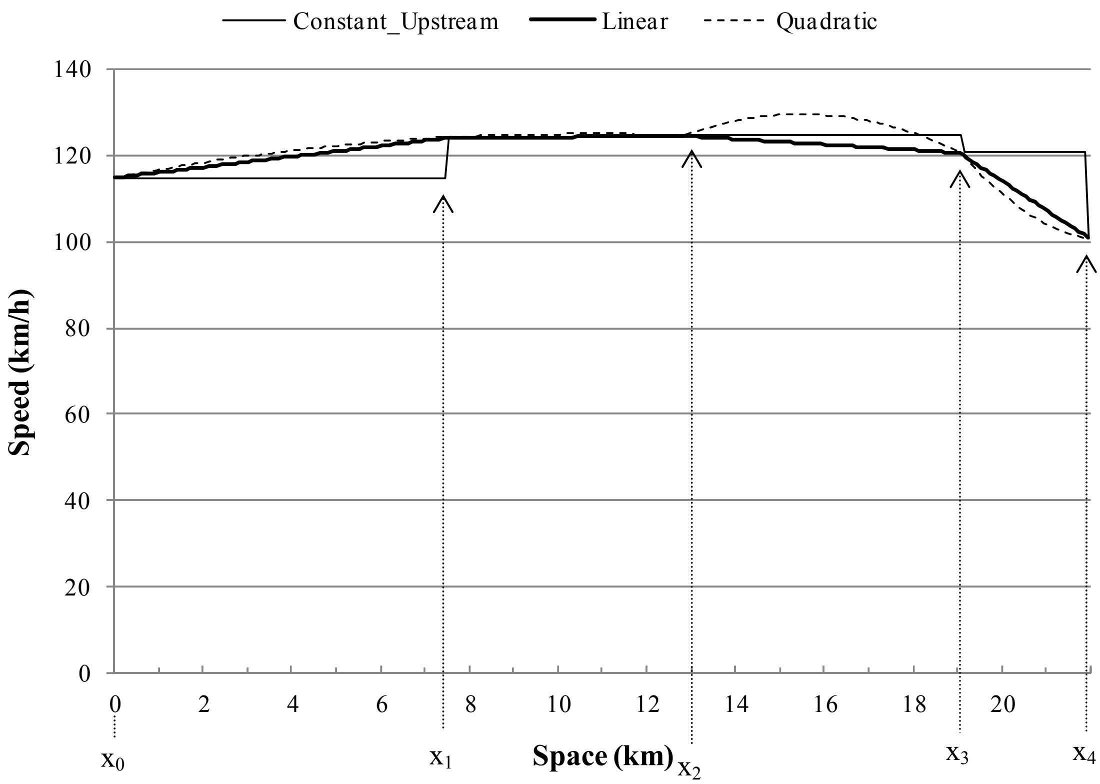
6. EVALUATION OF PROPOSED SPEED SPATIAL INTERPOLATION METHODS

Measured AVI travel times shown in Figure 12 are arrival based. Therefore, in order to evaluate the accuracy of different travel time estimations, based on different speed interpolation between point measurements, it is necessary to reconstruct the vehicles' trajectories. The trajectory reconstruction process detailed in Section 3 is applied to the speeds field given by the loop test data shown in Figure 11 and considering the space discretization presented in Figure 9 at a time step of three minutes. Once the trajectories have been reconstructed, the resulting travel times are also arrival based, and the comparison with measured travel times is appropriate.

Several methods of speed interpolation between detector sites have been analyzed: constant upstream, constant downstream, midpoint, linear and truncated quadratic (where "V_{min}"=10 km/h and "V_{max}"=130 km/h). These methods have been considered as representative of each category, being the results that are easily extrapolated to the remaining methods. By way of illustration, Figure 13 shows the speed profile over space on the test site resulting from the considered interpolation methods.

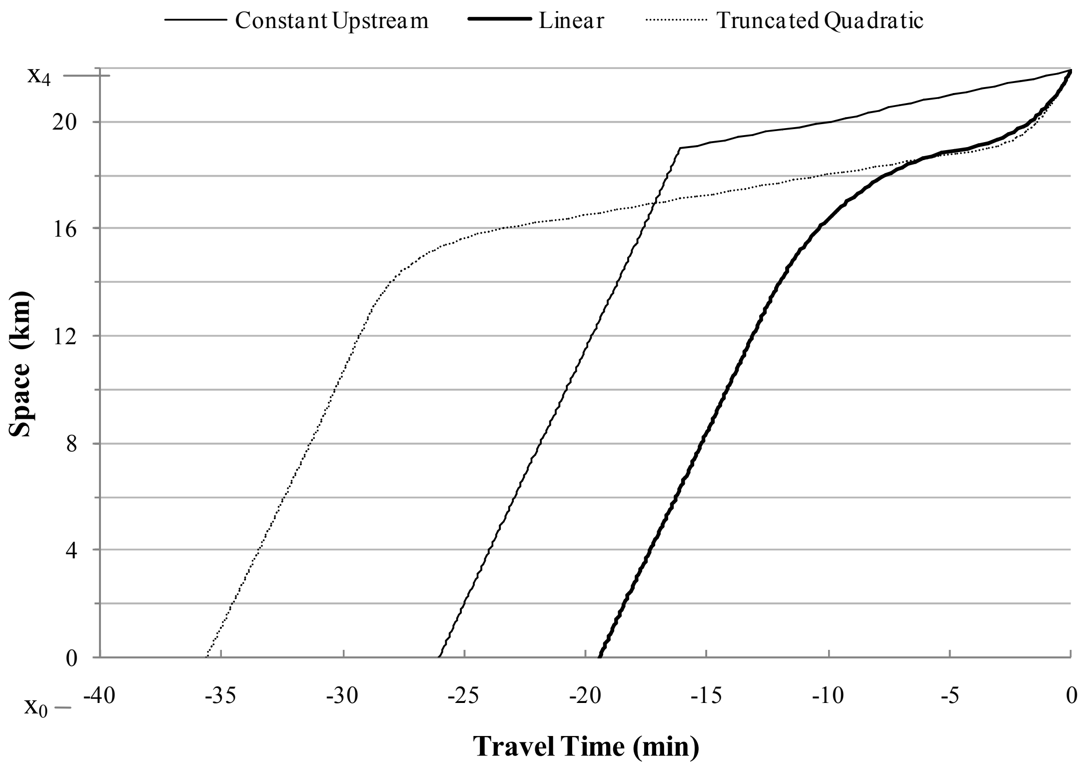


a)

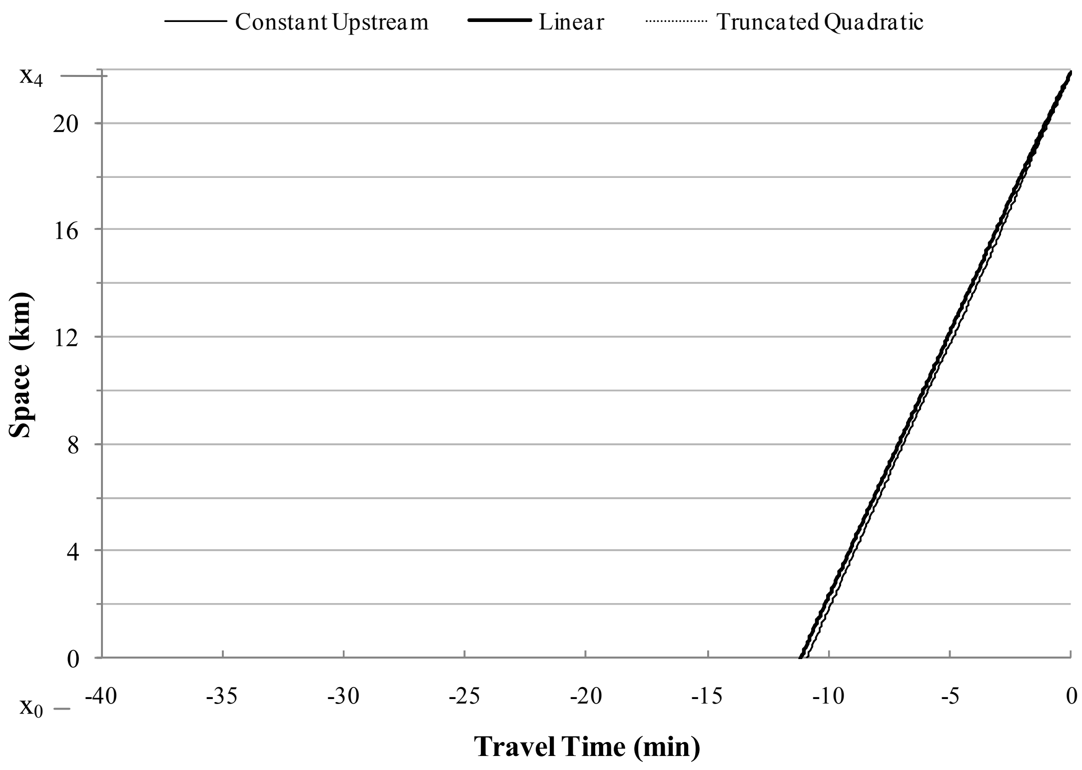


b)
FIGURE 13 Interpolated speed profiles on the test site, a) Arrival time 16:00h – partially congested stretch, b) Arrival time 20:18h – free flowing stretch.

Obviously, the absolute differences between speed profiles are smaller, as are the differences in the measured speeds. At the limit, they would converge to the same measured speed for the whole corridor. This is reflected in the resulting virtual vehicle reconstructed trajectory, as can be seen in Figure 14. In addition Figure 14 aims to show the behavior of the reconstructed trajectories, for instance piecewise linear in the case of piecewise constant speeds.



a)



b)

FIGURE 14 Reconstructed trajectories between "x_4" and "x_0" from different speed interpolations, a) Arrival time 16:00h – partially congested trip, b) Arrival time 20:18h – free flow trip.

The stretch between "Cardedeu junction" ("y_3") and the loop 4 location ("x_4") (see Figure 9 for details) is selected as the evaluation section. This selection responds for several reasons: firstly, ground truth travel times are available for this stretch. Secondly, "y_3" is nearby loop 3 location ("x_3") so that travel time estimations would clearly depend only on speed measurements of loops 3 and 4, making the interpretation of results easier. Finally, afternoon congestion on the test site grows along the whole section so that free flowing traffic, congestion onset, fully congested traffic and congestion dissolve episodes are identifiable. Figure 15 plots the comparison between measured and estimated travel times.

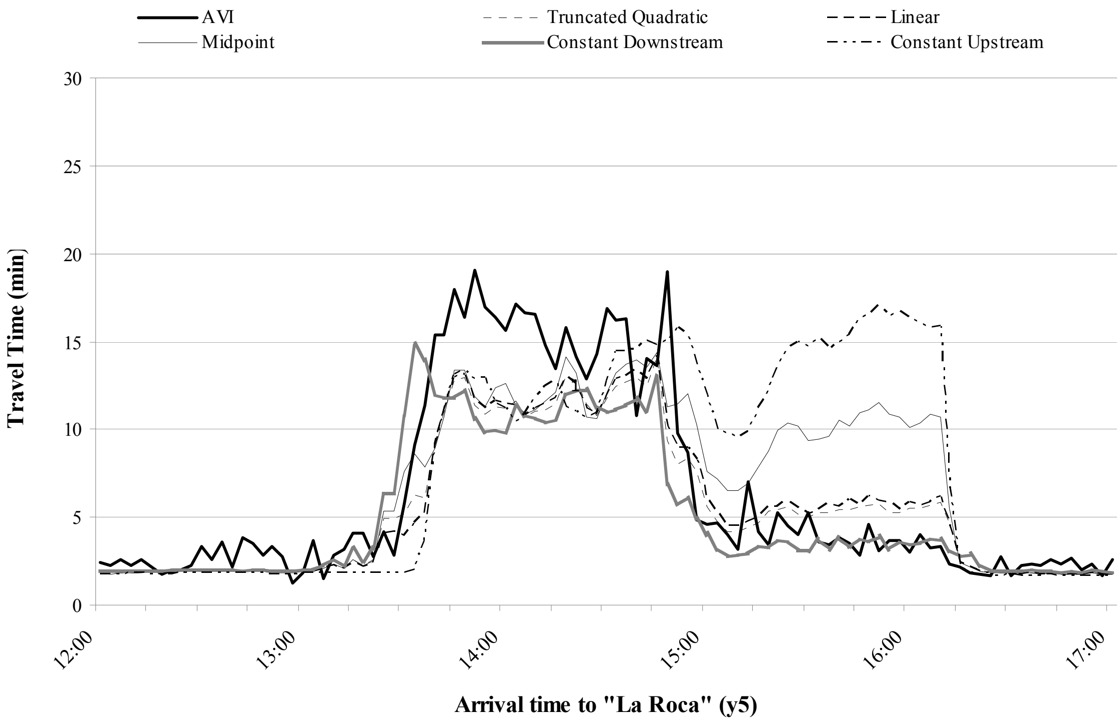


FIGURE 15 3 minute average travel time estimations from reconstructed trajectories on different speed space interpolation assumptions (section between "y_3" and "x_4"; 3.22 km).

From Figure 15, and in accordance with traffic evolution and loop speed measurements presented in Figure 10 and 11 respectively, several episodes in relation to the spanning of congestion over the section can be identified:

1. Free flowing traffic for vehicles finishing their trip at "La Roca" between 11:00 and 12:00, among others.
2. Congestion onset between 13:27 and 13:39 (arrival times at "La Roca"), when the queue was growing upstream at an approximated speed of 8.3 km/h. Note from Figure 11 that it took up to 21 minutes for the queue to grow between "x_4" to "x_3", a 2.9 km section.
3. Congested traffic on the whole stretch between 13:42 and 14:48.
4. Congestion dissolve between 14:51 and 15:00, when the queue was dissolving from downstream due to the removal of the patrol roadblock at an approximated speed of 14.5 km/h. Note from Figure 11 that it took up to 12 minutes for the queue to dissolve between "x_4" to "x_3".

5. Partially congested stretch due to a lane closure (resulting from the truck breakdown) nearby the upstream end of the stretch between 15:24 and 16:06.

Within episodes “a”, “c” and “e”, queues do not evolve with time in the evaluation stretch. Small travel time variations are only due to speed variance among drivers in free flowing conditions or to stop&go oscillations in congested traffic. In this case, the average error in the period is a good performance indicator of each travel time estimation method. The average error can be decomposed as a bias (the mathematical expectation of the error) and a residual (the standard deviation of the error), as formulated in the following equations:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{T}_i - T_i)^2} \tag{27}$$

$$Bias = \overline{\hat{T}} - \overline{T} = \frac{1}{n}\sum_{i=1}^{n}\hat{T}_i - \frac{1}{n}\sum_{i=1}^{n}T_i \tag{28}$$

$$RRE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left[(\hat{T}_i - \overline{\hat{T}}) - (T_i - \overline{T})\right]^2} \tag{29}$$

Where: “$\hat{T}_i$” is the estimated travel time over the section
“T_i” is the measured travel time over the section
“$\overline{\hat{T}}$” and “$\overline{T}$” are their respective arithmetic averages

“*RMSE*” stands for the Root Mean Squared Error and “*RRE*” for the Root Residual Error, where:

$$RMSE^2 = bias^2 + RRE^2 \tag{30}$$

Table 1 presents these performance indicators for each travel time estimation method in stationary traffic states: free flowing conditions, totally congested stretch and partially congested stretch (i.e. episodes “a”, “c” and “e”).

TABLE 1 Numerical Differences Between Measured and Estimated Travel Times for an Aggregation Period of 3 Minutes on the section between “y_3” and “x_4” (3.22 km) – Stationary Conditions

a)		**AVI measured travel time**		**Constant upstream estimation**		**Constant downstream estimation**	
		Mean	Dev.	Bias	RRE	Bias	RRE
a) Free flowing	minutes	2,11	0,62	-0,38	0,59	-0,20	0,59
	%	-	-	-18%	28%	-10%	28%
c) Totally congested	minutes	15,67	1,97	-3,08	2,50	-4,66	2,73
	%	-	-	-20%	16%	-30%	17%
e) Partially congested	minutes	3,75	0,66	11,95	1,22	-0,27	0,77
	%	-	-	318%	33%	-7%	20%

b)		Midpoint estimation		Linear estimation		Truncated quadratic estimation	
		Bias	RRE	Bias	RRE	Bias	RRE
a) Free flowing	minutes	-0,29	0,59	-0,29	0,59	-0,27	0,59
	%	-14%	28%	-14%	28%	-13%	28%
c) Totally congested	minutes	-3,34	2,51	-3,69	2,45	-3,87	2,52
	%	-21%	16%	-24%	16%	-25%	16%
e) Partially congested	minutes	6,66	1,00	1,97	0,74	1,68	0,74
	%	178%	27%	52%	20%	45%	20%

Qualitatively in Figure 15 and quantitatively in Table 1, it is shown that the performance of all methods is almost the same in episodes where uniform traffic conditions span for the whole section (episodes "a" and "c").

All methods perform well in free flowing conditions, as the average error is of the same order of magnitude as the travel time standard deviation. However, a systematic underestimation is observed. This results from a speed overestimation at loop detector sites in free flowing conditions in relation to the average speed of vehicles across the turnpike section. Two reasons explain this bias: One, the computations of time mean speeds instead of the lower space mean speeds, as stated in the introduction of the paper; Two, loop detectors are installed on privileged spots of the highway, far from problematic sections where speed drops off, like junctions and weaving sections.

On the contrary, all methods are not accurate enough when congestion spans for the whole stretch, reporting average errors at approximately twice the standard deviation of travel times in these episodes. Again, systematic underestimation is observed. Computing space mean speeds would slightly improve this bias, but it must be noted that the main reason for this underestimation is the biased speed measurement of loop detectors in stop&go situations, when only the "go" part of the movement is measured.

A different behavior of the methods appears in the case of a partially congested stretch (where the part covered by the queue does not evolve with time). Note that episode "e" corresponds to a situation where the queue only spans for a few upstream meters of the section, but stepping on the loop spot. These results in travel times are similar to the free flowing situation, but very low speeds are measured at the upstream loop location. It should be clear in this situation why constant downstream estimation outperforms all the other methods, and why constant upstream produces completely flawed travel times. Obviously, midpoint, linear and truncated quadratic estimations are in between. It is interesting to note that linear and truncated quadratic estimations beat midpoint estimations. Recall from Section 4.2 that piecewise constant approaches are equivalent to travel times resulting from a constant weighted harmonic average speed where the weighting factors are the relative coverage of each piece. Linear and truncated quadratic estimations could also be seen as piecewise constant approaches, with infinitesimally small pieces. One can easily realize that, while the arithmetic average of speeds would be approximately equal in all three approaches, harmonic ones are not, due to a higher influence of lower speeds. This is the reason why midpoint travel time estimations are significantly higher than linear and quadratic ones.

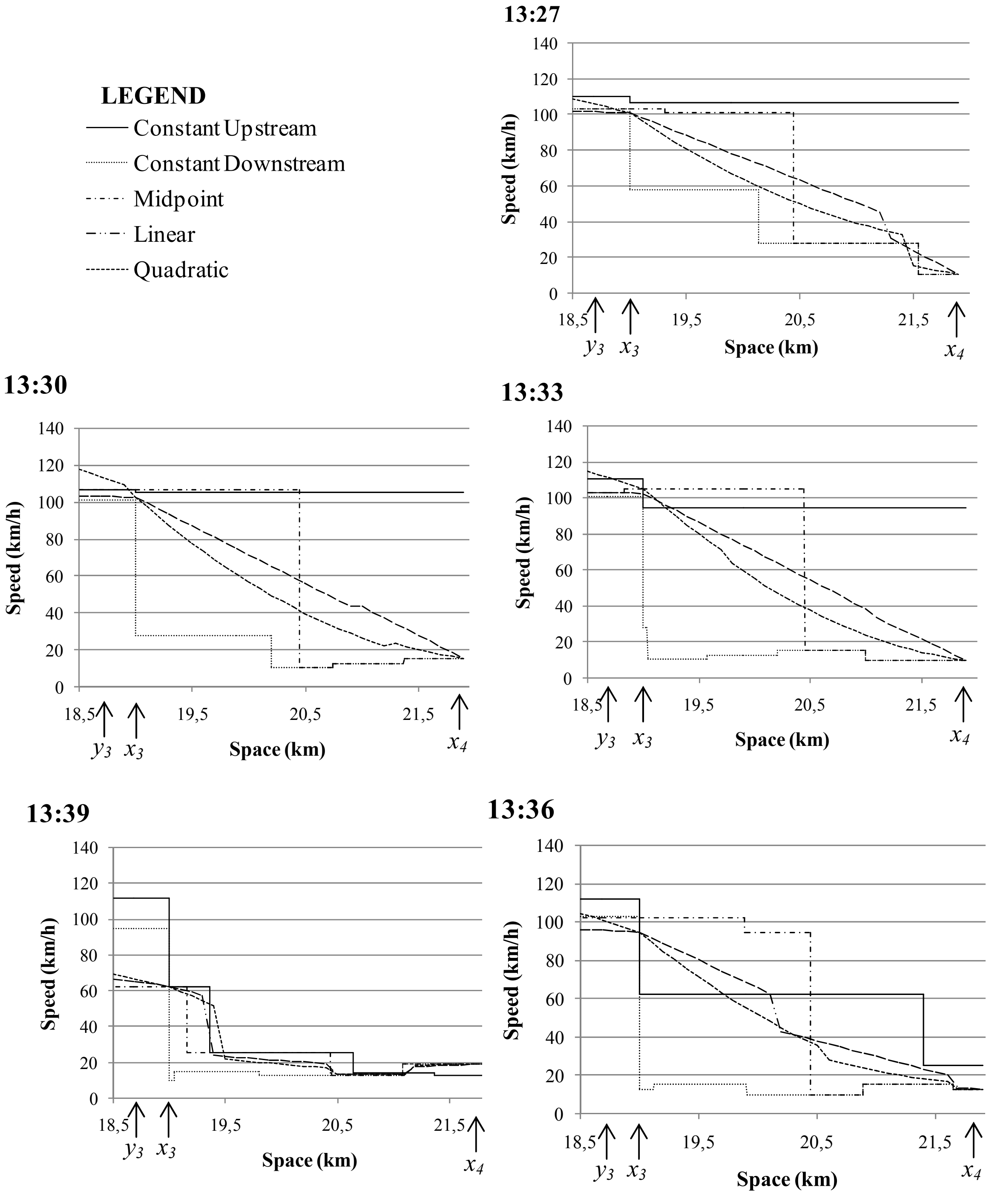
This does not mean that linear and truncated quadratic approaches outperform the midpoint algorithm in partially congested situations. It would be true in a case where the queue covers a small part of the section (like the situation analyzed here), but it would be the inverse if the queue spans for almost the whole length of the stretch.

What is evident from the presented results is that none of the methods are intrinsically better than the other in the case of partially congested situations. The best one is dependent on matching the method assumptions with the queue coverage of the section.

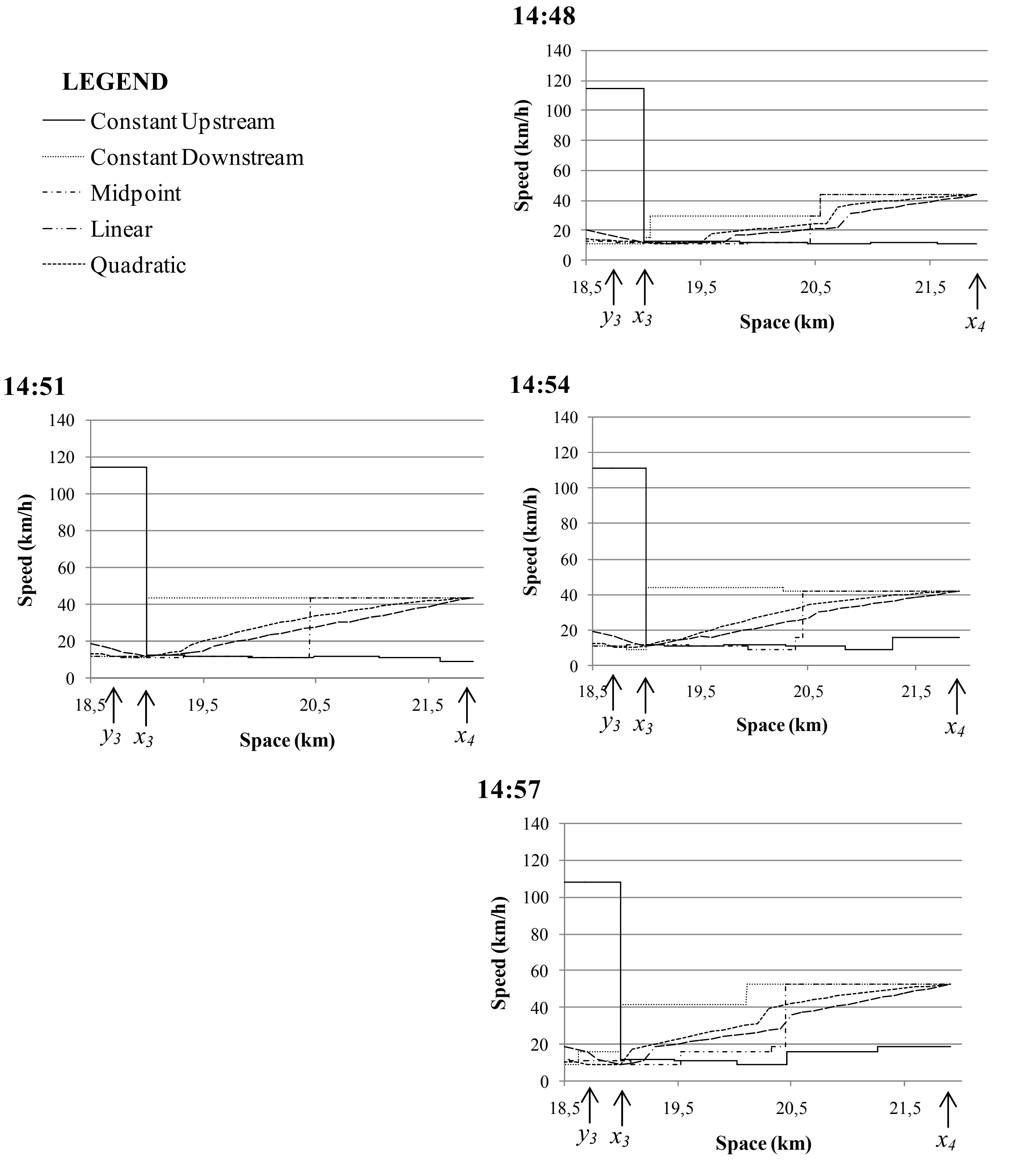
TABLE 2 Numerical Differences Between Measured and Estimated Travel Times for an Aggregation Period of 3 Minutes on the section between "y_3" and "x_4" (3.22 km) – Evolving Conditions

	Arrival time at "La Roca"	AVI measure (min)	Absolute error (minutes)				
			Constant upstream	Constant downstream	Midpoint	Linear	Truncated quadratic
b) Congestion onset	13:27	2,87	-1,07	3,49	2,50	1,27	2,06
	13:30	5,73	-3,89	5,05	1,91	-1,76	-0,54
	13:33	9,12	-7,11	5,68	-0,50	-4,46	-2,87
	13:36	11,34	-7,69	2,60	-3,46	-6,08	-5,24
	13:39	15,39	-6,36	-3,42	-6,42	-6,20	-6,08
d) Congestion dissolve	14:48	18,96	-3,94	-12,13	-7,72	-8,76	-9,57
	14:51	9,77	5,99	-4,06	1,71	-0,86	-1,71
	14:54	8,72	6,61	-2,65	3,35	0,26	-0,34
	14:57	4,89	8,80	-0,04	5,41	3,37	2,75

This evidence is also seen in situations when queues evolve with time within the section. Episode "b" corresponds to a congestion onset from downstream (congestion grows against traffic direction) while during episode "d" congestion dissolves also from downstream (against traffic direction). Table 2 presents numerical results of the absolute errors committed with each estimation approach, and shows how the error decreases when traffic state approaches the assumptions of the method. The logic of the absolute error behavior is clear, although some values may seem on the wrong side, resulting from the systematic overestimation of average speeds in congested conditions (see Table 1). Figure 16 may help in the interpretation of numerical values in Table 2.


LEGEND
Constant Upstream
Constant Downstream
Midpoint
Linear
Quadratic
13:27
13:30
13:33
13:39
13:36
Speed (km/h)
Space (km)
0
20
40
60
80
100
120
140
18,5
19,5
20,5
21,5
y_3
x_3
x_4

a)



b)

FIGURE 16 Reconstructed speed profiles on the section between "y_3" and "x_4". a) Congestion onset episode. b) Congestion dissolve episode.

7. PROPOSED NAÏVE METHOD FOR ONLINE TRAVEL TIME ESTIMATION

From the previous section it was concluded that any travel time estimation method based on speed interpolation between measurement points which obviate traffic dynamics, does not perform better than the others. In general, all of them provide highly underestimated travel times in fully congested stretches, while performance on partially congested sections is case specific. Although all methods result in similar average errors over long time periods, if one wants to avoid extreme punctual errors, constant upstream and constant downstream methods should be avoided. Among the others, midpoint algorithm stands out due to its simplicity.

However, other implications must be considered in an online context (i.e. real time information systems), where the "goodness" of the estimation takes another meaning besides the accuracy of the measurement. Note that when a driver receives travel time information at an instant "t", he wishes to obtain his expected travel time. Recall from Section 2 that this is a departure based travel time (DTT) for the time period (t_i, t_{i+1}) containing "t". In fact, this represents future information. In contrast, the available measured information could be an arrival based travel time (ATT) for the previous time interval (t_{i-1}, t_i) if direct travel time measurements were available or a true travel time estimation (TTT) from speed measurements at loop detector sites, again for the time interval (t_{i-1}, t_i). In this context, "goodness" of the estimation should be defined as how the measured ATT or the estimated TTT approach the expected DTT at the next time period. Therefore, the quality of the online travel time information does not only depend on the accuracy of the measurement but also on its forecasting capabilities. In this situation, and taking into account that ATT represents outdated information (see Section 2), it should not come as a surprise that rough TTT estimations could provide a better approach to DTT than precisely measured ATT, as TTT avoids the delay in the information resulting from trajectory based measurements, which need the vehicles to have finished their trip in order to obtain the travel time information.

Corridor TTT estimations result from simply adding up section travel times at the same time interval. This provides the desired immediacy in reporting any travel time variation, but only solves half the problem as traffic will evolve from the instant of the measurement to the time the traveler actually undertakes the journey in the next time period. Instant corridor travel times constructed from the addition of accurate link level travel time estimations must be seen as the best real time estimations one can obtain without falling under the uncertainties of forecasting. In spite of this, one has to bear in mind the discouraging fact that any accuracy improvement may be swamped out by the evolution of the traffic state until the forecasting horizon.

Deprived of the time smoothing process which constituted the trajectory reconstruction, and taking into account that the time interval "Δt" must be small in order to reduce the horizon of the DTT prediction, provide a frequent update of the information and not smooth out travel time variations, TTT estimation suffers from fluctuations, resulting from the stochastic nature of driver's behavior. Obviously, the variance of the time mean speed estimation is reduced with the increase of the number of observations (see Equation 31). Therefore, as small "Δt" imply fewer observations, higher average speed variance results, and this leads to volatile TTT estimations during small time intervals.

Severe travel time fluctuations over consecutive time intervals damage the drivers' perceived credibility of the information system. Therefore, a time smoothing process is necessary. Moving average or exponential smoothing methods are usually proposed [Cortés et al., 2002; Treiber and Helbing, 2002; Kothuri et al., 2008;]. However, these standard smoothing processes imply a delay in the detection of speed changes which are not attributed to a fluctuation but to a passage of a shockwave. This implies the loss of the immediacy benefits of TTT.

7.1. Average speed fluctuations smoothing process

An intelligent smoothing process is proposed which smoothes out travel time fluctuations while preserving the immediacy in the detection of significant speed changes. The method determines whether a speed variation is a fluctuation and therefore must be smoothed, or whether it is a consequence of a change in the traffic state and therefore must not be smoothed in order to maintain the immediacy of the information.

Then, a speed variation is considered a change in a traffic state and will not be smoothed in the case of:

- A large speed variation after another large speed variation of the same sign, showing a tendency.
- A small variation after a large speed change. A change in the mean speed has occurred.

On the contrary, a speed variation is considered a fluctuation and will be smoothed if:

- Consecutive small speed variations.
- Consecutive and opposite sign large variations.
- Sharp speed change after a small variation.

Note that in this last situation it is not possible, until the next time step, to determine if the sharp increase or decrease of speed responds to a large fluctuation or to a significant variation. The default assumption is to smooth these data, waiting for the next time step measurement to reach a conclusion. Therefore, priority is given to smoothing in relation to immediacy, but it could also be the other way around.

If necessary, the smoothing process consists of a moving average over the last 15 minutes (5 time intervals in the present application) or since the last large speed variation in case a transition occurs.

There remains the need to define what is considered a small speed change in relation to a large variation. Consider vehicular speed as a random variable "V" whose mathematical expectation is "θ" and variance "σ_v^2" over a stationary period. The arithmetic mean of the speed observations within "Δt" is an unbiased estimator of "θ". It can be shown that the variance of the sample mean estimator is given by:

$$Var(\bar{v}) = \frac{\sigma_v^{\ 2}}{n} \tag{31}$$

Where "$\bar{v}$" is the sample mean and "*n*" the number of observations within "*Δt*". Considering the central limit theorem, the sample mean random variable is normally distributed, and therefore the absolute error of the estimation, "ε_v", can be expressed as:

$$\varepsilon_v = (prob.level) \cdot \frac{\sigma_v}{\sqrt{n}} \tag{32}$$

With a 99% confidence interval on the estimation and with a normal distribution, the probability level takes approximately a value of 3. Then, the maximum relative error, "e_v", in the estimation of the time mean speed of a stationary traffic stream is:

$$e_v = 3 \cdot \frac{C.V.}{\sqrt{n}} \tag{33}$$

where "*C.V.*" stands for the speed coefficient of variation, $C.V. = \sigma_v / \theta$. Taking into account the values of the speed C.V. reported in [Soriguera and Robusté, 2011], of approximately 0.25, the maximum errors on the average speed estimation in each loop detector and every time period can be computed.

Finally, a speed variation between consecutive time intervals is considered large if the intersection of the respective confidence intervals is null.

