NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF RURAL AND SURVEYING ENGINEERING
DEPARTMENT OF INFRASTRUCTURE AND RURAL DEVELOPMENT

Towards data–driven microscopic traffic simulation models

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

by

VASILEIA PAPATHANASOPOULOU
Graduate Surveying Engineer of NTUA

Supervisor: Constantinos Antoniou
Associate Professor NTUA

LABORATORY OF TRANSPORTATION ENGINEERING
Athens, April 2018

National Technical University of Athens
School of Rural and Surveying Engineering
Department of Infrastructure and Rural Development
Laboratory of Transportation Engineering

Towards data–driven microscopic traffic simulation models

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

by

VASILEIA PAPATHANASOPOULOU

Supervisor: Constantinos Antoniou
Associate Professor NTUA

Approved by:
Assoc. Prof., Constantinos Antoniou, NTUA (Doctoral research supervisor)
Prof. Haris Koutsopoulos, Northeastern University (Supervising committee member)
Assoc. Prof. Vincenzo Punzo, University of Naples Federico II (Supervising committee member)
Prof. George Yannis, NTUA
Assist. Prof. Eleni Vlahogianni, NTUA
Assist. Prof. Ioanna Spyropoulou, NTUA
Prof. Francisco Pereira, DTU

The approval of the doctoral dissertation by the School of Surveying Engineering of National Technical University of Athens does not imply acceptance of the author's views (N. 5343/ 1932, article 202).

Athens, April 2018

National Technical University of Athens
School of Rural and Surveying Engineering
Department of Infrastructure and Rural Development
Laboratory of Transportation Engineering

Acknowledgements

First of all, I would like to thank my supervisor Professor Constantinos Antoniou for his constant guidance, encouragement and support during all my studies.

I wish to thank all the members of the supervising committee for their valuable comments and suggestions. I would also like to thank Prof. Vincenzo Punzo from the University of Napoli – Federico II for kindly providing the Naples data and enhanced NGSIM data, as well as Prof. Tomer Toledo from Technion- Israel Institute of Technology for kindly making the Indian data used in this research freely available.

I am grateful to the Special Fund for Research Grants of NTUA, the Onassis Foundation Scholarship, and the EMPARCO project, for the financial support provided for my doctoral research.

I am pleased to thank my fellow researchers at NTUA, Ioulia, George, Christine, Harris and Thanassis, for their cooperation and most of them for their help in my study and daily life. I would also like to thank my friends, Themis, Maria, Stavroula and Panos, for their support during my studies.

Last but not least, I thank with all my heart my family, my father George, my mother Katerina and my sister Maria, for their endless motivation, encouragement and support.

Athens, 12 April 2018

V.P.

Abstract

The objective of this research is to develop more accurate, robust and reliable microscopic models. An integrated methodological framework based on non–parametric approaches is proposed for estimation of data–driven microscopic traffic simulation models. The methodology is implemented using different machine learning techniques such as clustering, classification, locally weighted regression, spline fitting, Gaussian processes, Kernel support vector machines and neural networks. The methodology is demonstrated using real trajectory data from three different sources and specifically an experiment from Naples, NGSIM data and non–lane disciplinary trajectory data from India. The focus is given on car–following models and Gipps' model, one of the most extensively used car–following models, is calibrated against the same data in order to be used as a reference benchmark. Many parameters affect driving behavior and it is explored how the performance of the models is improved by including more explanatory variables. Then, a practical and simple approach is developed and motivated for the online calibration of microscopic traffic simulation models, which considers dynamic parameters for individual drivers, in time and space. The model adapts to driving behavior in a rolling horizon and leads to less than 10% error in speed prediction even for ten steps into the future. This research also examines the feasibility and the benefits of using data–driven models on mixed traffic trajectory data, including non–lane discipline and heterogeneity in vehicle types, common characteristics in cities in developing countries. Although typical car–following models are theoretically justified, data–driven approaches are more flexible and allow the easy incorporation of additional information to the process of speed estimation. The results indicate that data–driven models could ensure reliability and improvement in estimation of microscopic models.

Keywords

Data–driven models, microscopic traffic simulation models, traffic modeling, machine learning, on–line calibration, NGSIM data, non–lane discipline

Περίληψη

Στόχος της έρευνας είναι η ανάπτυξη πιο αξιόπιστων μικροσκοπικών κυκλοφοριακών προτύπων. Αναπτύσσεται μια ολοκληρωμένη μεθοδολογία για την εκτίμηση προτύπων κυκλοφοριακής προσομοίωσης με τη χρήση καινοτόμων και ευέλικτων μεθόδων μηχανικής μάθησης, όπως η ταξινόμηση, η ομαδοποίηση, η τοπικά σταθμισμένη παλινδρόμηση (*loess*), οι καμπύλες *splines*, οι *Gaussian* διαδικασίες, οι διανυσματικές μηχανές υποστήριξης και τα νευρωνικά δίκτυα. Τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα αυτή περιλαμβάνουν δεδομένα από τρεις διαφορετικές πηγές, δεδομένα από τη Νάπολη, τα *NGSIM* δεδομένα και δεδομένα από την Ινδία. Δίνεται έμφαση στα πρότυπα ακολουθίας οχημάτων και για τα ίδια δεδομένα εφαρμόζεται το μοντέλο του *Gipps*, ένα γνωστό μοντέλο ακολουθίας οχημάτων που χρησιμοποιείται ως μοντέλο αναφοράς στην παρούσα έρευνα. Επειδή πολλοί παράγοντες επηρεάζουν τη συμπεριφορά του οδηγού, εξετάζεται κατά πόσο βελτιώνεται το μοντέλο ενσωματώνοντας περισσότερες μεταβλητές. Επιπλέον, εξετάζεται η δυναμική βαθμονόμηση κυκλοφοριακών προτύπων λαμβάνοντας υπόψη τη δυναμική μεταβολή των παραμέτρων για κάθε οδηγό, στον χρόνο και το χώρο. Οι παράμετροι μεταβάλλονται σε έναν κυλιόμενο χρονικό ορίζοντα και επιτυγχάνεται πρόβλεψη της ταχύτητας έως 10% για δέκα βήματα στο μέλλον. Διερευνάται η χρήση μοντέλων καθοδηγούμενων από τα δεδομένα σε συνθήκες μεικτής κυκλοφορίας χωρίς τήρηση των λωρίδων κυκλοφορίας και με μεγάλη ποικιλία ως προς τον τύπο των οχημάτων, κοινά χαρακτηριστικά των αναπτυσσόμενων χωρών. Αν και τα κλασσικά πρότυπα ακολουθίας οχημάτων είναι θεωρητικά τεκμηριωμένα, τα πρότυπα βασισμένα σε δεδομένα προσφέρουν μεγαλύτερη ευελιξία και επιτρέπουν την εύκολη ενσωμάτωση νέων μεταβλητών. Τα αποτελέσματα υποδεικνύουν ότι τα πρότυπα που βασίζονται σε δεδομένα μπορούν να συμβάλλουν στην εκτίμηση πιο αξιόπιστων μικροσκοπικών προτύπων.

Λέξεις κλειδιά

Μοντέλα βασισμένα σε δεδομένα, μικροσκοπικά πρότυπα κυκλοφοριακής προσομοίωσης, μοντελοποίηση κυκλοφορίας, μηχανική μάθηση, δυναμική βαθμονόμηση, δεδομένα *NGSIM*, μη τήρηση λωρίδων κυκλοφορίας

Summary

Motivation

The rapid development of technology has contributed to the availability of high–quality traffic data, leading the way for the development of more advanced simulation models. As a result of an explosive increase of the data that are being generated and collected, data–driven modeling is emerging as a fast developing field of transportation research. In big data era, it is important to be able to handle the available information to increase accuracy and reliability of traffic models. The emergence of new transportation modes (services and technologies), also new challenges for modeling traffic system, have created a need for more robust and advanced traffic simulation models. In particular, traffic simulation models need to capture the operations and interactions among new traffic systems. In a future world with a cooperative vehicle to vehicle and vehicle to infrastructure communication, all traffic modes and conditions need to be modeled. It is important to be able to offer solutions on–line and provide information and guidance back to drivers.

Objective

Transportation is experiencing a period of great development potential and changes, including new modes and new data sources. In an era of big data and autonomous vehicles, traffic simulation models need to adapt to these new challenges. The objective of this research is to provide an alternative modeling approach for traffic simulation models and develop more accurate, robust and reliable microscopic models. The proposed methodology aims to increase model flexibility and provide the opportunity for incorporation of additional parameters without the need to resort to cumbersome reformulations of conventional models functional form. The proposed modeling approach can take advantage of a wide range of available data, and is therefore suitable to implementation in the context of intelligent transportation systems.

Methodology

In this research an integrated methodological framework based on non–parametric approaches is proposed for estimation of data–driven microscopic traffic simulation models. The methodology is implemented using different computational approaches such as locally weighted regression, spline fitting, Gaussian processes,

Kernel support vector machines and neural networks. These methods are effectively employed with clustering and classification techniques, so as more detailed models to be produced. The methodology is demonstrated using real trajectory data from three different sources and specifically an experiment from Naples, NGSIM data and non–lane disciplinary trajectory data from India. The focus is given on car–following models and Gipps' model, one of the most extensively used car–following models, is calibrated against the same data in order to be used as a reference benchmark. The performance of all the models presented in this research is evaluated using several goodness–of–fit measures so as a comprehensive and objective assessment of prognostics performance to be provided. In addition, a policy evaluation methodology based on distributions rather than single aggregate measures is applied, too. Then, a comparison among the different models is presented and comparative benefits as well as limitations of each one are identified.

Many parameters affect driving behavior and in this research it is explored how the performance of the models is improved by including more parameters. Machine learning techniques that allow the incorporation of additional information, such as traffic density, vehicle type of both the leader and following vehicles, may lead to more detailed models, and are very difficult to integrate within conventional, analytical models. The effects of different predictor variables on the models are explored through a quantitative and qualitative analysis. For a more in depth analysis, a meta–model is developed to evaluate the magnitude of the effect of the considered predictor variables on the models. Then, a practical and simple approach is developed and motivated for the online calibration of microscopic traffic simulation models, which considers dynamic parameters for individual drivers, in time and space. The model adapts to driving behavior in a rolling horizon and leads to less than 10% error in speed prediction even for ten steps into the future, in all considered datasets. After the validation of data–driven models on data of vehicles characterized by lane discipline, this research examines the feasibility and the benefits of using data–driven models on mixed traffic trajectory data, which are characterized by non–lane discipline and heterogeneity in vehicle types, common characteristics in cities in developing world.

Conclusions and research contributions

Modeling driving behavior plays a fundamental role in traffic management, safety research and the development of Intelligent Transportation Systems. This research makes several contributions to the state–of–the–art of microscopic traffic simulation:

- A methodological framework based on non–parametric approaches has been developed for simulation of driving behavior. Microscopic conventional models represent individual vehicles and their interactions and are capable of simulating traffic to a high level of detail, but they do require a long execution time, as their successful application depends on the effectiveness of calibration process. On the other hand, the proposed methodology offers great flexibility and there is no need for time consuming calibration process. Data–driven models result in better fit, comparing to the traditional models, and thus could be a plausible substitute for theory–based models.

- Computational approaches allow the easy incorporation of additional parameters. Conventional models do not allow the easy incorporation of additional variables without labor undue, since cumbersome reformulations of functional form should be performed. Data–driven microscopic models have been proposed in this research as a way to overcome these limitations and capture driving behavior in an efficient way taking into account various variables. Additional variables, such as traffic density, have been incorporated into the proposed models and more detailed models have been developed.

- The use of various machine learning techniques for estimation of microscopic models is explored. The question of which machine learning technique could be the most appropriate one for traffic simulation models has not been answered conclusively. This research provides some more input into this ongoing active research field.

- The impact of various predictor variables on the models is investigated. A meta–model is developed to evaluate the magnitude of the effect of the considered predictor variables on driving behavior.

- Data–driven models are validated on non-disciplinary trajectory data with heterogeneous mixture of vehicle types and are proved to be a promising perspective for microscopic traffic simulation in the developing world, where these conditions occur in a common basis.

- Data–driven models and on–line calibration of microscopic models provide a robust solution to autonomous driving. Aiming at safety, reliability and convenience, an autonomous vehicle should adapt to user preferences and simulate human driving reactions naturally, preventing abrupt acceleration and jerk. Undoubtedly, in this context, this research contributes significantly into learning driving styles and realizing autonomous driving.

- Data–driven estimation of microscopic models appears to be a promising tool that could offer considerable benefits if integrated into microscopic traffic simulators, resulting in higher accuracy and reliability of model outputs.

Typical car following models are relied on mathematical formulas and are theoretically justified, though they are more restrictive. On the other hand, machine learning approaches may not provide as much insight into traffic flow theory as the traditional models do, though they are more flexible and allow the easy incorporation of additional information to the process of speed estimation. The results indicate that computational approaches could ensure reliability and improvement in estimation of microscopic models. Data–driven models could provide a robust solution to autonomous driving and be incorporated in traffic microsimulators. Conclusively, better representation of driving behavior contributes substantially into the development of Intelligent Transportation Systems, which are directly related to the concept of smart cities. This contribution could be translated to sustainable transportation solutions, reduced costs in terms of safety, time, money, energy and environmental impact, and by extension to benefits of the humanity.