7.2. Section travel time smoothing process due to wave propagation

The smoothing process could also be considered a naïve first approach to add traffic dynamics in travel time estimation from punctual speed measurements, particularly in traffic transitions. This can be done by applying the moving average smoothing even when a real traffic state transition is detected, under certain conditions. Note that, maybe in an abuse of simplification, congestion onset always grows from downstream and against the traffic flow direction (omitting the rare moving bottleneck episodes). On the contrary, congestion dissolves from upstream in the same direction as traffic in the case of a reduction of demand, and from downstream against traffic in the case of an increase in the bottleneck capacity.

These concepts could be applied in the travel time smoothing process. It is only needed to detect congestion onset / dissolve within the link. This is easily achieved by comparing the average speeds measured at the loop detector sites which define the link (see Table 3).

Once the nature of an evolving traffic episode is detected, and given the expected direction of the evolution of the transition, an "intelligent" moving average smoothing can be applied. The smoothing should only be applied when the transition episode evolves toward the highway section steered (in terms of travel time information) by the loop detector being considered. Obviously, this would depend on the travel time interpolation method considered.

TABLE 3 Detection of Traffic Evolving Conditions

	Loop "*j*" (downstream end)	Loop "*j-1*" (upstream end)
Congestion onset	$v_j^{(t)} + \varepsilon_{v_j}^{(t)} < v_j^{(t-1)} - \varepsilon_{v_j}^{(t-1)}$	
Congestion dissolve from downstream	$v_j^{(t)} - \varepsilon_{v_j}^{(t)} > v_j^{(t-1)} + \varepsilon_{v_j}^{(t-1)}$	$v_j^{(t-1)} - \varepsilon_{v_j}^{(t-1)} < v_{j-1}^{(t)} - \varepsilon_{v_{j-1}}^{(t)} < v_j^{(t-1)} + \varepsilon_{v_j}^{(t-1)}$ *or* $v_j^{(t-1)} - \varepsilon_{v_j}^{(t-1)} < v_{j-1}^{(t)} + \varepsilon_{v_{j-1}}^{(t)} < v_j^{(t-1)} + \varepsilon_{v_j}^{(t-1)}$
Congestion dissolve from upstream	$v_{j-1}^{(t-1)} - \varepsilon_{v_{j-1}}^{(t-1)} < v_j^{(t)} - \varepsilon_{v_j}^{(t)} < v_{j-1}^{(t-1)} + \varepsilon_{v_{j-1}}^{(t-1)}$ *or* $v_{j-1}^{(t-1)} - \varepsilon_{v_{j-1}}^{(t-1)} < v_j^{(t)} + \varepsilon_{v_j}^{(t)} < v_{j-1}^{(t-1)} + \varepsilon_{v_{j-1}}^{(t-1)}$	$v_{j-1}^{(t)} - \varepsilon_{v_{j-1}}^{(t)} > v_{j-1}^{(t-1)} + \varepsilon_{v_{j-1}}^{(t-1)}$

Note: Superscript stands for time interval of calculation.

In the case of the Midpoint algorithm (see Figure 6a), moving average should be applied in link "*j(b)*" travel times if congestion onset or congestion dissolve from downstream is detected at loop "*j*". Smoothing should also apply to link "*j(a)*" in case detector "*j-1*" detects congestion dissolve from upstream.

The only remaining issue is how long the moving average should span. The answer to this question is specific of each application, and it must account for the length of the application of each speed measurement (in the midpoint case half of the link length) and on the characteristic wave speed on the link. Again, this must be considered to be a naïve approach, as the characteristic wave speed represents the speed of the shockwave between any two congested states, but not between congested and free flowing states or vice versa. The characteristic wave speed is a maximum speed enfolding all wave speeds in the section. For the layout presented in Figure 9, considering a length of the link of 3 km, and a characteristic wave speed of 14.5 km/h, the smoothing should span for approximately 6 minutes. This is 2 time periods of a moving average.

7.3. Results of the smoothing process on the test site

The proposed intelligent smoothing has been applied to the test data. Results are shown in Figure 17, which represents the same scenario as Figure 15. Note that in Figure 15 all travel times were arrival based. On the contrary, in Figure 17, AVI measured arrival based travel times are shown only as a reference. Recall that the objective of the online estimation is to approach the AVI measured departure based travel times which would not be available on an online context, as this would be future information. Delay of arrival based travel times in the detection of the congestion onset is evident. The benefits in immediacy of using true travel times, for instance using the Midpoint algorithm are also clear. However, one must realize that true travel times resulting from punctual speed measurements preserve their intrinsic lack of accuracy. This leads to some paradoxical situations where the accurate but delayed measured arrival based travel times (in case they are available) are a better estimation of the expected travel times in relation to the more updated but excessively flawed true travel times. This commonly happens in the case of low loop surveillance densities.

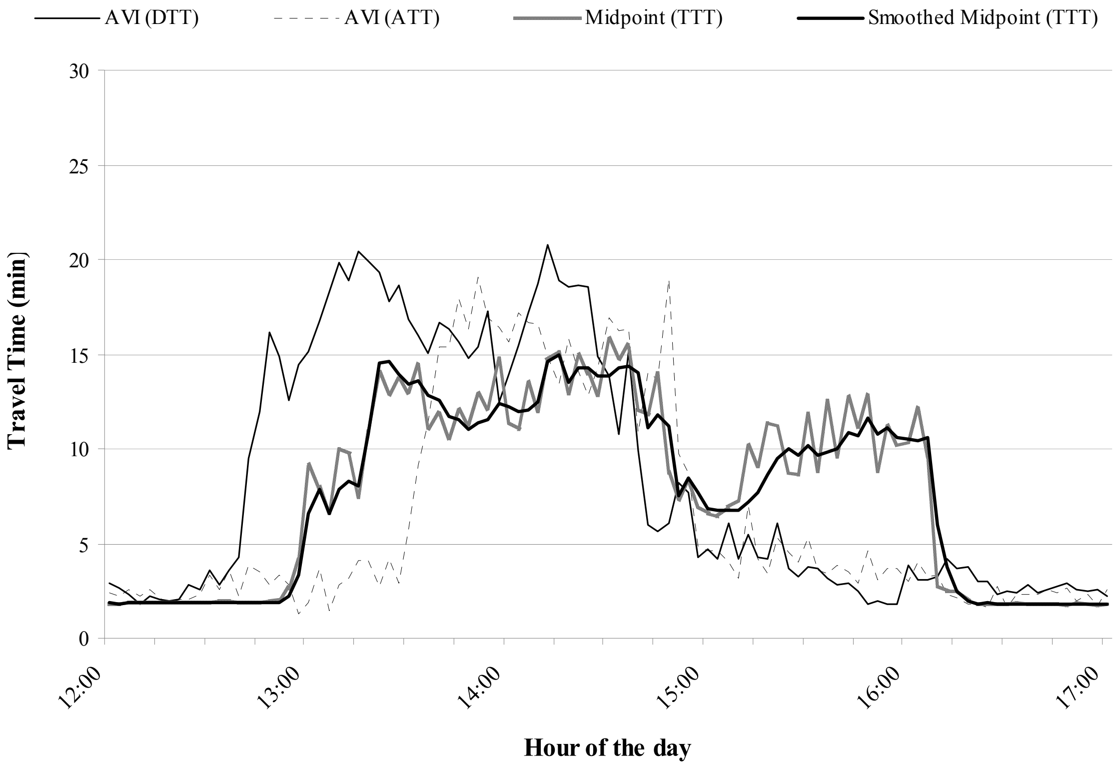


FIGURE 17 Results of the intelligent smoothing process applied to Midpoint TTT (3 minute average; section between "y_3" and "x_4", 3.22 km).

Finally, Figure 17 also shows the results of the intelligent smoothing of true travel times, achieving a reduction from 0.15 to 0.05 in the coefficient of variation of travel times in congested episodes, while preserving the immediacy in the detection of significant traffic stream transitions.

8. CONCLUSIONS AND FURTHER RESEARCH

There is a need for travel time information in metropolitan freeways, where in most cases solely loop detector surveillance is available. Several methods have been developed in order to estimate travel times from speed measurements at loop detector sites whose main differences lie in the speed interpolation approach between point measurements. In fact, the ignorance of speed evolution between measurement points represents a major drawback for these types of methods.

The present paper demonstrates conceptually and with an accurate empirical comparison resulting from accurate travel time definitions, that travel time estimation methods based on mathematical speed interpolations between measurement points, which do not consider traffic dynamics and the nature of queue evolution, do not contribute in an intrinsically better estimation, independently of the complexity of the interpolation method. All of them show a similar performance when a unique traffic state covers the whole target stretch. It can be concluded that all methods perform well in free flowing conditions in spite of a slight systematic underestimation. In contrast, all methods provide highly underestimated estimations in completely congested sections, resulting in unrealistic travel times. The main reason for this bad performance is the inability of loop detectors to capture the speed oscillations produced by stop&go traffic, resulting in inaccurate and systematic overestimated average speed estimates. The improvement of average speed loop measurements under congested situations should be considered as an issue for further research.

Major errors can arise in the case of partially congested sections, resulting from a congestion onset (or dissolve) episode or due to the activation of a bottleneck within the section (either recurrent or incident related). In this situation, different methods provide significantly different estimations. The quality of each method relies on the fitness of the method assumptions of the real evolution of the queue along the section. In practice, the relative benefits between commonly used interpolation assumptions, like constant speed over the whole stretch, arbitrary location for the speed change (e.g. midpoint algorithm), or speed profiles resulting from mathematical interpolations blind to traffic dynamics (e.g. linear or truncated quadratic approaches) are site specific, and in the case of indiscriminate use, depend on chance. Avoiding methods based on only one loop detector measurement (e.g. constant upstream or constant downstream) will prevent the highest punctual errors. Therefore, under no better approach, midpoint, linear or truncated quadratic methods are recommended, from which midpoint algorithm stands out for its simplicity. The key issue for an efficient method should be the estimation of queue length within the section so that each speed measurement could be assigned to an adequate length of the highway stretch.

The absolute magnitude of these estimation errors directly depend on the loop detector spacing. Lower surveillance densities result in higher probability of a traffic transition within the section defined by two consecutive loops. In addition, the estimation error would be propagated over a longer length, resulting in higher absolute errors. Being aware of this situation, and realizing the difficulties of locating traffic state transitions in long sections, most traffic agencies have chosen to invest in higher loop densities (typically 1 loop detector every 500m), in order to obtain realistic travel times from their measurements. In this configuration, based on intensive surveillance, the selected travel time estimation method does not matter. However, the frequent detector failure and malfunctioning, which translates to temporally increased detector spacing, should also be taken into account.

The conclusions of this paper should not be taken as suggesting not to use the existing schemes in surveillance configurations where this intensive monitoring is not available, but that they should be used only carefully and not be taken as perfect. For instance, most of the methods can be used to provide an upper and lower bound on the travel time, using one end or the other of a link.

In a real-time context, the main advantage of travel times estimated from loop measurements is the possibility of obtaining true travel times; this is to obtain a virtual measurement for a vehicle travel time before the end of its journey. This provides benefits in the immediacy of the detection of transitions in the traffic stream state. However, these benefits are usually obscured by the lack of accuracy of these measurements and their excessive fluctuations. This last issue is addressed in the paper, proposing a simple method to smooth the fluctuations while preserving the immediacy.

From the paper it is concluded that on the one hand, directly measuring travel times provides accuracy benefits, although delayed information. On the other hand, the indirect estimation from speed measurements provides immediacy in exchange for a loss of accuracy. It seems that both types of measurements should be complementary. A data fusion scheme capable of taking the better of each one remains an issue for further research.

REFERENCES

Abdulhai, B. and S.M. Tabib. (2003) Spatio-temporal inductance-pattern recognition for vehicle reidentification. *Transportation Research Part C* **11**(3-4), 223-239.

Claudel, C.G., A. Hofleitner, N. Mignerey and A. M. Bayen. (2009) Guaranteed bounds on highway travel times using probe and fixed data. *Paper presented in the 88th Transportation Research Board Annual Meeting*. Washington D.C.

Coifman, B. (2001) Improved velocity estimation using single loop detectors. *Transportation Research Part A* **35**(10), 863-880.

Coifman, B. (2002) Estimating travel times and vehicle trajectories on freeways using dual loop detectors. *Transportation Research Part A* 36(4), 351-364.

Coifman, B. and E. Ergueta. (2003) Improved vehicle reidentification and travel time measurement on congested freeways. *ASCE Journal of Transportation Engineering* **129**(5), 475-483.

Coifman, B. and S. Krishnamurthya. (2007) Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C* **15**(3), 135-153.

Coifman, B., and M. Cassidy. (2002) Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research Part A* **36**(10), 899-917.

Coifman, B., S. Dhoorjaty and Z. Lee. (2003) Estimating the median velocity instead of mean velocity at single loop detectors. *Transportation Research Part C* **11**(3-4), 211-222.

Cortés, C.E., R. Lavanya, J. Oh, and R. Jayakrishnan. (2002) General-purpose methodology for estimating link travel time with multiple-point detection of traffic. *Transportation Research Record: Journal of the Transportation Research Board* **1802**, 181-189.

Dailey D.J. (1999) A statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B* **33**(5), 313-322.

Dailey, D.J. (1993) Travel time estimation using cross-correlation techniques. *Transportation Research Part B* **27**(2), 97-107.

Edie, L.C. (1965) Discussion of traffic stream measurements and definitions. *Proc. 2nd International Symposium on the Theory of Traffic Flow*, J. Almond, ed., OECD, Paris, 139-154.

Fariello, B. (2002) *ITS America RFI travel time projects in North America.* San Antonio TransGuide Program. Texas Department of Transportation.

Hellinga, B. (2002) Improving freeway speed estimates from single loop detectors. *ASCE Journal of Transportation Engineering* **128**(1), 58-67.

Kothuri, S.M., K.A. Tufte, E. Fayed and R.L. Bertini. (2008) Toward understanding and reducing errors in real-time estimation of travel times. *Transportation Research Record: Journal of the Transportation Research Board* **2049**, 21-28.

Kothuri, S.M., K.A. Tufte, S. Ahn and R.L. Bertini. (2007) Using archived ITS data to generate improved freeway travel time estimates. *Paper presented in the 86th Transportation Research Board Annual Meeting*. Washington D.C.

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.

Li, R., G. Rose and M. Sarvi. (2006) Evaluation of speed-based travel time estimation models. *ASCE Journal of Transportation Engineering* **132**(7), 540-547.

Lin, W., J. Dahlgren and H. Huo. (2004) Enhancement of vehicle speed estimation with single loop detectors. *Transportation Research Record: Journal of the Transportation Research Board* **1870**, 147-152.

Lucas, D.E., P.B. Mirchandani and N. Verma. (2004) Online travel time estimation without vehicle identification. *Transportation Research Record: Journal of the Transportation Research Board* **1867**, 193-201.

Mikhalkin, B., H. Payne and L. Isaksen. (1972) Estimation of speed from presence detectors. *Highway Research Board* **388**, 73-83.

Nam, D. H., and D.R. Drew. (1996) Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. *ASCE Journal of Transportation Engineering* **122**(3), 185-191.

OECD/JTRC (2010) *Improving reliability on surface transport networks*. OECD Publishing, Paris.

Oh, J.S., R. Jayakrishnan and W. Recker. (2003) Section travel time estimation from point detection data. *Paper presented in the 82nd Transportation Research Board Annual Meeting*. Washington D.C.

Palen, J. (1997) The need for surveillance in intelligent transportation systems. *Intellimotion*, 6:1, pp. 1-3, 10. University of California PATH, Berkeley, CA.

Petty, K.F., P. Bickel, M. Ostland, J. Rice, F. Schoenberg, J. Jiang and Y. Rotov. (1998) Accurate estimation of travel time from single loop detectors. *Transportation Research A* **32**(1), 1-17.

Pushkar, A., F. Hall and J. Acha-Daza. (1994) Estimation of speeds from single-loop freeway flow and occupancy data using cusp catastrophe theory model. *Transportation Research Record: Journal of the Transportation Research Board* **1457**, 149-157.

Rakha H. and W. Zhang. (2005) Estimating traffic stream space-mean speed and reliability from dual and single loop detectors. *Transportation Research Record: Journal of the Transportation Research Board* **1925**, 38-47.

Soriguera, F. and F. Robusté. (2011) Estimation of traffic stream space-mean speed from time aggregations of loop detector data. *Transportation Research Part C* **19**(1), 115-129.

Soriguera, F., D. Rosas and F. Robusté. (2010) Travel time measurement in closed toll Highways. *Transportation Research Part B* **44**(10), 1242–1267.

Sun, L., J. Yang and H. Mahmassani. (2008) Travel time estimation based on piecewise truncated quadratic speed trajectory. *Transportation Research Part A* **42**(1), 173-186.

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

Turner, S. M., W. L. Eisele, R. J. Benz, and D.J. Holdener. (1998) *Travel time data collection handbook*. Research Report FHWA-PL-98-035. Texas Transportation Institute, Texas A&M University System, College Station, Tex.

van Lint, J. and N. van der Zijpp. (2003) Improving a travel-time estimation algorithm by using dual loop detectors. *Transportation Research Record: Journal of the Transportation Research Board* **1855**, 41-48.

Wang, Y. and L.H. Nihan. (2000) Freeway traffic speed estimation with single-loop outputs. *Transportation Research Record: Journal of the Transportation Research Board* **1727**, 120-126.

Wang, Y. and L.N. Nihan. (2003) Can single-loop detectors do the work of dual-loop detectors? *ASCE Journal of Transportation Engineering* **129**(2), 169-176.

APPENDIX A3

Travel Time Measurement in Closed Toll Highways

ABSTRACT

Travel time for a road trip is a drivers' most appreciated traffic information. Measuring travel times on a real time basis is also a perfect indicator of the level of service in a road link, and therefore is a useful measurement for traffic managers in order to improve traffic operations on the network. In conclusion, accurate travel time measurement is one of the key factors in traffic management systems.

This paper presents a new approach for measuring travel times on closed toll highways using the existing surveillance infrastructure. In a closed toll system, where toll plazas are located on the on/off ramps and each vehicle is charged a particular fee depending on its origin and destination, the data used for toll collection can also be valuable for measuring mainline travel times on the highway. The proposed method allows estimating mainline travel times on single sections of highway (defined as a section between two neighboring ramps) using itineraries covering different origin – destinations. The method provides trip time estimations without investing in any kind of infrastructure or technology. This overcomes some of the limitations of other methods, like the information delay and the excess in the travel time estimation due to the accumulation of exit times (i.e. the time required to travel along the exit link plus the time required to pay the fee at the toll gate).

The results obtained in a pilot test on the AP-7 toll highway, near Barcelona in Spain, show that the developed methodology is sound.

Keywords: Highway travel time measurement, toll highways, toll ticket data.
The accuracy

1. INTRODUCTION

There is common agreement among drivers, transportation researchers and highway administrations that travel time is the most useful information to support trip decisions (users) and to assess the operational management of the network (administrators), (Palen, 1997).

In response to these needs for accurate road travel time information, researchers and practitioners from all over the world have worked hard in this direction. During the last two decades, research efforts have been focused on the indirect estimation of road travel times, using the fundamental traffic variables, primarily each vehicle's speed observed at discrete points in the freeway. The prominence of this approach results from the fact that, for ages, these have been the unique available traffic data, as provided by inductance loop detectors. Advances in this research area have been huge as it demonstrates a vast related literature. The efforts made in improving the accuracy of speed estimations from single loop detectors should be emphasized (Coifman, 2001; Dailey, 1999; Mikhalkin et al., 1972; Pushkar et al., 1994). However, although accurate spot speed estimations have been obtained (as in the case of using double loop detectors), travel time estimates could still be flawed due to extrapolating spot measurements to a highway section, with the possibility of different traffic conditions (congested or not) along its length. Note that this problem is greater on highways with a low density of detection sites. One detector site every half kilometer of highway is desirable to reduce the effect of this problem (Hopkin et al., 2001). Several approaches have been published trying to overcome this limitation without falling into the enormous cost of intensive loop surveillance, proposing, for example, different methods for the reconstruction of vehicle trajectories between loop detectors (Coifman, 2002; Cortés et al., 2002; Li et al., 2006; van Lint and van der Zijpp, 2003). In addition to these problems, it must also be taken into account that the loop speed estimates in the case of stop and go traffic situations do not adequately represent the space mean speed of the traffic stream.

A different approach to the indirect estimation of link travel times using loop detectors consists of comparing the cumulative counts (N-curves) from consecutive loop detectors (instead of using the spot speed measurement at the detector site). In the case of all on and off ramps being monitored, the flow conservation equation can be applied to obtain the travel time in the stretch (Nam and Drew, 1996; van Arem, 1997). This method does not suffer from previous speed estimation limitations, but must account for loop detector drift that can jeopardize the accuracy of the results.

Lately, research on travel time estimation using loop detector data has focused on direct measurement, consisting of measuring the time interval that a particular vehicle takes to travel from one point to another. To achieve this goal several authors propose a smart use of loop detector data, on the basis of the re-identification of particular vehicles in consecutive loop detectors by means of characteristic length (Coifman and Ergueta, 2003; Coifman and Krishnamurthya, 2007) or particular inductive signature on the detector (Abdulhai and Tabib, 2003; Sun et al., 1998, 1999). An extended approach of these algorithms is the re-identification of particular features of vehicle platoons instead of individual vehicles (Coifman and Cassidy, 2002; Lucas et al., 2004). All these last contributions can, however, not be put into practice with the current common hardware and/or software loop configurations. Most operating highway agencies would have to upgrade their systems in the field to accomplish these objectives.

In another order of events, the deployment of ITS (Intelligent Transportation Systems) during the last decade has brought the opportunity of using more suitable traffic data to directly measure travel times (Turner et al., 1998). This is the case with AVI (Automated Vehicle Identification) data obtained, for instance, from the readings of vehicle toll tags or from video license plate recognition. By matching the vehicle ID at different locations on the highway, link travel times can be directly obtained if the clocks at each location are properly synchronized. Another approach for directly measuring travel times is to use automatic floating car data obtained from different technologies such as GPS (Global Positioning Systems) or the emerging cellular phone geo-location. Take as an example the Mobile Century field experiment performed recently in a Californian highway (Herrera et al., 2010). In these schemes travel times are obtained by the real time tracking of probe vehicles, being their number critical for the accuracy of the measurements. Results obtained by Herrera et al. (2010) suggest that a 2-3% penetration of GPS-enabled cell phones in the drivers' population is enough to provide accurate measurements of the velocity of the traffic stream.

The present paper focuses on the direct highway travel time measurement using AVI data from toll collection systems. Although this concept is not new (Davies et al., 1989), the contributions found in the literature primarily deal with the usage of ETC (Electronic Toll Collection) data to measure travel times. These systems identify the vehicles by means of on-vehicle electronic tags and roadside antennas located, sometimes ad hoc, on the main highway trunk. Under this configuration the basic problems are the level of market penetration of the electronic toll tags and how to deal with time periods when only small samples are available in order to obtain a continuous measurement of travel times (Dion and Rakha, 2006; SwRI, 1998). Surprisingly, very few contributions are found related to travel measurement using the primitive configuration of a closed toll system. The concept of a "closed" toll system refers to the fact that the toll a particular driver pays varies depending on the origin and destination of his trip and is approximately proportional to the distance traveled on the highway. In contrast, one has to bear in mind the "open" toll systems, where toll plazas are strategically located so that all drivers pay the same average fee at the toll booth.

This paper deals with travel time measurement in the typical closed toll system configuration, widely extended in Europe and Japan for a long time, and by the authors knowledge only discussed in Ohba et al. (1999) in the particular case of main highway trunk toll plazas and a single origin destination pair. Under the closed toll configuration, each vehicle entering the highway receives a ticket (traditionally a card with magnetic band or more recently a virtual ticket using an ETC device), which is collected at the exit. The ticket includes the entry point, and the exact time of entry. By cross-checking entry and exit data, the precise time taken by the vehicle to travel along the itinerary (route) can be obtained (obviously clocks at the entry and exit toll plazas are considered to be synchronized). Averages can be obtained from the measurements for all the vehicles traveling along the same itinerary in the network. In relation to using only ETC based travel time estimation, a particular advantage of this configuration is the huge amount of data, since all vehicles have their entry/exit ticket (real or virtual), solving the problem of the market penetration of the ETC devices. However, other problems arise from this configuration, which are discussed in the next section of the paper. The proposed solution involves the estimation of the single section travel time (i.e. the time required to travel between two consecutive ramps on the highway) and also the exit time for each ramp (i.e. the time required to travel along the exit link plus the time required to pay the fee at the toll gate). Combining both estimations makes it possible to calculate all the required route travel times.

This paper is organized as follows: first, in Section 2, the context of the problem and the solution approach are described. In Section 3, the concept of the algorithm and the basic notation and formulation are provided, keeping the data filtering process and the more mathematical expressions of the algorithm at the end of the paper in Appendices A, B and C. Section 4 presents some modifications of the algorithm for its implementation in real time or off-line configurations. Then in Section 5, results of the application of the model to the AP-7 highway in Spain are presented. Finally, general conclusions and issues for further research are discussed.

2. OBJECTIVE OF THE PROPOSED ALGORITHM

As stated above, travel time can be obtained by directly measuring the time taken for vehicles to travel between two points on the network, and this seems particularly easy in closed toll highways, where the data needed for the toll collection makes it possible to obtain itinerary travel times for all origin – destination relations on the highway. Despite this apparent simplicity, several problems arise.

In this configuration, travel time data is obtained once the vehicle has left the highway. All direct travel time measurements and also some indirect estimation algorithms (e.g. some application of the N-curves method), suffer from this drawback. This type of travel time measurement, which will be named measured travel time (MTT), represents a measurement of a past situation and involves a great delay in information in the case of long itineraries or congested situations. Another limitation of this data is that travel times are only valid for a particular origin – destination itinerary such that partial on route measurements cannot be obtained.

In order to reduce delay in travel time information (and still directly measure), it is necessary to estimate MTTs for itineraries as short as possible. In a closed toll highway context, this leads to measuring single section travel times (i.e. between consecutive junctions). By doing so, information delay is reduced to a single section travel time. Single section measurements also overcome the limitation of itinerary specific travel times, since travel time estimation for long trips (i.e., more than one single section) can be obtained by adding the different single section travel times that configure the route. This procedure provides valid information for all drivers who pass through the highway section (regardless of whether they have the same origin – destination itinerary or not), and could also enable incident detection applications by tracking down the conflictive highway sections.

The itinerary travel time resulting from the addition of the travel times spent in the single sections that form the route at the same instant will be named instantaneous trip travel time (ITT) and assumes that traffic conditions will remain constant in each section until the next travel time update. This estimation is a better approach to the predicted travel time (PTT), which represents an estimation of the expected travel time for a driver entering the highway at the present instant, than MTT. Note that the ITT is a virtual measure in the sense that in fact no driver has followed the trajectory from which this travel time comes.

By way of illustration, Figure 1 shows an example of the implications of different trip travel time constructions. The information delay in the case of trip MTT involves very negative effects in case of dramatic changes in traffic conditions during this time lag (e.g. an incident occurs). The construction of trip ITT by means of single section travel times reduces this information delay and the resulting travel time inaccuracies. As the traffic conditions do not remain constant until the next single section travel time update, the trip ITT also differs from the true PTT. Note that the intention in Figure 1 is to show the maximum differences that could arise between MTT, ITT and PTT. This happens in case of rapidly evolving traffic conditions, for instance when an incident happens. Due to the deliberate construction of Figure 1, MTT misses the onset of congestion, while the ITT is able to detect it. This is the case where benefits of using ITT as opposed to MTT would be maximized.

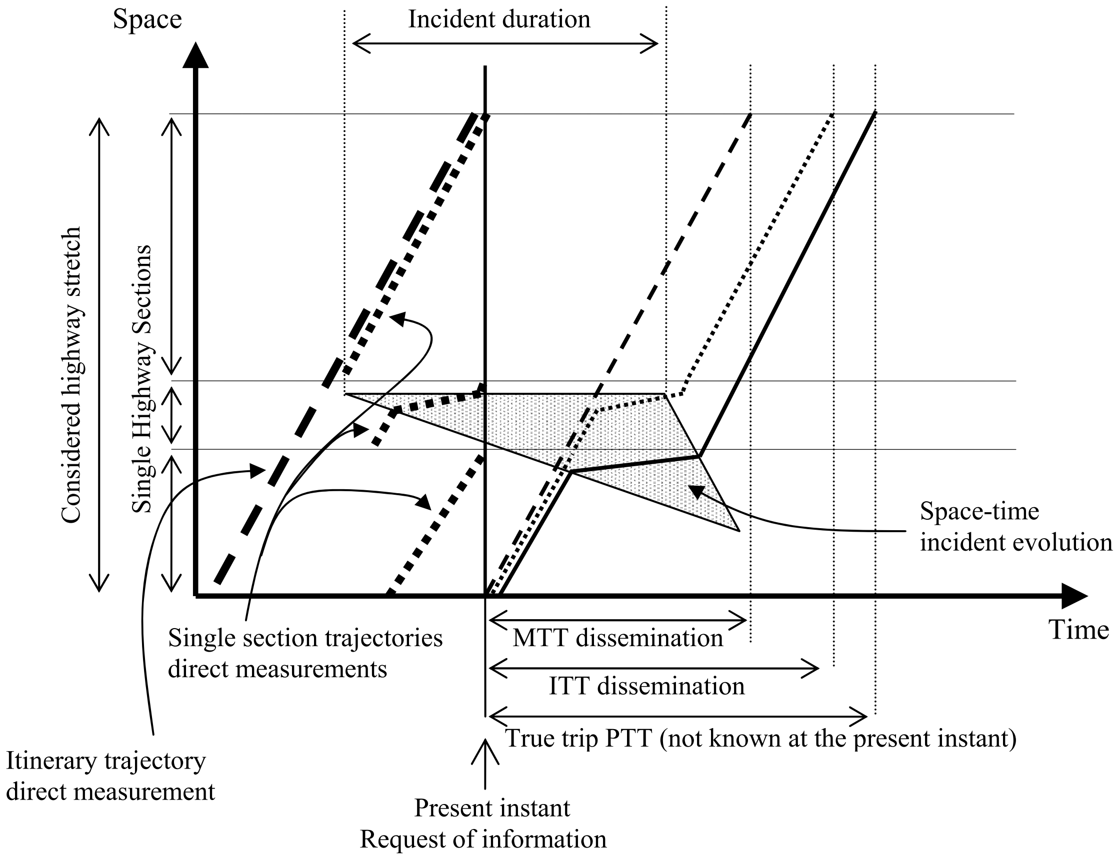


FIGURE 1 Travel time definitions and their possible implications in the dissemination of information.

In this context, the main goal of the algorithm proposed in the present paper is to obtain the required single section travel times from the available closed toll system data with no additional surveillance infrastructure. Obviously, a naïve method could be to only consider measurements between consecutive entry and exit ramps. This solution may reduce excessively the amount of available data in certain sections of the network, where the volume of traffic entering and leaving the highway at consecutive junctions is low, but there is a large volume of through traffic. Even in the case of interurban highways (which is the common case where highways are equipped with toll booths at each entrance and exit – closed highway systems) where consecutive junctions are many kilometers away and a significant number of drivers traveling a single section could be achieved, this naïve method does not account for the "exit time" (i.e. the time required to leave the highway) and the "entrance time" (i.e. the time required to enter the highway). It must be taken into account that the measurement points are located at the very end of the on/off ramps, sometimes a couple of kilometers away from the main highway trunk (see Figure 2). The exit time includes the time required to travel along the exit ramp (deceleration and overcoming the distance along the ramp until reaching the toll booth) plus the time required to pay the fee (perhaps with a small queue). In this situation, if the time to travel along a particular route, composed of several single sections, is calculated by simply adding the single itinerary travel times, the resulting travel time would be largely overestimated, because it would include as many exit and entrance times as there are single sections comprising the itinerary.

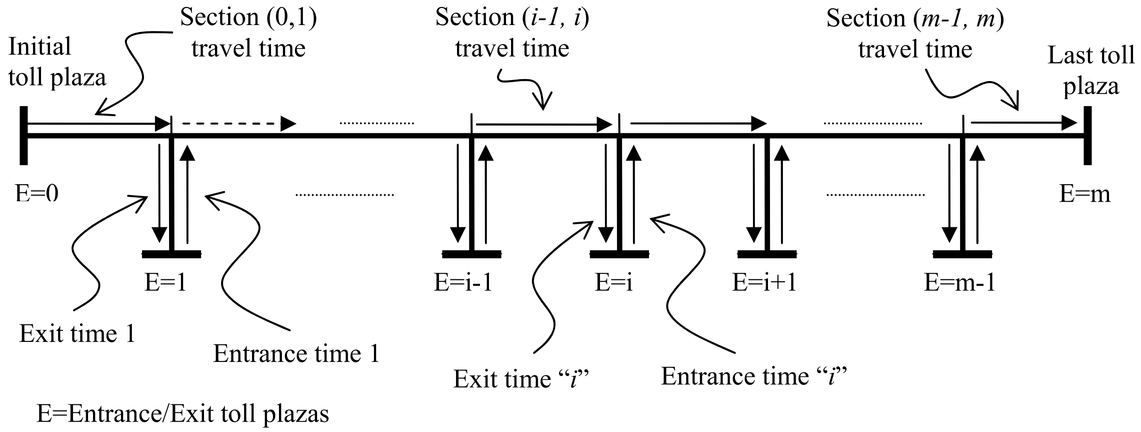


FIGURE 2 Closed highway network travel time elements.

Another solution to estimate single section travel times that would overcome the previous problems, is to install roadside beacons on the main highway trunk to detect vehicles equipped with an ETC system tag, and convert the traditional closed system into an ETC based travel time measurement system. However, and in addition to the high implementation costs (up to $100 000 per lane and measuring point if the overhead gantry is not available), the market penetration of the toll tags become a problem, as referenced in the introduction.

The algorithm presented here estimates the single section travel times without reducing excessively the amount of available data, and makes it possible to split this time into the main highway trunk travel time and the exit time, without any additional surveillance equipment. The exit time is a very useful collateral result for highway operators, as it is an indicator of the toll plazas' level of service.

3. ESTIMATION OF SINGLE SECTION TRAVEL TIMES: THE SIMPLE ALGORITHM'S UNDERLYING CONCEPT

For each particular vehicle "*k*" running along a highway with a closed tolling system, the travel time spent on its itinerary between origin "*i*" and destination "*j*", expressed as "$t_{i,j,k}$", can be obtained by matching the entry and exit information recorded on its toll ticket. As the toll ticket also includes the type of vehicle, only the observations corresponding to cars (including also motorbikes) are considered. Trucks are not considered, due to their lower speeds, which would bias the travel time estimations in free flow conditions. Of course, truck observations could be considered, as a family apart, if one is interested in obtaining specific free flow travel time for trucks. Note that the objective is to provide accurate travel time information for the driver of a car whose travel speed is considered as safe and comfortable under the existing traffic conditions by the average driving behavior.

The travel time information updating interval is defined as "Δt". Then, "$t_{i,j}^{(p)}$" refers to a representative average of the "$t_{i,j,k}^{(p)}$" data obtained in the "*p*" time interval (i.e. all the "*k*" vehicles that have exited the highway at "*j*", coming from "*i*" between the time instants "p-Δt" and "*p*"). To obtain this representative average "$t_{i,j}^{(p)}$" is not an easy task; in fact this is the key thing in the only ETC based travel time estimation systems. Problems arise from the nature of these data, with high variability in the number of observations (depending on the selected itinerary and time period) and different types of outliers to be removed to avoid producing erroneous travel time estimates. The range of solutions is huge, from the simple arithmetic mean, to the complex data-filtering algorithms developed by Dion and Rakha (2006), where a good overview of these filtering methods and current applications is also presented. The data-filtering algorithm used in the present approach is presented in Appendix A.

Once representative averages of travel times in all itineraries for the time interval just elapsed are obtained, the next step is to calculate the single section travel time and the exit time.

3.1. Basic Algorithm

Consider the highway stretch between entrance 0 and exit 1. The average travel time in this single itinerary, "$t_{0,1}$"[1] can be divided into two parts: the single section travel time "$t_{s(0,1)}$" and the exit time "$t_{ex(1)}$" (see Figure 2).

$$t_{0,1} = t_{s(0,1)} + t_{ex(1)} \tag{1}$$

[1] Note that to simplify the notation and clarify the concepts, the superscript "*(p)*" is omitted in sections 3.1 and 3.2. The original notation will be recovered in section 4.1, where "*(p)*" plays an important role.

By subtracting different travel times of selected itineraries, the single section travel times and the exit times can be obtained by canceling out the exit times[2] (see Figure 3). Then for the *(0,1)* itinerary:

$$t_{s(0,1)} \approx t_{0,2} - (t_{1,2} - t_{en(1)}) \approx t_{0,3} - (t_{1,3} - t_{en(1)}) \approx ... \tag{2}$$

$$t_{ex(1)} = t_{0,1} - t_{s(0,1)} \tag{3}$$

Where, "$t_{en(1)}$" is the entrance time at on-ramp 1 (i.e. the time required to travel along the entrance link)

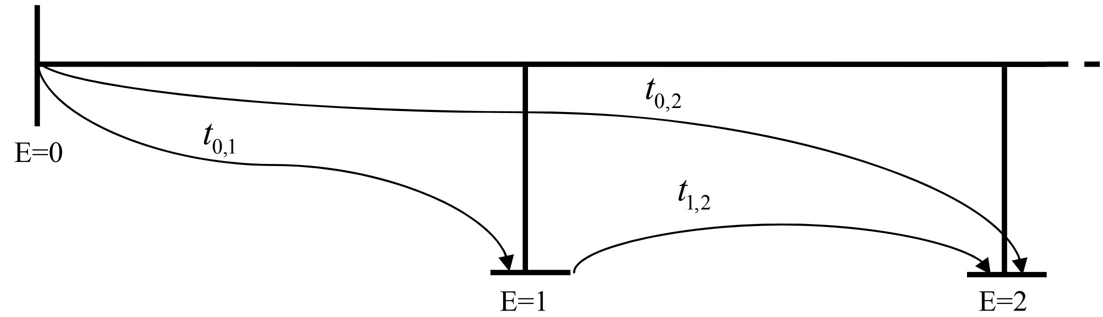


FIGURE 3 Section (0,1) travel time estimation.

In a general expression for itineraries with an entrance different from the initial toll plaza, located on the main trunk of the highway, Equation 1 should be rewritten as:

$$t_{i,i+1} = t_{en(i)} + t_{s(i,i+1)} + t_{ex(i+1)} \qquad \forall i = 1,...,m-1 \tag{4}$$

Where, "*m*" is the last toll plaza on the highway, located on the main highway trunk.

The entrance time "$t_{en(i)}$" can be estimated as a constant parameter for each entrance "*i*". It is calculated by considering a constant acceleration from the toll booth (where the vehicle is stopped) until the end of the entrance ramp (where it can be considered the vehicle is traveling at 90 km/h). The length of the entrance ramps are designed so that this speed could be achieved with typical vehicle acceleration rates. In addition, a reaction time of 5 seconds at the toll booth is added. Note the implicit assumption related to this constant entrance time: free access to the highway. There isn't any type of ramp metering scheme, neither any queue to join the main trunk traffic. This assumption generally holds in closed toll highways, because the ticketing at the toll booth smooth the entrance rate to the ramp, which, in general, is lower than the merging rate when entering the main highway traffic stream. However, in high demanded on-ramps and heavy congested highway main trunk, on-ramp queues should not be discarded, even though some of the queue is shifted upstream of the toll booth. Therefore constant entrance times should be considered as a simplification of the method that must be taken into account before each particular application.

[2]Equations 1, 3, 4, 6, 7, 9, 11, 21 and 29 are not exactly true. The lack of accuracy can be seen as a notation simplification that helps to clarify the concept, and is further detailed in Appendix B.

The solution to this limitation is simple: to measure or to estimate entrance times, between the toll booth and the effective incorporation to the mainline. In order to measure it some type of additional surveillance equipment at the incorporation is needed (e.g. a vehicle identification control point – ETC antenna, automatic license plate reading – or a loop detector). Then entrance times could be directly measured or indirectly estimated using one of the multiple methods described in the introduction of the paper (e.g. from difference between cumulative counts). If this additional surveillance equipment is not available, and as the objective is to provide travel time estimations without investing in any kind of infrastructure or technology (as long as the instant that each vehicle passes through the toll plaza is registered), an alternative solution could be proposed if entrance time variations are considered to be critical. This consists in the estimation of a maximum merging rate and to compare it with the entering flow at the on-ramp toll booth. Using deterministic queuing diagrams, the delay at the on-ramp could be easily obtained. The main problem here could be the estimation of the maximum merging rate, which depends on the traffic conditions on the main highway trunk. However, it has been experimentally found (Cassidy and Ahn, 2005) that the maximum rate at which vehicles can enter a congested freeway from an on-ramp is a fixed proportion of the downstream freeway flow.

Following the same logic as in Equation 2, the general expression to calculate the single section travel time and the exit time can be formulated as:

$$t_{s(i,i+1)} \approx (t_{i,i+2} - t_{en(i)}) - (t_{i+1,i+2} - t_{en(i+1)}) \approx (t_{i,i+3} - t_{en(i)}) - (t_{i+1,i+3} - t_{en(i+1)}) \approx \ldots \quad (5)$$

$$t_{ex(i+1)} = t_{i,i+1} - t_{s(i,i+1)} - t_{en(i)} \quad (6)$$

Equations 5 and 6 involve estimating single section travel times using different origin - destination itineraries. Note that in some traffic situations the use of different lanes is not independent of the destination. Therefore, if there is a significant speed gradient between these lanes, this implies that travel times in the highway section are not identical independent of the vehicle destination. This could lead to a significant error.

However these situations only significantly arise near off-ramps, for instance when the demand for exiting the tollway exceeds the capacity of the off-ramp. This results in a spill-back of the queue into the main highway trunk, congesting the rightmost lanes and reducing the capacity of the rest of the lanes due to the “friction” between congested and not congested lanes, and due to last minute lane changes. The situation would result in the rightmost lanes congested, and only composed of vehicles wishing to exit in the next off-ramp, while on the other lanes traffic stream would be uninterrupted, but not at free flow speed, and composed by drivers heading to all other destinations. Therefore this section travel times depend on the destination, but only on two groups of destinations, the next off-ramp and all the others. Empirical evidence shows (Muñoz and Daganzo, 2002) that after some few kilometers, even a wide multilane highway becomes FIFO. Therefore the lane effect only arises for a limited length. In case of interurban highways, which may be narrower and the sections between junctions longer, it is even more appropriate to assume that multi-pipe traffic states (i.e. non-FIFO congested regimes) will be confined within the single section defined by the off-ramp.

Then note that this does not imply any drawback to the proposed algorithm since the implicit assumption is that all vehicles traversing a freeway section heading to all destinations except the next one have similar travel times in the section. This results from the fact that in order to calculate a single section travel time, we operate with itineraries heading to the same destination (to cancel out exit times); therefore, this last section destination specific travel time is also cancelled out.

The only implication of this destination specific section travel time, being this section the last section of the vehicle itinerary, appears in the calculation of the exit time, when last sections of the itinerary are considered. The traffic situation exemplified above would result in an increased exit time for the off-ramp. This only implies that the exit time does not only take place in the off-ramp, but also queuing in the rightmost lanes of the main highway trunk in the last section.

The traffic situation described could also go the other way around: congestion in the main trunk while less demand than capacity for the off ramp. This would not result in significant non FIFO queues due to speed gradient across lanes in the section, since in common freeway configurations with a constant number of lanes, the main trunk congestion would block the off-ramp until the vehicles reach the exit point. However there exist some specific freeway configurations where an auxiliary lane is available only for the next off-ramp, preventing some vehicles to queue if they do not want to cross the bottleneck, and therefore reducing the extent of the queue. This would result in an uninterrupted flow for vehicles heading to the next exit, while congested in the mainline for all other destinations. The proposed algorithm in this case computes the congested travel time for the section, and a low or eventually even negative exit time for the off-ramp. This exit time would correspond to the sum of two effects: the true exit time and a (negative) time representing the time savings in the last freeway section in comparison to those drivers in the section not heading to the off-ramp. In conclusion, if one wants to disseminate the travel time to that specific off-ramp, by adding the congested single section travel time to the (negative) exit time, the correct itinerary travel time would be obtained.

An additional remark is that the proposed algorithm cannot split the travel time for the last section "$t_{m-1,m}$" into the main highway trunk travel time and the exit. Nevertheless, this lack of information is not so important, because the last toll plaza is usually located in the main highway trunk and all the vehicles traveling along the last section must go through this toll plaza. In such a way, the interesting information for the driver in this last stretch of the highway is the total aggregated travel time, including both the main trunk travel time and the exit time. In contrast, the exit time in the last toll plaza would be useful information for the highway operator to know the level of service of this last toll plaza.

From the above equations, it can be seen that there are "*m-(i+1)*" equivalent estimations for the "*i,i+1*" single section travel time. Each of these estimations results from different itinerary travel times, with its different associated lengths. Note that considering long trips as a possible alternative in the calculation of the single section travel times in Equation 5, implies an increase in the information delay (because the application of Equation 5 requires that all the considered vehicles had left the highway). In addition, travel time for long trips can be considered as less reliable as it can be increased by factors that are unrelated to traffic conditions, for example if some drivers stop for a break or re-fueling. This fact implies an increase in the standard deviation of the average itinerary travel time, due to the higher probability of stops on a long trip.

These considerations suggest that for some applications (e.g. real time application) the basic algorithm should be restricted to short trip data. This restriction is detailed in Section 4.1, corresponding to real time implementation.

Once all the valid alternatives for estimating the single section travel time are selected (this is different if working off-line with a complete database of events versus working in real time), some smoothing or averaging algorithm must be applied to calculate a unique value for the single section travel time. This smoothing algorithm is presented in Appendix C.

3.2. Extended Algorithm

Equations 5 and 6 show that a particular single section travel time is calculated from travel time observations of all the vehicles entering in the origin of the section, except those traveling only in the considered stretch, which are considered in calculating the exit time of the section. The basic algorithm does not consider the vehicles traveling along the stretch but that have entered at a previous entrance.

On certain stretches of the highway, particularly during night hours, the amount of available data may be insufficient to perform the described calculations because of the low flow in a particular itinerary of those considered. Although travel time information under low traffic conditions arouses lesser interest to drivers and highway administrations due to its easier predictability with historic information, in such cases it is possible to increase the amount of available data by considering alternative itineraries for the calculations. For example, a second order algorithm implies the estimation of a two-section (i.e. two consecutive single sections) travel time. If the second order algorithm between entrance 1 and exit 3 is considered, then:

$$t_{1,3} = t_{en(1)} + t_{s(1,3)} + t_{ex(3)} \tag{7}$$

Proceeding in the same way as in Equations 5 and 6, the section travel times and the exit times can be obtained for the (1,*3*) itinerary as (see Figure 4):

$$t_{s(1,3)} \approx (t_{1,4} - t_{en(1)}) - (t_{3,4} - t_{en(3)}) \approx (t_{1,5} - t_{en(1)}) - (t_{3,5} - t_{en(3)}) \approx ... \tag{8}$$

$$t_{ex(3)} = t_{1,3} - t_{s(1,3)} - t_{en(1)} \tag{9}$$

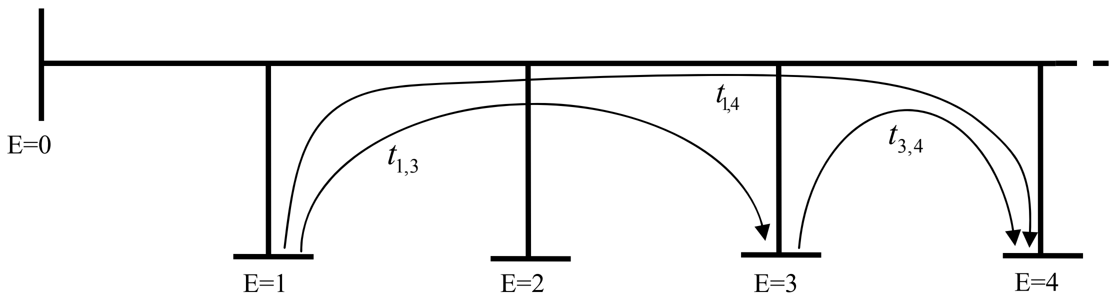


FIGURE 4 Section (1,3) travel time estimation with second order algorithm.

The general expression for a 2nd order algorithm can be written as:

$$t_{s(i,i+2)} \approx (t_{i,i+3} - t_{en(i)}) - (t_{i+2,i+3} - t_{en(i+2)}) \approx (t_{i,i+4} - t_{en(i)}) - (t_{i+2,i+4} - t_{en(i+2)}) \approx ... \qquad (10)$$

$$t_{ex2(i+2)} = t_{i,i+2} - t_{s(i,i+2)} - t_{en(i)} \qquad (11)$$

Where the subscript "*2*" in the exit time notation (Equation 11) and in the single section travel time notation (Equation 12), stands for the estimation using a 2nd order algorithm.

The two-section travel time can be seen as the addition of two consecutive single section travel times

$$t_{s(i,i+2)} = t_{s2(i,i+1)} + t_{s2(i+1,i+2)} \qquad (12)$$

To calculate these two addends:

$$t_{s2(i,i+1)} = t_{s(i,i+2)} - t_{s(i+1,i+2)} \qquad (13)$$

Where it is assumed that there is enough data to obtain an accurate estimation of the first order single section travel time corresponding to "$t_{s(i+1,i+2)}$". Finally, replacing the result of Equation 13 in Equation 12:

$$t_{s2(i+1,i+2)} = t_{s(i,i+2)} - t_{s2(i,i+1)} \qquad (14)$$

Equations 13 and 14 could be easily modified in the case that "$t_{s(i,i+1)}$" was the known addend of Equation 12. If both, "$t_{s(i,i+1)}$" and "$t_{s(i+1,i+2)}$" are known, any combination of Equations 13 and 14 or its modifications could be used (see Appendix C for details).

There is one remaining situation, when neither"$t_{s(i,i+1)}$" nor "$t_{s(i+1,i+2)}$" can be obtained from a 1st order algorithm. Then we simply assume the same average speed across both sections, and reach:

$$t_{s2(i,i+1)} = \frac{t_{s(i,i+2)} \cdot l_{s(i,i+1)}}{l_{s(i,i+2)}} \qquad (15)$$

$$t_{s2(i+1,i+2)} = \frac{t_{s(i,i+2)} \cdot l_{s(i+1,i+2)}}{l_{s(i,i+2)}} \qquad (16)$$

Where "$l_{s(i,j)}$" is the length of the highway stretch between junctions "*i*" and "*j*".

This strong assumption implies that the use of Equations 15 and 16 should be restricted to very particular situations where no other information is available.

There still remain some questions to be answered, for example when a first order single section travel time is considered accurate enough, or how to fuse single section travel times (one coming from a first order algorithm and the other from a second order). These aspects are further analyzed in Appendix C.

As a summary, the step by step implementation of the running phase of the algorithm (all the default values are already set) results as follows:

- Step 1: Compute all the itinerary travel times in the time interval "*p*", "$t_{i,j}^{(p)}$" using the filtering algorithm presented in Appendix A.
- Step 2: Estimate all the first order single section travel times using all the possible alternatives in Equation 5 and fuse them using the equations presented in Appendix C to find a first order estimation for the single section travel times of all single sections, "$t_{s(i,i+1)}^{(p)}$".
- Step 3: Compute the first order estimation of the exit time "$t_{ex(i+1)}^{(p)}$", using Equation 6.
- Step 4: Estimate all the second order single section travel times using all the possible alternatives in Equation 10 and fuse them using the equations presented in Appendix C, to find a second order estimation for the single section travel times of all single sections, "$t_{s2(i,i+1)}^{(p)}$".
- Step 5: Compute the second order estimation of the exit time "$t_{ex2(i+1)}^{(p)}$", using Equation 11.
- Step 6: Fuse the first and second order estimations of the single section travel times and of the exit times using the equations provided in Appendix C, to obtain the final travel time estimations.

4. MODIFICATIONS FOR THE REAL TIME AND OFF-LINE IMPLEMENTATIONS OF THE ALGORITHM

As stated in the first paragraph of Section 3, any single section travel time "$t_{s(i,i+1)}$" is related to a particular time period "*p*" (with a duration of "*Δt*"). Obviously, for the algorithm to be accurate, the average travel times (in particular the ones used in Equations 5, 6, 10 and 11) must result from observations of vehicles traveling along the same stretch of the highway in the same time period. This has some important implications on the time periods to consider in the different alternatives for estimating a single section travel time (of different itinerary lengths), and are different if working in a real time basis or in an off-line configuration where some future information is available.

4.1. Real Time implementation

For a real time implementation, Equation 5 should be rewritten as:

$$\begin{aligned} & t_{s(i,i+1)}^{(p)} \approx (t_{i,i+2}^{(p)} - t_{en(i)}) - (t_{i+1,i+2}^{(p)} - t_{en(i+1)}) \approx (t_{i,i+3}^{(p)} - t_{en(i)}) - (t_{i+1,i+3}^{(p)} - t_{en(i+1)}) \approx \dots \\ & \dots \approx (t_{i,i+n*}^{(p)} - t_{en(i)}) - (t_{i+1,i+n*}^{(p)} - t_{en(i+1)}) \end{aligned} \qquad (17)$$

In a real time application it is needed to obtain "*i+n**", the critical highway exit "*i+n*" ($n \geq 2$) which represents the maximum length of an itinerary whose vehicles leaving the highway at time interval "*p*", and having traveled along the itinerary "*i,* *i+n**", have traveled simultaneously with other vehicles traveling from "*i*" to "*i+2*" and leaving the highway at the same time interval "*p*". The simultaneous travel happens along the sections "*i, i+1*" and "*i+1, i+2*". The trajectories diagram sketched in Figure 5 may help to understand this concept.

In Figure 5, "*δt*" represents the required minimum time interval overlapping of trajectories to ensure simultaneous traveling, which is set at ⅓ of "*Δt*". Note that some additional trajectories in Figure 5 could reach exit "*i+n+1*" before "*p*" and still coincide with trajectories traveling along the itinerary (*i, i+2*). However, this coincide zone will elapse for a time period shorter than the minimum required, "*δt*". Then, the conditions to ensure enough simultaneous traveling can be derived as:

$$\begin{aligned} p - t_{i,i+n}^{(p)} \geq p - t_{i,i+2}^{(p)} - \Delta t + \delta t \rightarrow\ t_{i,i+n}^{(p)} \leq t_{i,i+2}^{(p)} + \Delta t - \delta t \\ p - t_{i,i+n}^{(p)} - \Delta t + \delta t \leq p - t_{i,i+2}^{(p)} \rightarrow\ t_{i,i+n}^{(p)} \geq t_{i,i+2}^{(p)} - \Delta t + \delta t \end{aligned} \quad (18)$$

"*n**" must be calculated as the maximum "*n*" for which Equation 18 holds. Note that in general $t_{i,i+2}^{(p)} \leq t_{i,i+n}^{(p)}$ (for $n \geq 2$; unless huge problems take place at exit ramp "*i+2*") and $\Delta t > \delta t$. This results in the first inequality of Equation 18 being usually the restrictive one.

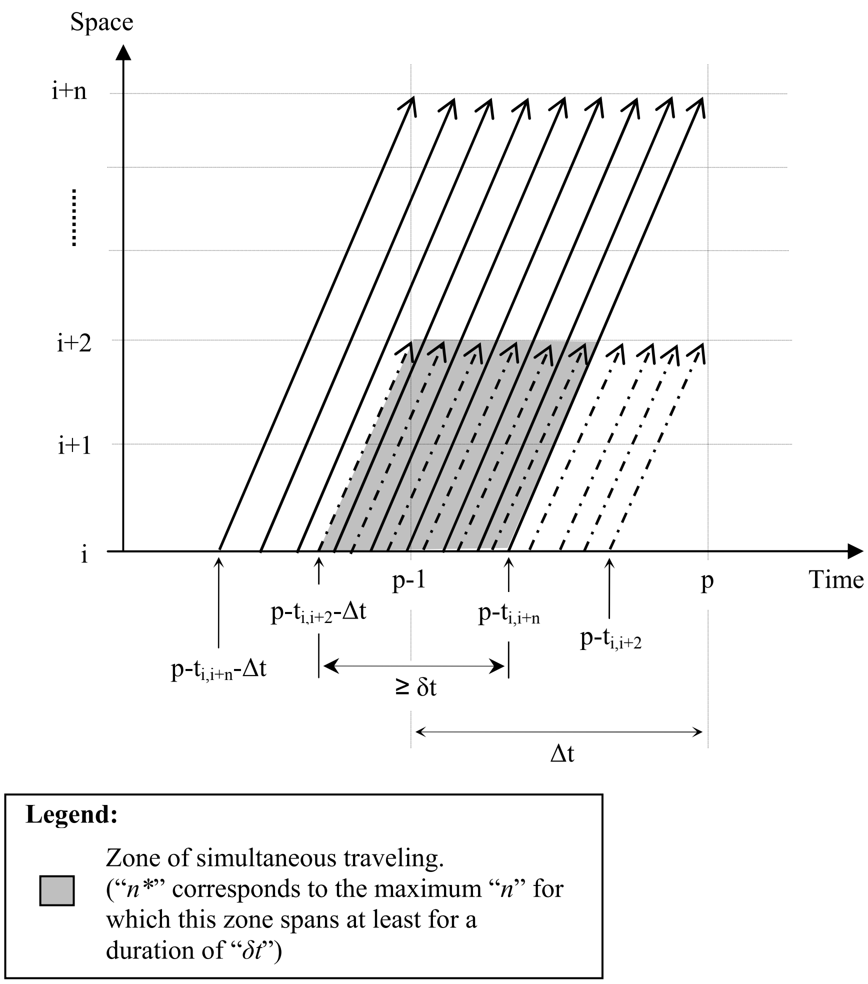


FIGURE 5 Vehicle trajectories considered in Equation 17.

Applying a similar construction to Equation 10 (2nd order equation), we would obtain the following conditions:

$$\begin{aligned} t_{i,i+n}^{(p)} &\le t_{i,i+3}^{(p)} + \Delta t - \delta t \\ t_{i,i+n}^{(p)} &\ge t_{i,i+3}^{(p)} - \Delta t + \delta t \end{aligned} \qquad (n \ge 3) \qquad \textit{(19)}$$

As in general $t_{i,i+3}^{(p)} \ge t_{i,i+2}^{(p)}$ (again unless huge problems take place at exit ramp "*i+2*"), Equation 19 holds for "n^*" if Equation 18 does. However, Equation 18 is defined for "$n \ge 3$", which means that the 2nd order travel times can only be considered simultaneously with 1st order travel times in a real time implementation if "$n^* \ge 3$".

Perhaps an excess of simplification, but helpful in visualizing the concept, "$t_{i,i+n}^{(p)}$" could be seen as "$t_{i,i+2}^{(p)} + t_{i+2,i+n}^{(p)}$" (the excess of simplification results from the rejection of "$t_{ex(i+2)}^{(p)}$" and "$t_{en(i+2)}$"). Substituting this simplification into Equation 18:

$$\begin{aligned} t_{i+2,i+n}^{(p)} &\le \Delta t - \delta t \\ t_{i+2,i+n}^{(p)} &\ge -\Delta t + \delta t \end{aligned} \qquad \textit{(20)}$$

Equation 20 represents the simplified condition to obtain "n^*", the largest value that can take "*n*" while not violating the restrictions. Note that the second inequality of this simplified condition is irrelevant as it always holds. "n^*" sets the possible alternatives in calculating the single section travel times from Equation 17, on a real time basis. These restrictions also have a direct implication in setting "Δt", the larger "Δt", the large "n^*" could be. This implies more alternatives in estimating the single section travel time, in addition to more observations within each itinerary. In contrast, large "Δt" implies a low updating frequency, which results in an increase of information delay, with disastrous consequences when facing rapidly evolving traffic conditions.

Another implication of these restrictions is the fact that "n^*" decreases when the travel time in the highway increases (i.e. in congested situations), in particular in highway sections from (*i+2, i+3*) and downstream. If the congested section is the section (*i, i+1*) or (*i+1, i+2*), there is no implication related to "n^*", but this results in an increase of the delay of the information (remember that toll tickets represent a MTT measure).

Finally, it must be considered that the exit time "$t_{ex(i+1)}$"expressed in Equation 6, is also related to a particular time interval "*p*". Then Equation 6 should be rewritten as:

$$t_{ex(i+1)}^{(p)} = t_{i,i+1}^{(p)} - t_{s(i,i+1)}^{(p)} - t_{en(i)} \qquad \textit{(21)}$$

For Equation 21 to be consistent, the vehicles considered in the calculation of "$t_{s(i,i+1)}^{(p)}$" must have traveled along the section "*i, i+1*" together with those whose average travel time is "$t_{i,i+1}^{(p)}$". Considering the trajectories involved in these calculations, the resulting condition is:

$$\begin{aligned} p - t_{i,i+2}^{(p)} \ge p - t_{i,i+1}^{(p)} - \Delta t + \delta t &\rightarrow t_{i,i+2}^{(p)} \le t_{i,i+1}^{(p)} + \Delta t - \delta t \\ p - t_{i,i+2}^{(p)} - \Delta t + \delta t \le p - t_{i,i+1}^{(p)} &\rightarrow t_{i,i+2}^{(p)} \ge t_{i,i+1}^{(p)} - \Delta t + \delta t \end{aligned} \qquad \textit{(22)}$$

Again, in an excess of simplification, "$t_{i,i+2}^{(p)}$" could be seen as "$t_{i,i+1}^{(p)} + t_{i+1,i+2}^{(p)}$". This can be considered approximately true if there's no problem (i.e. congestion) in the entrance or exit ramps at junction "*i+1*". Substituting this simplification into the first inequality of Equation 22:

$$t_{i+1,i+2}^{(p)} \leq \Delta t - \delta t \tag{23}$$

From this restriction it results that, on a real time information basis, the exit time at ramp "*i+1*" can only be obtained if Equation 23 holds. Otherwise, the exit time can only be obtained for a past time interval, using the off-line formulation that follows.

4.2. Off-line Implementation

In case the objective is not real time information, and the aim is to reconstruct the single section travel times of past situations (e.g. to obtain travel time templates), then the whole database is available. This means that "future" information (in relation to the time interval of calculation "*p*") can be used (i.e. there's no need to "wait" until the vehicle has left the highway to obtain its MTT). In this situation, real time restrictions could be modified to obtain a more robust algorithm. Equation 17 could be rewritten as (see Figure 6):

$$\begin{aligned} &t_{s(i,i+1)}^{(p)} \approx (t_{i,i+2}^{(p)} - t_{en(i)}) - (t_{i+1,i+2}^{(p)} - t_{en(i+1)}) \approx \ldots \approx (t_{i,i+n^*}^{(p)} - t_{en(i)}) - (t_{i+1,i+n^*}^{(p)} - t_{en(i+1)}) \approx \\ &(t_{i,i+n^*+1}^{(p+\Delta t)} - t_{en(i)}) - (t_{i+1,i+n^*+1}^{(p+\Delta t)} - t_{en(i+1)}) \approx \ldots \approx (t_{i,i+n^{(2)*}}^{(p+\Delta t)} - t_{en(i)}) - (t_{i+1,i+n^{(2)*}}^{(p+\Delta t)} - t_{en(i+1)}) \approx \ldots \end{aligned} \tag{24}$$

Using a variation of Equation 18, "$n^{(2)}*$" can be obtained as the maximum "$n^{(2)}$" that fulfills the following restriction:

$$\begin{aligned} t_{i,i+n^{(2)}}^{(p+\Delta t)} &\leq t_{i,i+2}^{(p)} + 2\cdot\Delta t - \delta t \\ t_{i,i+n^{(2)}}^{(p+\Delta t)} &\geq t_{i,i+2}^{(p)} - 2\cdot\Delta t + \delta t \end{aligned} \tag{25}$$

Equation 19 could be modified in the same way to obtain the condition to apply to the 2nd order algorithm.

Then, in a general expression, the single section travel time and the exit time can be calculated off-line as:

$$t_{s(i,i+1)}^{(p)} \approx (t_{i,i+k}^{(q)} - t_{en(i)}) - (t_{i+1,i+k}^{(q)} - t_{en(i+1)}) \qquad k = 2, \ldots, m \tag{26}$$

Where,

$$q = \begin{cases} p & if \;\; k \leq n^* \\ p + \Delta t & if \;\; n^* < k \leq n^{(2)}* \\ \vdots & \qquad \vdots \\ p + (r-1)\cdot\Delta t & if \;\; n^{(r-1)}* < k \leq n^{(r)}* \end{cases} \qquad r \in \mathbb{N} \tag{27}$$

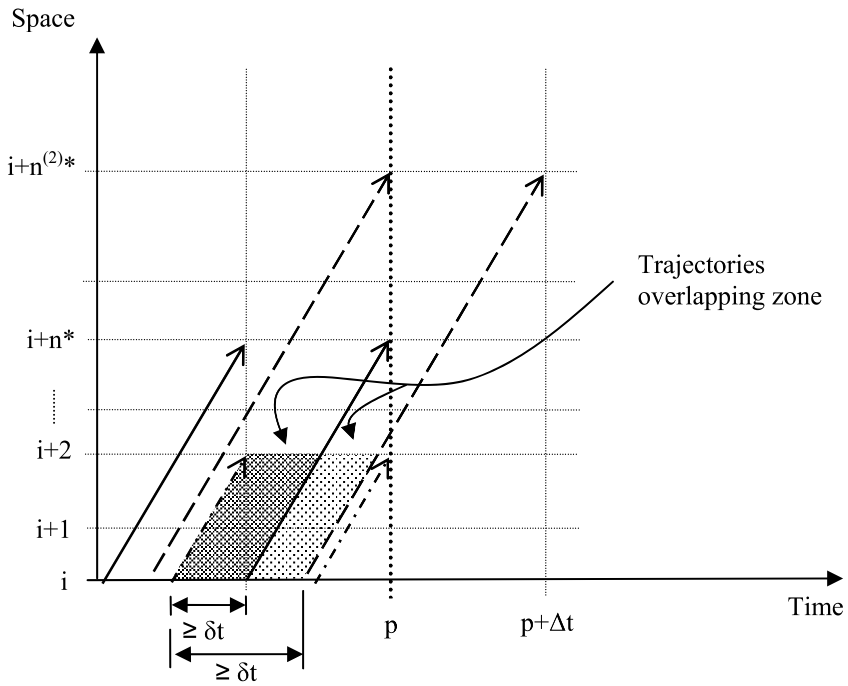


FIGURE 6 Vehicle trajectories considered in Equation 24.

For each "*k*" it is necessary to find the minimum "$r \in \mathbb{N}$" which satisfies Equation 27, where "$n^{(r)}*$" for a given "*r*", is the maximum "$n^{(r)}$" that satisfies the following restrictions:

$$\begin{aligned} t_{i,i+n^{(r)}}^{(p+(r-1)\cdot\Delta t)} &\leq t_{i,i+2}^{(p)} + (r-1)\cdot\Delta t - \delta t \\ t_{i,i+n^{(r)}}^{(p+(r-1)\cdot\Delta t)} &\geq t_{i,i+2}^{(p)} - (r-1)\cdot\Delta t + \delta t \end{aligned} \tag{28}$$

Finally, for the off-line calculation of the exit time "$t_{ex(i+1)}^{(p)}$", Equation 21 should be rewritten as:

$$t_{ex(i+1)}^{(p)} = t_{i,i+1}^{(q)} - t_{s(i,i+1)}^{(q)} - t_{en(i)} \tag{29}$$

Where "*q*" is obtained from Equation 27 with a value of "*k=2*".

5. APPLICATION TO THE AP-7 HIGHWAY IN SPAIN

This new approach for direct travel time measurement using existing toll infrastructure has been tested for the AP-7 toll highway in Spain. The AP-7 highway runs along the Mediterranean cost corridor, from the French border to the Gibraltar Strait. Nevertheless, the pilot test was restricted to the north east stretch of the highway from "La Roca del Vallès" toll plaza, near Barcelona, to the French border at "La Jonquera". This stretch is approximately 120 km long.

The first of the pilot tests was performed with the April 18th 2008 data. This was a very conflictive Friday in terms of traffic, as a fatal crash happened on the highway. The accident forced the closing of two of the three existing lanes in the southerly direction towards Barcelona, causing severe congestion and serious delays.