Future prospects

Directions for further research are outlined in this section.

- **Data–driven simulators**

  Data–driven microscopic models could be integrated in a traffic microsimulator to be used for real–time applications. The results presented in the research provide clear evidence that data–driven traffic approaches have the potential to contribute to improved modeling capabilities, in light of new data and emerging simulation needs. A network–wide validation using a microscopic traffic simulator would create a flexible environment. Implementation aspects should be carefully considered.

- **Clustering of sub–models**

  In Section 3.1 two methodological approaches are proposed for estimation of data–driven models. In this research the focus is given on the first methodological approach which is applied directly on the data. In the second methodological approach, more elaborate approach, data are divided into clusters before the model fitting. In such a way more detailed models could be developed and testing on data should be performed. Guidelines for the selection of one or the other approach and the best way of clustering should be given.

- **Space gaps in the optimization algorithm**

  In this research speeds have been used in the optimization algorithm so that the model minimizes the difference between observed and simulated speeds. It is proposed as a future prospect that space gaps instead of speeds can be used for the model optimization.

- **Incorporation of additional variables in the models**

The proposed methodological framework is flexible, less time-consuming and allows the incorporation of additional variables that may influence driving behavior (such as density of the surrounding area, vehicle type, drivers' age, road infrastructure etc.). In addition, the proposed methodology could be employed with flexible data-driven models which allow incorporation of further variables moving towards an integrated solution for the simulation of mixed traffic. In Section 4.5 the incorporation of additional information is feasible and density of the surrounding area is explored as an extra variable. However, there are many variables, the influence of which on driving behavior has not been explored, yet. Further information on various predictor variables (such as weather, lighting, road geometry, percent of autonomous vehicles etc.) could be also added to the model and tested if it is significant.

- **Parameters evolution**

  In Chapter 5, the prediction of the dynamic parameters was simple, in the sense that the dynamically calibrated parameters were assumed as the best available estimate for the short–term values of these parameters. Further research could consider secondary models that would actually aim at predicting the evolution of these parameters, as well, e.g. via autoregressive, polynomial or other statistical forecasting specifications.

- **Model calibration separately for each vehicle type**

  In case study for mixed traffic conditions (chapter 6) the calibration for the Gipps model as well as for the data–driven model is implemented using a representative sample from each vehicle type category. For a more in depth analysis, different models for each vehicle type category could be calibrated in order to develop more detailed models.

- **Vehicle–dependent models**

  Vehicle-dependent models need to be developed, as the drivers of vehicles with unequal dimensions tend to have different driving behaviors; furthermore, different vehicle types are characterized by varying vehicle kinematics. Especially, in case of heterogeneous traffic vehicle type plays a significant role as it is indicated in Section 6.3.3. The performance of a model seems to be differentiated as per the vehicle type. The best performance is achieved for cars and light commercial vehicles, while higher RMSN are observed for other vehicle types, especially for trucks and auto-rickshaws. Vehicle type should be incorporated as a categorical variable in the process. Thus, it is foreseen that further exploration into this could open up opportunities to understand and simulate driving behavior in non–lane discipline conditions with heterogeneity of vehicle types.

- **Exploration of longitudinal and lateral movement separately for mixed traffic conditions**

  In this research speed for mixed conditions is explored as a sum of longitudinal and lateral speed. However, longitudinal and lateral speed could be explored separately in order to investigate the model efficiency for each direction.

- **Mixed traffic**

  Crowd simulation and swarm–like models could be also used for modeling mixed traffic conditions due to weak–lane discipline characteristics.

- **Deep learning and tree–based modeling**

  In this research various machine learning techniques have been used. However, other methods, such as deep learning and tree–based algorithms, should also be applied in order to offer an overall comparison of machine learning techniques for the estimation of microscopic traffic simulation models.

- **Integrated behavior models**

  Car–following and lane–changing behaviors should be incorporated into one data–driven model, as there is interaction between these two behaviors.

- **Further experimental analysis**

  The estimation of data–driven models as well as the dynamic calibration and multiple time step prediction have been successfully demonstrated using actual data from a variety of facilities. However, additional testing on richer data and further applications in different networks need to be performed.

Εκτενής Περίληψη

Κίνητρο για έρευνα

Με το επιστημονικό ενδιαφέρον να στρέφεται στα αυτόνομα οχήματα, τα μοντέλα ακολουθίας οχημάτων πρέπει να αντικατοπτρίζουν την ετερογένεια στη συμπεριφορά των οδηγών. Τα μικροσκοπικά μοντέλα περιγράφουν τις αλληλεπιδράσεις μεταξύ των οχημάτων ή μεταξύ των οχημάτων και του οδικού δικτύου. Λόγω της μεγάλης λεπτομέρειας που προσφέρουν, θεωρούνται κατάλληλα για την ανάπτυξη «ευφυών συστημάτων αναφοράς» και τον έλεγχο των κυκλοφοριακών συστημάτων. Οι ολοένα αυξανόμενες απαιτήσεις για μεγαλύτερη ακρίβεια και ευελιξία στη προσομοίωση της συμπεριφοράς των οδηγών έχουν οδηγήσει στην ανάπτυξη πολλών μικροσκοπικών μοντέλων κατά τις τελευταίες δεκαετίες. Αρκετά από αυτά αδυνατούν να παρέχουν μια αξιόπιστη εκτίμηση χωρίς κατάλληλη βαθμονόμηση των παραμέτρων τους. Στόχος της έρευνας είναι η ανάπτυξη μιας ολοκληρωμένης μεθοδολογίας για την εκτίμηση αξιόπιστων προτύπων κυκλοφοριακής προσομοίωσης σε ποικίλες κυκλοφοριακές συνθήκες με τη χρήση καινοτόμων και ευέλικτων μεθόδων μηχανικής μάθησης, όπως η τοπικά σταθμισμένη παλινδρόμηση (*loess*), οι καμπύλες *splines*, οι *Gaussian* διαδικασίες, οι διανυσματικές μηχανές υποστήριξης και τα νευρωνικά δίκτυα. Οι μέθοδοι μηχανικής μάθησης παρά το γεγονός ότι αδυνατούν να εξηγήσουν ποιοτικά τις συσχετίσεις που μοντελοποιούν, παρουσιάζουν μεγαλύτερη ευελιξία, καθώς επιτρέπουν την εισαγωγή περισσότερων παραμέτρων. Οι *Antoniou and Koutsopoulos* (2006) και οι *Antoniou et al.* (2013) έχουν προτείνει τη χρήση μεθόδων μηχανικής μάθησης για μακροσκοπική και μεσοσκοπική προσομοίωση. Στο πλαίσιο της έρευνας αυτής το ενδιαφέρον στρέφεται σε μικροσκοπικό επίπεδο.

Αντικείμενο έρευνας

Τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα αυτή περιλαμβάνουν δεδομένα από τρεις διαφορετικές πηγές, δεδομένα από τη Νάπολη, τα *NGSIM* δεδομένα και δεδομένα από την Ινδία. Αρχικά για την εκπαίδευση των μοντέλων εισάγονται δεδομένα, όπως ταχύτητα και απόσταση που αφορούν το προπορευόμενο και το ακόλουθο όχημα. Στη συνέχεια εξετάζεται η περαιτέρω βελτίωση του μοντέλου από την εισαγωγή δεδομένων από τις παρακείμενες λωρίδες και έπειτα η μεθοδολογία εφαρμόζεται σε δεδομένα με βασικά χαρακτηριστικά την έλλειψη λωρίδων κυκλοφορίας και την μεγάλη ποικιλία ως προς τον τύπο των οχημάτων. Σε αυτές τις συνθήκες για τα διαθέσιμα δεδομένα η προτεινόμενη μεθοδολογία υπερτερεί του μοντέλου του *Gipps* και επίσης φαίνεται ότι ο ρόλος του οχήματος είναι καθοριστικός στη συμπεριφορά των οδηγών. Επιπλέον, εξετάζεται η δυναμική βαθμονόμηση κυκλοφοριακών μοντέλων και αναγνωρίζονται τα οφέλη από τη χρήση δυναμικής έναντι στατικής βαθμονόμησης στην μοντελοποίηση της κυκλοφορίας. Καθώς οι κυκλοφοριακές συνθήκες τροποποιούνται δυναμικά με το χρόνο, προτείνεται η δυναμική εναλλαγή των παραμέτρων των μοντέλων, έτσι ώστε να προσομοιώνεται όσο το δυνατόν καλύτερα η συμπεριφορά των οδηγών στις εκάστοτε συνθήκες.

Μεθοδολογία

Η προτεινόμενη μεθοδολογία βασίζεται στη χρήση τεχνικών μηχανικής μάθησης. Με τη χρήση των τεχνικών αυτών αναγνωρίζονται πρότυπα και συσχετίσεις στα διαθέσιμα δεδομένα (εκπαίδευση μοντέλου). Με αυτόν τον τρόπο προκύπτει ένα μικροσκοπικό μοντέλο που βασίζεται στα δεδομένα εκπαίδευσης. Στη συνέχεια χρησιμοποιείται μια νέα σειρά δεδομένων, στην οποία εφαρμόζεται το πρότυπο που δημιουργήθηκε με σκοπό να γίνει εκτίμηση κυκλοφοριακών μεγεθών (π.χ. ταχύτητα) που αφορούν τα νέα δεδομένα (δοκιμαστικά δεδομένα). Αναμφισβήτητα τα δεδομένα εκπαίδευσης πρέπει να περιλαμβάνουν επαρκές πλήθος παρατηρήσεων, ώστε το πρότυπο που προκύπτει να είναι αντιπροσωπευτικό. Η προτεινόμενη μεθοδολογία εφαρμόζεται χρησιμοποιώντας πέντε διαφορετικές τεχνικές μηχανικής μάθησης. Συγκεκριμένα, μέθοδοι μηχανικής μάθησης, όπως τοπικά σταθμισμένη παλινδρόμηση (*loess*), καμπύλες *splines*, *Gaussian* διαδικασίες, διανυσματικές μηχανές υποστήριξης και νευρωνικά δίκτυα δοκιμάστηκαν και φαίνεται να παρουσιάζουν παρόμοια συμπεριφορά παρέχοντας αξιόλογες εναλλακτικές λύσεις. Παράλληλα για τα ίδια δεδομένα εφαρμόζεται το μοντέλο του *Gipps*, ένα γνωστό μοντέλο ακολουθίας οχημάτων που χρησιμοποιείται στο *AIMSUN* λογισμικό προσομοίωσης και χρησιμοποιείται ως μοντέλο αναφοράς στην παρούσα έρευνα.

Συμπεράσματα και συνεισφορά έρευνας

Στο πλαίσιο της έρευνας αυτής εξετάστηκαν οι μέθοδοι μηχανικής μάθησης ως εναλλακτικές μέθοδοι για την εκτίμηση προτύπων ακολουθίας οχημάτων. Η προτεινόμενη μεθοδολογία παρουσιάζει μεγάλη ευελιξία και ταχύτητα ως προς τη διαχείριση των δεδομένων, καθώς δεν απαιτείται ειδική βαθμονόμηση αρκετών παραμέτρων όπως συνήθως χρειάζεται στα παραδοσιακά μοντέλα. Επίσης, παρέχουν τη δυνατότητα εξέτασης νέων παραμέτρων που επιδρούν στη συμπεριφορά των οδηγών και δεν περιορίζονται σε αυτές που ορίζουν οι μαθηματικοί τύποι των παραδοσιακών μοντέλων. Για παράδειγμα επιτρέπουν την εύκολη ενσωμάτωση πρόσθετων μεταβλητών, όπως π.χ. της πυκνότητας των οχημάτων στις παρακείμενες λωρίδες. Παρά το γεγονός ότι οι τεχνικές μηχανικής μάθησης στερούνται θεωρητικής τεκμηρίωσης παρέχουν τη δυνατότητα για πιο εξειδικευμένα και λεπτομερή μοντέλα μικροσκοπικής προσομοίωσης. Η καινοτομία στην παρούσα έρευνα έγκειται στη χρήση μεθόδων μηχανικής μάθησης για την ανάπτυξη ευέλικτων και αξιόπιστων μικροσκοπικών μοντέλων που επιτρέπουν την προσομοίωση της συμπεριφοράς των οδηγών με ακρίβεια παρέχοντας παράλληλα τη δυνατότητα εύκολης ενσωμάτωσης πρόσθετων μεταβλητών καθώς και αξιοποίησης πληθώρας πληροφοριών που είναι σήμερα διαθέσιμες. Επιπλέον, η προσομοίωση κυκλοφορίας σε συνθήκες οδήγησης χωρίς τήρηση των λωρίδων κυκλοφορίας, όπως συμβαίνει στον αναπτυσσόμενο κόσμο, είναι υπό εξέλιξη και η παρούσα έρευνα συνεισφέρει σημαντικά προς την κατεύθυνση αυτή.