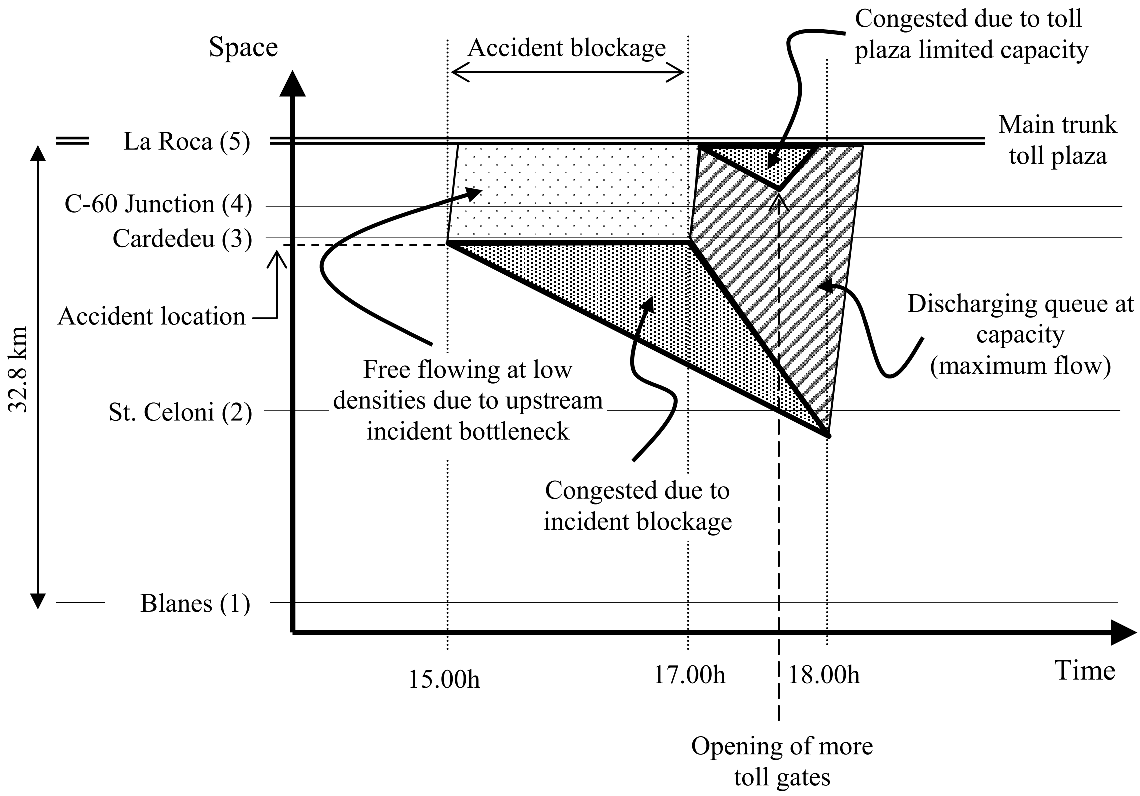
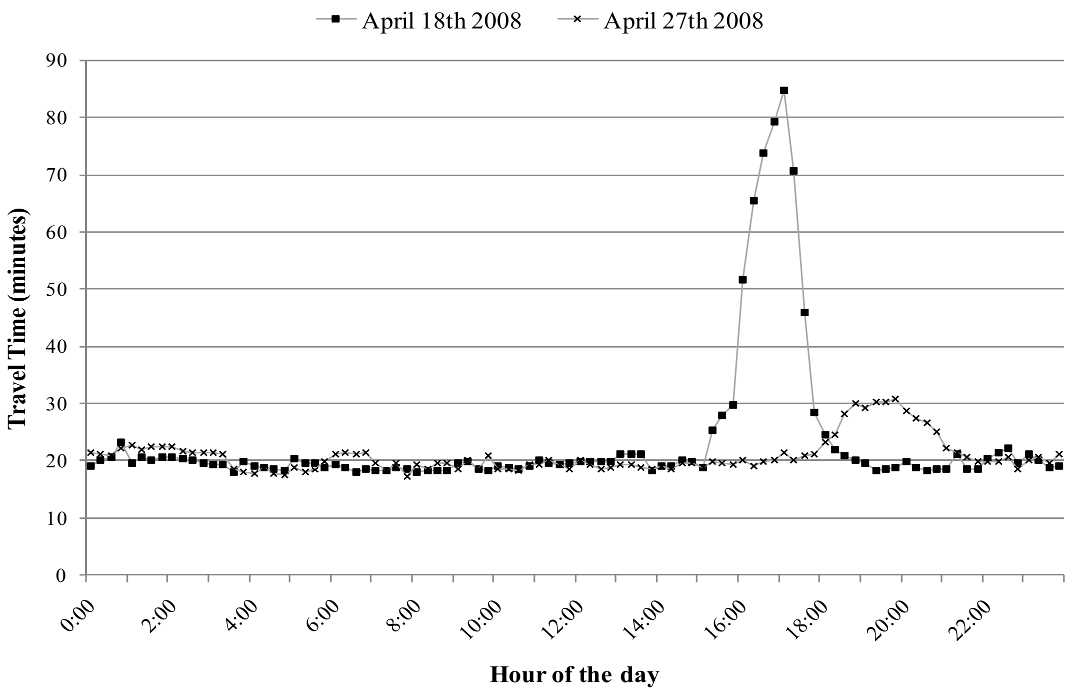


FIGURE 7 Space-time evolution of traffic states in the highway (April 18th 2008).



Note: These travel times correspond to "$t_{1,5}^{(p)}$", the representative average of the MTT for the itinerary "*1,5*" every 15 minutes directly obtained from toll ticket data.

FIGURE 8 Travel time from Blanes to La Roca.

Figure 7 shows the physical configuration of the test site and the spatial and temporal evolution of the traffic congestion on the highway (shock-wave analysis). From the diagram, it can be seen that the accident happened around 3.00 pm near the "Cardedeu" junction, causing a huge bottleneck at this location. The accident could not be cleared until 5.00 pm, causing long queues to grow (longer than 13 km). After the accident was cleared (and the corresponding bottleneck removed), the queue discharged at capacity, with a flow in excess of the maximum service rate of the available toll gates at the "La Roca" toll plaza, located on the main highway trunk. This caused a small queue to grow just upstream of the toll barrier, which started to dissipate when the capacity of the toll plaza was increased by the opening of more toll gates. The whole incident implied a maximum delay of 1 hour for the vehicles travelling from "Blanes" to "La Roca" (a distance of 32.8 km), as can be seen in Figure 8. This conflictive situation with rapidly evolving conditions represents a perfect environment for testing the proposed algorithm, as travel time information is crucial and the delay in reporting the information is disastrous. In addition, in order to fully test the algorithm for different conditions on the highway, in particular for more "normal" conditions, a second pilot test is presented with April 27th 2008 data, a usual Sunday in the highway with recurrent evening slight delays on the southbound direction (see Figure 8), resulting from the massive return to Barcelona after spending a day or weekend on the coast.

The present application of the algorithm will be performed in an on-line basis. This selection is due to the fact that the on-line application is more restrictive than the off-line, and the contribution of the method is more relevant, because in addition to provide detailed decomposition of itinerary travel time in single sections and exit times (like in the off-line application), it also provides an increased immediacy in reporting travel time information to the drivers, crucial in real-time information systems.

5.1. Selection of the sampling duration "*Δt*"

The selection of the sampling duration "*Δt*" is, in practice, very much relevant, as it determines how well the method performs with data typically available. Two goals must be pursued in the selection of "*Δt*". On the one hand "*Δt*" should be as short as possible in order to provide a frequent update and an accurate tracking of travel time evolution, avoiding averaging and smoothing the possible rapidly changing traffic conditions. On the other hand, "*Δt*" should be large enough to include enough travel time observations for the estimation to be statistically significant. From basic estimation theory it can be stated that, given a desired statistical significance or probability level of the estimation, "$\alpha \in (0,1)$", the resulting maximum absolute error "*ε*" in the estimation of an average travel time "$t_{i,j}^{(p)}$" from individual measurements, assumed independent, is related to the number of observations "$N_{i,j}^{(p)}$" by:

$$\varepsilon_{(t_{i,j}^{(p)})} = \frac{Z_{(1+\alpha)/2} \cdot \sigma_{(t_{i,j}^{(p)})}}{\sqrt{N_{i,j}^{(p)}}} \tag{30}$$

Where "$\sigma_{(t_{i,j}^{(p)})}$" stands for the standard deviation of travel time observations and where applying the standard normal cumulative distribution function to "$Z_{(1+\alpha)/2}$" we obtain the cumulative probability of "$(1+\alpha)/2$".

A minimum "*N*" must be achieved in the itinerary travel time estimation in order to delimit the maximum error, given a statistical significance. In general, for a given demand on the highway, the maximum error will be lower for longer "*Δt*"s, as more observations will be available. In addition, for a given "*Δt*", the error will be lower in those situations with high demand and low travel time variance.

The error formulated in Equation 30, is not directly the absolute error in the estimation of single section travel times from the proposed algorithm. Note that, in the on-line application of the algorithm, the single section travel time results from the average of "*n*-1*" subtractions between two different itinerary travel times (see Equation 17). In addition, if "*n*≥ 3*", "*n*-2*" additional estimations are available from a 2nd order algorithm (see Equation 19). Then it can be stated that the maximum absolute error in the estimation of a single section travel time, "$t_{s(i,j)}^{(p)}$", is:

$$\begin{aligned} \varepsilon_{(t_{s(i,i+1)}^{(p)})} &= \frac{\sqrt{2}\cdot\varepsilon_{(t_{i,j}^{(p)})}}{\sqrt{n^*-1}} \qquad if\ n^*=2 \\ \varepsilon_{(t_{s(i,i+1)}^{(p)})} &= \frac{\sqrt{2}\cdot\varepsilon_{(t_{i,j}^{(p)})}}{\sqrt{2n^*-3}} \qquad if\ n^*\geq 3 \end{aligned} \tag{31}$$

Longer "*Δt*"s imply an increase in the number of the itineraries considered in the calculation of "$t_{s(i,j)}^{(p)}$" in the case of an on-line application (i.e. an increase in "*n**", see Equation 18), and therefore a second source for the error reduction when increasing "*Δt*". For the off-line application the number of considered itineraries is maximum and independent of "*Δt*" (i.e. "*n*=m*"). Then, in general, for a same "*Δt*" the obtained maximum error of the single section travel time off-line estimation will be lower or equal than in the on-line case.

In the selection process for "*Δt*", four options are considered: 1, 3, 5 and 15 minutes. 1 minute is considered to be a lower bound as it does not seem necessary to increase further the updating frequency because traffic conditions do not evolve so quickly. In addition, and as it will be seen next, the proposed algorithm may not be suitable to work with this fine resolution. 15 minutes is considered as an upper bound because it is considered to be the minimum updating frequency acceptable in order to be used as a real time information system. Table 1 shows the expected maximum absolute error obtained for different "*Δt*"s and considering the typical demand patterns in the test site.

If travel time information is disseminated every two single sections (e.g. using variable message signs), from Table 1 it can be seen that the expected average cumulative error in two sections for an updating interval of 5 minutes is below the one minute error threshold for at least half of the calculation periods in almost all demand patterns. Although the selection of a 5 minutes "*Δt*" could be used for all situations it may also be useful the selection of a longer interval for free flowing periods, when travel time variations are unlikely. This may reduce the fluctuations in the estimations.

TABLE 1 Maximum Expected Estimation Errors for Various Demand Patterns and Different "Δt"

Demand Pattern(i)	Δt	Itinerary travel time standard deviation "σ" (minutes)		N		$n*$		Maximum expected single section travel time absolute error(ii) (minutes)	
		Median	25 – 75% Percentile	Median	25 – 75% Percentile	Median	25 – 75% Percentile	Median	25–75% Percentile
A	1 min.	1.11	0.63 – 1.70	2	1 – 3	2	2 – 2	1.22	0.52 – 2.40
	3 min.	1.29	0.92 – 1.85	3	1 – 6	2	2 – 2	1.00	0.51 – 2.62
	5 min.	1.37	0.98 – 1.87	4	1 – 9	4	3 – 4	0.45	0.21 – 1.33
	15 min.	1.55	1.55 – 1.87	22	11 – 33	4	4 – 4	0.21	0.14 – 0.36
B	1 min.	1.22	0.92 – 1.56	4	1 – 9	2	2 – 2	0.90	0.44 – 2.20
	3 min.	1.24	0.90 – 1.47	5	2 – 22	2	2 – 2	0.81	0.27 – 1.61
	5 min.	1.22	0.87 – 1.49	7	2 – 30	4	3 – 4	0.30	0.10 – 0.94
	15 min.	1.39	1.30 – 1.54	114	58 – 151	4	4 – 4	0.08	0.07 – 0.13
C	1 min.	0.99	0.73 – 1.26	7	2 – 12	2	2 – 2	0.53	0.29 – 1.38
	3 min.	0.99	0.73 – 1.24	8	2 – 34	2	2 – 2	0.49	0.18 – 1.36
	5 min.	1.00	0.69 – 1.26	9	2 – 52	4	3 – 4	0.21	0.06 – 0.72
	15 min.	1.18	1.02 – 1.28	181	114 – 218	4	4 – 4	0.06	0.04 – 0.08
D	1 min.	3.09	2.09 – 3.95	6	1 – 10	2	2 – 2	1.84	0.92 – 4.83
	3 min.	2.76	1.15 – 3.65	7	2 – 28	2	2 – 2	1.47	0.31 – 3.65
	5 min.	2.46	0.93 – 3.61	9	2 – 42	3	3 – 3	0.68	0.12 – 1.93
	15 min.	3.57	3.09 – 4.06	134	110 – 167	4	4 – 4	0.2	0.15 – 0.25
E	1 min.	2.19	1.48 – 3.38	4	1 – 7	2	2 – 2	1.55	0.79 – 4.78
	3 min.	2.02	1.24 – 2.93	5	2 – 18	2	2 – 2	1.28	0.41 – 3.21
	5 min.	1.98	1.24 – 2.85	7	2 – 28	4	3 – 4	0.48	0.15 – 1.64
	15 min.	2.60	2.14 – 4.21	89	74 – 104	4	4 – 4	0.17	0.13 – 0.31

Note:
i) Data obtained from April 18th and 27th, 2008.

A. Night hours (low demand). Average itinerary travel time considering all itineraries between junctions (1) and (5) = 6.54 minutes.
B. Off-Peak hours (moderate demand). Average itinerary travel time = 6.40 minutes.
C. Peak hours (free flowing high demand). Average itinerary travel time = 5.91 minutes.
D. Recurrent congested periods (high demand). Average itinerary travel time = 11.17 minutes.
E. Incident conditions (moderate demand + congestion). Average itinerary travel time = 12.37 minutes.

ii) Applying Equations 30 and 31, considering a probability level of $\alpha = 0.68$ and the average values for "N" and "σ" shown in this table.

5.2. Accuracy of the Algorithm

To check the accuracy of the algorithm, the ground truth travel times experienced by the drivers travelling from "Blanes" to "La Roca", "$t_{1,5}^{(p)}$" obtained as an accurate average of the itinerary travel times resulting from the toll ticket data of only those vehicles travelling in that particular itinerary and reaching (5) in any "p" time interval (this data is plotted in Figure 8, considering "Δt = 15 minutes"), will be compared with "$\tilde{t}_{(1,5)}^{(p)}$" the travel time resulting from the proposed algorithm (i.e. the addition of single section travel times, entrance time and exit time) along the same itinerary and exiting at the same time interval "p".

In order to evaluate the performance of the algorithm in terms of accuracy, the travel times to compare must be of the same nature. Then, as "$t_{1,5}^{(p)}$" is an MTT, "$\widetilde{t}_{(1,5)}^{(p)}$" must be so. This means that "$\widetilde{t}_{(1,5)}^{(p)}$" does not result from the simple addition of single section travel times at time interval "*p*", but from a backwards trajectory reconstruction process. This is:

$$\widetilde{t}_{(1,5)}^{(p)} = t_{en(1)} + \sum_{i=1}^{4} t_{s(i,i+1)}^{(q_i)} + t_{ex(5)}^{(p)} \tag{32}$$

Where "q_i" stands for the time period to consider in the trajectory reconstruction, as it will vary taking into account that the time taken up by the virtual vehicle in travelling along consecutive sections must be considered when trying to estimate a trajectory based travel time (e.g. MTT or PTT). In contrast, in case of estimating an ITT, which is not trajectory based, these "q_i"s time periods have to be considered all equal to "*p*". In the present case, a backwards reconstruction is needed and "q_i"s have to be calculated iteratively starting from downstream.

Although the single section travel times that configure the itinerary will only vary every "Δt", the headway between launched virtual vehicles can be as short as desired. In the present application, one virtual vehicle was launched every minute.

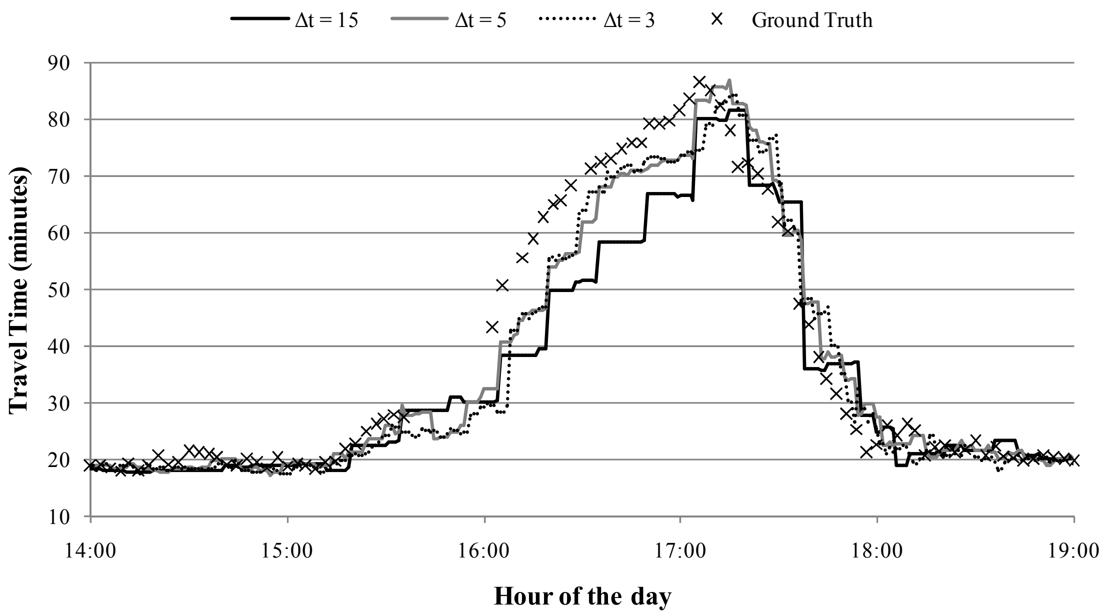
TABLE 2 Accuracy of the Algorithm for Various Demand Patterns and Different "Δt" considering the itinerary (1,5)

Demand Pattern[(i)]	Test Day	Period of the day	Δt	Itinerary travel time absolute error [(ii)] (relative) Mean	Max.
A	April 18th, 2008	0am – 6am 9pm – 12pm	3 min.	1.44 (0.07)	6.25 (0.25)
			5 min.	1.18 (0.06)	3.14 (0.17)
			15 min.	1.03 (0.05)	2.48 (0.13)
	April 27th, 2008	0am – 8am	3 min.	1.74 (0.09)	4.19 (0.24)
			5 min.	1.66 (0.08)	4.49 (0.19)
			15 min.	1.62 (0.08)	3.61 (0.16)
B	April 18th, 2008	6am – 7am 9am – 3pm 7pm – 9pm	3 min.	0.95 (0.05)	4.59 (0.20)
			5 min.	1.07 (0.06)	3.53 (0.21)
			15 min.	0.94 (0.05)	2.30 (0.12)
	April 27th, 2008	8am – 4pm 10pm – 12pm	3 min.	1.09 (0.06)	3.43 (0.17)
			5 min.	1.14 (0.06)	3.26 (0.17)
			15 min.	0.89 (0.05)	3.34 (0.17)
C	April 18th, 2008	7am – 9am	3 min.	1.30 (0.07)	3.32 (0.18)
			5 min.	1.04 (0.06)	2.87 (0.13)
			15 min.	0.75 (0.04)	1.08 (0.06)
	April 27th, 2008	4pm – 5pm	3 min.	0.87 (0.04)	2.89 (0.14)
			5 min.	0.65 (0.03)	1.56 (0.08)
			15 min.	0.76 (0.04)	1.08 (0.05)
D	April 27th, 2008	5pm – 10pm	3 min.	1.47 (0.06)	5.07 (0.16)
			5 min.	1.28 (0.05)	4.25 (0.14)
			15 min.	1.06 (0.04)	2.30 (0.09)
E	April 18th, 2008	3pm – 7pm	3 min.	3.62 (0.09)	14.72 (0.34)
			5 min.	3.59 (0.08)	11.58 (0.23)
			15 min.	5.84 (0.11)	18.09 (0.30)

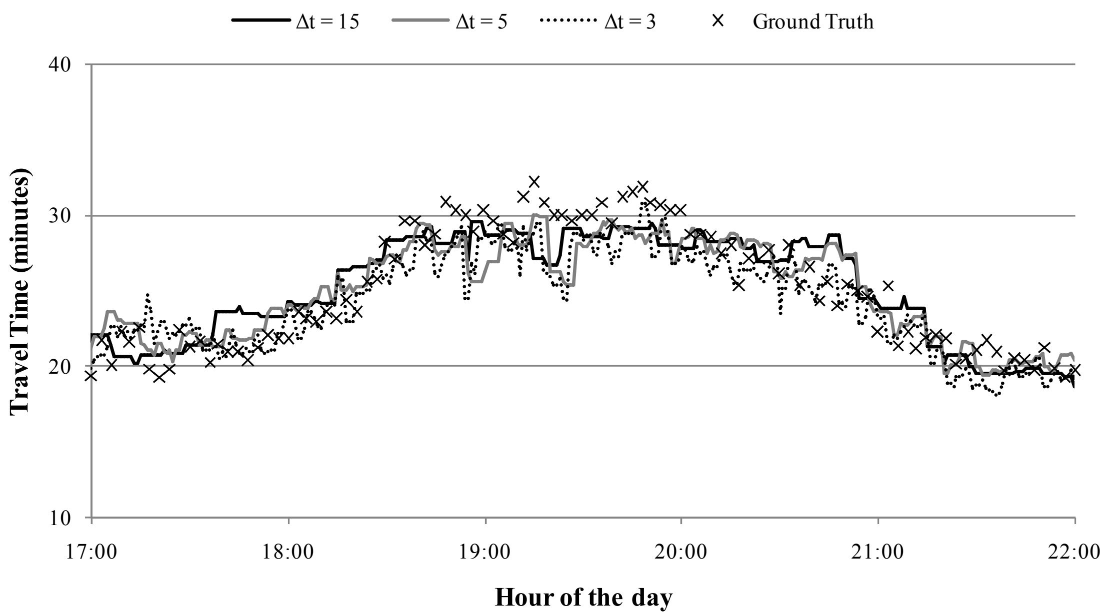
Note: i) Periods defined as in Table 1.
ii) In minutes.

Table 2 shows the error committed in the estimation of "$\tilde{t}_{(1,5)}^{(p)}$" in relation to the ground truth "$t_{1,5}^{(p)}$" for several demand patterns from different days, and considering various "*Δt*". Note from the numerical values that, in general, the obtained error follows the logic detailed in the previous section for the expected single section estimation error. However it is worth to notice that for the rapidly evolving incident related traffic conditions (i.e. scenario E in Table 2) the obtained errors for "*Δt=15*" are significantly larger than expected. This is a clear consequence of the lack of independence between travel time individual measurements in a 15 minute period during rapid congestion onset or dissolve.

From the above results it can be confirmed that for free-flowing conditions (i.e. demand scenarios A-C) and even for moderate recurrent congestion (i.e. scenario D) an updating interval of 15 minutes provides the best accuracy results. However, this long interval is not capable of tracking the rapidly changing conditions of incident related congestion (i.e. scenario E), particularly in the congestion onset. For this last situations, "*Δt* = 5 minutes" is adequate. Figure 9 provides graphical evidence of the algorithm behavior in congested conditions.



a)



b)
FIGURE 9 Accuracy of the algorithm for different updating periods. a) Incident related congestion (April 18th, 2008) b) Moderate recurrent congestion (April 27th, 2008).

An updating interval of 5 minutes can be selected as a compromise solution for all the demand scenarios. In this case, the algorithm is accurate enough to provide travel time information to drivers and road administrations, with a mean absolute relative error below 10% in all situations, and in particular, in critical situations with huge and rapid variations in travel times.

5.3. Value as a Real Time Information System

One of the main advantages of the estimation of single section travel times is the reduction of the information delay for long itineraries when providing real time information to drivers. This improvement, resulting from the dissemination of an ITT for the itinerary (instead of a MTT with a greater delay), is only relevant when traffic conditions evolve in a time horizon equal to the travel time (i.e., congestion onset and congestion dissolve). Otherwise, both estimations would lead to the same result.

Table 3 shows the travel times to be disseminated in real time for the same analyzed itinerary (*1, 5*) in the case of using the information directly obtained from toll ticket data (MTT) or the ITT obtained by the addition of single section travel times resulting from the proposed algorithm. Both are compared with the real travel time of drivers *entering* the highway during the next time period, who are able to receive the information (note that this is the target information – a PTT – which is not known at the instant the information is disseminated. This target travel time is only known after the vehicles have left the highway).

TABLE 3 Real Time Travel Time dissemination in relation to the available information

Type of available information	Travel time disseminated at "*p*" (Min:Sec)	
	April 18th 2008 "*p*"=17.30h (scenario E) Congestion dissolve	April 27th 2008 "*p*"=18.15h (scenario D) Congestion onset
Only itinerary information (MTT) $t_{1,5}^{(p)}$	46:04	24:32
Single section travel time information (ITT) $t_{en(1)} + \sum_{i=1}^{4} t_{s(i,i+1)}^{(p)} + t_{ex(5)}^{(p)}$	37:18	27:16
Real travel time for those vehicles entering the highway between instants "*p*" and "*p+1*" (PTT)	24:47	30:14

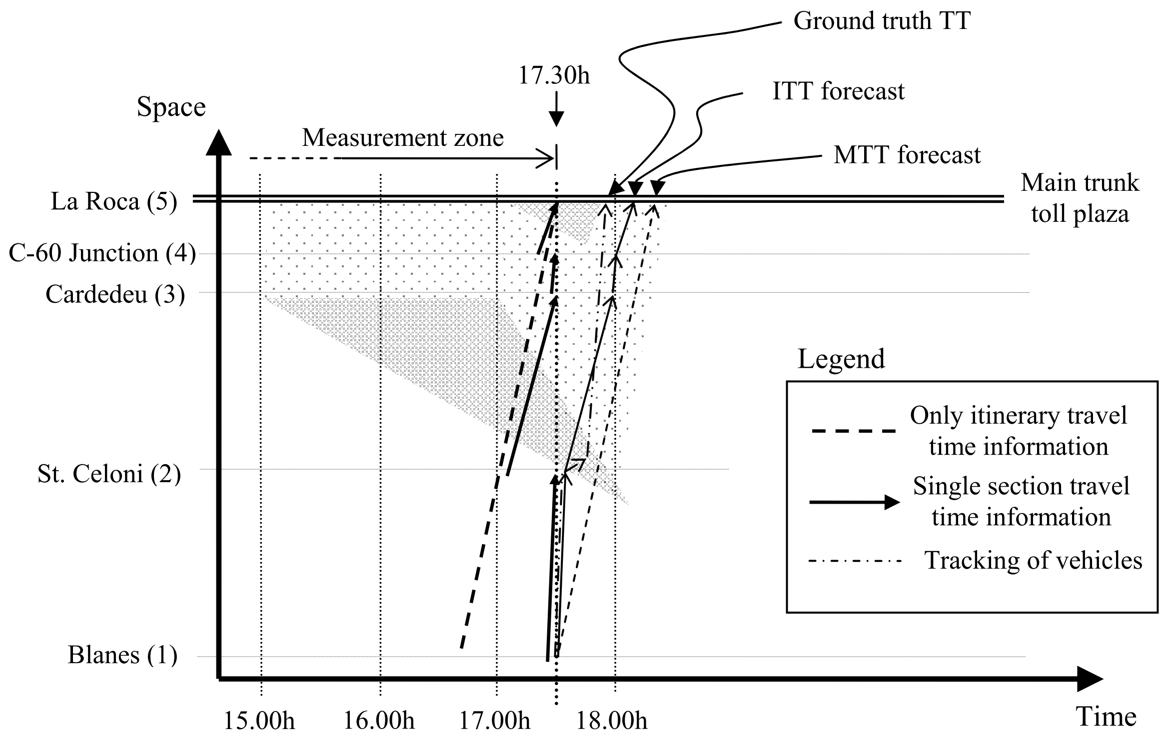


FIGURE 10 Effect of the available information in the dissemination of travel times (April 18th, 2008).

As can be seen from results in Table 3 from data corresponding to both test days and different evolving traffic conditions, ITT performs well in recurrent and non-recurrent traffic conditions. Generally performs better than MTT, particularly under non-recurrent and rapidly changing conditions, where the benefits of reducing the information delay are bigger. However, the same quickly evolving conditions also imply big differences between ITT and PTT. Therefore one could say that while the performance of

MTT in supporting real time information systems is bad, the performance of ITT is not as bad, but still flawed due to the null forecasting capabilities of a real time measurement. This is the most that can be done without sinking in the uncertainties of forecasting and modeling. These concepts are clarified in Figure 10, using a trajectories diagram.

5.4. Exit Time Information

The proposed algorithm for estimating single section travel times in a closed highway system provides, as a collateral result, the highway exit time. This information is of great interest to highway operators, since drivers' value of this particular time is greater than the traveling time, strictly speaking. This means that a little delay in the payment at the toll gate is a big nuisance to drivers, who attribute a low level of service to the highway trip. Obviously these delays for payment greatly penalize the operators' reputation.

For this reason, highway operators are especially sensitive to keep this exit time as low as possible, while maintaining costs. This trade-off is achieved with an accurate schedule of toll gates (i.e. number of open toll gates, and direction of operation, as some of the toll gates are reversible). In fact, the experience achieved during years of operation on the AP-7 highway has given the operators an accurate knowledge of the demand pattern at each junction, so that in most of the situations all the exit times are limited to between 1 and 3 minutes, which represent very short delays for exiting the highway. However, in some incident situations, with unexpected demand, or some problems at the toll gates, queues could grow, causing delays to be of more consideration (see Figure 11).

Under these conditions, a quantitative estimation of the exit times at the junction can alert the highway operations center in order to reschedule the toll gates. In addition, information could be disseminated to drivers to allow them to select an alternative junction to exit the highway, avoiding a delay in the trip and helping to alleviate the problem.

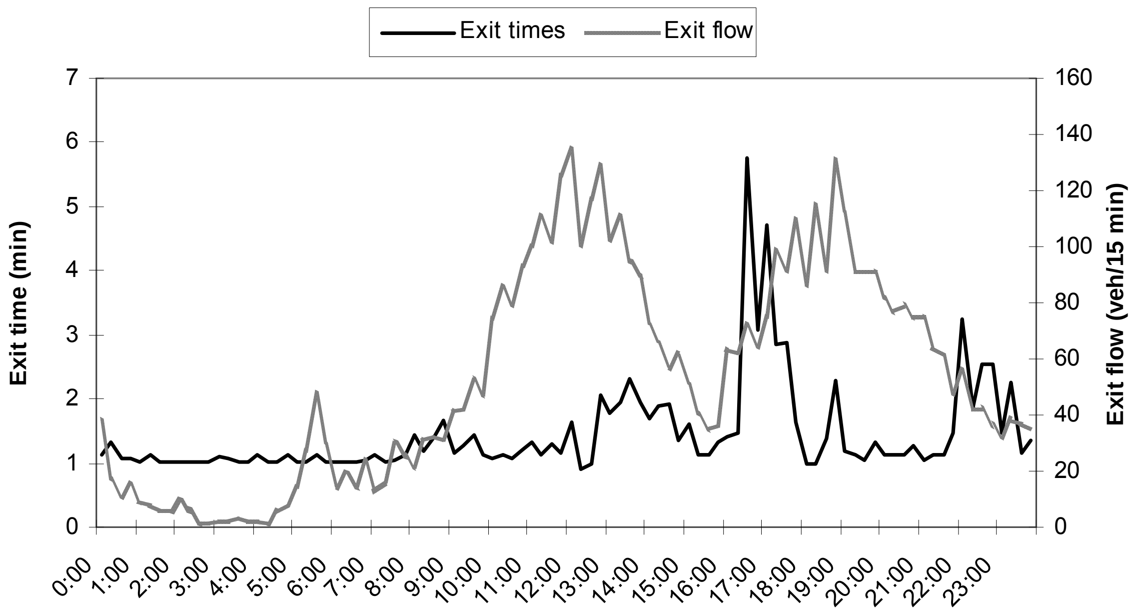


FIGURE 11 Exit times at the Cardedeu (*3*) junction, (April 6th 2008).

6. CONCLUSIONS AND FURTHER RESEARCH

Link travel time is the most appreciated information for road users. The new approach in this paper for calculating the travel time on highways using toll infrastructure is a simple one and can be easily put into practice with the existing infrastructure. The scheme uses data obtained from the toll tickets on highways with a closed tolling system. Rather than simply calculating the itinerary travel time by comparing the entry and exit times on the ticket, the approach presented here could be used to increase the information available from toll ticket data.

The proposed method is capable of estimating single section travel times (i.e. time required to travel between two consecutive junctions on the main trunk of the highway) and also the exit time at each junction (i.e. the time required to travel along the exit link plus the time required to pay the fee at the toll gate). Combining both estimations it is possible to calculate all the required itinerary travel times, even those with very few observations where direct measurement would be problematic, and avoiding the information delay for real time application.

The results of the pilot test carried out on the AP-7 highway in Spain indicate the suitability of the method for the link travel time estimation in a closed toll highway system.

Moreover, the accuracy in the travel time estimation should make the development of a robust incident detection system possible, by comparing the real time estimations to the recurrent travel times. Since the method supplies exit time information, it is also possible to detect whether the incident was on the main highway or in the ramp toll plaza. This is valuable information for the highway operators, enabling them to deal with the incident.

Further research may consider data fusion with other sources of data, such as loop detectors, to obtain information within the inner section (between junctions). Travel time prediction (to improve the forecasting capabilities of a real time measurement) on the basis of the present scheme is also a key factor for future research.

REFERENCES

Abdulhai, B., Tabib, S.M., (2003). Spatio-temporal inductance-pattern recognition for vehicle reidentification. *Transportation Research Part C* **11**(3-4), 223-239.

Cassidy, M.J., Ahn, S., (2005). Driver turn-taking behavior in congested freeway merges. *Transportation Research Record* **1934**, 140-147.

Coifman, B., (2001). Improved velocity estimation using single loop detectors. *Transportation Research Part A* **35**(10), 863-880.

Coifman, B., (2002). Estimating travel times and vehicle trajectories on freeways using dual loop detectors. *Transportation Research Part A* **36**(4), 351-364.

Coifman, B., Cassidy, M., (2002). Vehicle reidentification and travel time measurement on congested freeways. *Transportation Research Part A* **36**(10), 899-917.

Coifman, B., Ergueta, E., (2003). Improved vehicle reidentification and travel time measurement on congested freeways. *ASCE Journal of Transportation Engineering* **129**(5), 475-483.

Coifman, B., Krishnamurthya, S., (2007). Vehicle reidentification and travel time measurement across freeway junctions using the existing detector infrastructure. *Transportation Research Part C* **15**(3), 135-153.

Cortés, C.E., Lavanya, R., Oh, J., Jayakrishnan, R., (2002). General-purpose methodology for estimating link travel time with multiple-point detection of traffic. *Transportation Research Record* **1802**, 181-189.

Dailey, D.A., (1999). Statistical algorithm for estimating speed from single loop volume and occupancy measurements. *Transportation Research Part B* **33**(5), 313-322.

Davies, P., Hill, C., Emmott, N., (1989). Automatic vehicle identification to support driver information systems. *Proc. of IEEE Vehicle Navigation and Information Systems Conference*, 11–13 September, A31-35.

Dion, F. Rakha, H., (2006). Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transportation Research Part B* **40**(9), 745-766.

Herrera, J.C., Work, D.B., Herring, R., Ban, X., Jacobson, Q., Bayen, A.M., (2010). Evaluation of traffic data obtained via GPS-enabled mobile phones: the Mobile Century field experiment. *Transportation Research Part C* **18**(4), 568-583.

Hopkin, J., Crawford, D., Catling, I., (2001). *Travel time estimation*. Summary of the European Workshop organized by the SERTI project, Avignon, November 2001.

Li, R., Rose, G., Sarvi, M., (2006). Evaluation of speed-based travel time estimation models. *ASCE Journal of Transportation Engineering* **132**(7), 540-547.

Lucas, D.E., Mirchandani, P.B., Verma, N., (2004). Online travel time estimation without vehicle identification. *Transportation Research Record* **1867**, 193-201.

Mikhalkin, B., Payne, H., Isaksen, L., (1972). Estimation of speed from presence detectors. *Highway Research Record* **388**, 73-83.

Muñoz, J.C., Daganzo, C.F., (2002). The bottleneck mechanism of a freeway diverge. *Transportation Research Part A* **36**(6), 483-505.

Nam, D. H., Drew, D.R., (1996). Traffic dynamics: method for estimating freeway travel times in real time from flow measurements. *ASCE Journal of Transportation Engineering* **122**(3), 185-191.

Ohba, Y., Ueno, H., Kuwahara, M., (1999). Travel time calculation method for expressway using toll collection system data. *Proc. of the 2nd IEEE Intelligent Transportation Systems Conference*, 471-475.

Palen, J., (1997). The need for surveillance in intelligent transportation systems. *Intellimotion* **6**(1), 1-3.

Pushkar, A., Hall, F., Acha-Daza, J., (1994). Estimation of speeds from single-loop freeway flow and occupancy data using cusp catastrophe theory model. *Transportation Research Record* **1457**, 149-157.

Sun, C., Ritchie, S.G., Tsai, K., (1998). Algorithm development for derivation of section-related measures of traffic system performance using inductive loop detectors. *Transportation Research Record* **1643**, 171-180.

Sun, C., Ritchie, S.G., Tsai, K., Jayakrishnan, R., (1999). Use of vehicle signature analysis and lexicographic optimization for vehicle reidentication on freeways. *Transportation Research Part C* **7**(4), 167-185.

SwRi, (1998). *Automatic vehicle identification model deployment initiative – System design document*. Report prepared for TransGuide, Texas Department of Transportation, Southwest Research Institute, San Antonio, TX.

Turner, S.M., Eisele, W.L., Benz, R.J., Holdener, D.J., (1998). *Travel time data collection handBook*. Research Report FHWA-PL-98-035. Federal Highway Administration, Office of Highway Information Management, Washington, DC.

van Arem, B., van der Vlist, M.J.M., Muste, M.R., Smulders, S.A., (1997). Travel time estimation in the GERIDIEN project. *International Journal of Forecasting* **13**(1), 73-85.

van Lint, J.W.C., van der Zijpp, N.J., (2003). Improving a travel-time estimation algorithm by using dual loop detectors. *Transportation Research Record* **1855**, 41-48.

APPENDIX A3-A: OBTAINING A REPRESENTATIVE AVERAGE OF ITINERARY TRAVEL TIMES IN A "Δt" TIME INTERVAL

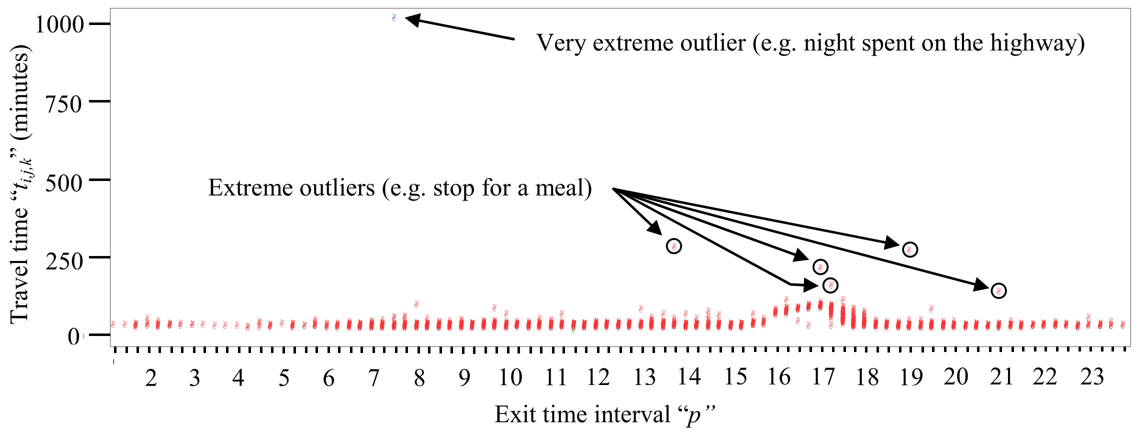
The source data to calculate the single section travel times using the proposed algorithm is the "$t_{i,j}^{(p)}$", symbolizing a representative average of the individual itinerary travel time observations "$t_{i,j,k}^{(p)}$" obtained in the "*p*" time interval (i.e. all the "*k*" vehicles that have exited the highway at "*j*", coming from "*i*" between the time instants "*p-Δt*" and "*p*"). As introduced in Section 3, the estimation of this average can be tricky. In fact, it represents the main goal in the ETC-only based travel time estimation systems, where vehicles are identified at several control points on the main highway trunk, and the measured travel times are directly single section travel times.

Problems in the estimation of a representative average of itinerary travel times arise from the following characteristics of these data (see Figure A1):

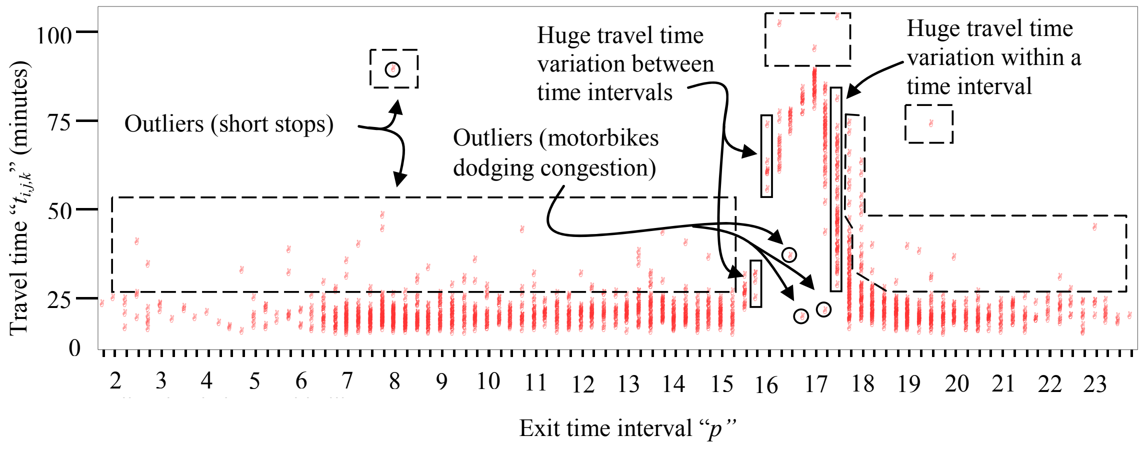
- Possibility of high variability in travel time observations within a "*Δt*" time interval (higher variability as "*Δt*" is higher), and high variability in travel time observations between consecutive time intervals.
- Few observations for some combinations of itinerary and time interval.
- Presence of outliers, whose measured travel time is not related to traffic conditions. There are two main types of outliers: travel time is in excess due to stops on the highway during the trip, and travel time is overly low resulting from motorbikes dodging traffic jams.

Obviously, outliers must not be considered in the estimation of the average travel time. However, it is not obvious to decide on whether an observation is an outlier or not (unless it is very extreme). This is particularly difficult in the case of very few observations, and taking into account the possibility of high variability in travel times (e.g. an accident has happened).

In this situation, standard methods for outlier identification are not suitable, and a smart data filtering process is necessary, taking into account the traffic state. The data filtering method presented in this appendix uses two alternatives, depending on the number of observations in the set. On the one hand, if the set contains enough observations, the median represents a good average and only a smoothing process is required. On the other hand, for sets with few observations, a more careful filtering is needed.



a)



b)

FIGURE A1 Toll ticket travel time observations from Blanes to La Roca, "$t_{1,5,k}$" (April 18th 2008). a) All the observations. b) Truck observations and extreme outliers removed.

A.1. Elimination of Extreme Outliers

Extreme outliers or measurement errors are easily detected (see Figure A1), and therefore can be easily eliminated from the observations database. An observation "$t_{i,j,k}$" is considered as an extreme outlier if:

$$\begin{aligned} & t_{i,j,k} < 0 \\ & t_{i,j,k} > 5\ hours \\ & i = j \\ & \frac{l_{i,j}}{t_{i,j,k}} > 200\ km/h \end{aligned} \qquad (A1)$$

Where "$l_{i,j}$" is the total length of the itinerary (considering the on and off ramps).

For a maximum length of itinerary of about 120 km, all the situations captured by equation A1 represent measurement errors or stops. However, not all the measurement errors or stops are captured by equation A1.

A.2. Data filtering in Sets With Enough Data

Assume that a set of itinerary travel times, defined by an origin "*i*", a destination "*j*" and a time interval "*p*", has "$N_{i,j}^{(p)}$" observations. This set has enough data if "$N_{i,j}^{(p)}$" is bigger than a threshold value "$N_{i,j}^{*(p)}$", which must be set for the selected "Δt" using Equation 30, data in Table 1, and defining a maximum acceptable error given a statistical significance for the estimation. For instance, in the present application for "Δt = 5 minutes", considering a maximum acceptable error of 1 minute in the average itinerary travel time estimation with a statistical significance "$(1-\alpha)$" of 0.9, the threshold value "$N_{i,j}^{*(p)}$" equals 8 itinerary travel time observations. In this case where the set has enough data, "$\hat{t}_{i,j}^{(p)}$" is defined as the median of the "$t_{i,j,k}$" observations, where "$k = 1, \ldots, N_{i,j}^{(p)}$". The interquartile range of the set is also computed, "$I\hat{Q}R_{i,j}^{(p)} = \hat{q}_3 - \hat{q}_1$", where "$\hat{q}_3$" and "$\hat{q}_1$" are the 75% and 25% percentiles respectively. In the present case, with enough observations, it is not necessary to explicitly eliminate the outliers, as the problem is solved by considering the median statistic (instead of the mean), which is highly insensitive to a few extreme values.

The representative average to use in the single section travel time estimation algorithm is an exponentially smoothed value of "$\hat{t}_{i,j}^{(p)}$". The exponential smoothing is carried out on a logarithmic scale (Equation A2) to account for the log-normal distribution of travel times (Li et al., 2007). Note that travel time distribution is skewed to the right, reflecting the fact that travel times do not significantly decrease below a free flowing travel time (vehicles travelling around the posted speed limit), while significantly higher travel times are possible, especially if congestion builds up. This suggests that the lower "Δt" is, the lower will be this skewness (for small "Δt"s it is less likely that different traffic states arise), and therefore, travel time distribution would tend to normal for small time intervals of measurement.

$$
\begin{aligned}
t_{i,j}^{(p)} &= \exp\left[\alpha_{i,j}^{(p)} \ln\left(\hat{t}_{i,j}^{(p)}\right) + \left(1-\alpha_{i,j}^{(p)}\right)\cdot \ln\left(t_{i,j}^{(p-1)}\right)\right] \\
q1_{i,j}^{(p)} &= \exp\left[\alpha_{i,j}^{(p)} \ln\left(\hat{q}1_{i,j}^{(p)}\right) + \left(1-\alpha_{i,j}^{(p)}\right)\cdot \ln\left(q1_{i,j}^{(p-1)}\right)\right] \\
q3_{i,j}^{(p)} &= \exp\left[\alpha_{i,j}^{(p)} \ln\left(\hat{q}3_{i,j}^{(p)}\right) + \left(1-\alpha_{i,j}^{(p)}\right)\cdot \ln\left(q3_{i,j}^{(p-1)}\right)\right] \\
IQR_{i,j}^{(p)} &= q3_{i,j}^{(p)} - q1_{i,j}^{(p)}
\end{aligned}
\qquad (A2)
$$

The statistical significance of the average itinerary travel time estimation, "$\alpha_{i,j}^{(p)} \in (0,1)$", is selected as the smoothing factor. "$\alpha_{i,j}^{(p)}$" depends on the number of observations in the set and on the acceptable absolute error "ε", which may be used as a calibration parameter. In the present application "ε" has been set to 1 minute, As there are fewer observations, the average will be less reliable, "$\alpha_{i,j}^{(p)}$" will be lower and greater weight will be given to past observations. "$\alpha_{i,j}^{(p)}$" could be seen as a reliability index for "$t_{i,j}^{(p)}$", taking a value of one for perfect information given the acceptable error and decreasing as the information becomes less reliable. A minimum weight of 0.5 is given in the case where there is at least one valid observation. More formally:

$$
\alpha_{i,j}^{(p)} = \max\left\{2\cdot F^{-1}\left(\frac{\varepsilon\cdot\sqrt{N_{i,j}^{(p)}}}{\sigma_{(t_{i,j}^{(p)})}}\right) - 1, \quad 0.5\right\} \qquad (A3)
$$

Where"$F^{-1}(z)$" is the inverse cumulative distribution function of a standard normal probability distribution and "$\sigma_{(t_{i,j}^{(p)})}$" is the standard deviation of the average itinerary travel time estimation. Approximate average data for this standard deviation is provided in Table 1.

In the present section, dealing with groups with enough observations, "$\alpha_{i,j}^{(p)}$" is defined by the first condition of Equation A3, and the result is approximately equal to one (perfect information). In this case, Equation A2 is of little use as "$t_{i,j}^{(p)} \approx \hat{t}_{i,j}^{(p)}$".

A.3. Data Filtering in Sets With Few Data

In case the set does not have enough data ("$0 < N_{i,j}^{(p)} < N_{i,j}^{*(p)}$"), then the median of these observations cannot be considered a representative average of the itinerary travel time for that time interval. The data filtering process in this case, tries to decide if an observation could be an outlier or not. If the answer is positive, then the observation is eliminated from the database. The median of the remaining valid observations is set as "$\hat{t}_{i,j}^{(p)}$", the input required for the exponential smoothing process detailed in Equations A2 and A3.

To determine if an observation could be an outlier, two confidence intervals are defined: $\left(t^{(p)}_{\min i,j}, t^{(p)}_{\max i,j}\right)$ and a broader $\left(t^{(p)}_{MIN\,i,j}, t^{(p)}_{MAX\,i,j}\right)$, where:

$$
\begin{aligned}
t^{(p)}_{\max i,j} &= \exp\left[\ln\left(q^{(p-1)}_{3\,i,j}\right) + \gamma \cdot \ln\left(IQR^{(p-1)}_{i,j}\right)\right] \\
t^{(p)}_{MAX\,i,j} &= \exp\left[\ln\left(q^{(p-1)}_{3\,i,j}\right) + \rho^{(p-1)}_{i,j} \cdot \ln\left(IQR^{(p-1)}_{i,j}\right)\right] \\
t^{(p)}_{MIN\,i,j} &= \exp\left[\ln\left(t_{i,j,0}\right) - \gamma \cdot \ln\left(IQR_{i,j,0}\right)\right] \\
t^{(p)}_{\min i,j} &= \begin{cases} \exp\left[\ln\left(q^{(p-1)}_{1\,i,j}\right) - \gamma \cdot \ln\left(IQR^{(p-1)}_{i,j}\right)\right] & if \ \geq t^{(p)}_{MIN\,i,j} \\ t^{(p)}_{MIN\,i,j} & o.w. \end{cases}
\end{aligned}
\tag{A4}
$$

These confidence intervals for an itinerary "*i,j*" and time interval "*p*" are defined as an extra time above the third quartile (maximum) or below the first quartile (minimum) of travel time distribution in the previous time interval. These extensions depend on the interquartile range of travel times in the previous time interval. This responds to the fact that when some change starts to develop in a "*p*" time period, the interquartile range increases. Then the confidence intervals for the possible acceptance of an observation for the next time period will be broader, with larger variations than usual in travel times. The lowest threshold of these confidence intervals is limited by "$t^{(p)}_{MIN\,i,j}$", computed as the free flowing travel time "$t_{i,j,0}$" (defined as the first quartile of the travel time distribution in free flowing conditions) minus a fraction of the free flowing conditions interquartile range "$IQR_{i,j,0}$". Obviously, none of these thresholds can exceed the extreme values set in Equation A1. Only as an order of magnitude, "$IQR_{i,j,0}$" results approximately from a variation interval of 20 km/h around the average free flow speed of approximately 110 km/h. The absolute magnitude of "$IQR_{i,j,0}$" depends on the length of the "*i,j*" itinerary.

The amplitudes of the confidence intervals defined in equation A4, depend on a proportionality constant of "γ", calibrated to 0.5, and on "$\rho^{(p-1)}_{i.j}$" for the "*MAX*" threshold. "$\rho^{(p-1)}_{i.j}$" is defined as:

$$
\rho^{(p-1)}_{i.j} = \lambda \cdot \left[2 - \left(\alpha^{(p-1)}_{i.j}\right)^{\max(N^{(p-1)}_{out\ i,j},\ s^{(p-1)}_{i,j})}\right]
\tag{A5}
$$

Where "$\alpha^{(p-1)}_{i.j} \in (0.5\,,\,1)$", "$N^{(p-1)}_{out\ i,j}$" is the number of observations that have been considered outliers in the "*p-1*" time interval, and "$s^{(p-1)}_{i,j}$" is the number of consecutive intervals without any observation before the time interval "*p-1*". The larger these last two variables are, the less reliable the previous travel time average is considered to be, as it is not tracking accurately the travel time evolution. As a consequence, broader confidence intervals will be considered for the target time interval "*p*". "λ" is the default value for perfect previous information and is set to a value of 3.

In this context, four situations can be defined, given a travel time observation "$t^{(p)}_{i,j,k}$":

a) $t^{(p)}_{i,j,k} \in \left[t^{(p)}_{\min i,j}, t^{(p)}_{\max i,j}\right]$; Then "$t^{(p)}_{i,j,k}$" is considered a valid observation.

b) $t_{i,j,k}^{(p)} \notin \left[t_{MIN\,i,j}^{(p)}, t_{MAX\,i,j}^{(p)}\right]$; Then "$t_{i,j,k}^{(p)}$" is considered an outlier and is eliminated from the database.

c) $t_{i,j,k}^{(p)} \in \left[t_{\max\,i,j}^{(p)}, t_{MAX\,i,j}^{(p)}\right]$; Then "$t_{i,j,k}^{(p)}$" is considered a doubtful observation.

d) $t_{i,j,k}^{(p)} \in \left[t_{MIN\,i,j}^{(p)}, t_{\min\,i,j}^{(p)}\right]$; Then "$t_{i,j,k}^{(p)}$" is considered a doubtful observation.

A.3.1. Deciding on $t_{i,j,k}^{(p)} \in \left[t_{\max\,i,j}^{(p)}, t_{MAX\,i,j}^{(p)}\right]$ Doubtful Observations

An itinerary travel time observation falling in the excess doubtful zone can result from two situations:

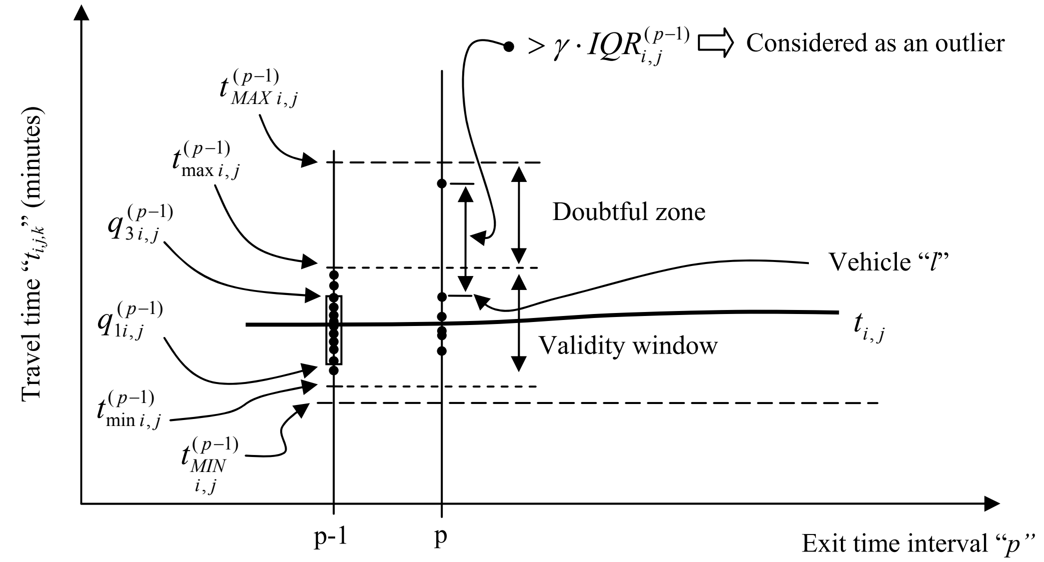
- The vehicle has stopped for a short time (e.g. for a quick refuel). Then the observation should be considered as an outlier and eliminated from the database.
- There is a sudden travel time increase in the highway (e.g. due to an incident). Then the observation should be considered as valid.

To decide which is the cause of this doubtful observation, two contrasts are developed. Firstly, the difference between the doubtful observation and the other existing valid observations in the set is analyzed (overtaking contrast). If this difference is considered small, the observation is accepted. Otherwise, if the previous contrast cannot be applied (e.g. there are few valid observations in the set) or the difference is considered to be rather large within a time interval, then the traffic state (i.e. congested or not) is assessed. If traffic is considered to be congested, the observation is considered as valid, otherwise, the observation is an outlier, and it is eliminated from the database.

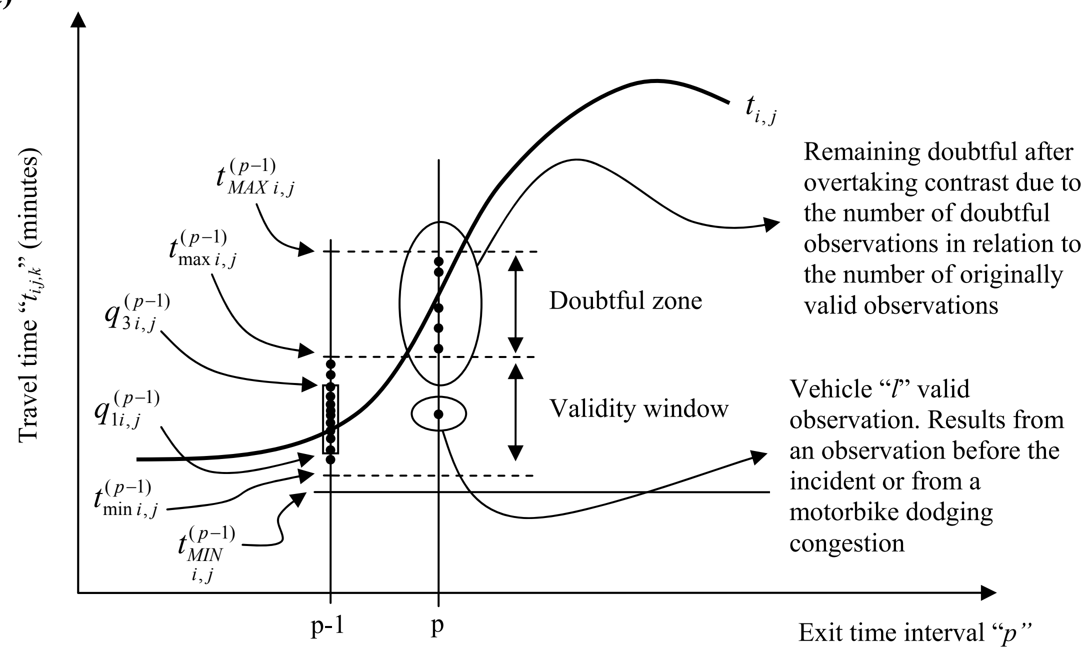
A.3.1.1. Overtaking Contrast This contrast stands for the fact that if some vehicles are capable of achieving significantly lower travel times for the same time interval and itinerary as the doubtful observation, then these vehicles are overtaking the "doubtful" vehicle whose large travel time is not related to general traffic conditions but to the specific behavior of this vehicle. Specifically, assume that at least one valid observation exists in the set of observations where the doubtful "$t_{i,j,k}^{(p)}$" is contained. Name vehicle "*l*" the vehicle whose itinerary travel time, "$t_{i,j,l}^{(p)}$" is the maximum within the valid observations, then "$t_{i,j,k}^{(p)}$" is also considered valid if:

$$t_{i,j,k}^{(p)} \leq t_{i,j,l}^{(p)} + \gamma \cdot IQR_{i,j}^{(p-1)} \qquad (A6)$$

If Equation A6 does not hold and the number of originally valid observations in the set exceeds "$N_{i,j}^{(p)}/2$" (i.e., half of the total number of observations, then "$t_{i,j,k}^{(p)}$" is considered an outlier and is eliminated. Otherwise, the congestion contrast must be applied to decide. This last restriction in the number of valid observations stands for the fact that vehicle "*l*" could also be a motorbike dodging the congestion (see Figure A2).



a)



b)

FIGURE A2 Sketch of the overtaking contrast. a) Elimination of an outlier. b) Overtaking contrast does not make it possible to decide.

A.3.1.2. Congestion Contrast In case the overtaking contrast is not meaningful and "$t_{i,j,k}^{(p)}$" remains as a doubtful observation, the congestion contrast has the last word. As there does not exist a vehicle that has overtaken the vehicle "*k*" in a valid travel time, showing the evidence of the possibility of travelling within the validity window, the only possibility to decide if the observation results from a voluntary stop or from congestion in the itinerary is by estimating the traffic state within every highway section. If there is a high probability of congestion within some stretch of the itinerary, the observation is considered as valid. In contrast, if free flowing conditions are estimated, then the observation is considered an outlier.

Toll tickets are not the best data source to decide whether there is congestion or not in the highway in real time, due to the MTT nature of these data. Loop detector data would be more suitable for this objective (e.g. assessing the occupation at each loop detector), opening a gap for data fusion schemes. However, from toll tickets the origin destination matrix can be obtained on a reconstructed basis (i.e. once the vehicles have left the highway). Therefore, an approximation to the flow in each section can be obtained for the previous time interval (Daganzo, 1997). Then, in the congestion contrast it is assumed that the probability of congestion in the present time interval depends on the traffic flow in the previous time interval. If this flow exceeds 75% of the capacity of the infrastructure, a high probability of congestion is given to the highway section, and the observation is considered valid.

Highway capacities can be estimated using the HCM (Transportation Research Board, 2000), but note that the capacity of an infrastructure is a dynamic variable and should be modified in case of an incident or bad weather conditions. Then, an accurate application of the data filtering process requires some kind of information input to modify default capacities in incident conditions. In addition, not only main trunk capacities must be assessed, also the exit toll gate capacities, as congestion can arise in the off-ramp due to limited exit capacity. These capacities can be easily obtained by taking into account the number of open gates and the type of these gates (i.e. manual payment-230 veh/h, automatic credit card payment-250 veh/h and non-stop ETC systems-700 veh/h).

A.3.2. Deciding on $t_{i,j,k}^{(p)} \in \left[t_{MIN\,i,j}^{(p)}, t_{\min\,i,j}^{(p)}\right]$ Doubtful Observations

An itinerary travel time observation falling in the lower doubtful zone can result from two situations:

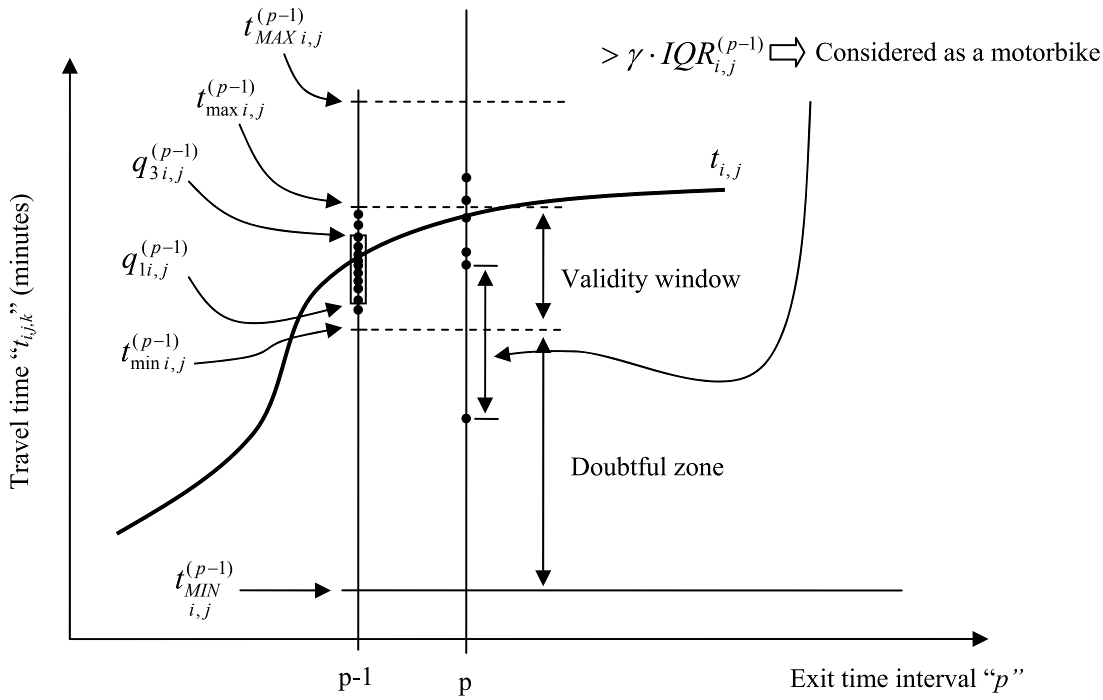
- A motorbike dodging the traffic jam by wriggling between cars. This type of observation should be considered as an outlier.
- There is a sudden travel time decrease in the highway due to congestion dissipation. Then the observation should be considered as valid.

This type of outlier is not as problematic as the ones in excess, for two reasons. Firstly, there are few outliers of this nature. And secondly, in congestion dissipation, traffic flows at capacity, and it is not usual to find itineraries with few observations. On this basis, the doubtful observation is considered to be a motorbike if the difference between its travel time and the other existing valid observations in the set is considered large.

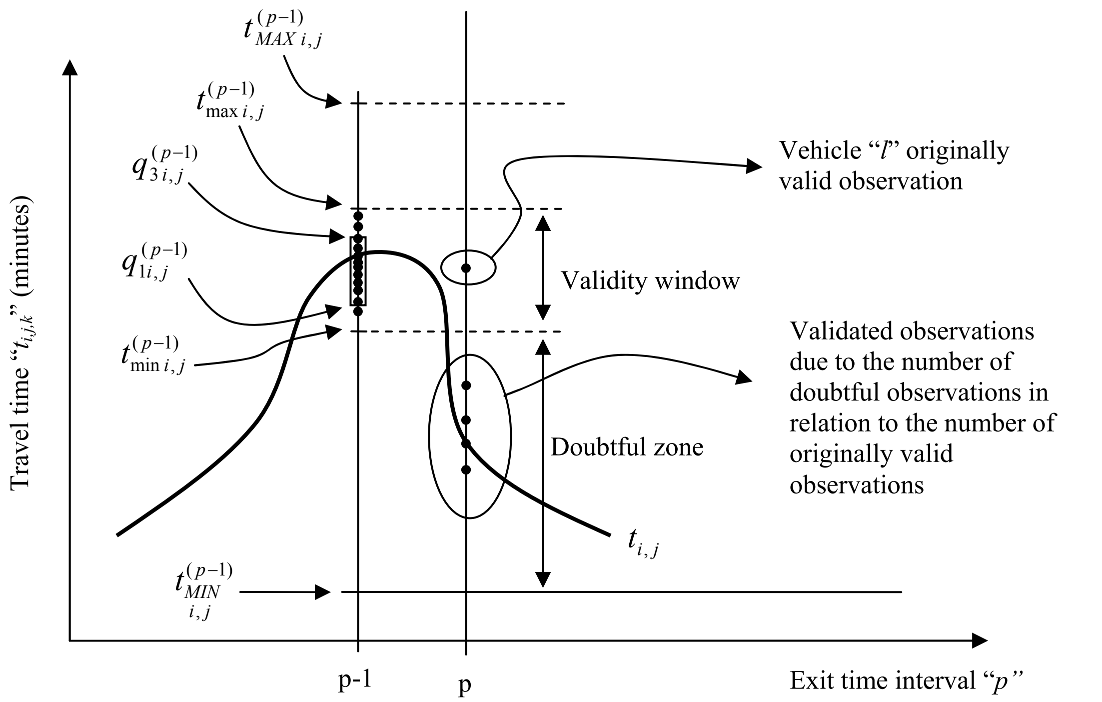
More formally, assuming there exists at least one valid observation in the set of observations where the doubtful "$t_{i,j,k}^{(p)}$" is contained, and vehicle "*l*" is the vehicle whose itinerary travel time, "$t_{i,j,l}^{(p)}$" is the minimum within the valid observations, then "$t_{i,j,k}^{(p)}$" is considered a motorbike if:

$$t_{i,j,k}^{(p)} \le t_{i,j,l}^{(p)} - \gamma \cdot IQR_{i,j}^{(p-1)} \qquad (A7)$$

In case equation A7 holds, the observation is only considered as an outlier and eliminated if the number of originally valid observations in the set exceeds "$N_{i,j}^{(p)}/2$". Otherwise, the observation is considered as valid (see Figure A3).



a)



b)

FIGURE A3 Deciding on lower travel time doubtful observations. a) Elimination of an outlier. b) Validation of observations in a congestion dissipation situation.