Η μοντελοποίηση της συμπεριφοράς των οδηγών παίζει καθοριστικό ρόλο στη διαχείριση της κυκλοφορίας, την οδική ασφάλεια και την ανάπτυξη και εξέλιξη των "Εξυπνων Συστημάτων Μεταφοράς". Η έρευνα αυτή συνεισφέρει σημαντικά στον κλάδο της μικροσκοπικής κυκλοφοριακής προσομοίωσης και συγκεκριμένα:

- Προτείνεται μια μεθοδολογία βασίζομενη σε μη παραμετρικές μεθόδους για την προσομοίωση της συμπεριφοράς των οδηγών. Τα κλασσικά μικροσκοπικά μοντέλα περιγράφουν τις αλληλεπιδράσεις των οχημάτων σε μεγάλη λεπτομέρεια. Ωστόσο, η βαθμονόμηση των παραμέτρων τους είναι απαραίτητη και απαιτεί πολύ χρόνο. Αντιθέτως η προτεινόμενη μεθοδολογία προσφέρει μεγάλη ευελιξία χωρίς να απαιτείται η χρονοβόρα διαδικασία της βαθμονόμησης πολλών παραμέτρων. Επιπλέον, με την προτεινόμενη μεθοδολογία εξασφαλίζεται καλύτερη απόδοση.

- Επιτυγχάνεται η εύκολη ενσωμάτωση πρόσθετων μεταβλητών με τις τεχνικές μηχανικής μάθησης. Στα κλασσικά μοντέλα η ενσωμάτωση πρόσθετων μεταβλητών μπορεί να οδηγήσει σε προβήματα πολύπλοκων μαθηματικών σχέσεων. Η προτεινόμενη μεθοδολογία προσφέρεται ως εναλλακτική λύση για να ξεπεραστούν οι περιορισμοί των κλασσικών μοντέλων. Στην παρούσα έρευνα αναπτύσσεται ένα ευέλικτο μοντέλο ακολουθίας οχημάτων που περιλαμβάνει πληροφορίες από τις παρακείμενες λωρίδες και πιο συγκεκριμένα πληροφορίες για τον αριθμό των οχημάτων. Με αυτόν τον τρόπο συνυπολογίζονται πιθανές αλληλεπιδράσεις μεταξύ των οχημάτων σε περιβάλλον πολλάπλων λωρίδων κυκλοφορίας.

- Εξευρευνήθηκε η χρήση ποικίλων τεχνικών μηχανικής μάθησης. Αν και η επιλογή της καταλληλότερης μεθόδου δεν είναι ξεκάθαρη, η έρευνα αυτή παρέχει κάποιες κατευθύνσεις για το ερώτημα αυτό. Συγκεκριμένα, η προτεινόμενη μεθοδολογία υλοποιήθηκε με τη χρήση ποικίλων τεχνικών μηχανικής μάθησης, όπως τοπικά σταθμισμένη παλινδρόμηση (*loess*), καμπύλες *splines*, *Gaussian* διαδικασίες, διανυσματικές μηχανές υποστήριξης και νευρωνικά δίκτυα. Οι *Gaussian* διαδικασίες απαιτούν αρκετό χρόνο για την εκπαίδευση του μοντέλου σε σύγκριση με τις υπόλοιπες τεχνικές, αλλά είναι αρκετά γρήγορες κατά την εφαρμογή τους σε νέα δεδομένα για την πρόβλεψη τιμών των μεταβλητών. Η μέθοδος λοεσς συνδυάζει ταχύτητα, απλότητα και ακρίβεια. Ωστόσο, όλες οι τεχνικές που δοκιμάστηκαν παρέχουν αξιόπιστες εναλλακτικές λύσεις.

- Διερευνάται η επίδραση των διαφόρων μεταβλητών στο μοντέλο με τη χρήση ενός μετα–μοντέλου που αναπτύχθηκε για να αξιολογήσει την επίδραση κάθε μεταβλητής στη συμπεριφορά του οδηγού. Για χαμηλές ταχύτητες κυκλοφορίας, η απόσταση από το προπορευόμενο όχημα φαίνεται να είναι σημαντική για την πρόβλεψη ταχύτητας. Ωστόσο, για υψηλότερες ταχύτητες η μεταβλητή αυτή δεν φαίνεται να είναι σημαντική. Αυτό μπορεί να αποδωθεί στο γεγονός ότι σε χαμηλές ταχύτητες οι αποστάσειςμεταξύ των οχημάτων τείνουν να είναι μικρότερες και κατά συνέπεια πιο καθοριστικές για τη συμπεριφορά του οδηγού.

- Διερευνάται η χρήση μοντέλων καθοδηγούμενων από τα δεδομένα σε συνθήκες μεικτής κυκλοφορίας χωρίς τήρηση των λωρίδων κυκλοφορίας και με μεγάλη ποικιλία ως προς τον τύπο των οχημάτων. Τα μοντέλα αυτά μπορούν να συμβάλλουν στην κυκλοφοριακή προσομοίωση σε συνθήκες του αναπτυσσόμενου κόσμου. Πιο συγκεκριμένα αναπτύχθηκε μια μεθοδολογία που περιλαμβάνει αναγνώριση του κρίσιμου προπορευόμενου οχήματων σε συνθήκες πολλαπλών προπορευόμενων οχημάτων, καθορισμό προσωρινών νοητών λωρίδων κυκλοφορίας και εκπαίδευση μικροσκοπικών προτύπων από εμπειρικά δεδομένα παρόμοιων κυκλοφοριακών συνθηκών. Επιτυγχάνεται εκτίμηση της ταχύτητας για τις επόμενες χρονικές στιγμές και για συγκερκιμένες κατηγορίες οχημάτων με αιχμή της καμπύλης πυκνότητα των σφαλμάτων μικρότερη από 10%.

- Τα προτεινόμενα μοντέλα καθώς και η δυναμική βαθμονόμηση των μικροσκοπικών μοντέλων παρέχουν βοηθητικές λύσεις προς την κατεύθυνση των αυτόνομων οχημάτων. Στοχεύοντας στην ασφάλεια, την αξιοπιστία και την άνεση, ένα αυτόνομο όχημα πρέπει να μπορεί αν προσαρμόζεται στις προτιμήσεις του χρήστη και να προσομοιώνει τη συμπεριφορά του οδηγού με φυσικό τρόπο εμποδίζοντας απότομες επιταχύνσεις και επιβραδύνσεις. Αναμφισβήτητα, η παρούσα έρευνα συνεισφέρει σημαντικά στην κατανόηση της οδηγικής συμπεριφοράς και στην αποκωδικοποίηση διαφορετικών τρόπων οδήγησης, συμπεριλαμβανομένων διαφορετικών τύπων οχημάτων, οδηγών, συνθηκών και οδικών δικτύων.

Προτάσεις για το μέλλον

Κατευθύνσεις για μελλοντική έρευνα μπορεί να είναι οι εξής:

- Προσομοίωση με χρήση μοντέλων καθοδηγούμενων από τα δεδομένα

  Μοντέλα καθοδηγούμενα από δεδομένα μπορούν να ενσωματωθούν σε κυκλοφοριακούς προσομοιωτές και να χρησιμοποιηθούν για εφαρμογές σε πραγματικό χρόνο. Τα αποτελέσματα της διατριβής συνιστούν ότι τα μοντέλα αυτά μπορούν να συμβάλλουν στη βελτίωση της μοντελοποίησης, δεδομένων των νέων αναγκών προσομοίωσης και των διαθέσιμων δεδομένων. Η επικύρωση των μοντέλων αυτών σε επίπεδο δικτύου σε περιβάλλον προσομοιωτή μπορείνα παρέχει ευέλικτες λύσεις κυκλοφοριακής προσομοίωσης.

- Ταξινόμηση και δημιουργία υπο–μοντέλων

  Στο κεφάλαιο 3.1 παρουσιάζονται δύο μεθοδολογικές προσεγγίσεις. Στην έρευνα αυτή δόθηκε έμφαση στην πρώτη μεθοδολογική προσέγγιση που εφαρμόζεται απευθείας στα δεδομένα. Στη δεύτερη μεθοδολογική προσέγγιση, τα δεδομένα χωρίζονται σε ομάδες πριν την εκπαίδευση του μοντέλου και στη συνέχεια προκύπτει ένα μοντέλο για κάθε ομάδα. Με αυτόν τον τρόπο μπορούν να προκύψουν πιο εξειδικευμένα μοντέλα. Ωστόσο, απαιτούνται κατευθύνσεις για την καλύτερη δυνατή ταξινόμηση καθώς και για την επιλογή της πρώτης ή δεύτερης μεθοδολογικής προσέγγισης.

- Χρήση αποστάσεων αντί ταχυτήτων στον αλγόριθμο βελτιστοποίησης

  Ο αλγόριθμος βελτιστοποίησης που χρησιμοποιείται στην παρούσα έρευνα περιλαμβάνει την ελαχιστοποίηση της διαφοράς μεταξύ παρατηρούμενων και προσομοιωμένων τιμών ταχύτητας. Εναλλακτικά προτείνεται να χρησιμοποιηθεί η ελαχιστοποίηση της διαφοράς μεταξύ παρατηρούμενων και προσομοιωμένων τιμών απόστασης.

- Ενσωμάτωση πρόσθετων μεταβλητών στο μοντέλο

Στην παρούσα έρευνα επιτυγχάνεται η ενσωμάτωση πρόσθετων μεταβλητών στο μοντέλο. Ως παράδειγμα χρησιμοποιείται η πυκνότητα στις παρακείμενες λωρίδες κυκλοφορίας. Ωστόσο, υπάρχουν πολλές μεταβλητές που μπορούν να εξεταστούν, όπως η ηλικία του οδηγού, οι καιρικές συνθήκες, ο φωτισμός, το ποσοστό των αυτόνομων οχημάτων κλπ.).

- Εξέλιξη παραμέτρων

  Στην έρευνα αυτή διερευνήθηκε η δυναμική βαθμονόμηση των παραμέτρων για την πρόβλεψη κυκλοφοριακών μεγεθών σε μικρό χρονικό ορίζοντα. Θα ήταν σκόπιμη η ανάπτυξη δευτερευόντων μοντέλων που θα προβλέπουν την εξέλιξη των παραμέτρων αυτών σε μεγαλύτερο χρονικό ορίζοντα.

- Μοντέλα εξαρτώμενα από τον τύπο οχήματος

  Κρίνεται σκόπιμη η ανάπτυξη μοντέλων εξαρτώμενων από τον τύπο του οχήματος, καθώς οι οδηγοί εμφανίζονται να έχουν διαφορετική συμπεριφορά ανάλογα με τον τύπο του οχήματος. Επίσης, διαφορετικοί τύποι οχήματος τείνουν να έχουν διαφορετικές κινηματικές δυνατότητες.

- Μεικτά μοντέλα

  Η συμπεριφορά ακολουθίας οχημάτων σε συνδυασμό με τη συμπεριφορά αλλαγής λωρίδας θα ήταν σκόπιμο να περιγράφονται από το ίδιο μοντέλο, καθώς υπάρχουν αλληλεπιδράσεις ανάμεσα στις δύο αυτές συμπεριφορές. Επιπλέον, η εξέταση των ταχυτήτων ή των αποστάσεων ξεχωριστά κατά μήκος και κατά πλάτος μπορεί να παρέχει περισσότερες πληροφορίες για τη σύνθετη κίνηση σε περιβάλλον μεικτών συνθηκών.

- Χρήση άλλων τεχνικών μηχανικής μάθησης

  Στο πλαίσιο της έρευνας αυτής χρησιμοποιήθηκαν αρκετές τεχνικές μηχανικής μάθησης. Ωστόσο, εφαρμογές της βαθιάς μάθησης ή δενδροειδών μοντέλων πρέπει επίσης να εξεταστούν ώστε να προκύψει μια συνολική σύγκριση για τις βέλτιστες εναλλακτικές επιλογές με σκοπό την ανάπτυξη μοντέλων καθοδηγούμενων από τα δεδομένα.

- Περαιτέρω πειραματική ανάλυση

  Η εκτίμηση μη παραμετρικών μοντέλων και η δυναμική βαθμονόμηση εφαρμόστηκαν επιτυχώς σε πειραματικά δεδομένα. Ωστόσο, κρίνεται σκόπιμη η περαιτέρω πειραματική ανάλυση σε περισσότερα δεδομένα.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and overview

Traffic simulation models have been used for several decades to conduct detailed analysis of traffic networks. Traffic simulation models play an important role in traffic management and safety research (Barceló et al., 2010). Modeling traffic behavior has also contributed significantly in the development of Intelligent Transportation Systems (Koutsopoulos and Farah, 2012). Zhang et al. (2011) have expressed the need for a shift from a conventional technology–driven system into a more powerful multifunctional data–driven intelligent transportation system. Microscopic models describe driving behavior and interactions among vehicles. Moving towards autonomous driving, more and more reliable and accurate models are required. Nowadays, there has been an increasing interest in self–driving or autonomous vehicles. Aiming at safety, reliability and convenience, an autonomous vehicle should adapt to user preferences and simulate human driving reactions naturally, preventing abrupt acceleration and jerk (Kuderer et al., 2015).

Nowadays, technological advances have significantly improved our traffic data collection capabilities and increasing volumes of potentially useful data are readily available in low-cost opportunistic sensors. Methods such as differential GPS and real time kinematic allow the collection of high fidelity traffic data (Ranjitkar et al., 2005) and consequently may improve the accuracy of traffic simulation models. On the other hand, ubiquitous sensors (e.g. accelerometers and gyroscopes) from regular smartphones could provide a much richer sample of heterogeneous data, which could allow much richer calibration, e.g. utilizing distributions rather than point values (Antoniou et al., 2014b). For a review of novel data collection techniques and their applications to traffic management applications see Antoniou et al. (2011).