A.3.3. Estimation of the Interquartile Range in Sets with Few Data

Once the data filtering process has decided on the validity of the observations in the set, and "$\hat{t}_{i,j}^{(p)}$" has been calculated as the median of the valid observations, the only remaining variable to estimate in order to set the confidence intervals for the next time interval (using Equations A.2 and A.4) is the interquartile range "$I\hat{Q}R_{i,j}^{(p)}$". In the considered set with few data, the statistical calculations of "$\hat{q}_{3\,i,j}^{(p)}$" and "$\hat{q}_{1\,i,j}^{(p)}$" are not meaningful. Therefore, it is assumed that the interquartile range does not vary from the last time interval. Under this assumption, the 25% and the 75% percentiles of the travel time distribution for the considered "*p*" time interval can be calculated as:

$$\begin{aligned} \hat{q}_{1\,i,j}^{(p)} &= \hat{t}_{i,j}^{(p)} - (t_{i,j}^{(p-1)} - q_{1\,i,j}^{(p-1)}) \\ \hat{q}_{3\,i,j}^{(p)} &= \hat{t}_{i,j}^{(p)} + (q_{3\,i,j}^{(p-1)} - t_{i,j}^{(p-1)}) \end{aligned} \qquad (A8)$$

A.4. Sets with No Data

It can happen, particularly at nighttime, that there is no travel time observation for a particular itinerary ("$N_{i,j}^{(p)}$=0"). Note that in this case "$\alpha_{i,j}^{(p)}$=0" The process to obtain "$t_{i,j}^{(p)}$" in this situation is simple. Travel times are assumed to maintain a linear constant evolution from the two last time intervals. This assumption is considered valid if at least one of these last time intervals has some observations. Otherwise, travel time is set to a default free flow travel time. The resulting formulation for sets with no data is:

$$t_{i,j}^{(p)} = \begin{cases} t_{i,j}^{(p-1)} + \left(\dfrac{\alpha_{i,j}^{(p-1)} + \alpha_{i,j}^{(p-2)}}{2} \right) \cdot \left(t_{i,j}^{(p-1)} - t_{i,j}^{(p-2)} \right) & \text{if} \quad \alpha_{i,j}^{(p-1)} > 0 \quad \text{or} \quad \alpha_{i,j}^{(p-2)} > 0 \\ t_{i,j,0} & \text{o.w.} \end{cases} \qquad (A9)$$

In relation to the interquartile range, if "$\alpha_{i,j}^{(p-1)} > 0$" or "$\alpha_{i,j}^{(p-2)} > 0$" then it is calculated in the same way as in Section A.3.3. Otherwise the interquartile range is set to the default value "$IQR_{i,j,0}$".

Taking into account the fact of the related values of "*0*" for the reliability indicator "$\alpha_{i,j}^{(p)}$" of the measurement, does not give any specific weight to these measurements in the smoothing equation for the next time intervals.

APPENDIX A3-A: REFERENCES

Daganzo, C. F., 1997. *Fundamentals of transportation and traffic operations.* Pergamon. Elsevier Science Ltd.

Li, R., Rose, G., Sarvi, M., 2007. Using automatic vehicle identification data to gain insight into travel time variability and its causes. *Transportation Research Record* 1945, 24-32.
Transportation Research Board, 2000. *Highway capacity manual*, Special Report 209, National Research Council, Washington, DC.

APPENDIX A3-B: ACCURATE FORMULATION OF THE BASIC ALGORITHM

Figure B.1 represents a zoom of the "*i, i+1*" section of Figure 2. From this figure it can be seen that Equation 4 and by extension Equations 1, 3, 6, 7, 9, 11, 21, and 29 are not accurate.

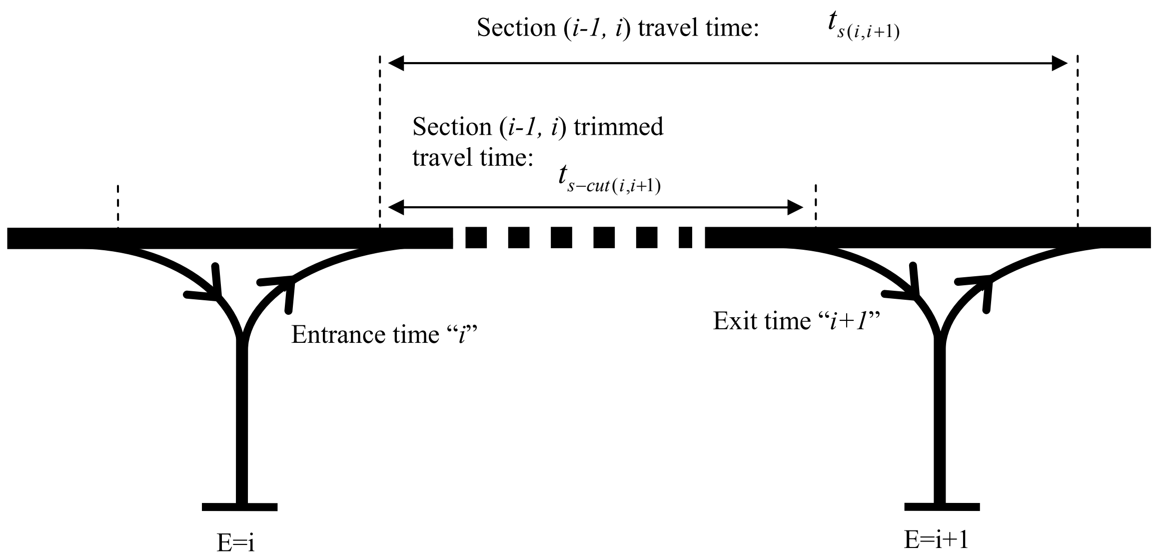


FIGURE B1 Detailed sketch of a highway junction.

To be accurate, Equation 6 should be rewritten as:

$$t_{ex(i+1)} = t_{i,i+1} - t_{s-cut(i,i+1)} - t_{en(i)} \tag{B.1}$$

Where the single section trimmed travel time can be obtained as:

$$t_{s-cut(i,i+1)} = \frac{t_{s(i,i+1)} \cdot l_{s-cut(i,i+1)}}{l_{s(i,i+1)}} \tag{B.2}$$

With "$l_{s(i,i+1)}$" and "$l_{s-cut(i,i+1)}$" the section length and the trimmed section length respectively.

Applying the same modification to the two-section travel time, Equation 11 should be rewritten as:

$$t_{ex2(i+2)} = t_{i,i+2} - t_{s-cut(i,i+2)} - t_{en(i)} \tag{B.3}$$

$$t_{s-cut(i,i+2)} = \frac{t_{s(i,i+2)} \cdot \left(l_{s(i,i+1)} + l_{s-cut(i+1,i+2)}\right)}{l_{s(i,i+1)} + l_{s(i+1,i+2)}} \qquad (B.4)$$

These modifications, which in practice have little consequence due to the magnitude of the section length in relation to the inner junction highway length, can also be applied to Equations 3, 9, 21 and 29.

This trimmed single section travel time can also be applied in the reconstruction of vehicle itineraries. Note that when reconstructing a trajectory by adding single section travel times, the last section travel time to consider must be a trimmed one.

APPENDIX A3-C: FUSION OF DIFFERENT ESTIMATIONS OF SINGLE SECTION TRAVEL TIMES.

From the basic and extended algorithm, different estimations of the single section travel times are obtained. Specifically, in an off-line application of the algorithm the following can be obtained:

- "$m-(i+1)$" first order estimations of single section travel times "$t_{s(i,i+1)}^{(p)}$"
- "$m-(i+2)$" second order estimations of two section travel times "$t_{s(i,i+2)}^{(p)}$", which result in "$2\cdot\left(m-(i+2)\right)$" single section second order travel times "$t_{s2(i,i+1)}^{(p)}$".

In case of real time application, "*m*", the ordinal number identifying the last exit in the highway should be replaced by its limitation "$i+n_i^{*(p)}$" (see Section 4.1).

C.1. Fusion of First Order Single Section Travel Time Estimations

The objective of this data fusion is to obtain a representative average of the different estimations of a particular single section travel time. To do so, a simple weighted average is applied. The two aspects to consider in the determination of the weighting factor for each one of the estimations are:

- As the length of the itineraries used to estimate a single section travel time increase, the reliability of the estimation decreases. This accounts for the greater variance of the travel time distribution in longer itineraries, and for the decrease of the trajectories overlapping zone (see Section 4.1).
- An accuracy indicator of each single section travel time estimation can be obtained as the minimum of the statistical significances in the travel time itineraries that take part in the calculation, "$\min\left(\alpha_{i,j}^{(p)}, \alpha_{i+1,j}^{(p)}\right)$".

Under these assumptions, the fused first order single section travel time is obtained as[1]:

$$t_{s1(i,i+1)}^{(p)} = \frac{\sum_{j=i+2}^{m}\left\{\left[\frac{1+\min\left\{\alpha_{i,j}^{(q)}\ ,\ \alpha_{i+1,j}^{(q)}\right\}}{\ln(j-i)}\right]\cdot\left[(t_{i,j}^{(q)}-t_{en(i)})-(t_{i+1,j}^{(q)}-t_{en(i+1)})\right]\right\}}{\sum_{j=1+2}^{m}\left[\frac{1+\min\left\{\alpha_{i,j}^{(q)}\ ,\ \alpha_{i+1,j}^{(q)}\right\}}{\ln(j-i)}\right]} \tag{C.1}$$

Recall from Section 4.2 that "*q*" equals "*p*" for the real time application.

Again, the real time application version of Equation C.1 is obtained by simply replacing "$i + n_i^{*(p)}$" instead of "*m*" and "*p*" instead of "*q*".[2]

Taking into account the conc[e]pts detailed in section 3.1 and appendix B, the first order exit time is obtained as:

$$t_{ex1(i+1)}^{(p)} = t_{i,i+1}^{(p)} - t_{s1-cut(i,i+1)}^{(p)} - t_{en(i)}^{(p)} \tag{C.2}$$

Finally, an accuracy indicator of these fused first order estimations is defined in Equation C.3. These accuracy indicators only consider the first two estimations of the single section travel time, which are obtained from the shortest itineraries.

$$\begin{aligned} \alpha_{s1(i,i+1)}^{(p)} &= \max\left\{\min\left\{\alpha_{(i,i+2)}^{(q)},\ \alpha_{(i+1,i+2)}^{(q)}\right\},\ \min\left\{\alpha_{(i,i+3)}^{(q)},\ \alpha_{(i+1,i+3)}^{(q)}\right\}\right\} \\ \alpha_{ex1(i+1)}^{(p)} &= \min\left\{\alpha_{(i,i+1)}^{(p)},\ \alpha_{s1(i,i+1)}^{(p)}\right\} \end{aligned} \tag{C.3}$$

C.2. Fusion of Second Order Single Section Travel Time Estimations

Proceeding in the same way and under the same assumptions of the previous section, the second order fused two-section travel time is obtained as:

$$t_{s2(i,i+2)}^{(p)} = \frac{\sum_{j=i+3}^{m}\left\{\left[\frac{1+\min\left\{\alpha_{i,j}^{(q)}\ ,\ \alpha_{i+2,j}^{(q)}\right\}}{\ln(j-i)}\right]\cdot\left[(t_{i,j}^{(q)}-t_{en(i)})-(t_{i+2,j}^{(q)}-t_{en(i+2)})\right]\right\}}{\sum_{j=i+3}^{m}\left[\frac{1+\min\left\{\alpha_{i,j}^{(q)}\ ,\ \alpha_{i+2,j}^{(q)}\right\}}{\ln(j-i)}\right]} \tag{C.4}$$

[1]Note the change in notation in Appendix C in relation to section 3.1 of the main text. "$t_{s1(i,i+1)}^{(p)}$" stands for the fused first order single section travel time. Likewise, "$t_{s2(i,i+1)}^{(p)}$" stands for the fused single section travel time coming from the second order algorithm. Finally "$t_{s(i,i+1)}^{(p)}$" is the notation for the fused first and second order single section travel time. This is the information to be disseminated. Same notation criteria applies to the exit times. Section 3.1 notation is maintained for the clarity of the concepts.

[2] This modification applies for all the remaining equations of this appendix. The default equations are presented in its off-line version.

To obtain the related fused second order single section travel time, recall Equations 13, 14, 15 and 16:

$$\left.\begin{aligned} t^{(p)}_{s2(i,i+1)} &= t^{(p)}_{s2(i,i+2)} - t^{(p)}_{s1(i+1,i+2)} \\ t^{(p)}_{s2(i+1,i+2)} &= t^{(p)}_{s2(i,i+2)} - t^{(p)}_{s2(i,i+1)} \end{aligned}\right\} \quad if \quad \alpha^{(p)}_{s1(i+1,i+2)} \geq \alpha^{(p)}_{s1(i,i+1)} > 0$$

$$\left.\begin{aligned} t^{(p)}_{s2(i+1,i+2)} &= t^{(p)}_{s2(i,i+2)} - t^{(p)}_{s1(i,i+1)} \\ t^{(p)}_{s2(i,i+1)} &= t^{(p)}_{s2(i,i+2)} - t^{(p)}_{s2(i+1,i+2)} \end{aligned}\right\} \quad if \quad \alpha^{(p)}_{s1(i,i+1)} > \alpha^{(p)}_{s1(i+1,i+2)} > 0 \qquad (C.5)$$

$$\left.\begin{aligned} t^{(p)}_{s2(i,i+1)} &= \frac{t^{(p)}_{s2(i,i+2)} \cdot l_{s(i,i+1)}}{l_{s(i,i+2)}} \\ t^{(p)}_{s2(i+1,i+2)} &= \frac{t^{(p)}_{s2(i,i+2)} \cdot l_{s(i+1,i+2)}}{l_{s(i,i+2)}} \end{aligned}\right\} \quad if \quad \alpha^{(p)}_{s1(i,i+1)} = \alpha^{(p)}_{s1(i+1,i+2)} = 0$$

Finally, an accuracy indicator of these fused second order estimations of the single section travel times can also be defined as:

$$\begin{aligned} \alpha^{(p)}_{s2(i,i+1)} &= \alpha^{(p)}_{s2(i+1,i+2)} = \min\left\{\alpha^{(p)}_{(i,i+3)}, \alpha^{(p)}_{(i+2,i+3)}, \max\left\{\alpha^{(p)}_{s1(i,i+1)}, \alpha^{(p)}_{s1(i+1,i+2)}\right\}\right\} \\ \alpha^{(p)}_{ex2(i+2)} &= \min\left\{\alpha^{(p)}_{(i,i+2)}, \alpha^{(p)}_{s2(i+1,i+2)}\right\} \end{aligned} \qquad (C.6)$$

C.3. Fusion of First and Second Order Single Section Travel Times

To obtain the final estimation of the single section travel time "$t^{(p)}_{s(i,i+1)}$", it is only necessary to calculate a weighted average of first and second order estimations. The weighting factors are the accuracy indicators of each one of the estimations.

$$t^{(p)}_{s(i,i+1)} = \frac{\alpha^{(p)}_{s1(i,i+1)} \cdot t^{(p)}_{s1(i,i+1)} + \alpha^{(p)}_{s2(i,i+1)^+} \cdot t^{(p)}_{s2(i,i+1)^+} + \alpha^{(p)}_{s2(i,i+1)^-} \cdot t^{(p)}_{s2(i,i+1)^-}}{\alpha^{(p)}_{s1(i,i+1)} + \alpha^{(p)}_{s2(i,i+1)^+} + \alpha^{(p)}_{s2(i,i+1)^-}} \qquad (C.7)$$

Where the superscripts (+) and (-) in the second order estimations refer to the two possibilities of obtaining a single section travel time from a two-section travel time (i.e. "$t^{(p)}_{s2(i,i+1)}$" can be obtained from "$t^{(p)}_{s1(i,i+2)}$" and from "$t^{(p)}_{s1(i-1,i+1)}$").

Applying a similar weighted average to the exit times:

$$t^{(p)}_{ex(i+1)} = \frac{\alpha^{(p)}_{ex1(i+1)} \cdot t^{(p)}_{ex1(i+1)} + \alpha^{(p)}_{ex2(i+1)} \cdot t^{(p)}_{ex2(i+1)}}{\alpha^{(p)}_{ex1(i+1)} + \alpha^{(p)}_{ex2(i+1)}} \qquad (C.8)$$

Equations C.7 and C.8 are valid for "$i \in (0,...,m-1)$" (see Figure 2). Note that for "*i=0*" and for "*i=m-1*", some of the terms in equations C.7 and C.8 are not defined (i.e. "$t^{(p)}_{s2(i,i+1)^+}$" is not defined for "*i=m-1*", "$t^{(p)}_{s2(i,i+1)^-}$" and "$t^{(p)}_{ex2(i+1)}$" are not defined for "*i=0*").

In these situations, the related accuracy indicators have a value of zero, and equations C.7 and C.8 still hold.

APPENDIX A4

Highway travel time accurate measurement and short-term prediction using multiple data sources

ABSTRACT

The development of new traffic monitoring systems and the increasing interest of road operators and researchers in obtaining reliable travel time measurements, motivated by society's demands, have led to the development of multiple travel time data sources and estimation algorithms. This situation provides a perfect context for the implementation of data fusion methodologies to obtain the maximum accuracy from the combination of the available data.

This paper presents a new and simple approach for the short term prediction of highway travel times, which represent an accurate estimation of the expected travel time for a driver commencing on a particular route. The algorithm is based on the fusion of different types of data that come from different sources (inductive loop detectors and toll tickets) and from different calculation algorithms. Although the data fusion algorithm presented herein is applied to these particular sources of data, it could easily be generalized to other equivalent types of data.

The objective of the proposed data fusion process is to obtain a fused value more reliable and accurate than any of the individual estimations. The methodology overcomes some of the limitations of travel time estimation algorithms based on unique data sources, as the limited spatial coverage of the algorithms based on spot measurement or the information delay of direct travel time itinerary measurements when disseminating the information to the drivers in real time. The results obtained in the application of the methodology on the AP-7 highway, near Barcelona in Spain, are found to be reasonable and accurate.

In short, the travel time data fusion algorithm presented in this paper tries to be as simple as possible and yet still improve the existing naïve approaches.

Keywords: travel time estimation, data fusion, loop detectors, toll ticket data, Bayesian combination.The accuracy

1. INTRODUCTION

Most developed countries, unable to carry on with the strategy of expanding transportation infrastructures once they become saturated, are now focusing their efforts on the optimization of infrastructure usage by means of operational and management improvements. This policy results from environmental, budget and land occupancy limitations, the latter being especially restrictive in metropolitan areas where high population density is combined with increasing mobility needs of society.

The availability of accurate and reliable travel time information appears to be the key factor for an improved management of road networks, since it allows an effective estimation of traffic states and provides the most valuable and understandable information for road users [Palen, 1997]. This evidence has not gone unnoticed by some European countries (Spain, France, Denmark, Italy, Finland, United Kingdom, Sweden, the Netherlands, Norway and Germany) grouped under the Trans-European Road Network (TERN), which are currently developing travel time estimation projects [Hopkin et al., 2001].

This interest expressed by transportation agencies and highway operators in addition to the development of ITS (Intelligent Transportation Systems) has led to a new framework in traffic data management and has increased the variety of reliable, precise and economically viable road surveillance technologies [Klein, 2001; Skesz, 2001; Martin et al., 2003; US DOT, 2006]. In addition, the appearance of ATIS (Advanced Traveler Information Systems) has made possible a simple and efficient dissemination of information addressed to the road user.

This context results in a new situation where it isn't rare that for a particular highway stretch to have several traffic measurements coming from different available sources (primarily in congested metropolitan highways). In addition, different algorithms or methodologies have been developed to obtain travel time estimations from these traffic measurements, obtaining a remarkable accuracy, not without an extensive research effort in recent years. Turner et al. (1998) gives a comprehensive overview of these travel time estimation methods. However, each of these methods is flawed by the intrinsic characteristics of the original measurement. The availability of different travel time estimations from different data sources usually results in complementary flaws of the estimations. This opens up new horizons for data fusion techniques applied to different estimations of road trip travel time.

Researchers' interest in travel time data fusion techniques has been increasing since the late 90s. In the USA Palacharla and Nelson (1999) studied the application of fuzzy logic to travel time estimation, evaluating which hybrid system was more effective (i.e. the fuzzy based on a neural network or on an expert system). They conclude that the neural network hybrid system is more precise, increasing the quality of the results obtained with classical travel time estimation methods. A similar methodology was tested in 2006 by the Austrian Department of Traffic, Innovation and Technology [Quendler et al., 2006], whose objective was to obtain reliable travel times and to determine the congestion level of the road network using multiple data sources (inductive loop detectors, laser sensors and floating taxi cars). A reduction of 50% in the number of mistakes in the estimation of traffic state is claimed using this methodology known as ANFIS (Adaptable Neural Fuzzy Inference System). Later, Sazi-Murat (2006) relied on ANFIS to obtain delay times in signalized intersections, achieving better results than the Highway Capacity Manual (2000), mainly in heavily congested situations. Also Lin et al. (2004) and Tserekis (2006) deal with the short-term prediction of travel time in arterials, decomposing the total delay into link delay and intersection delay. The authors propose a simple model and prove a reasonable degree of accuracy under various traffic conditions and signal coordination levels.

Researchers in Singapore and China have tried to obtain predictions of traffic stream states using Bayesian inferences on a neural structure from a unique source of data. The results improve those using simple neural networks in 85% of the situations [Weizhong et al, 2006]. Park and Lee (2004), both Koreans, have obtained travel time estimations in urban areas by implementing neural networks and Bayesian inferences, both independently, and using data from inductive loops and floating cars. In both cases the results are considered promising.

In France, researchers have developed conceptually simple data fusion techniques. The best examples are the works of El Faouzi (2005a, 2005b) and El Faouzi and Simon (2000) in the evidential Dempster-Shafer inference, which could be considered a generalization of Bayesian theory, improving the results of classical Bayes theories in pilot test runs on a highway near Toulouse. Two sources of data were used: license plate matching and inductive loop detectors. These experiences are being used by French highway operators for the estimation of travel times in their corridors [Ferré, 2005; AREA, 2006; Guiol and Schwab, 2006].

Swedish and Scottish road operators (SRA - Sweden Road Administration and Transport for Scotland) have since 2001 been analyzing the implantation of data fusion systems to obtain road travel times in their networks. The Scottish pilot test on the A1 motorway in the surroundings of Edinburgh uses up to 4 data sources: tracking of cellular phones, inductive loop detectors, floating car data and license plate matching. Surprisingly, the cellular phone tracking, despite its lack of location accuracy in dense urban environments, stands out for its reliability [Peterson, 2006; Scott, 2006].

In the Netherlands, van Lint et al. (2005) use neural networks for the prediction of travel times with gaps in the data, obtaining satisfactory results in spite of this partial information. Recently, van Hinsbergen and van Lint (2008) propose a Bayesian combination of travel time short-term prediction models in order to improve the accuracy of the predictions for real time applications of this information. Results are promising, but further research is recommended increasing the number and diversity of the models to combine.

In this context, the present paper proposes a new data fusion approach for travel time estimation in order to provide real time information to drivers entering a highway.

Note that to accomplish this objective, not only accurate measurement is necessary, but also short term prediction of travel times [Rice and van Zwet, 2001]. A two level fusion process is proposed. On the one hand, the first fusion level tries to overcome the spatial limitations of point measurements, obtaining a representative estimation for the whole stretch. On the other hand, the second level of fusion tries not only to measure but also to predict travel times in order to achieve a more reliable estimation for the real time dissemination of the information.

This methodology results in a simple travel time data fusion algorithm, which when implemented on top of existing data collection systems, allows us to exploit all the available data sources and outperforms the two most commonly used travel time estimation algorithms. The results of a pilot test on the AP-7 highway in Spain are outlined in the paper and show that the developed methodology is sound.

The paper is organized as follows: section two describes the different natures of travel time measurements and highlights the main objective and contribution of the paper. In section three, the basic and simple algorithms used to obtain the source travel time data to fuse are presented. Section four describes the methodology used for the development of the two level data fusion system. Section five presents the results of the application of the model from the AP-7 highway in Spain. Finally, some general conclusions and issues for further research are discussed in Section six.

2. TRAVEL TIME DEFINITIONS

There are two main methodologies used to measure travel time on a road link: the direct measurement and the indirect estimation. The direct travel time measurement relies on the identification of a particular vehicle on two points of the highway. These control points define the travel time target stretch. By simply cross-checking the entry and exit times of the identified vehicles, the travel time from one point to the other is obtained. The data collection techniques used in this first approach are defined by the identifying technology. Identification by means of license plates or using the toll tags ID (i.e. on-board electronic devices to pay the toll at turnpikes equipped with electronic toll collection –ETC– systems) are commonly used. These technologies are grouped under the AVI (Automated Vehicle Identification) systems. Usual problems in this direct travel time measurement are obtaining a representative number of identifications (this problem is particularly severe when the identification tag given to the vehicles has a low market penetration; take as an example the case of the TransGuide system in San Antonio, Texas, where the AVI tags were given solely for the purpose of estimating travel times [SwRI, 1998]) or the elimination of frequent outliers (e.g. a driver stops for a break). In addition, and probably the most important shortcoming of direct measurements for real time applications, is the information delay. Travel time measurements are obtained once the vehicle has finished its itinerary. Therefore, a direct travel time measurement results in what will be named an MTT (i.e. Measured Travel Time) and represents a measurement of a past situation involving a delay in the real time dissemination of information (equal to the travel time). Obviously this drawback turns out to be more severe as travel time increases (i.e. long itineraries or congestion episodes).

The alternative is the indirect travel time estimation. It consists of measuring any traffic flow characteristic (usually the fundamental variables- flow, density and speed) on some particular points of the target stretch, and applying some algorithm in order to obtain the travel time estimation. Under this configuration, every portion of the length of the stretch is assigned to one of the measurement spots, and it is assumed that the point measurement reflects the homogeneous traffic conditions of the whole portion. Obviously the main problem with these types of methods is the lack of fulfilment of this last assumption, mainly in heavy traffic conditions when traffic conditions can vary dramatically within the assigned portion of highway. To limit the effects of this drawback, a high density of detection sites are necessary, reducing the length of highway assigned to each detector. Other limitations of these methods are the lack of accuracy of the detection technologies used to measure traffic variables, considering the inductive loop detectors as the most widely used.

In this indirect travel time estimation, the itinerary travel time (i.e. the total travel time in the target stretch) is obtained by the addition of the travel times in the portions of highway that configure the stretch. This total itinerary measure, named as ITT (i.e. Instantaneous Travel Time) in the present paper, uses only the last available data and assumes that traffic conditions will remain constant in each section indefinitely. The main advantage of ITT is the immediacy in travel time data, reflecting the very last events on the highway. Usually, this “last minute” information is considered as the best estimation of future traffic evolution. This leads ITT estimation to be used as a naïve approach to the predicted travel time (PTT), which represents an estimation of the expected travel time for a driver entering the highway at a particular instant.

Note that in fact the ITT is a virtual measurement in the sense that no driver has followed a trajectory from which this travel time results. Particularly interesting are the works of van Lint and van der Zijpp (2003) and Li et al. (2006) where the authors propose and evaluate different algorithms in order to reconstruct the real vehicle trajectories from only ITT measurements in order to obtain an estimation of the MTT. The key question in this trajectory offline reconstruction (i.e. the information delay does not matter, only accuracy, as real time information is not the objective) is how the spot speeds are generalized over space. Constant, linear or smoothed approximations are evaluated in the referenced papers.

By way of illustration, Figure 1 shows an example of the implications of different trip travel time constructions. The information delay in the case of trip MTT involves very negative effects in the case of dramatic changes in traffic conditions during this time lag (e.g. an incident happens). The construction of trip ITT by means of section travel times reduces this information delay and the resulting travel time inaccuracies. As the traffic conditions do not remain constant until the next single section travel time update, the trip ITT also differs from the true PTT.

Figure 1 only aims to show the benefits of ITT immediacy in travel time prediction and the drawbacks of the MTT delayed information. In practice, the usual lack of accuracy of section travel times added up to obtain the itinerary ITT, can spoil all the benefits, resulting in situations where MTT would be just as good/bad as ITT. In addition it is also possible that some particular evolutions of traffic state result in paradoxical situations where, although an accurate ITT estimation being available, this does not imply being a better approach to PTT in relation to MTT.

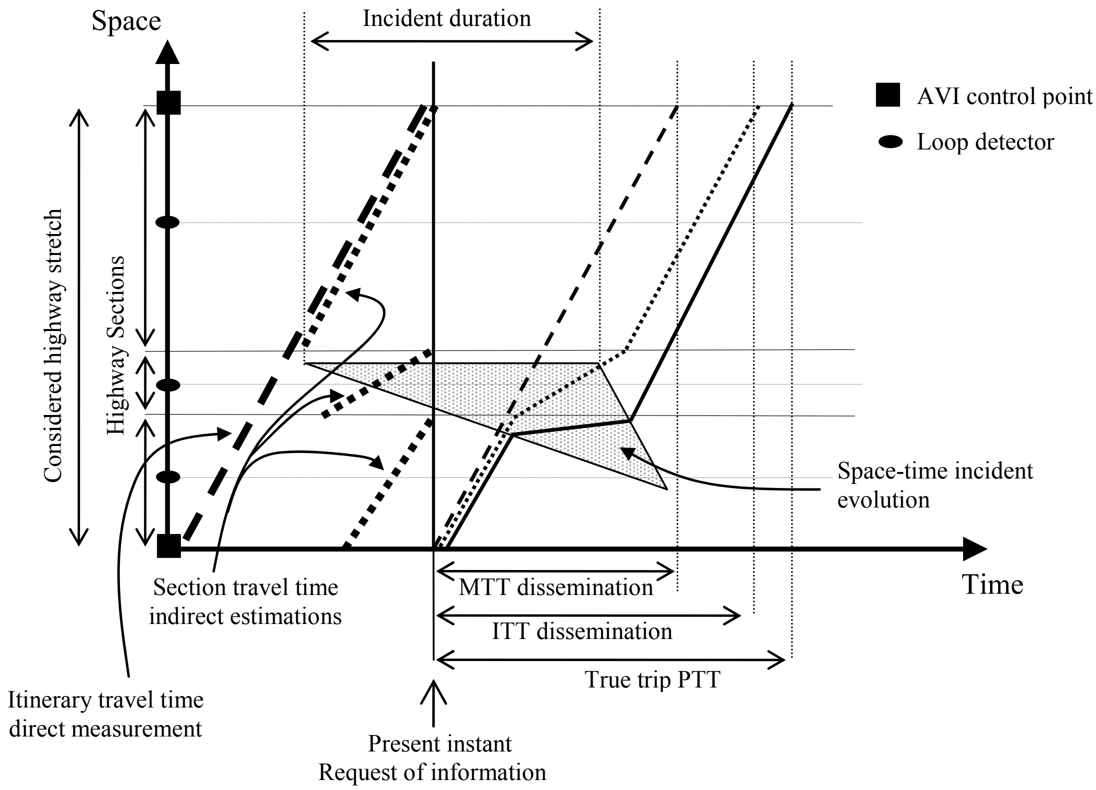


FIGURE 1 Travel time definitions and its implications in the dissemination of the information.

3. NAÏVE TRAVEL TIME ESTIMATION ALGORITHMS

Although not being the objective of this research, it is necessary for the complete understanding of the process to describe how the original data to fuse are obtained. These original data consist of two different ITT estimations and one MTT measurement. It is worthwhile to highlight the fact that the algorithms to obtain these first travel time estimations and the technologies that provide the source measurements presented in this section are in no way limiting, and could be substituted for any other procedure or technology that provide the same type of result (i.e. two ITT estimations and one MTT measurement).

In this context, this section presents an algorithm meant to obtain travel times from spot speed measurements (obtained from dual loop detectors), an algorithm to estimate travel times from loop detector traffic counts and a procedure to obtain travel time measurements from toll ticket data. The accuracy of these algorithms is not the objective of the research. Therefore, the proposed algorithms only explore the concept of estimation and are applied in their simplest version. Several improvements of these algorithms could be found in the literature, but these are not considered here. Recall that the objective is to increase the accuracy of these original estimations (whatever they are) by means of combining them.

3.1. Spot speed algorithm for travel time estimation

As stated earlier, this method is based on the speed measurement on a highway spot by means of dual electromagnetic loop detectors. Travel time could then be obtained by simply applying the following equation:

$$T_{1(i,t)} = \frac{l_i}{\bar{v}_{(i,t)}} \tag{1}$$

Where: $T_{1(i,t)}$ is the average travel time in the highway section "i" and time interval (t-$1,t$), obtained from algorithm 1 (spot speed algorithm – 1st ITT estimation)

l_i is the length of section "i", the portion of the highway stretch considered to be associated with loop detector "i" (see Figure 3).