As a result of an explosive increase of the data that are being generated and collected, data–driven modeling is emerging as a fast developing field of transportation research. The use of machine learning methods in transportation is still limited. There are several successful demonstrations of machine learning algorithms in the field of intelligent autonomous vehicles. At a macroscopic level, Antoniou and Koutsopoulos (2006a) have proposed estimation of traffic dynamics models using machine learning approaches instead of the classic way of speed–density relationships. In the context of mesoscopic traffic simulation models, Antoniou et al. (2013) have developed a methodology based on data–driven models for the identification and short-term prediction of traffic state and local speed. The results are promising and the introduction of data–driven models into microscopic traffic simulation needs to be also investigated.

Recently, more and more researchers tend to adopt the concept that drivers behave differently in different traffic conditions (Yang and Koutsopoulos, 1996; Ahmed, 1999; Toledo et al., 2003; Wang et al., 2005; Koutsopoulos and Farah, 2012). In this case, sub-phases can be recognized, such as free-flowing, approaching, close-following, car-following, emergency braking, and stop-and-go. This has led to the development of multi-regime car following models, according to which different rules are adopted under different traffic states, so that driving behavior can be best captured. A generalization of such multi-regime approaches is an attractive perspective. However, a large number of regimes may result to overly complex models and developing the equations to model them can become cumbersome. Furthermore, incorporating additional measurement data to these models is very complicated. These limitations have motivated us to suggest within this research an alternative methodology for the estimation of car-following models, combining flexible, data-driven components.

Machine learning techniques may allow for robust and reliable representation of driving behavior and contribute into the development of flexible microscopic models, anticipating future needs. Moving towards autonomous vehicles, models should be able to reflect additional heterogeneity in driving behavior and traffic networks. Undoubtedly, in this direction, machine learning techniques could play a key role in learning driving styles and realizing Autonomous Driving. Machine learning methods can capture driving behavior in an efficient way taking into account various explanatory variables. In contrast, conventional car–following models based on a mathematical formula may not allow the incorporation of all these variables because of the high number of parameters (Papathanasopoulou and Antoniou (2015b); Antoniou and Koutsopoulos (2006a)).

Traffic simulation models have been formulated for lane–based conditions. However, simulation of mixed traffic flow in non lane based heterogeneous conditions is still a challenge. In recent years, there has been an increasing interest in modeling driving behavior in developing countries where conditions, such as non-lane discipline and heterogeneity in vehicle types, prevail. Traffic flow in developing countries is very complex in nature and safety issues arise. Data-driven models may offer a reliable alternative for simulation in mixed traffic conditions.

1.2 New modeling challenges and data opportunities

Transportation is experiencing a period of great development potential and changes, including new modes and new data sources. In an era of big data and autonomous vehicles, traffic simulation models need to adapt to these new challenges. The objective of this research is to provide an alternative modeling approach for traffic simulation models. This modeling approach can take advantage of a wide range of available data, and is therefore suitable to implementation in the context of intelligent transportation systems.

1.2.1 New modeling requirements

The emergence of new transportation modes (services and technologies), also new challenges for modeling traffic system, have created a need for more robust and advanced traffic simulation models. In particular, traffic simulation models need to capture the operations and interactions among new traffic systems.

New systems impact different aspects of traffic demand and driving behavior. Autonomous vehicles for example, under active development, are expected to be gradually introduced in the market. Therefore, autonomous driving constitutes one of the future modes that should be modeled, as well as the interaction of autonomous vehicles and classical vehicles. In addition, new modes such as car–sharing (Barth et al., 2004) may not be easily modeled using the current models. These modes share attributes of both private and public transport. These modes are expected to be more popular, as they offer a plausible alternative solution to severe parking problems in metropolitan areas (Xu and Lim, 2007). Furthermore, there has been an increasing interest in modeling driving behavior in developing countries where conditions, such as non-lane discipline and heterogeneity in vehicle types, prevail. Traffic flow in developing world is very complex in nature and safety issues arise.

The focus here is on traffic moving from the local level to the network level. Towards sustainable mobility, vehicle–to–vehicle and vehicle–to–infrastructure architectures should be employed in real–time applications to develop effective traffic solutions based on real–time data. Data-driven models may offer a reliable alternative for simulation of all these new modes that require the evolution of traffic simulation models including data fusion of various data sources.

1.2.2 New data sources

The rapid development of technology has contributed to the availability of high–quality traffic data, leading the way for the development of more advanced simulation models. As a result of an explosive increase of the data that are being generated and collected, data–driven modeling is emerging as a fast developing field of transportation research. Zhang et al. (2011) have expressed the need for a shift from a conventional technology–driven system into a more powerful multifunctional data–driven intelligent transportation system.

On the demand side, social media networks provide a huge volume of data including temporal, spatial, and textual information that could be exploited in the transportation field (Chaniotakis et al., 2016). In the era of big data, it is important to be able to handle the available information to increase accuracy and reliability of traffic models.

On the supply side, technological advances have significantly improved our traffic data collection capabilities and increasing volumes of potentially useful data are readily available from low-cost opportunistic sensors. Other sources of data (such as cameras, GPS, cell phone tracking, and probe vehicles) are increasingly used as supplementary measurement systems (El Faouzi et al., 2011). Methods such as differential GPS allow the collection of high fidelity traffic data (Ranjitkar et al., 2005) and consequently may improve the accuracy of traffic simulation models. On the other hand, ubiquitous sensors (e.g. accelerometers and gyroscopes) from regular smartphones could provide a much richer sample of heterogeneous data, which could facilitate both, the development of a new generation of models and their calibration, e.g. utilizing distributions rather than point values (Antoniou et al., 2014b). For a review of novel data collection techniques and their applications to traffic management applications see (Antoniou et al., 2011). Drones could also be a future option for data collection. Drones equipped with video cameras have been used for the acquisition of accurate vehicle tracking profiles (Guido et al., 2016).

1.2.3 Future challenges

Kaisler et al. (2013) define “Big Data” as the amount of data just beyond technology’s capability to store, manage, and process efficiently. Advances in information technology are likely to offer new opportunities for transportation and generate changes in speed and efficiency.

New data sources can help us to optimize the transportation networks and improve the balance between demand and supply. Providing accurate traffic information is becoming a major challenge for road traffic management and the deployment of intelligent transportation systems. The focus should be placed on travel time and cost minimization, as well as environmental challenges.

New transportation modes need to be integrated in the cities and a new balance between new and old systems needs to be found, especially as the penetration level of autonomous vehicles keeps changing. In a future world with a cooperative vehicle to vehicle and vehicle to infrastructure communication, all traffic modes and conditions need to be modeled. It is important to be able to offer solutions on–line and provide information and guidance back to drivers.

1.3 Problem statement and research objectives

The objective of this research is the representative modeling of traffic and the estimation of traffic simulation models under varying traffic conditions using innovative and flexible methods. The research contributes significantly to the optimal planning, management of transport networks and autonomous driving. The proposed methodology aims to increase model flexibility and provide the opportunity for incorporation of additional explanatory variables without the need to resort to cumbersome reformulations of conventional models functional form.

Microscopic models describe driving behavior which is affected by many parameters. They often comprise different detailed models, including car-following, lane-changing and gap-acceptance models and their successful application is closely related to calibration process of their parameters. Simulation models do not always adequately reflect field conditions outside of the time period for which they have been calibrated (Balakrishna et al., 2007a; Daamen et al., 2014; Henclewood et al., 2012). Furthermore, incorporation of additional variables into models may lead to cumbersome mathematical relationships and they are very difficult to integrate within conventional, analytical models.

Simulating driving behavior in high accuracy allows short–term prediction of traffic parameters, such as speeds and travel times, which are basic components of Advanced Traveler Information Systems (ATIS). Models with static parameters are often unable to respond to varying traffic conditions and simulate effectively the corresponding driving behavior. It has therefore been widely accepted that the model parameters vary in multiple dimensions, including across individual drivers, but also spatially across the network and temporally. Furthermore, traffic simulation models have been formulated for lane–based conditions. In developing countries conditions, such as non-lane discipline and heterogeneity in vehicle types, prevail and traffic simulation is still a challenge in mixed traffic. Due to lack of lane discipline, it is difficult to identify leader–follower pairs and thus a methodology on temporary virtual lanes is proposed.

In this research a comparative analysis is attempted based on the following alternatives:

- fixed models versus new adaptive alternatives
- conventional versus data–driven models
- simplified models versus models including additional variables
- static versus dynamic models
- lane-based versus non-lane based models

1.4 Dissertation outline and research plan

The remainder of this dissertation is organized as follows. Chapter 2 provides a literature review for both conventional and data–driven microscopic models followed by a qualitative comparison between them.

Chapter 3 analyzes the proposed methodological framework including the integrated methodology for estimation of data–driven microscopic models. Methodological components and evaluation methods are analyzed.

In chapters 4 –6 the experimental setup of this research is outlined. The proposed methodology is demonstrated using real trajectory data from three different sources and specifically an experiment from Naples, NGSIM data and non–lane disciplinary trajectory data from India. Traffic characteristics of the available data are presented, followed by a deeper exploration of the field data. In chapters 4 –5 two different approaches of data–driven models are proposed.

In Chapter 4, a case study is conducted using the available data and the results are presented. Various machine learning techniques are employed for implementation of the proposed methodology. The focus is given on longitudinal behavior.

In Chapter 5, online calibration for microscopic traffic simulation and dynamic multi-step prediction of traffic speed are implemented. The proposed modeling process is data–driven as the estimation of model parameters is originated and produced by the data from the previous time instants.

In Chapter 6, mixed traffic behavior is also explored. A methodological framework for simulation of mixed traffic is developed. The methodology consists of two parts: the identification of follower–leader pairs and the determination of virtual lane changes.

Finally, in Chapter 7 conclusions are drawn and directions for future work are provided. Directions on integrated behavior models are provided. Preliminary work on using the distributions of the field data for the calibration of microscopic traffic simulation software is described, too.

By the end of the thesis, an integrated methodological framework is developed that can simulate driving behavior under various conditions with flexibility and better accuracy. Figure 1.1 illustrates the structure and the design of this research.

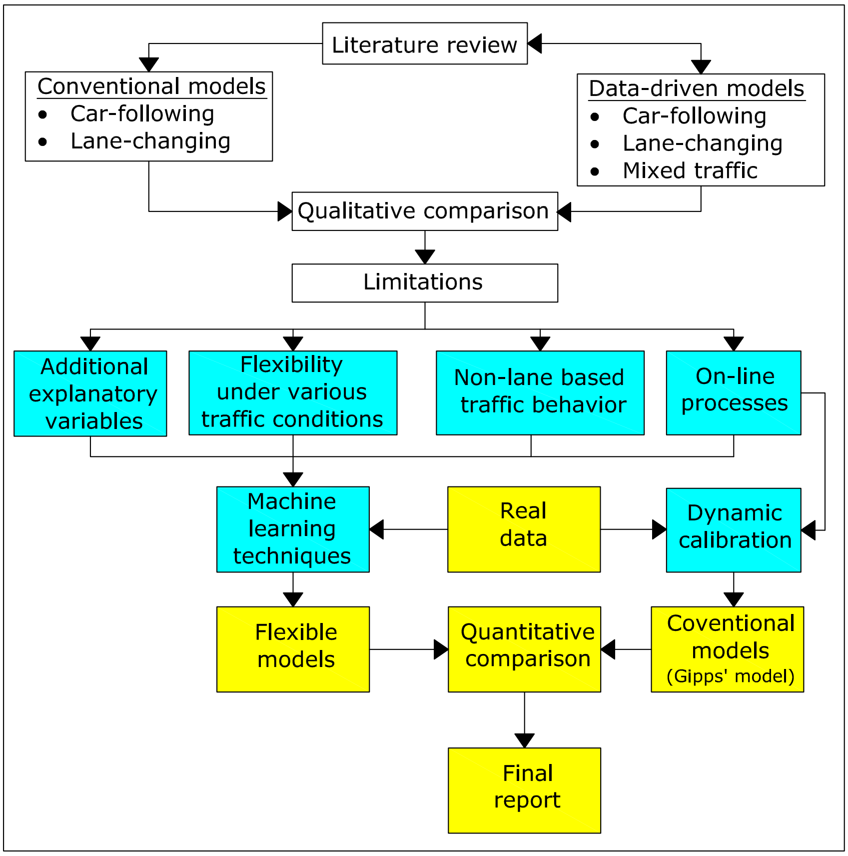


Figure 1.1: Research plan: literature review (white), model improvement (cyan), model improvement–contributions (yellow)

Chapter 2

Synthesis of the state of the art

Data-driven techniques provide a new paradigm for modeling driving behavior by extracting the correlations directly from the data. This section reviews classical microscopic models, in particular car following, lane changing models and some attempts on integrated behavior models, followed by a qualitative comparison between conventional and data–driven models in terms of modeling philosophy, accuracy and flexibility. Limitations of existing models are identified and research gaps are highlighted.

2.1 Heterogeneity in driving behavior

Microscopic traffic models are developed to simulate driving behavior, which can be influenced by many driving characteristics. Saifuzzaman and Zheng (2014) have summarized the main driving characteristics as follows:

- Socio–economic characteristics (e.g., age, gender, income, education, etc.).
- Reaction time.
- Estimation errors: spacing and speeds cannot be estimated with high accuracy by drivers.
- Perception threshold: human cannot perceive small changes as distinctive abilities are limited.
- Temporal anticipation: drivers can predict traffic situation for the next few time instants.
- Spatial anticipation: drivers consider not only the immediate leader but also further vehicles ahead.
- Context sensitivity: traffic conditions may affect driving behavior.

- Imperfect driving: for the same condition driving behavior may be different at different times.
- Aggressiveness or risk averseness.
- Driving capabilities.
- Driving needs.
- Distraction.
- Desired speed.
- Desired spacing.
- Desired time headway.

Traffic models should capture driving heterogeneity in order to be realistic. However, it is complicated to incorporate all these factors in a model or to create different sub–models for different situations. As George Box famously put it, "All models are wrong, but some are useful" (Box, 1979).

2.2 Traffic simulation models

Simulation models are used as the fundamental tool of traffic management and safety research, as they allow the evaluation of traffic plans before their implementation (Barceló et al., 2010). Depending on the level of detail, simulation models are classified into microscopic, mesoscopic and macroscopic models. In **microscopic** models vehicles are described individually and interactions between vehicles or between vehicles and the road network are included (Bellemans et al., 2002). Each vehicle is described by parameters such as its origin, destination, desired speed, acceleration and deceleration, the type of vehicle and the driver's characteristics (Bellemans et al., 2002). **Macroscopic** traffic models use aggregated variables to describe traffic phenomena. Such models simulate the movement as a continuous flow, using theories often inspired by the fluid dynamics. Macroscopic measurements include speed, traffic flow and traffic density (Boxill and Yu, 2000). Finally, **mesoscopic** models provide an intermediate situation, in which they model individual vehicles but at an aggregate level, usually using speed–density relationships and queuing models to model vehicle dynamics. Thus, mesoscopic models share common characteristics with both macroscopic and microscopic models (Boxill and Yu, 2000) and aim to combine the benefits of both, while overcoming their respective limitations.

An ongoing debate among traffic modelers relates to the relative benefits of each level of simulation models. Microscopic models provide the highest level of detail for advanced transport applications and Intelligent Transportations Systems (Antoniou and Koutsopoulos, 2006a). While microscopic traffic simulation models have a higher computational complexity, compared to mesoscopic or macroscopic models, they are more suited to the evaluation and operation of ITS, as they can model in detail more complex aspects of such systems. For example, it would be harder to model managed lanes, vehicle actuated traffic control systems and public transport priority systems with a mesoscopic or macroscopic model. Furthermore, microscopic models are appropriate for modeling mixed traffic as heterogeneity in vehicle types does not allow assumptions of homogeneity that are common in macro– and mesoscopic models. In addition, interactions between surrounding vehicles are critical in mixed traffic conditions and are described in detail in microscopic models. Developments of Advanced Traveler Information Systems (ATIS) rely significantly on the capability to perform accurate estimates of the current traffic state and short–term predictions of driving behavior and traffic characteristics, such as speed (Vlahogianni et al., 2005a; van Lint et al., 2005; Vlahogianni et al., 2008). Due to a number of practical, data and computational considerations, during the past two decades, ATIS applications have been mostly supported by mesoscopic or macroscopic traffic simulation models. Data collection and computational advances are making it possible to consider more detailed, microscopic models for this kind of applications. Naturally, such models introduce a number of complications, and therefore their adoption should be clearly motivated and justified. In this research, emphasis is placed on microscopic models and especially on car–following and lane–changing behavior.

2.3 Overview of traffic simulators

Road traffic tends to become a priority concern, especially for urban cities, where heavy congestion problems are met. Simulation is necessary for understanding traffic problems, assessing the effects of incidents and finding effective solutions. In general, simulation is a dynamic representation of some part or aspects of the real world through time. Traffic simulation is widely used as a tool to test or evaluate a traffic plan of action before its implementation. Currently, there are several traffic simulation softwares. A brief overview and comparative analysis is performed for some of the most used and most mature traffic simulation softwares.

- AIMSUN (Advanced Interactive Microscopic Simulator for Urban and Non-Urban Networks) microscopic traffic simulator was developed by Barceló and Casas (2005) and is available from TSS–Transport Simulation Systems (Spain). The car–following models implemented in AIMSUN are based on the Gipps model (Gipps, 1981, 1986). AIMSUN combines microscopic and mesoscopic modeling.

- ARCHISIM (Bonte et al., 2006), developed by the team Modeling and Simulation of the National Institute for Research on Transport and Safety, is a behavioral model based on multi–agent concepts.

- CORSIM (CORridor SIMulation) has been developed in United States by the Federal Highway Administration (FHWA) and is appropriate for exploring geometric configurations, incident zone impacts, ramp metering options and traffic control (Boxill and Yu, 2000). However, may not capture effectively European traffic behavior. For instance, the "keep right rule" is not included. Furthermore, relaxation phenomenon has not been considered and driver classes are limited (Oud, 2016).

- MATSim (Multi–Agent Transport Simulation), developed by the Polytechnic of Zurich, is an agent–based simulation tool that relies on activity-based approach rather than origin–destination matrices to generate traffic demand.

- MITSIMLab (MIcroscopic Traffic SIMulation Laboratory) (Yang and Koutsopoulos, 1996), developed at MIT (Massachusetts Institute of Technology), is an open source application that supports evaluation of advanced systems for traffic management and guidance.

- Paramics (Parallel Microscopic Simulation), marketed by Quadstone Paramics (UK), illustrates the psychophysical car-following model developed by Fritzsche (1994). The basic concept is that the car-following phase space is divided into five regions representing different modes of car–following. It offers good visual display including 3-D visualization.

- SimTraffic (Husch and Albeck, 2000), marketed by Trafficware (United States), is user–friendly and simulates individual vehicle movements using algorithms for driving behavior and vehicle performance. The disadvantage is that SimTraffic does not provide automated statistical analysis and detailed information of vehicle variables (Saidallah et al., 2016)

- SUMO (Simulation of Urban Mobility) is a free microscopic traffic flow simulation software developed at the German Aerospace Centre. It includes the safe distance Krauss car–following model (Krauß, 1998), an extension of the Gipps model (Gipps, 1981), and the Krajzewicz lane–changing model (Krajzewicz, 2009). The problem lays in the insufficient capacity of the network and it is attempted to increase capacity through parameter calibration (Maciejewski, 2010).

- TRANSIMS (TRansportation ANalysis and SIMulation System) (Smith et al., 1995), developed at Los Alamos National Laboratory (USA), is a free simulation system. It is based on cellular automaton to conduct analysis of a regional transportation system. It integrates a new modeling paradigm of individual travelers and multi-modal transport, advances, providing advances in travel forecasting process. Despite of considered simplifications, it is computationally effective and is appropriate for large networks (Maciejewski, 2010).

- TransModeler, marketed by Caliper Corporation (USA), can simulate all types of networks and complex multimodal systems. It also supports modeling of high occupancy vehicle (HOV) lanes, bus lanes, parking lots, tolls, evacuation scenarios and inter–vehicle interactions.

- VISSIM is provided for microscopic traffic simulation by PTV. It is based on psychophysical models reported by Wiedemann (1994, 1991) and a rule-based lane selection model reported by (Wiedemann, 1991). It is usually used for transit signal priority and interchange design. Although VISSIM offers great flexibility, it is hard to be calibrated due to the large set of parameters (Oud, 2016).

Table 2.1: Traffic simulators overview

Traffic simulators	Model			Category		Visualization	
	Micro	Meso	Macro	Open source	Commercial	2D	3D
AIMSUN	x	x	x		x	x	x
ARCHISIM	x				x	x	
CORSIM	x				x	x	x
MATSim	x			x		x	
MITSIMLab	x			x		x	
Paramics	x				x	x	x
SimTraffic	x				x	x	x
SUMO	x			x		x	
TRANSIMS	x	x		x		x	
TransModeler	x	x	x		x	x	x
VISSIM	x				x	x	x

In order to achieve a realistic representation of the real world, models that could capture effectively heterogeneity of driving behavior under various conditions need to be integrated in traffic simulation softwares.

2.4 Conventional microscopic traffic simulation models

Microscopic models include gap-acceptance, speed adaptation, lane changing, ramp merging, overtaking, and car-following models (Olstam and Tapani, 2004).

Car–following and lane–changing manoeuvres constitute the most common driving behaviour on urban roads and highways. Therefore, in this research the focus is placed on to modeling these two driving situations. A lane-changing situation causes the immediate deviation of driving behavior from the common car–following models. Thus, there is a shift towards the development of an integrated model to imitate both car-following and lane–changing behaviors. Three types of models are further explored:

- Car–following models
- Lane–changing models
- Integrated behavior models

2.4.1 Car–following

Car–following models describe the longitudinal movements of vehicles. The concept of car–following was first introduced by Reuschel (1950) and Pipes (1953). Car–following models have been studied with many diverse approaches for decades. Car following models typically inspect driving behavior with respect to the leading vehicle in the same lane. A vehicle is limited by the movement of the leading vehicle, because driving at the desired rate may lead to a crash (Olstam and Tapani, 2004). According to Bonsall et al. (2005), the main parameters that are involved in the majority of car–following models are the following:

- Desired speed is the speed at which the driver wishes to travel.
- Desired headway is the minimum safe time or distance between two successive cars that the follower vehicle is unwilling to compromise even when at rest.
- Reaction time is the time delay required by any driver in order to respond to any stimulus and take an action.
- Normal Acceleration is the acceleration that the driver wishes to acquire in a normal following situation
- Normal Deceleration is the braking that the driver wishes to apply in a non-emergency situation.

Initially, car following models were developed to represent a single state of traffic, such as the traffic state, where the subject vehicle reacts to the actions of the leading vehicle (Reuschel, 1950; Pipes, 1953). Moreover, many of the earlier car following models, including the General Motors models (Chandler et al., 1958; Gazis et al., 1961; Chandler et al., 1958) refer to low–speed situations and may not be suitable for high–speed networks. Recently, more and more researchers tend to adopt the concept that drivers behave differently in different traffic conditions (Koutsopoulos and Farah, 2012; Yang and Koutsopoulos, 1996; Ahmed, 1999; Toledo, 2003; Wang et al., 2005; Koutsopoulos and Farah, 2012). In this case, sub–phases can be recognized, such as free-flowing, approaching, close–following, car-following, emergency braking, and stop–and–go. This has led to the development of multi–regime car following models, according to which different rules are adopted under different traffic states, so that driving behavior can be best captured. A generalization of such multi–regime approaches is an attractive perspective. However, a large number of regimes may result to overly complex models and developing the equations to model them can become cumbersome. Furthermore, incorporating additional measurement data to these models is very complicated.

A historical review of car-following models is provided by Braskstone and McDonald (2000). They have classified car-—following models into five groups: Gazis-–Herman—Rothery (GHR) model, collision–avoidance model (CA), linear model, psychophysical or action–point model (AP), and fuzzy logic–based model. However, more models have been developed since then and some scientists ((Panwai and Dia, 2005) and (Wei, 2014)) have included them into additional groups. Taking into consideration the abovementioned classifications, classical car-following models are classified into the following groups depending on the utilized logic.

Gazis–Herman–Rothery (GHR) Model

The first version of GHR model was proposed in 1958 at the General Motors Research Laboratory in Detroit (Chandler et al., 1958) and was later further researched by (Gazis et al., 1961). Due to ineffectiveness for both low and high densities applying the same formula, several extensions to the GM framework were proposed (for instance by (Tordeux et al., 2010)). The model is based on Equation 2.1, which relates follower's acceleration to the speed of the follower, the speed difference between follower and leader, spacing between the subject vehicles and driver reaction time.

$$a_n(t) = cv_n^m \frac{\Delta v(t-T)}{\Delta x^l(t-T)} \tag{2.1}$$

where:

$a_n(t)$ is the acceleration of vehicle n at time instant t

v_n is the speed of vehicle n

Δx is the relative distance between vehicle n and n-1

Δv is the relative speed between vehicle n and n-1

T is the driver reaction time

l,m,c are constants

Drivers respond to the speed difference as a major stimulus (Wei, 2014). The application of this model requires calibration of model parameters l and m for a particular network. Braskstone and McDonald (2000) explained that the GHR model is now being used less frequently because of the large number of contradictory findings for parameter values. Moreover, the assumption of the GM model that if the speed of the leading vehicle is higher than the following one, then the driver of the following driver will accelerate is re-examined and should be revised (Koutsopoulos and Farah, 2012).