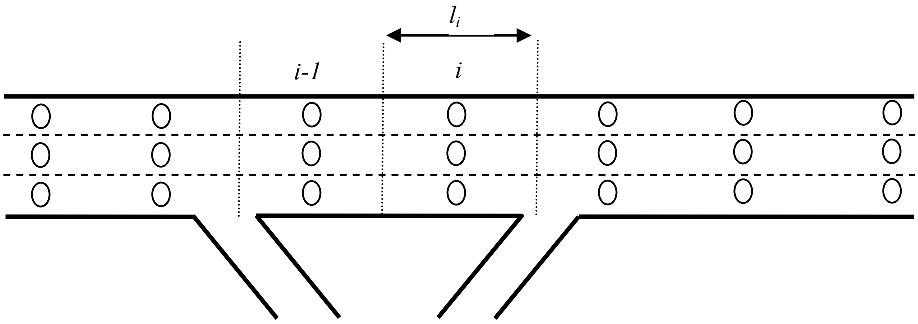
$\bar{v}_{(i,t)}$ is the spatial mean speed measured in loop detector "i" and time interval (t-$1,t$), calculated as the harmonic mean of the "$n_{(i,t)}$" individual vehicles' speed "$v_{k(i,t)}$" measured during (t-$1,t$).

$$\bar{v}_{(i,t)} = \frac{n_{(i,t)}}{\sum_{k=1}^{n_{(i,t)}} \frac{1}{v_{k(i,t)}}} \tag{2}$$

The hypothesis considered in the application of this algorithm is that traffic flow characteristics stay constant on the whole stretch and throughout the whole time period. To limit the dramatic effects of this first source of error, a high density of surveillance and a frequent actualization of variables is needed. As an order of magnitude, this algorithm provides reasonably accurate estimations by itself when there is one loop detector every 500m, and the updating interval is less than 5 minutes.

In addition, in congestion situations with frequent stop-and-go traffic, the measured spatial mean speed can be very different from the real mean speed of traffic flow, as detectors only measure the speed of the vehicles when they are moving, and do not account for the time that vehicles are completely stopped. As a result of this flaw, travel time estimations using this algorithm in congested situations can be largely underestimated. To overcome this problem, different smoothing schemes could be applied. For example averaging the measured speed in loop detector "i" and time interval (t-$1,t$) with previously measured speeds in time and space. These evolutions of the algorithm are not considered here, where the lack of accuracy of spatial mean speed in congested situations is taken into account increasing the margin of error of this travel time estimation.

Finally, in the real-world application of this type of algorithm, usually a third source of error arises. This results from the fact that the common practice at Traffic Management Centres is to compute the time mean speed of traffic stream (i.e. arithmetic average of individual vehicle speeds) instead of the space mean speed (i.e. harmonic average detailed in Equation 2), the one that relates distance with average travel time (Equation 1). A local time mean speed structurally over estimates the space mean because faster observations are overrepresented [Daganzo, 1997]. Therefore average travel times computed in this situation will be slightly underestimated. Space mean speeds are considered to be available in the rest of the paper.



O: loop detectors every 500 m.

FIGURE 2 Required surveillance configuration to apply only the spot speed algorithm.

3.1.1. Expected error of the spot speed algorithm

As will be described in the next section, an important element in the first fusion level is the expected error of each one of the individual values to fuse. Hence, it is necessary to define the expected error of the ITT algorithms.

As stated before, the expected error in the spot speed algorithm for the travel time estimation arises mainly from two reasons: the spatial generalization of a point measurement and the lack of accuracy of this point measurement. The magnitude of the first source of error is estimated by altering the assigned portion of highway of each detector to the most unfavorable situation. In addition, it is considered that a loop detector can overestimate the mean speed of traffic in 100% in stop&go situations (i.e. half the time moving and half the time stopped). Stop&go traffic is considered likely to occur when the measured mean speed at the detector site falls below 80 km/h. Equations 3 to 5 together with the sketch in Figure 3 describe this margin of error.

$$T_{1(i,t)}^{\min} = \frac{l_i^{(-)} \cdot}{\max(\bar{v}_{(i-1,t)}, \bar{v}_{(i,t)})} + \frac{l_i^{(+)} \cdot}{\max(\bar{v}_{(i,t)}, \bar{v}_{(i+1,t)})} \tag{3}$$

$$T_{1(i,t)}^{\max} = \frac{l_i^{(-)} \cdot}{\min(\bar{v}_{(i-1,t)}^{\min}, \bar{v}_{(i,t)}^{\min})} + \frac{l_i^{(+)} \cdot}{\min(\bar{v}_{(i,t)}^{\min}, \bar{v}_{(i+1,t)}^{\min})} \tag{4}$$

Where:

$$\bar{v}_{(i,t)}^{\min} = \begin{Bmatrix} \bar{v}_{(i,t)} & if\ \bar{v}_{(i,t)} \geq 80\, km/h \\ 0.5 \cdot \bar{v}_{(i,t)} & if\ \bar{v}_{(i,t)} < 80\, km/h \end{Bmatrix} \tag{5}$$

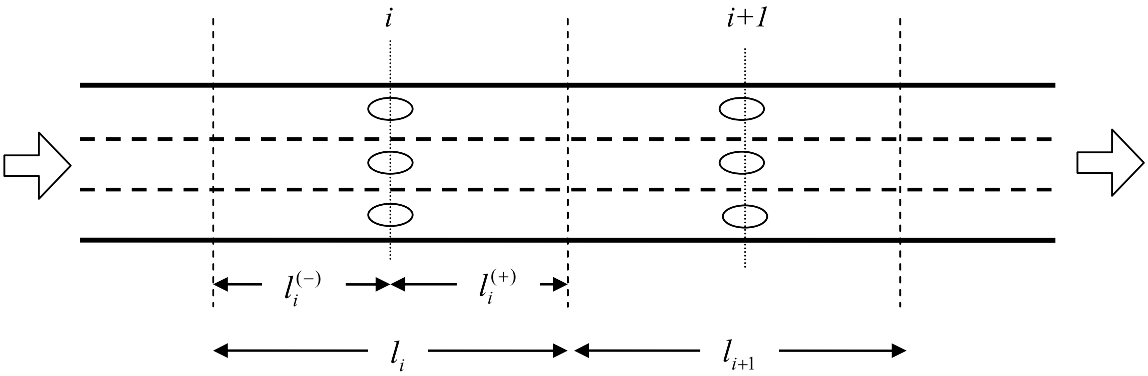


FIGURE 3 Determining the expected error of the spot speed algorithm.

The error incurred in case of using time mean speeds (instead of space mean speeds) in congested conditions, is included in the "measurement" error of loop detectors in stop&go situations. In free flowing conditions, this situation implies a slight underestimation of travel times (approximately 2%), which can be considered negligible in relation to the spot speed generalization error.

3.2. Cumulative flow balance algorithm for travel time estimation

An alternative exists for estimating travel times from loop detector data in highway stretches that do not benefit from the required surveillance density needed for the application of the spot speed algorithm. The cumulative flow balance algorithm estimates travel time directly from loop detector traffic counts, without the previous calculation of speed. This solves the problem of the lack of accuracy in the mean speed estimation in congested situations, and allows the usage of single loop detectors with the same accuracy as the double loop in the traffic counts, but they are unable to accurately estimate vehicle speeds. The algorithm uses the entrance and exit flows in the highway stretch in order to calculate the travel time by using a simple flow balance method. To apply this algorithm, all the highway ramps must be equipped with loop detector units. The surveillance scheme required is displayed in Figure 4, where "$n^t_{(i,t)}$" is the input traffic count of the main highway trunk in the time interval (*t-1,t*), "$n^t_{(i+1,t)}$" is the trunk output count and "$n^r_{(i,t)}$" and "$n^r_{(i+1,t)}$" are the entering and exiting counts through the ramps comprised between detectors "*i*" and "*i+1*".

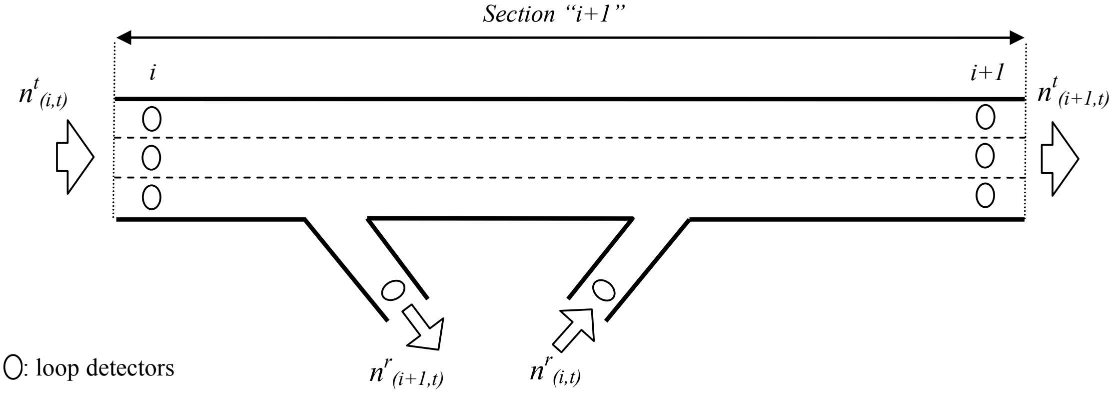


FIGURE 4 Required surveillance configuration to apply the cumulative flow balance algorithm.

Then the total entering and exiting flows are:

$$n_{(i,t)} = n_{(i,t)}^{t} + n_{(i,t)}^{r}$$
$$n_{(i+1,t)} = n_{(i+1,t)}^{t} + n_{(i+1,t)}^{r} \qquad (6)$$

The travel time estimation from these traffic counts is depicted in the N-curve diagram (cumulative counts vs. time) sketched in Figure 5, where:

$N_{(i,t)}$ is the cumulative traffic count that have entered the highway section at time "*t*".

$N_{(i+1,t)}$ is the cumulative traffic count that have exited the highway section at time "*t*".

$$N_{(i+1,t)} = \sum_{j=1}^{t} n_{(i+1,j)} \qquad (7)$$

$S_{(i+1,t)}$ is the vehicles accumulation in the Section "*i+1*" of the highway at time "*t*".

$t_{n(i+1,t)}$ is the travel time of the vehicle counting the number "n" at detector "*i+1*", that exits at time "*t*"

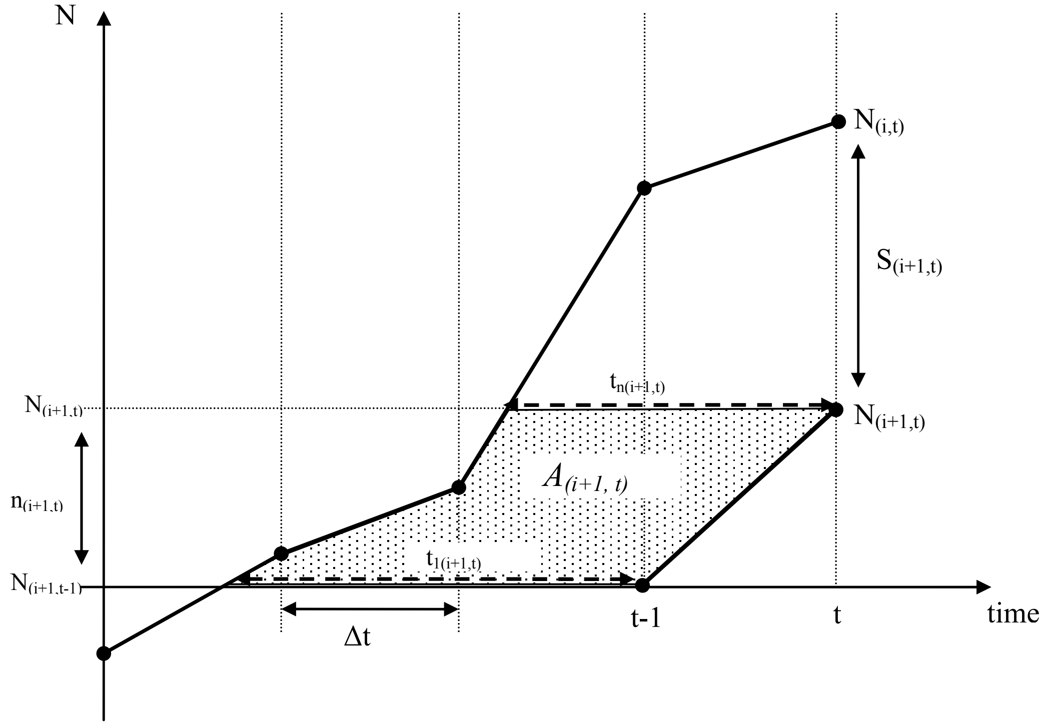


FIGURE 5 N-Curve diagram depicting the cumulative flow balance algorithm.

The average travel time in the time interval (*t-1, t*) can be estimated by calculating the shadowed area in Figure 5 and dividing it by the number of vehicles that have exited the section within this time interval.

$$T_{2(i+1,t)} = \frac{A_{(i+1,t)}}{n_{(i+1,t)}} \tag{8}$$

The subscript "2" in "$T_{2(i+1,t)}$" refers to the fact that this average travel time has been obtained using algorithm 2 (cumulative flow balance algorithm – 2nd ITT estimation).

Again, this is one of the simplest formulations of the algorithm, and several modifications can be applied, for instance to take into account the actual vehicle accumulation at time "t" ($S_{(i+1,t)}$). A literature review of these methods is given in Nam and Drew (1996).

The main problem with this type of formulations is the lack of exact accuracy in the detector counts, or specifically, the error in the relative counts of two consecutive detectors (i.e. all the vehicles entering the section must exit given enough time, and therefore the differences in cumulative counts between consecutive detectors should tend to zero when flows reduce to almost zero). This phenomenon is known as loop detector drift, and its effects are greatly magnified when using relative cumulative counts in consecutive detectors, due to the accumulation of the systematic drift with time. A simple drift correction has been applied to account for this fact, requiring that all the vehicles entering the section during a complete day must have exited on the same day. The drift correction factor is formulated as:

$$\delta_{(i+1)} = \frac{N_{(i,24h)}}{N_{(i+1,24h)}} \tag{9}$$

Then the corrected count at detector "i+1" is equal to:

$$N_{(i+1,t)} = N_{(i+1,t-1)} + \delta_{(i+1)} \cdot n_{(i+1,t)} \tag{10}$$

3.2.1. Expected error of the cumulative count algorithm

The main source of error in this travel time estimation algorithm is the detector drift. Despite using a "δ" factor to account for this flaw, this correction is based on "historic" measurements of the two detectors and considers the average detector drift over an entire day, which do not have to correspond exactly to the current drift of the detectors in a particular traffic state. Variations of +/- 0.5% in this drift correction factor should not be considered as rare. Then, the margin of error of the travel time estimations using this algorithm can be formulated as:

$$N^{\min}_{(i+1,t)} = N_{(i+1,t-1)} + (\delta_{(i+1)} - 0.005) \cdot n_{(i+1,t)} \tag{11}$$

$$N^{\max}_{(i+1,t)} = N_{(i+1,t-1)} + (\delta_{(i+1)} + 0.005) \cdot n_{(i+1,t)} \tag{12}$$

Finally,

$$T_{2(i+1,t)}^{\max} = \frac{A_{(i+1,t)}^{\max} \quad (use \ \ N_{(i+1,t)}^{\min})}{n_{(i+1,t)} \quad (use \ \ N_{(i+1,t)}^{\min})} \tag{13}$$

$$T_{2(i+1,t)}^{\min} = \frac{A_{(i+1,t)}^{\min} \quad (use \ \ N_{(i+1,t)}^{\max})}{n_{(i+1,t)} \quad (use \ \ N_{(i+1,t)}^{\max})} \tag{14}$$

3.3. Travel time estimation from toll ticket data

Travel time data can be directly obtained by measuring the time taken for vehicles to travel between two points on the network. On toll highways, the data needed for the fee collection system, can also be used for travel time measurement, obtaining an MTT.

On a highway with a "closed" tolling system, the fee that a particular driver has to pay at the toll plaza varies depending on his itinerary (origin-destination). In contrast, in an "open" highway system, toll plazas are strategically located so that all drivers pay the same average fee at the toll gate. In a closed highway system, each vehicle entering the highway receives a ticket (real -usually a card with magnetic band- or virtual -using an ETC device-), which is collected at the exit. The ticket includes the entry point, and the exact time of entry. By cross-checking entry and exit data, the precise time taken by the vehicle to travel along the itinerary (route) can be determined. A similar procedure can be applied in a highway with an open toll system, by identifying the vehicle at two consecutive payment sites.

Averages can be obtained from the measurements for all the vehicles travelling along the same itinerary in the network during a time interval (*t-1,t*). For each particular vehicle "*k*" travelling along a highway, the travel time spent on its itinerary between "*i*" (origin) and "*j*" (destination) expressed as "$T_{i,j,k,t}$" can be obtained by matching the entry and exit information recorded on its toll ticket. The average travel time for the itinerary in a particular time interval can be obtained by averaging the travel times of all vehicles that have exited the highway within this time period and have travelled along the same itinerary "(*i,j)*".

$$T_{3(i,j,t)} = median\,(T_{i,j,k,t}\,) \quad \forall k \ exiting \ the \ stretch \ (i,j), during \ (t-1,t) \tag{15}$$

Where: $T_{i,j,k,t}$ is the travel time for the itinerary "*i,j*" for a particular vehicle "*k*" that has exited the highway stretch within the time interval (*t-1, t*).

$T_{3(i,j,t)}$ is the average travel time for the itinerary "*i,j*" in a particular time period (*t-1, t*). Subscript "3" stands for the usage of algorithm 3: travel time from toll ticket data. This is an MTT estimation.

Note that the median is considered instead of the arithmetic mean to exclude the negative effects of outliers, which would result in an overestimation of the travel times. From these calculations, the MTT (measured travel times) are obtained.

The error in these measurements appear in the case of a great fraction of outliers (e.g. vehicles stopping for a break while travelling through the target stretch) in relation to the total identified vehicles. This situation only occurs when the total number of identified vehicles in the time interval is low. This can happen during late night hours or in the case of frequent updating of the information. This is the reason why the updating time interval of MTT is usually not lower than 15 minutes. In this situation, the error in these measurements can be omitted as its magnitude is much lower than the one obtained from point estimations. A detailed description of the travel time estimation process using toll ticket data can be found in Soriguera et al. (2010).

4. DATA FUSION METHODOLOGY

The travel time data fusion process proposed in this paper is a two level fusion. In a first level, the two ITT estimations are fused. As stated before, one of these indirect travel time measurements is obtained from speed measurements, and the other from traffic counts. Both measurements could be obtained from loop detectors. The first fusion tries to overcome the main limitations of these travel time estimation algorithms, which are on the one hand the spatial coverage limitations of the spot speed algorithm and on the other hand the lack of exact accuracy of traffic counts.

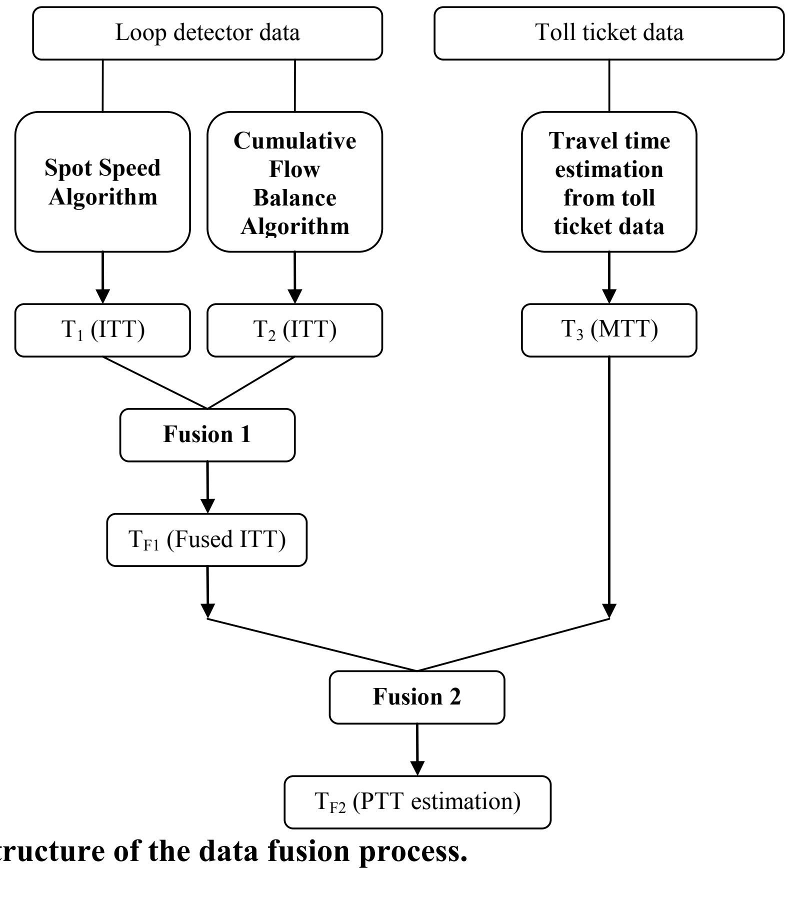


FIGURE 6 Structure of the data fusion process.

In the second fusion level, the resulting fused ITT from the first fusion is fused to an MTT, in the present application obtained from toll ticket data, which could also be obtained by any other identifying technology. The objective of this second fusion level is to increase the predictive capabilities of the accurately fused ITT by means of the information provided by the MTT, in order to obtain a better approximation to the true PTT. The data fusion structure can be seen in Figure 6.

4.1. First level data fusion

The first data fusion process fuses the two instantaneous travel times "T_1" (from a spot speed algorithm) and "T_2" (from a cumulative flow balance algorithm) to obtain "T_{F1}" a more accurate and uniform ITT in terms of spatial and temporal coverage. The objective of this first fusion is to reduce the flaws affecting each individual estimation using a more accurate fused estimation. Recall that these flaws are the spatial generalization and the accuracy of the speed measurements in the spot speed algorithm and the detector count drift in the cumulative flow balance algorithm.

The proposed fusion operator is a context dependent operator with constant mean behaviour [Bloch, 1996]. Since it is context dependent, it is necessary to define three contexts A, B and C. In each context the data fusion algorithm will follow a slightly different expression. Their definitions are:

- **A context**:

$$T^{\max}_{p(i,t)} \geq T^{\max}_{q(i,t)} \quad and \quad T^{\min}_{p(i,t)} \leq T^{\min}_{q(i,t)} \tag{16}$$

  where p, q *refers to the estimation algorithm* $\quad p, q = 1,2 \quad p \neq q$

- **B context**:

$$T^{\max}_{p(i,t)} \geq T^{\max}_{q(i,t)} \quad and \quad T^{\min}_{p(i,t)} \geq T^{\min}_{q(i,t)} \quad and \quad T^{\min}_{p(i,t)} \leq T^{\max}_{q(i,t)} \tag{17}$$

  where p, q *refers to the estimation algorithm* $\quad p, q = 1,2 \quad p \neq q$

- **C context**:

$$T^{\min}_{p(i,t)} \geq T^{\max}_{q(i,t)} \tag{18}$$

  where p, q *refers to the estimation algorithm* $\quad p, q = 1,2 \quad p \neq q$

Finally, the first level fused travel times are obtained applying the following fusion operator:

$$T_{F1(i,t)} = \begin{cases} T_{q(i,t)} & if\ A \\ \dfrac{T^{\max}_{q(i,t)} + T^{\min}_{p(i,t)}}{2} & if\ B\ or\ C \end{cases} \tag{19}$$

The analytic expression for the margin of error in this case is:

$$T_{F1(i,t)}^{\max} = \begin{cases} T_{\mathrm{q(i,t)}}^{\max} & \text{if A or B} \\ T_{\mathrm{p(i,t)}}^{\min} & \text{if C} \end{cases} \tag{20}$$

$$T_{F1(i,t)}^{\min} = \begin{cases} T_{q(i,t)}^{min} & if\ A \\ T_{p(i,t)}^{min} & if\ B \\ T_{q(i,t)}^{max} & if\ C \end{cases} \tag{21}$$

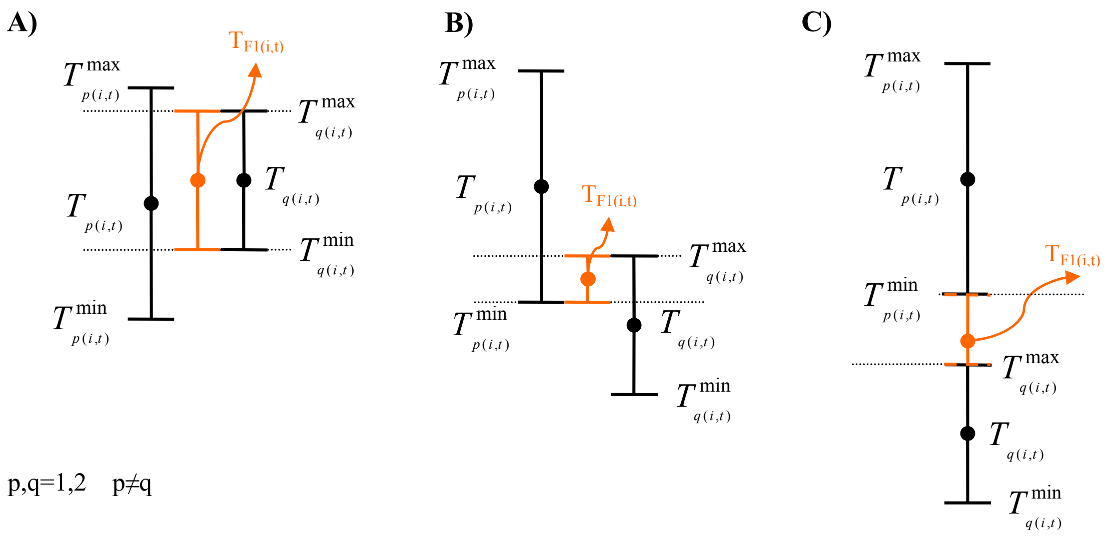


FIGURE 7 First level ITT fusion contexts.

In Figure 7 it is possible to observe that the operator is consistent and that its behaviour is clearly determined by the context. In the A and B contexts, the resulting error is smaller or equal to the smallest of the errors "ε_1" and "ε_2" (defined as the difference between the maximum and the minimum possible travel times) while in context C this could not be true. Context C must always be avoided since it represents flawed behaviour of the algorithm in the determination of "T_1", "T_2" or "ε_1" and "ε_2". This means that each original estimation algorithm needs a minimum accuracy and a correct estimation of the margin of error for the correct functioning of the fusion algorithm.

4.2. Second level data fusion

The second data fusion process starts with the fused ITT (T_{F1}) and the MTT from toll tickets (T_3). On the one hand, "T_{F1}" can be considered an accurate real time measurement of the travel time in the target stretch, and could be seen as a picture of the travel time in the highway stretch reflecting the traffic state in the (*t-1, t*) time interval. Recall that this complete picture is composed of several partial pictures, each one representing a portion of the stretch (Figure 1). The complete picture is obtained by simply joining these partial ones. On the other hand, the MTT, (T_3), results from direct travel time measurements, and due to its direct measurement nature it can also be considered as an accurate measurement (especially if the time interval lasts long enough to obtain a representative median).

The problem with these accurate travel time measurements is the time lag that exists between them and in relation to the final objective of the estimation: the predicted travel time (PTT) that will take a vehicle just entering the highway stretch. "T_{F1}" (ITT) is a real time picture (but nothing ensures that the traffic will remain constant during the next time interval), while "T_3" (MTT) results from the trajectories of vehicles that have recently travelled through the stretch, and therefore it is a measurement of a past situation, particularly in long itineraries. In this context the objective of the second fusion level is to gain knowledge of PTT from the two outdated accurate measurements.

4.2.1. Spatial and temporal alignment

Unlike in fusion 1, in the fusion 2 process the information is not provided by the same data source, and hence, the data will not be equally located in space and time. Therefore, a spatial and temporal alignment is needed before the data can be fused.

Usually the distance between AVI control points (the toll tag readers in this case) is greater than the spacing of loop detectors. In this situation, the spatial alignment of the measurements is directly given by the construction of the ITT from smaller section travel times (obtained from the fusion of point measurements) that match the MTT target stretch. This paper assumes this context and that the target highway stretch to estimate travel times fits with the stretch limited by these AVI control points (see Figure 1). In practice, these assumptions represent a common situation.

In relation to the temporal alignment, generally the updating frequency of ITT measurements is higher than the MTT one. This results from the fact that the accuracy of MTTs increases with the duration of the time intervals considered (which are the inverse of the updating frequency). The chosen dimension for the temporal alignment is the smallest time interval (corresponding to ITT), while the MTT is maintained constant until the next update (MTT remains constant during several ITT updates, due to its lower updating frequency). These concepts are described in more detail in Figure 8 and Table 1.

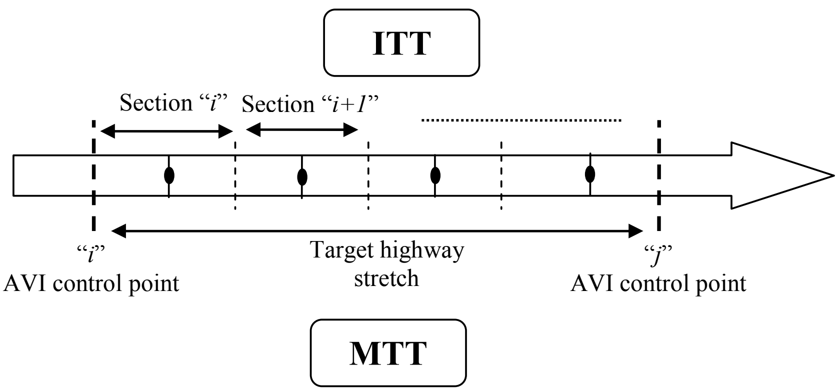


FIGURE 8 Spatial alignment.

TABLE 1 Original Data for the Spatial and Temporal Alignment

				Travel time data estimations				Spatial and temporal aligned data to fuse
				Space				
Time	MTT	ITT	ITT	Section "*i*"	Section "*i+1*"	Section "*i+2*"	Section "*i+3*"	
			MTT	Stretch (*i,j*)				
	ΔT	Δt	ITT	$T_{F1(i,t)}$	$T_{F1(i+1,t)}$	$T_{F1(i+2,t)}$	$T_{F1(i+3,t)}$	$T_{F1(i,t)}+T_{F1(i+1,t)}+T_{F1(i+2,t)}+T_{F1(i+3,t)}$
			MTT	$T_{3(i,j,T)}$				$T_{3(i,j,T)}$
		Δt	ITT	$T_{F1(i,t+1)}$	$T_{F1(i+1,t+1)}$	$T_{F1(i+2,t+1)}$	$T_{F1(i+3,t+1)}$	$T_{F1(i,t+1)}+T_{F1(i+1,t+1)}+T_{F1(i+2,t+1)}+T_{F1(i+3,t+1)}$
			MTT					$T_{3(i,j,T)}$
		Δt	ITT	$T_{F1(i,t+2)}$	$T_{F1(i+1,t+2)}$	$T_{F1(i+2,t+2)}$	$T_{F1(i+3,t+2)}$	$T_{F1(i,t+2)}+T_{F1(i+1,t+2)}+T_{F1(i+2,t+2)}+T_{F1(i+3,t+2)}$
			MTT					$T_{3(i,j,T)}$
		Δt	ITT	$T_{F1(i,t+3)}$	$T_{F1(i+1,t+3)}$	$T_{F1(i+2,t+3)}$	$T_{F1(i+3,t+3)}$	$T_{F1(i,t+3)}+T_{F1(i+1,t+3)}+T_{F1(i+2,t+3)}+T_{F1(i+3,t+3)}$
			MTT					$T_{3(i,j,T)}$
		Δt	ITT	$T_{F1(i,t+4)}$	$T_{F1(i+1,t+4)}$	$T_{F1(i+2,t+4)}$	$T_{F1(i+3,t+4)}$	$T_{F1(i,t+4)}+T_{F1(i+1,t+4)}+T_{F1(i+2,t+4)}+T_{F1(i+3,t+4)}$
			MTT					$T_{3(i,j,T)}$
		Δt	ITT	$T_{F1(i,t+5)}$	$T_{F1(i+1,t+5)}$	$T_{F1(i+2,t+5)}$	$T_{F1(i+3,t+5)}$	$T_{F1(i,t+5)}+T_{F1(i+1,t+5)}+T_{F1(i+2,t+5)}+T_{F1(i+3,t+5)}$
			MTT	$T_{3(i,j,T+1)}$				$T_{3(i,j,T+1)}$

Note: Δt is the time interval between ITT updates, while ΔT is the time interval between MTT updates. In this table it is assumed that $\Delta T=5\cdot\Delta t$.

4.2.2. Second level fusion operator

Once ITT (T_{F1}) and MTT (T_3) are spatially and temporally aligned it is possible to apply the second fusion algorithm. This operator uses the probabilistic logic [Bloch, 1996], based on Bayes' Theory:

$$p(PTT\mid ITT,MTT)=\frac{p(PTT,ITT,MTT)}{p(ITT,MTT)}=\frac{p(MTT\mid PTT,ITT)\cdot p(PTT,ITT)}{p(ITT,MTT)}=$$
$$=\frac{p(MTT\mid PTT,ITT)\cdot p(ITT\mid PTT)\cdot p(PTT)}{p(ITT,MTT)} \qquad (22)$$

As ITT and MTT are independent measurements, then results:

$$p(PTT\mid ITT,MTT)=\frac{p(MTT\mid PTT)\cdot p(ITT\mid PTT)\cdot p(PTT)}{p(ITT)\cdot p(MTT)} \qquad (23)$$

Where "$p(PTT\mid ITT,MTT)$" is the conditional probability of PTT given ITT and MTT, "$p(MTT\mid PTT)$" and "$p(ITT\mid PTT)$" are respectively the conditional probabilities of MTT and ITT given PTT and finally "$p(PTT)$", "$p(ITT)$" and "$p(MTT)$" are the individual probabilities of each estimation.

The probabilities in the right hand side of Equation 23 must be obtained by a statistical analysis of the calibration samples. This means that, in order to apply this second fusion, a previous period of "learning" of the algorithm is needed. This off-line period of learning with sample data, which can be seen as a calibration of the algorithm, needs the three travel time estimations at a time to calculate all the conditional and individual probabilities. Note that obtaining the true PTT in an off-line basis is not a problem, because a PTT is only a future MTT, available in an off-line context. Hence, PTTs and MTTs are the same values but with a time lag between observations equal to the travel time.

In the determination of these probabilities, ITT and MTT are rounded up to the next whole minute, in order to obtain more representative relations. This does not affect the quality of the results, because the user perception of travel time is never lower than this minute unit.

Once the "$p(PTT \mid ITT, MTT)$" probabilities are determined, a maximum posterior probability decision rule is chosen. This means that given "T_{F1}" and "T_3" (the original ITT and MTT data for the second level fusion), the selected PTT, "T_{F2}" is the one which maximizes the conditional probability. The running of the fusion algorithm is very simple because after the spatial and temporal alignment, it is only necessary to check the table of probabilities and to obtain the corresponding fused PTT.