Safety distance or collision avoidance (CA) Model

Collision-avoidance or safety distance model was introduced by Kometani and Sasaki (1958). Numerous other models have been also reported in the literature ((Benekohal and Treiterer, 1988)), (Gipps, 1981)). All these models are based on the idea that a safety distance should be maintained between the follower and the leader, so as a collision to be avoided with the leading vehicle. The safety distance is described by Equation 2.2 as a function of the speeds of the leader and the follower and driver's reaction time.

$$\Delta x(t-T) = av_{n-1}^2(t-T) + \beta_1 v_n^2(t) + \beta v_n(t) + \beta_0 \tag{2.2}$$

where:
v_n is the speed of vehicle n
v_{n-1} is the speed of vehicle $n-1$
Δx is the relative distance between vehicle n and n-1
T is the driver reaction time
β,β_1,β_0 are constants

According to Hidas (1998) several researchers (e.g. Chen et al., 1995 and Parker, 1996) have found that many drivers tend to adopt a "close following behavior" and the assumption of a safe distance is not obeyed. In the recent years, there is also the need for the deployment of car–following models in the field of intelligent autonomous vehicles. Van Arem et al. (2006) have developed a car–following model appropriate for automated vehicles. The drivers tend to adjust his speed to the leading vehicles while also maintaining constant time gap. The following paragraph provides some additional information on the Gipps model, which is widely adopted in micro-simulation and is still one of the most extensively used models Ciuffo et al. (2012). Therefore, it is also selected as the reference model for the framework developed in this research.

The Gipps model

The car–following model, used in Aimsun traffic simulator, is a safety distance model based on the model developed by Gipps (Gipps, 1981; Olstam and Tapani, 2004; Barceló et al., 2005). The model suggests that the speed of a vehicle (n) is subject to three constraints (Eq. 2.3). First, the vehicle speed does not exceed the driver's desired speed (V_n). Second, the vehicle accelerates rapidly until it approaches the desired speed and then the acceleration is reduced almost to zero. If two vehicles are far apart, they behave as in the free flow condition. These two conditions are summarized in the first part of Equation 2.3. The third condition is taken into account, when the vehicle is constrained by the vehicle in front. It is taken for granted that the following vehicle will adjust its velocity so as to keep a safe distance from the leading vehicle. This condition is described by the second part of Equation 2.3. Overall, according to the above restrictions, the speed of vehicle n at time ($t+\tau$) could be calculated by the following formula:

$$v_n[t+\tau]=min\begin{cases} v_n[t]+2.5\cdot a_n\cdot\tau\cdot(1-\frac{v_n[t]}{V_n}\cdot\sqrt{(0.025+\frac{v_n[t]}{V_n})} \\ b_n\cdot\tau+\sqrt{(b_n\cdot\tau)^2-b_n\cdot[2\cdot(x_{n-1}[t]-s_{n-1}-x_n[t])-v_n[t]\cdot\tau-\frac{v_{n-1}^2[t]}{\hat{b}}]} \end{cases} \tag{2.3}$$

where:

a_n : the maximum acceleration that the driver of vehicle n wishes to acquire (m/s^2).
b_n : the maximum braking that the driver of vehicle n wishes to apply in order to avoid a crash, b_n<0 (m/s^2).
$\hat{b}$: the estimated maximum braking that the driver of the leading vehicle (n-1) wishes to apply (m/s^2).
s_{n-1} = L_{n-1} + Safety, namely the size of the leading vehicle (n-1) including its length and the safety distance at which vehicle n is unwilling to compromise even when at rest (m).
V_n : the speed at which the driver of vehicle n wishes to travel (m/s).
x_n[t], x_{n-1}[t] : the location of the front side of the respective vehicle (n or n-1) at time t (m)
v_{n-1}[t] : the speed of the leading vehicle (n-1) at time t (m/s)
v_n[t] : the speed of the following vehicle (n) at time t (m/s)
τ : the apparent reaction time (a constant for all vehicles) (s)

Linear (Helly) Model

This model is rooted in the GHR models and was further improved by Helly (1959) who introduced the concept of desired following distance. The model calculates the acceleration of the follower as a function of desired following distance, speed of the following vehicle, relative distance and speed between follower and leader, and driver's reaction time.

$$\begin{aligned} a_n(t) &= c_1\Delta v(t-T)+c_2\Delta x(t-T)-D_n(t) \\ D_n(t) &= a+\beta v(t-T)+\gamma a_n(t-T) \end{aligned} \tag{2.4}$$

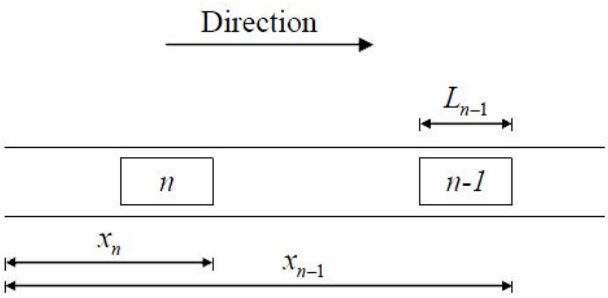


Figure 2.1: Notation for the Gipps model (Olstam and Tapani, 2004)

where:

$a_n(t)$ is the acceleration of vehicle n at time instant t

$D_n(t)$ is the desired following distance of vehicle n at time instant t

v is the speed of vehicle n at time instant t

Δx is the relative distance between vehicle n and n-1

Δv is the relative distance between vehicle n and n-1

T is the driver reaction time

α, β, γ, c_1, c_2 are constants

A similar model was used later by Yeo et al. (2008) to model an oversaturated freeway flow. Overall, the linear CF model fits well in a low acceleration pattern, but in high disturbance within the traffic flow the model overestimates headways (Braskstone and McDonald, 2000).

Psychophysical or AP Model

Psychological or AP models constitute another type of car–following models introduced by Michaels (1963) as they wanted to relax constraints of GM models. Drivers change their behavior at different regions which are determined by their psychological status. More specifically, drivers react to distance or speed difference between pairs of vehicles, when a threshold is reached. Different equations are applied for each region, as the behavior is differentiated. Thresholds are expressed as a function of speed difference and distance. Representative examples of psycho-physical car–following models are visual angle models (Michaels, 1963) and the ones developed by Leutzbach and Wiedemann (1986), Wiedemann and Reiter (1992) and Fritzsche (1994). Wiedemann and Reiter (1992) proposed that two vehicles moving in sequence may interact under four traffic states: free flowing, approaching, car following and emergency situation. These models depend on the human perception of motion. However, the ability to perceive speed differences and distances varies among drivers and thus estimation of individual thresholds is a difficult task.

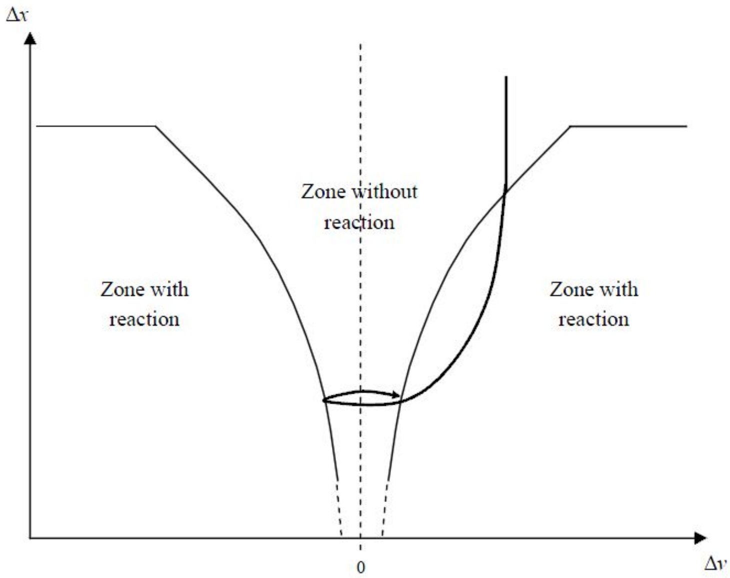


Figure 2.2: A psycho–physical car–following model (Olstam and Tapani, 2004)

Fuzzy–logic–based models

The application of fuzzy-logic principles to the GHR model was introduced by (Kikuchi and Chakroborty, 1992). The fuzzy system of car–following models describes a follower's response to the change of relative speed and headway to that of the leader's. Variables of the model are divided into a number of overlapping sets associated with a particular term such as "close" and "too close". Fuzzy sets according to the follower's desired relative speed and headway are used and are interpreted in logical operators or rules like if "too close" then use emergency deceleration. Another typical fuzzy rule would be: If Distance Divergence is "too Far" and relative speed is "closing" then the driver's response is "No Action". The application of the models and their effectiveness depends on the determination of membership functions (Panwai and Dia, 2005).

Cellular automata model

This type of model was first introduced by Nagel and Schreckenberg (1992). Their model has been extended by several others in the last few years (Blue et al., 1996; Rickert et al., 1996; Bham and Benekohal, 2004). In cellular automata, space and time are discrete and physical quantities take discrete values. A cellular automaton consists of a regular uniform lattice with discrete variables at the various sites. The state of a cellular automaton is completely specified by the values of the variables at each cell. The variables are updated simultaneously, based on the values of the variables in their neighborhood at the preceding time step according to a definite set of "local rules".

Optimal velocity model

Optimal Velocity model was introduced by Bando et al. (1994, 1995) and the governing equation is the following:

$$\begin{aligned} \frac{dx_i}{dt} &= v_i \\ \frac{dv_i}{dt} &= \alpha[V(x_{i+1} - xi) - v_i] \end{aligned} \tag{2.5}$$

where:

x_i positions of vehicles

v_i velocities of vehicles a a sensitivity parameter denoting the speed of driver's response.

The basic principles of optimal velocity model are the following. A vehicle obtains the optimal velocity determined by the distance from the leading car. At any vehicle density there is only one optimal velocity. Driver's reaction time and sensitivity to the vehicle's environment depend on whether the road is congested or not.

Desired spacing models

In the desired spacing models (Parker, 1996; Hidas, 1998, 2005), a desired spacing criterion is defined as a linear function of the speed. The desired spacing is an individual characteristic and varies among drivers and acceleration or deceleration states. Each driver adjusts his acceleration in order to achieve the desired spacing without attempting to explain behavioral tasks or to estimate reaction time. This model is proposed only for urban streets with low speeds by Hidas (1998).

Intelligent driver model

The Intelligent driver model was proposed by Treiber et al. (2000). It is a deterministic continuous car–following model. The acceleration is given as a continuous function of the speed, the gap and the speed difference from the leader vehicle. The model is given by

$$\begin{aligned} a_{IDM}(s, v, \Delta v) = \frac{dv}{dt} &= a[1 - (\frac{v}{v_0})^\delta - (\frac{s^*(v, \Delta v)}{s})^2] \\ s^*(v, \Delta v) &= s_0 + vT + \frac{v\Delta v}{2\sqrt{ab}} \end{aligned} \tag{2.6}$$

where:

v is the velocity of the subject vehicle

s is the current net space headway of the subject vehicle

Δv is the relative distance between the vehicle and the leader vehicle

$s^*(v, \Delta v)$ is the desired space gap

a is the maximum acceleration
s_0 is the minimum distance or the jam distance
v_0 is the desired speed of the subject vehicle
b is the desired deceleration
δ is the free acceleration exponent
T is the safe time headway

The advantage of the model is that it is collision free. However, it is formulated to describe traffic dynamics in one lane only and when the actual gap is significantly lower than the desired gap, e.g. in lane–changing situations, the model leads to unrealistic behavior. An enhanced model, called "ACC model", has been developed by Kesting et al. (2010) in order to prevent unnaturally strong braking reactions due to lane changes.

Capacity drop and traffic hysteresis

This model has been proposed by Zhang and Kim (2005). Capacity drop corresponds to a specific density, while traffic hysteresis reflects a stop and go situation. Multiphase traffic situations, such as steady state phase and congestions and transition phase simultaneously, are considered. The model provides more insight to the macroscopic level rather than microscopic level. This model lead to more realistic fundamental diagrams (flow–density).

2.4.2 Lane–changing

Lane–changing models describe the lateral movements of vehicles and often incorporate lane selection and lane–changing decision, recognizing acceptable conditions (lead and lag gaps) and lane-changing maneuver. Most models classify lane changes into two categories, mandatory and discretionary, based on the execution process (Ben-Akiva et al., 2009; Mathew, 2014). A **mandatory lane change (MLC)** occurs when a driver must change lane to follow a certain path. For instance, if a driver wants to turn right at the next intersection, he must change lane to follow the right most lane. On the other hand, a **discretionary lane change (DLC)** occurs when a driver changes to a lane perceived to offer better driving conditions, for instance if he tries to achieve his desired speed or to avoid trucks in the same lane, etc.

The execution of a lane change is modeled using gap acceptance models. Gaps are considered either in terms of time or space and in some models there is distinction between the lead gap and the lag gap (see Figure 2.3). Gap acceptance models are usually defined as binary problems, in which drivers accept or reject the available gap by comparing it with the critical gap (minimum acceptable gap) (Toledo et al., 2003).

In congested traffic conditions, other types of lane-changing mechanisms have

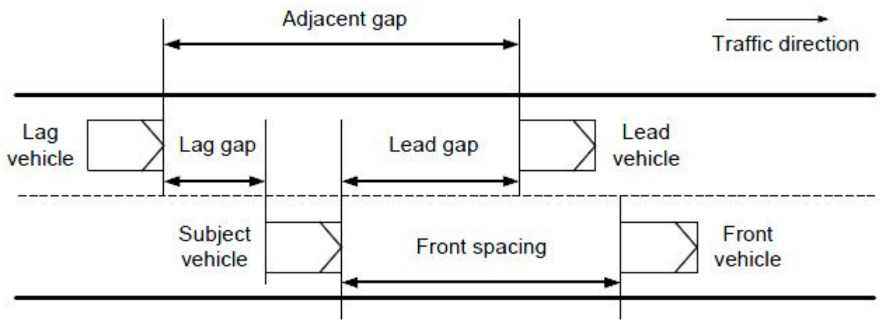


Figure 2.3: Vehicles and gaps within a lane–changing manoeuvre (Toledo et al., 2003)

been found in the literature and include forced and cooperative gap acceptance. **Forced merging** is executed if the gap on the target lane is not acceptable and the lag vehicle on the target lane is forced to decelerate until the gap is acceptable. **Cooperative merging** occurs when a driver changes lane through courtesy and cooperation of the lag vehicles on the target lane that decelerate in order to facilitate the lane change (Ben-Akiva et al., 2009; Mathew, 2014).

Many lane-changing models have been developed in the last few decades. These models often have different base principles, such as thresholds or utility functions. A critical review is provided by Moridpour et al. (2010) and Toledo (2007). Each model is based on a decision making process (Bonsall et al., 2005). The general structure of these lane–changing models is outlined in Figure 2.4 (Ben-Akiva et al., 2009).

Most lane change models assume that gap–acceptance is either a function of distance and speed difference, or deceleration is determined by a car-following model. However, in the real world, drivers tend to apply small decelerations and accept smaller time headways (Laval and Leclercq, 2008). This is the **relaxation phenomenon** which takes place whenever a lane change occurs at a short spacing and is not in accordance with the fundamental diagram (Leclercq et al., 2007).

Gipps (Gipps, 1986) introduced the first lane-changing decision model intended for microsimulation tools. Ahmed (1999) and Ahmed et al. (1996) have proposed a lane–changing model that captures both mandatory and discretionary lane changes. Lane–changing process is divided into three steps: decision to consider a lane change, choice of a target lane, and acceptance of gaps in the target lane. If the driving conditions in the current lane are not satisfactory, the driver compare them with conditions in adjacent lanes and selects a target lane. Lane utilities depend on the speeds of the lead and lag vehicles in target lane in comparison with the current and desired speeds of the vehicle. The model takes into account differences in driving behavior of heavy vehicles, too. Wiedemann and Reiter (1992) developed a lane changing decision model that considers the

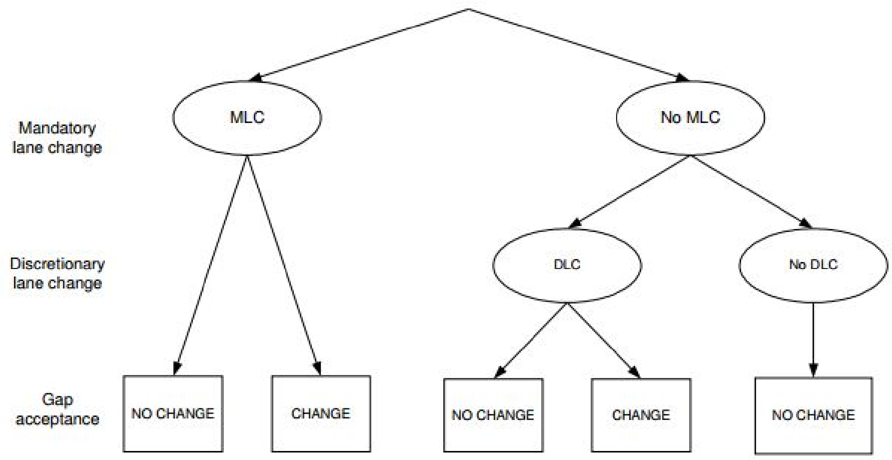


Figure 2.4: A generic structure of lane–changing models (Ben-Akiva et al., 2009)
* MLC: Mandatory lane change, DLC: Discretionary lane change

driver's perception of surrounding vehicles. Different drivers have different characteristics such as different driving capabilities and perception abilities. Wei et al. (2000) proposed a model of lane selection for drivers that turn into two–lane urban arterials. They identified another category of lane changes, preemptive lane changes. A preemptive lane change is performed by a driver when he is going to turn after some intersections. The proposed model is outlined in Figure 2.7. Hidas (2002, 2005) introduced the concept of driver's courtesy of the lag vehicle in the target lane. However, lane–changing is simulated as an instantaneous action. Kesting et al. (2007) incorporated in their model the consequences of lane change for the surrounding vehicles. Ben-Akiva et al. (2009) developed a model that integrates mandatory and discretionary lane changes in a single framework. Various types of lane-changing mechanisms, such as cooperative lane changing and forced merging, are included and heterogeneity in driving behavior has been taken into consideration. This model is presented in Figure 2.6 and has been implemented in MITSIMLab microscopic traffic simulator.

These works focus on the decision-making process for lane-changing, but they ignore the process of lane-changing execution as well as the speeds of other potentially involved vehicles. The conventional lane–changing approaches are based on predefined logic rules, which explain driving behavior to some extent. However, simplicity and inflexibility of such rules may lead to unrealistic lane–changing simulation (Bi et al., 2016), as they do not incorporate the inconsistencies and uncertainties of drivers' perception and decisions (McDonald et al., 1997). Further details are provided for some lane-changing models below.

The Gipps model

The Gipps model (Gipps, 1986) has been proposed to explain lane–changing decisions both on freeways and in urban driving conditions, including traffic signals, heavy vehicles and obstructions that affect drivers' behavior and decisions. The modeling approach is based on some simplified assumptions. Drivers' decision to proceed to a lane change depends on three factors, **safety**, **necessity** and **desirability** of lane changes.

During a lane–changing manoeuvre, the model identifies three zones based on the distance from the point of the intended turn. In the **remote–zone**, the intended turn is far away. There is no effect on drivers' lane–changing decision and the drivers focus on their desired speed. In the **middle–zone**, the intended turn is a middle distance away. The driver steers the vehicle towards his destination or the target lane ignoring opportunities to take speed advantage. In the **near–zone**, the distance from the exit point is too close and the drivers try to follow the correct lane without speed considerations.

The Gipps lane changing model provides a good and reasonable explanation on how a driver decides to execute a lane–change. However, according to Zheng (2014), the limitation of the model is that the vehicles change lanes only if there is a safe and adequate headway. The model does not take into consideration heterogeneity in driving behavior and the assumption on the required safety gap may not be applicable in congested traffic conditions that the required gaps may not be available.

The Hidas' model

Hidas (2002, 2005) have developed a lane changing decision model which incorporates the simulation of driver's courtesy in target lane (Figure 2.5). Lane–changing decisions are classified into three types: free, forced and cooperative. During a free lane change manoeuvre, there are no significant changes in lead and lag gaps. In a forced lane change, headways decrease before the lane change execution and increase after that, while in cooperative lane changes the opposite is observed. Cooperative lane–changing depends on the willingness and feasibility of the lag vehicle driver to decelerate in order to provide a sufficient space gap for the lane change. The main assumption of the model is that a lane change is feasible if there is a sufficient gap in the target lane.

The target lead gap and the target lag gap in a free lane changing manoeuvre are given by Equation 2.7. The lag gap in the target lane and the minimum acceptable target lead and lag space gaps in a cooperative or forced lane changing manoeuvre are calculated by Equations 2.8 and 2.9 respectively.

$$\begin{aligned} g_l &= g_{0l} - (v_s - b_s/2) + v_l \\ g_f &= g_{0f} - (v_f - b_f/2) + (v_s + b_S/2) \end{aligned} \tag{2.7}$$

$$g_f = g_{0f} - (v_f D_t - b_f/2D_t^2) + v_s D_t \tag{2.8}$$

$$\begin{aligned} g_{l,min} &= g_{min} + \begin{cases} c_l(v_s - v_l), v_s > v_l \\ \\ 0, otherwise \end{cases} \\ g_{f,min} &= g_{min} + \begin{cases} c_f(v_s - v_f), v_f > v_s \\ \\ 0, otherwise \end{cases} \end{aligned} \tag{2.9}$$

where g_l is the target lead gap
g_f is the target lag gap
g_{0l} is the target lead gap at the start of lane change
g_{0f} is the target lag gap at the start of lane change
v_s is the speed of the subject vehicle
v_f is the speed of the lag vehicle
v_l is the speed of the lead vehicle
b_s is the deceleration of the subject vehicle
b_f is the deceleration of the lag vehicle
D_t=D_v/b_f is the time of deceleration period, D_v is the speed decrease of the subject vehicle
c_l, c_f are constants

The Mobil model

The Mobil model ("Minimizing Overall Braking Induced by Lane changes") was developed by Kesting et al. (2007). It is an acceleration–based model that takes into consideration the consequences of a lane change for the followers in the origin and target lane. The parameters of the model determine how much the driver weighs the consequences for his followers. A safety criterion ensures that after the lane change the deceleration $\tilde{a_n}$ of the upstream vehicle in the target lane does not exceed a given safe limit b_{safe}.

$$\tilde{a_n} \geq b_{safe} \tag{2.10}$$

Kesting et al. (2007) have also formulated an asymmetric lane-changing criterion for two-lane freeways taking for granted that the right lane is the default lane. Two European traffic rules are described by the following formulas.

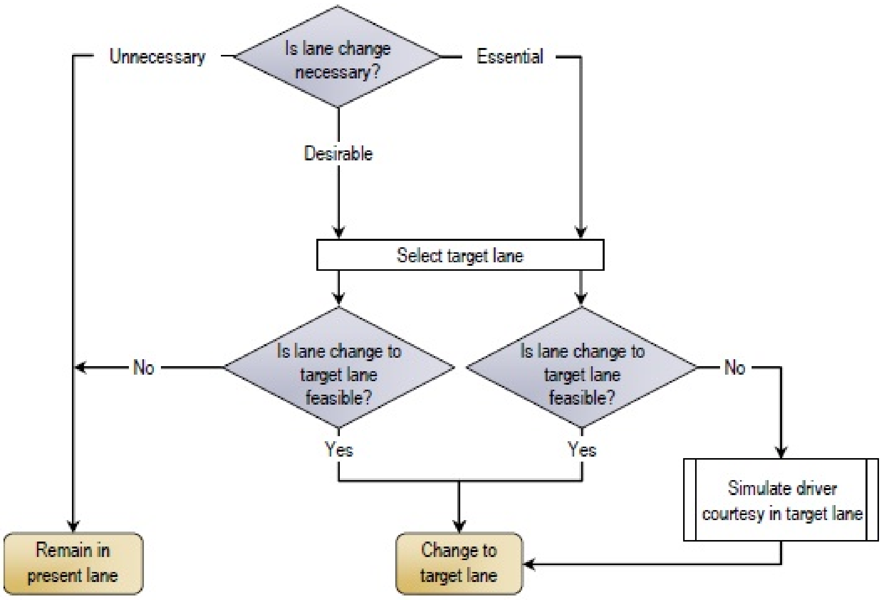


Figure 2.5: General structure of Hidas' lane–changing model (Hidas, 2002)

According to the first rule, overtaking from the right lane is forbidden, unless congested conditions prevail Equation 2.11. The second rule is summarized in Equations 2.12 and 2.13, whether the vehicle moves from left to right or from right to left respectively.

$$a_c^{eur} = \begin{cases} min(a_c, \tilde{a}_c), v_c > \tilde{v}_{lead} > v_{crit} \\ \\ a_c, otherwise \end{cases} \tag{2.11}$$

$$\tilde{a}_c - a_c^{eur} + p(\tilde{a}_n - a_n) > \Delta a_{th} - \Delta a_{bias} \tag{2.12}$$

$$\tilde{a}_c - a_c^{eur} + p(\tilde{a}_0 - a_0) > \Delta a_{th} - \Delta a_{bias} \tag{2.13}$$

where:
a_c: the current acceleration of the vehicle
$\tilde{a}_c$: the current acceleration of the vehicle after the lane change
v_c: the current speed of the vehicle
$\tilde{v}_{lead}$: the speed of the leader in the target lane
v_{crit}: the minimum speed of the traffic that can be considered as free-flow
p: politeness factor
Δa_{th}: threshold level of the advantages to avoid fluctuations
Δa_{bias}: additional bias to motivate the traffic to keep right
a_0: the current acceleration of the follower in the origin lane
$\tilde{a}_0$: the acceleration of the follower in the origin lane after the lane change

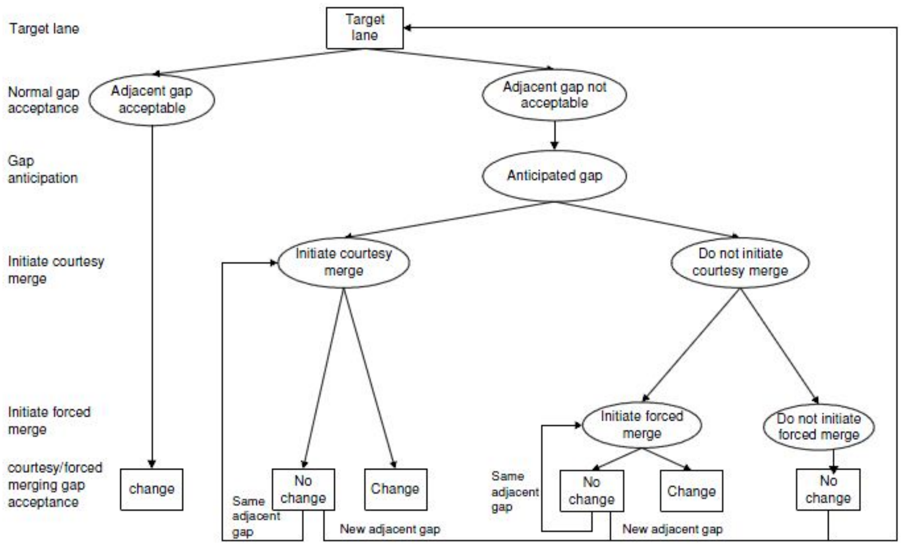


Figure 2.6: Structure of a model that integrates courtesy and forced merging (Ben-Akiva et al., 2009)

a_n: the current acceleration of the follower in the new lane
$\tilde{a}_n$: the acceleration of the follower in the new lane after the lane change

2.4.3 Integrated behavior models

An integrated behavior model incorporates both car–following and lane–changing behavior into one model. As these two behaviors are closely related to each other, there have been a few attempts to develop an integrated model. One of these is the Toledo' s model (Toledo, 2003).