The decision to leave a result void is taken if the probability value does not overcome a threshold defined by the user of the system. This situation denotes little probability that ITT and MTT values coincide in the same section and time interval (e.g. it is slightly probable that ITT= 1 min. and MTT=15 min.). A great number of voids in the running phase of the second fusion reveal great weaknesses of the original travel time estimation algorithms.

From Bayes' Theory it is also possible to obtain the accuracy of the result. Since when multiplying conditional probabilities part of the sample information gets lost, the uncertainty of the result (I) related with a pair of ITT and MTT, could be defined as:

$$I(ITT, MTT) = 1 - p(PTT \mid ITT, MTT) \qquad (24)$$

The goal of any travel time estimation system should be the reduction of this uncertainty, as this parameter is a good reliability indicator of the final result.

5. APPLICATION TO THE AP-7 HIGHWAY IN SPAIN

The data fusion technique proposed in this paper was tested on the AP-7 toll highway in Spain. The AP-7 highway runs along the Mediterranean coast corridor, from the French border to the Gibraltar Strait. Nevertheless, the pilot test was restricted to the north east stretch of the highway between the "La Roca del Vallès" and the "St. Celoni" toll plazas, near Barcelona. This stretch is approximately 17 km long.

The surveillance equipment installed on this stretch of the highway consists of 4 double loop detectors (i.e. approximately an average of 1 detector every 4 km). Moreover, the tolling system installed on the highway allows the direct measurement of travel times in the stretch. The duration of the loop detector data updating interval is 3 minutes, while the MTT are only updated every 15 minutes.

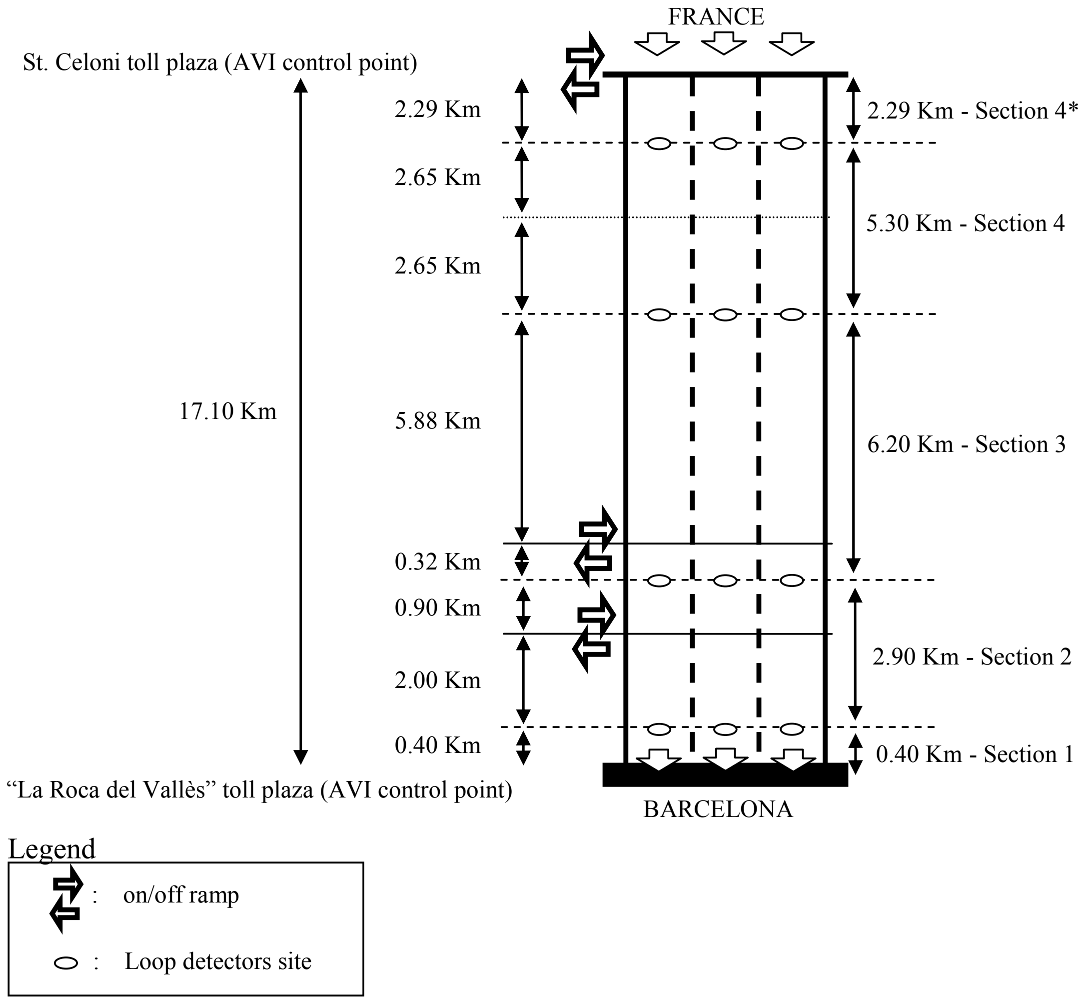


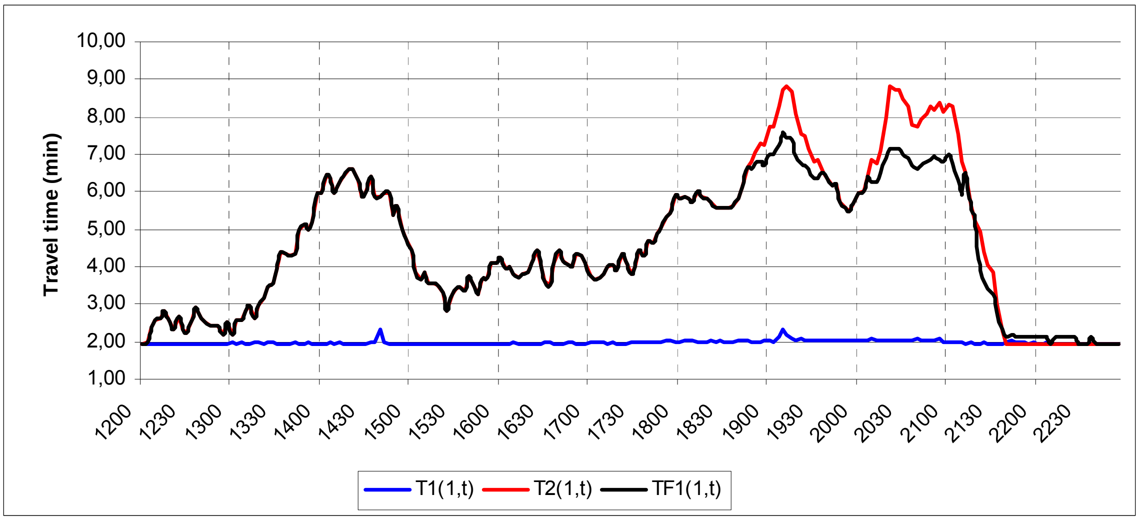
FIGURE 9 Surveillance equipment installed on the test site.

The pilot test was performed with the June 4th 2007 afternoon and evening data in the southbound direction towards Barcelona. This was a very conflictive period in terms of traffic, as it was a sunny holiday Monday in June, a time when a lot of people use this stretch of the AP-7 highway to return to Barcelona after a long weekend on the coast. The learning of the second fusion algorithm was carried out with data of a similar period from Sunday May 27th 2007.

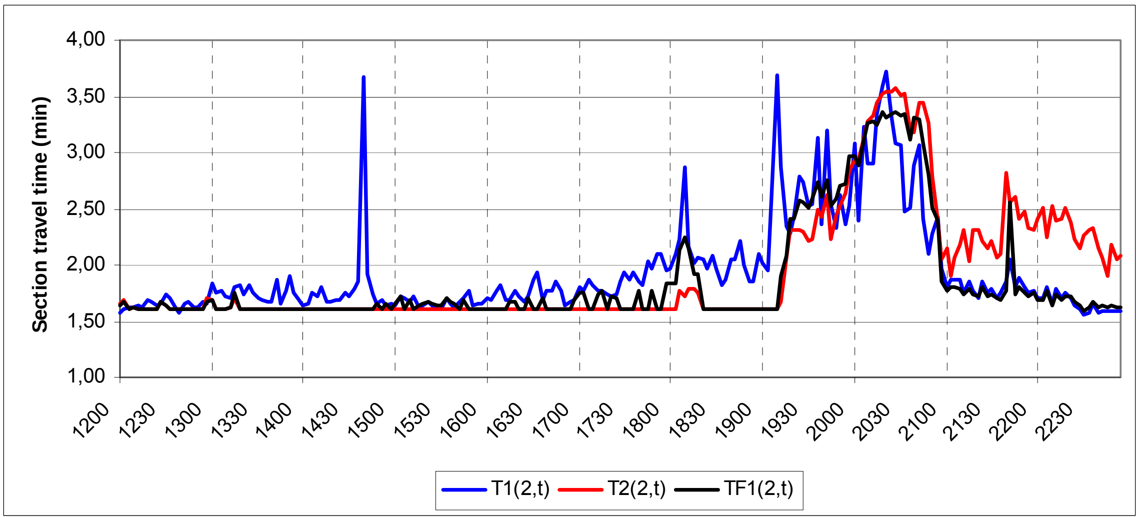
5.1. First level fusion results

Figure 10 (a to e) shows the results of the spot speed travel time estimation algorithm “$T_{1(i,t)}$” (in the figure notation), the cumulative flow balance algorithm, “$T_{2(i,t)}$” and the results of applying the first level fusion operator, “T_{F1}”, to these pairs of data in each section of the target stretch. Recall that all the information used in this level comes only from the speed and traffic count measurements at loop detector sites. Note that travel time estimations for section 4* are only available using the spot speed algorithm

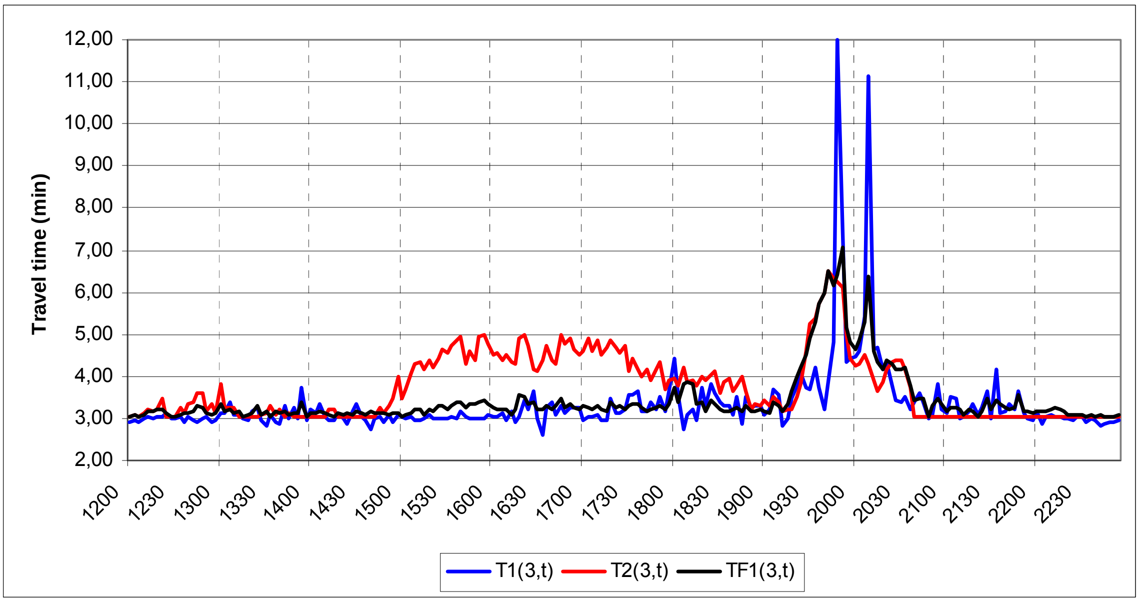
To evaluate the accuracy of the fusion operator it is necessary to compare these fused travel times to the real travel times, only available for the total stretch. Note that these real travel times are in fact the final objective of the estimation (i.e. the travel time of a vehicle obtained when the vehicle is entering the highway), which could not be obtained in real time application of the algorithm, as would correspond to future information. In an offline application (like the present evaluation) these real travel times are solely the MTT (from toll ticket data) moved backwards in time a time lag equal to the experienced travel time. Only for this particular evaluation purpose, the MTT used to represent the true travel time were obtained on a three minute basis. This is shown in Figure 11, where the travel time in the stretch, resulting from the reconstructed trajectories from ITT estimations in every section (from each algorithm alone and from the fused one) are compared to the real travel times in the stretch.



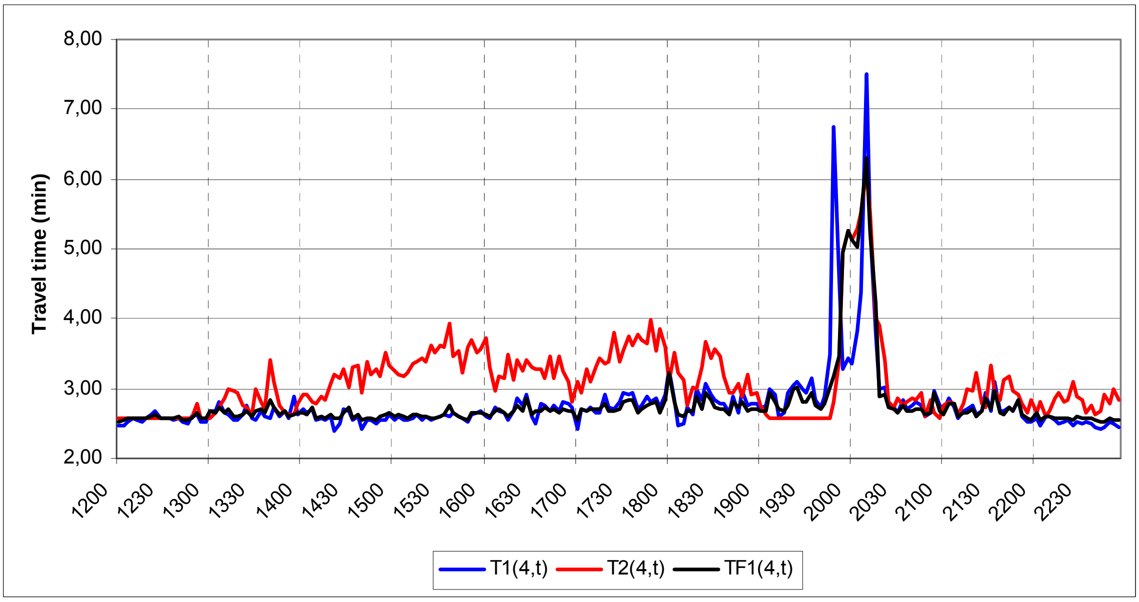
a)



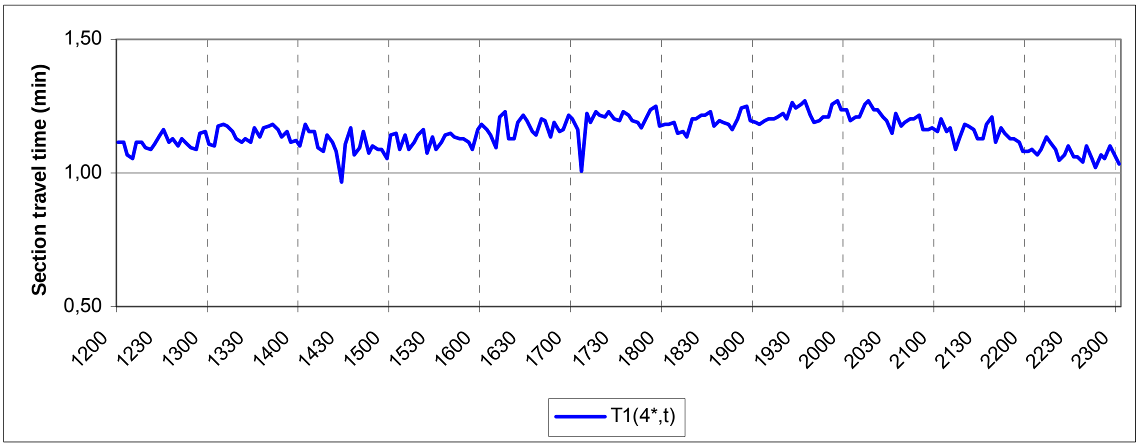
b)


Travel time (min)
12,00
11,00
10,00
9,00
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1200
1230
1300
1330
1400
1430
1500
1530
1600
1630
1700
1730
1800
1830
1900
1930
2000
2030
2100
2130
2200
2230
T1(3,t)
T2(3,t)
TF1(3,t)

c)


Travel time (min)
8,00
7,00
6,00
5,00
4,00
3,00
2,00
1200
1230
1300
1330
1400
1430
1500
1530
1600
1630
1700
1730
1800
1830
1900
1930
2000
2030
2100
2130
2200
2230
T1(4,t)
T2(4,t)
TF1(4,t)

d)



e)

FIGURE 10 First level fusion results on the AP-7 highway, June 4th 2007 data. a) Section 1 travel times. b) Section 2 travel times. c) Section 3 travel times. d) Section 4 travel times. e) Section 4* travel times.

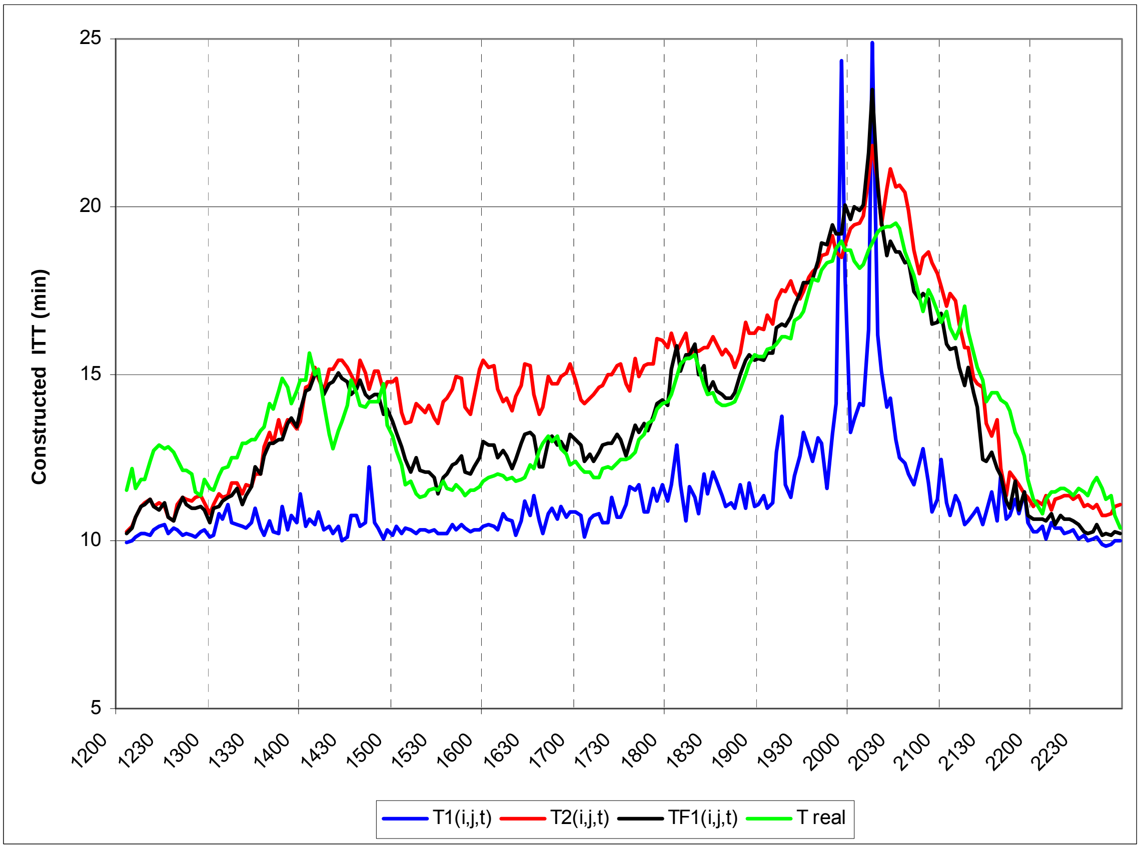


FIGURE 11 First level fusion results on the AP-7 highway from "St. Celoni" to "La Roca del Vallès", Spain, June 4th 2007 data.

Several comments arise from Figures 10 a) to e). Firstly, the inability of the spot speed algorithm to accurately describe the travel time variations resulting from the spatial evolution of jammed traffic can be clearly seen. A clear example is part a) of Figure 10. In this first section of the stretch, vehicles stop at "La Roca del Vallès" toll plaza to pay the toll fee. From 13.00h until 21.45h there were long queues to cross the toll gates. These queues of stopped traffic were not long enough to reach the detector site, 400 m upstream of the toll plaza. Take into account that the three lanes of the highway turn to more than 20 at the toll plaza, in order to achieve the necessary service rate and enough storage capacity near the plaza to avoid the growth of the queues blocking the on/off ramp 2.4 km upstream. This situation results in a great underestimation of travel times if using only the spot speed algorithm.

This same drawback of the spot speed algorithm cause the sharp travel time increases that can be seen in parts b) to d) of Figure 10. Using this algorithm, travel times remain next to the free flow travel times, unaware that the congestion is growing downstream and within the assigned section of highway and obviously underestimating the travel times, until the jam reaches the detector site, when the speed falls abruptly and the travel time sharply increases. But in this situation, the algorithm considers that the whole section is jammed (when upstream of the detector traffic could be flowing freely). This results in an overestimation of travel times.

The cumulative flow balance algorithm exhibits a smoother behaviour. Recall that the problems in this case arise due to the detector drift. Although a correction for the drift is applied taking into account the historic drift between each pair of detectors, it seems that this algorithm overestimates travel times in some periods. This is due to a higher drift in some periods of the day tested in relation to the historical observations.

From Figure 11 and numerical results in Table 2 it can be stated that great improvements were achieved with the first fusion operator with reductions of the mean estimation errors throughout the day. However, the maximum errors in a particular time slice slightly increased. This probably occurs due to an overestimation of the lower threshold of the spot speed algorithm margin of error, which should be lower. Note that this minimum travel time in congestion situations does not account for the possibility of congested conditions on two consecutive detector sites but traffic is free flowing in some portion of highway between them. This punctual and circumstantial increase of the maximum error does not jeopardize the great improvements achieved in global with the first fusion operator.

TABLE 2 Accuracy of the First Level of Fusion

Algorithm	Mean relative error	Mean absolute error (min)	Max. absolute error (min)	
			in excess	Lower
Spot speed	19.09 %	2.80	5.97	6.83
Cumulative	10.19 %	1.34	3.63	2.73
First fusion	6.01 %	0.81	4.60	2.93

6.2. Second level fusion results

Figure 12 shows the results of applying the second fusion operator to the test data. Recall that this fusion process considers only the whole target stretch for which MTT observations are available. The original data fused in this case are on the one hand the ITT resulting from the first fusion process "$T_{F1(i,j,t)}$", and on the other hand the MTT resulting from the toll ticket data"$T_{3(i,j,t)}$", updated only every 15 minutes. Both sets of data are rounded to the closest whole minute.

The results of the second fusion operator "$T_{F2(i,j,t)}$" are compared with the real PTT (Predicted Travel Time), that will suffer the drivers entering the stretch at that particular time. Recall that these real PTT could not be obtained in real time application of the algorithm, as they would correspond to future information.

Again, the results of this second fusion operator are promising, and better in terms of the overall functioning than in punctual estimations. The main criticism of this second level of fusion is the negative effects of the rounding to whole minutes, which causes sudden changes in the predicted travel time. This rounding is acceptable in terms of the diffusion of the information, where one minute accuracy is normally enough. However this rounding can vary the relative differences between the data to fuse in 1 minute, modifying the probabilities to consider and the corresponding result. As an issue for further research, a modified maximum posteriori probability decision rule should be analyzed. This decision rule should consist of taking into account (for instance as a weighted average) the occurrence probabilities of the two PTT values adjacent to the most likely one. This decision rule could diminish the negative effects of this necessary rounding.

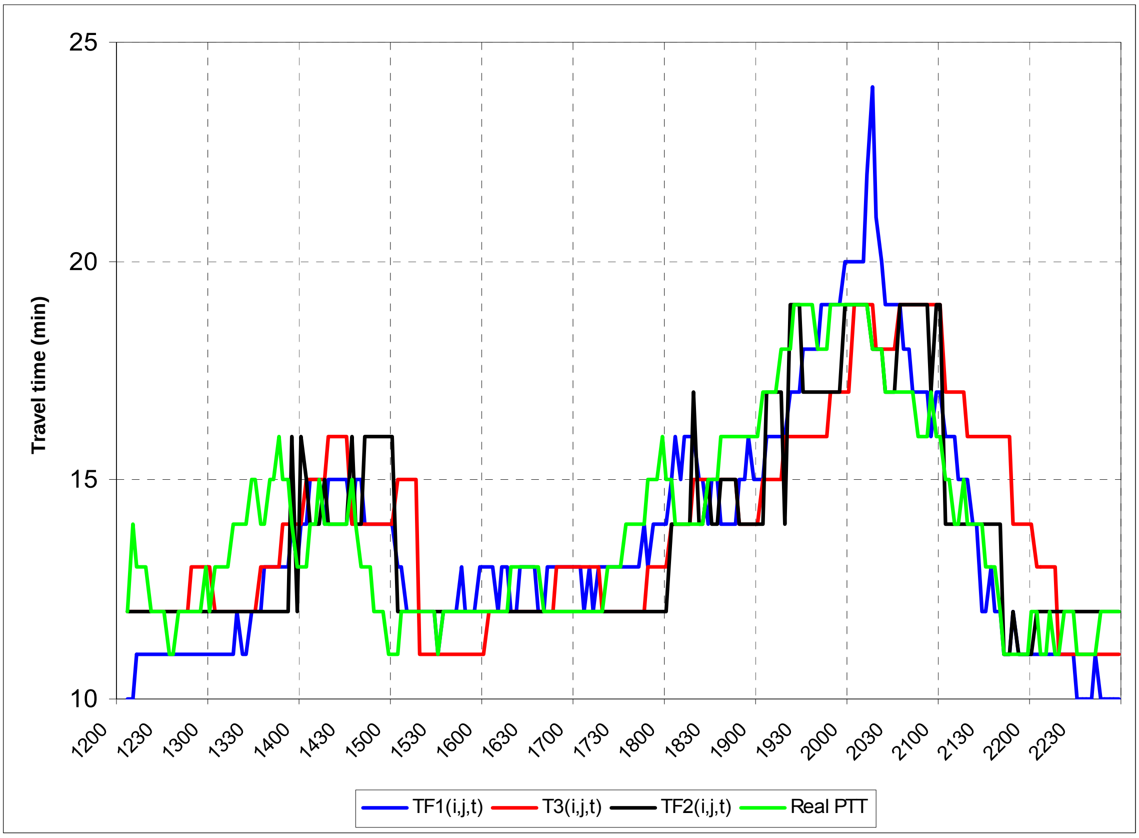


FIGURE 12 Second level fusion results on the AP-7 highway from "St. Celoni" to "La Roca del Vallès", Spain, June 4th 2007 data.

TABLE 3. Accuracy of the Second Level of Fusion

Algorithm	Mean relative error	Mean absolute error (min)	Max. absolute error (min)	
			in excess	Lower
F1 ITT	7.69 %	1,08	6	4
Toll Ticket MTT	10.33 %	1,43	5	5
Second Fusion	6.83 %	0.97	5	4

6. CONCLUSIONS AND FURTHER RESEARCH

This paper presents a simple approach for reliable road travel time estimation and short term prediction, using data fusion techniques. The objective is to obtain an accurate estimation of the travel time on a highway itinerary at the instant the driver enters the stretch. Therefore, short term forecasting is needed. The developed system can be easily put into practice with the existing infrastructure, and is able to use data obtained from any kind of sensor in any type of road link.

The proposed methodology needs several point estimations of travel times (obtained from loop detectors in the present application) and direct measurements of travel times in the target highway section (for example obtained from toll ticket data). The algorithms to obtain this original travel time data to fuse, although not the main objective of the paper, are also presented and discussed as an intermediate result.

The fusion algorithm is a two level process using both fuzzy logic and a probabilistic approach which implements the Bayes rule. The fused travel times are found to be more reliable than the initial ones and more accurate if the learning process is carefully developed.

The results of the pilot test carried out on the AP-7 highway in Spain indicate the suitability of the data fusion system for a better usage of the different surveillance equipment already installed on the roads.

Further developments are possible with the model, for example to analyze the effects of improving the accuracy of the source data to fuse in the final estimation, or defining the requirements in the learning process to improve the probabilistic fusion.

REFERENCES

AREA, (2006). Système d'information temps de parcours sur le réseau autoroutier AREA. *Asecap PULA 2006*, 105-115.

Bloch, I., (1996). Information combination operators for data comparative review with classification. *IEEE Transactions on Systems, Man and Cybernetics – Part A* **26**(1), 52-67.

Daganzo, C.F., (1997). *Fundamentals of transportation and traffic operations*. New York: Elsevier.

El Faouzi, N.E. and C. Simon, (2000). Fusion de données pour l'estimation des temps de parcours via la théorie de l'evidence. *Recherche transports sécurité* **68**,15-27.

El Faouzi, N.E., (2005). Bayesian and evidential approaches for traffic data fusion: Methodological issues and case study. *Proceedings of the 85th Transportation Research Board Annual Meeting*. Washington D.C.

El Faouzi, N.E., (2005). Fusion de données en I.T. Estimation multisource du temps de parcours. *1er Séminaire TISIC,* Paris, INRETS.

Ferré, J., (2005). *Quality control and fusion of road and weather data at SAPN* [online]. Centrico Briefing Note. Available from:

http://www.centrico.de/documents/briefing%20notes/20_Data_quality_control.pdf [Accessed 18 March 2007].

Guiol, R. and N. Schwab, (2006). Vers une information «temps de parcours» toujours plus performante et adaptée aux besoins de la clientèle: L'expérience ASF-ESCOTA. *Asecap PULA 2006*, 97-104.

Hopkin, J., D. Crawford, and I. Catling, (2001). *Travel time estimation*. Summary of the European Workshop organized by the SERTI project, Avignon.

Klein, L.A., (2001). *Sensor technologies and data requirements for ITS*. Boston: Artech House Publishers.

Li, R., G. Rose and M. Sarvi, (2006). Evaluation of speed-based travel time estimation models. *ASCE Journal of Transportation Engineering* **132**(7), 540-547.

Lin, W.H., A. Kulkarni, and P. Mirchandani, (2004). Short-term arterial travel time prediction for advanced traveller information systems. *Journal of Intelligent Transportation Systems* **8**(3), 143-154.

Martin, P.T., Y. Feng and X. Wang, (2003). *Detector technology evaluation*. Department of Civil and Environmental Engineering, University of Utah sponsored by US Department of Transportation.

Nam, D. H., and D.R. Drew, (1996). Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. *ASCE Journal of Transportation Engineering* **122**(3), 185-191.

Palacharla, P.V. and P.C. Nelson, (1999). Application of fuzzy logic and neural networks for dynamic travel time estimation. *International Transactions in Operational Research* **6**(1), 145-160.

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

Park, T. and S. Lee, (2004). *A Bayesian approach for estimating link travel time on urban arterial road network*. A. Laganà et al. (Eds.), Springer-Verlag Berlin Heidelberg.

Peterson, A., (2006). Travel times – A multi-purpose issue. *Proceedings of the 4th Conference of the Euro-regional Projects, I2Tern*. Barcelona.

Quendler, E., I. Kristler, A. Pohl, P. Veith, E. Beranek, H. Kubleka and J. Boxberger, (2006). Fahrerassistenzsystem LISA intelligente infrastruktur. V*ernetzungsworkshop des BMVIT* **3**(I2).

Rice, J. and E. van Zwet, (2001). A simple and effective method for predicting travel times on freeways. *Proceedings of the IEEE Conference on Intelligent Transportation Systems*. Oakland, CA, 227-232.

Sazi Murat,Y., (2006). Comparison of fuzzy logic and artificial neural networks approaches in vehicle delay modeling. *Transportation Research Part C* **14**(5), 316-334.

Scott, G., (2006). Scottish nacional journey system and development of data fusion. *Proceedings of the 4th Conference of the Euro-regional Projects, I2Tern*. Barcelona.

Skesz, S.L., (2001). *"State-of-the-art" report on non-traditional traffic counting methods*. Arizona Department of Transportation.

Soriguera, F., D. Rosas and F. Robusté. (2010) Travel time measurement in closed toll highways. *Transportation Research Part B* **44**(10), 1242–1267.

SwRI, (1998). *Automated vehicle identification tags in San Antonio: Lessons learned from the metropolitan model deployment initiative.* Science Applications International Corporation, McLean, VA.

TRB (Transportation Research Board), (2000). *Highway capacity manual.* National Research Council, Washington D.C.

Tsekeris, T., (2006). Comment on "Short-term arterial travel time prediction for advanced traveller information systems" by Wei-Hua Lin, Amit Kulkarni and Pitu Mirchandani. *Journal of Intelligent Transportation Systems* **10**(1), 41-43.

Turner, S.M., W.L. Eisele, R.J. Benz and D.J. Holdener, (1998). *Travel time data collection handBook*. Research Report FHWA-PL-98-035. Texas Transportation Institute, Texas A&M University System, College Park.

US Department of Transportation, (2006). *T-REX project: ITS unit costs database* [online]. Research and Innovative Technology Administration (RITA). Available from: http://www.itscosts.its.dot.gov [Accessed 18 March 2007].

van Hinsbergen, C.P.I. and J.W.C. van Lint, (2008). Bayesian combination of travel time prediction models. *Proceedings of the 87th Transportation Research Board Annual Meeting*, Washington D.C.

van Lint, J.W.C. and N. J. Van der Zijpp, (2003). Improving a travel time estimation algorithm by using dual loop detectors. *Transportation Research Record* **1855**, 41-48.

van Lint, J.W.C., S.P. Hoogendoorn and H.J van Zuylen, (2005). Accurate freeway travel time prediction with state-space neural networks under missing data. *Transportation Research Part C* **13**(5-6), 347-369.

Weizhong Z., D.H. Lee and Q. Shi, 2006. Short-term freeway traffic flow prediction: Bayesian combined neural network approach. *Journal of Transportation Engineering* **132**(2), 114-121.