This model consists of three main parts: the short–term goal, the short–term plan and the driver's actions. The short–term goal is the target lane of the driver. Then, the driver decides on a short term plan and chooses a target gap for the lane change. Finally, the driver takes action by adapting the acceleration and changes lane when his requirements for space gap are satisfied. If no lane change is necessary, the driver remains in the same lane and tries to get or to maintain the desired speed.

The Toledo's model structure is outlined in Figure 2.8. A four–level decision making is implemented based on target lane, gap acceptance, target gap and acceleration. Driver's actions are observable and are represented by squares in Figure 2.8. Decisions on short–term goal and short–term plan are latent and are shown as oval in Figure 2.8. Acceleration is represented as a continuous function, while lane changes as a discrete function. It is assumed that only one lane change could be executed during one time interval. The model includes several techniques and sub–models (for instance models utility of target lane, critical gap, etc.) to capture variable driver characteristics and inter–dependencies between

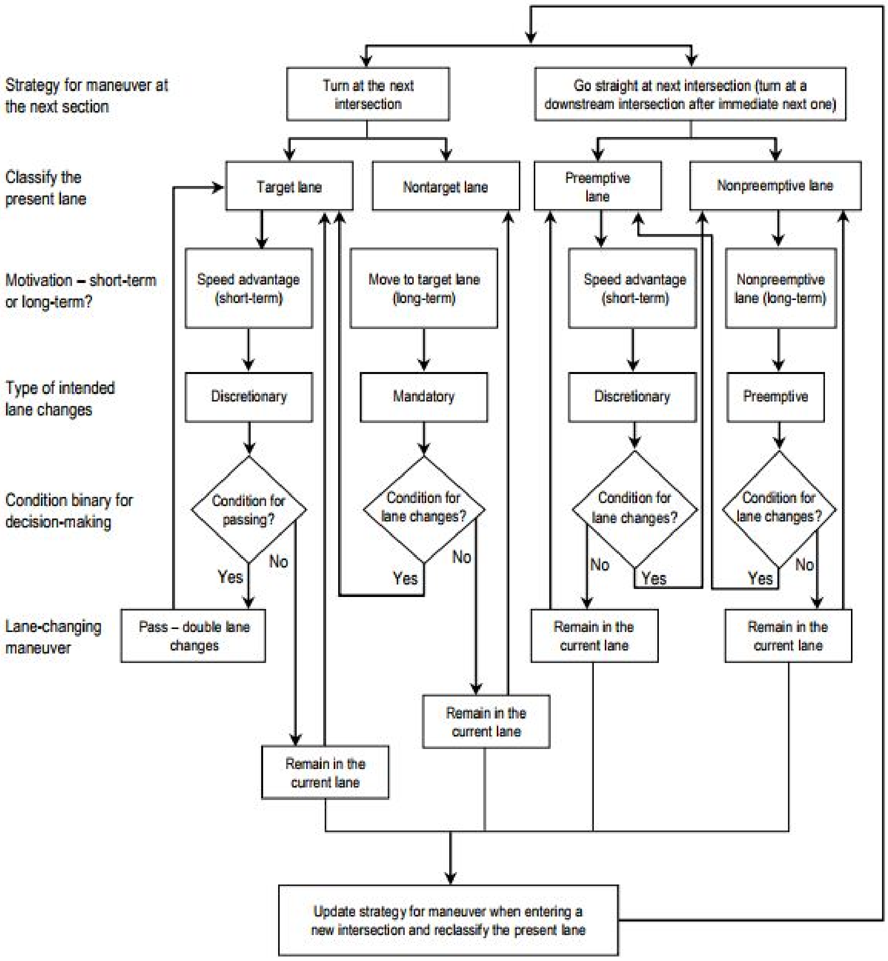


Figure 2.7: Structure of the model developed by Wei et al. (2000)

the four–level decisions. However, drivers may need to reconsider short–term goals and plans per time step, as traffic conditions change dynamically. This leads to the assumption that all state dependencies are captured by the explanatory variables of the model. Further details on the model are provided by Toledo (2003).

2.5 Data–driven microscopic traffic simulation models

Nowadays, the rapid development of technology has contributed to the availability of high–quality traffic data, leading the way for the development of more advanced microscopic models. Limitations of conventional models have been the motivation to explore alternative approaches for the estimation of microscopic models, combining flexible data–driven components. Such methods have been used in several transport–related applications. Various machine learning techniques have been used in transportation research in recent years. More than ten years ago, Antoniou and Koutsopoulos (2006c) developed a framework for speed esti-

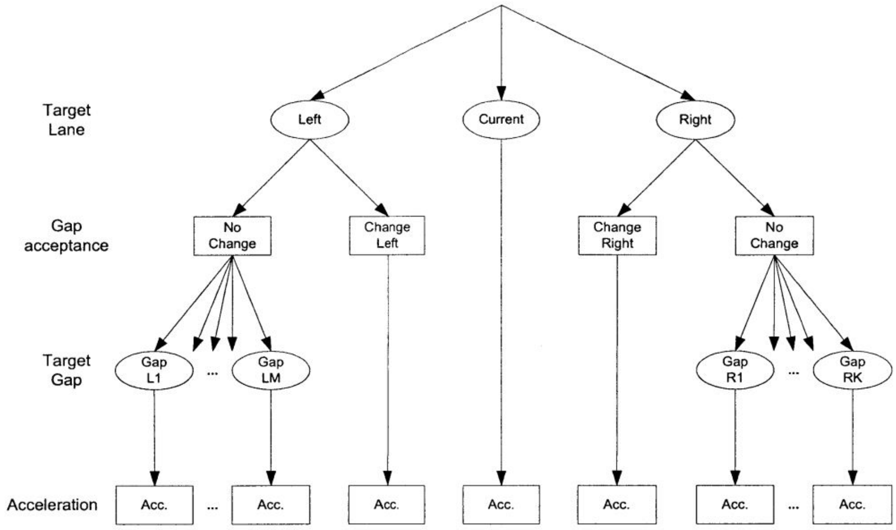


Figure 2.8: Integrated behavior model (Toledo, 2003)

mation using machine learning concepts, including clustering algorithms and locally weighted regression. Antoniou and Koutsopoulos (2006b) compared a number of machine learning techniques for speed estimation, including loess, support vector regression, and neural networks. Other data–driven methods, including neural networks (Huval et al., 2015), Gaussian processes (Chen et al., 2014) and Kernel methods offering similar capabilities (Karlaftis and Vlahogianni, 2011). Antoniou et al. (2013) developed a framework for dynamic traffic state estimation and prediction using machine learning methods. Kleyko et al. (2015) have compared three machine learning techniques, specifically logistic regression, neural networks, and support vector machines, for a vehicle classification problem and have indicated that logistic regression provided the best results. Jenelius and Koutsopoulos (2013) have presented a statistical models for travel time estimation for urban road network travel time estimation using low frequency probe vehicle data. Jenelius and Koutsopoulos (2018) used probabilistic component methods for traffic state prediction. Lv et al. (2015) and Huang et al. (2014) have used deep learning for traffic flow prediction.

Focusing on microscopic data–driven models, the available background literature is still limited. A brief overview of them is presented in the following sections and the focus is placed on car–following, lane–changing behavior and driving in mixed traffic conditions.

2.5.1 Car–following models

Innovative ways for modeling car–following behavior are based on data–driven methods. Zhang et al. (2011) have suggested and implemented the use of machine learning approaches to support a shift from conventional technology-driven systems into data–driven intelligent transportation system. Data–driven approaches have already been used in developing a fully adaptive cruise control system (Simonelli et al., 2009; Bifulco et al., 2013b) and in modeling car–following behavior via artificial neural networks (Colombaroni and Fusco, 2014; Chong et al., 2013; Zheng et al., 2013). Simonelli et al. (2009) have applied neural networks to develop a real–time learning model for capturing car-following behavior taking into consideration individual drivers' characteristics. Bifulco et al. (2013b) extended the work of Simonelli et al. (2009) into a framework for reproducing spacing in adaptive cruise control applications. Furthermore, Panwai and Dia (2007) developed a car following model based on neural networks and fuzzy neural networks. They tried different types of neural networks and validated their model using field data from two vehicles equipped with radar detectors. The results were promising as their models outperformed Gipps' model. Zheng et al. (2013) proposed a model based on neural networks, too. The difference of their model is that they used a two–level neural network structure. The first level is used to estimate the dynamic reaction delay, while the other to predict the acceleration of the following vehicle. While most data–driven studies adopt a neural network approach, there are several methods that either have not been adequately explored or have not been compared on the same data with other methods in order to obtain a better understanding on how the algorithm choice could influence the results.

2.5.2 Lane–changing models

To overcome the limitations of conventional lane-changing models, the scientific interest has shifted towards data—-driven traffic simulation using machine learning techniques. Kumar et al. (2013) have proposed a learning–based approach, using Support Vector Machine and Bayesian filtering, for online lane–change intention prediction. Their model predicts driver intention to change lanes about 1.3 seconds in advance. Ding et al. (2013) have explored the ability of a neural network to learn and identify the uncertainties and perceptions in human behavior from real driving data in order to predict a lane-changing trajectory. Hou et al. (2014) have developed a lane changing assistance system that advises drivers for safe gaps and if it is safe or unsafe to execute a mandatory lane–change. The model is validated on NGSIM data and predicts whether a driver will merge or not as a function of certain input variables using Bayes Classifier and Decision Trees. Bi et al. (2016) have developed a data-driven model to simulate the process of lane–changing in traffic simulation using RANDOM forests and back-propagation neural network algorithms. However, they do not take driver heterogeneity into account. Wang et al. (2017) have modeled various merging behaviors at expressway on-ramp bottlenecks using support vector machine (SVM) models. They have considered four merging behaviors with different degrees of merging risk. In comparison with other models including discrete choice model, Bayesian network and classification and regression tree, SVM achieves the best prediction results.

2.5.3 Mixed traffic conditions

Asaithambi et al. (2016) review driver behavior models under mixed traffic conditions and have pointed out limitations of current models, arguing that the main limitation is that they do not explicitly consider the wider range of situations that drivers in mixed traffic face. Munigety and Mathew (2016) have identified that due to weak lane discipline, drivers maneuvering in mixed traffic streams exhibit some peculiar patterns such as maintaining shorter headways, swerving, and filtering. They have also proposed that the lane should be divided into small strips in order to handle virtual lane movements. Li et al. (2015) have proposed a car–following model that considers the effect of two–sided lateral gaps and have they have shown that their model has larger stable region compared to a car–following model that captures the impacts from the lateral gap on only one side. In addition, Parsuvanathan (2015) has used proxy lanes between the main lanes. It is assumed that free space is perceived as lanes by small vehicles. However, distribution and types of vehicles could affect the width of the lanes. A grid-based modeling approach akin to cellular automata (Gundaliya et al., 2008) and a strip-based modelling method (Mathew et al., 2013) have also been proposed. Mathew et al. (2013) have based their idea on portions of traffic queues instead of regular main lane queues. Kanagaraj et al. (2013) have evaluated the performance of different car following models under mixed traffic conditions. However, they have not taken into account the fact that a vehicle may not be exactly in line with its leading vehicle due to weak lane discipline in mixed traffic. Metkari et al. (2013) have modified an existing car-following model in order to take into account lateral movements and include mixed traffic conditions. Choudhury and Islam (2016) have developed a latent leader acceleration model.

2.6 Qualitative comparison of data–driven and conventional models

A SWOT analysis in using data–driven models versus classical models is shown in Table 2.2. Data–driven models allow the easy incorporation of additional variables avoiding complex reformulations of fixed formulas. Conventional car following models rely on mathematical formulas and are derived from traffic flow theory; a property that often makes them more restrictive. Furthermore, machine learning techniques are non–parametric methods and the calibration of their hyperparameters is not so time–consuming as the calibration process of conventional models is. Data–driven models are based on correlation(Wei, 2014). Using classical models some assumptions on driving behavior are made and then the model is improved through parameters calibration. Instead, data–driven models are generated from data itself and could identify correlations that scientists could not even imagine. Causal inference is self-taught from learning experiences. In such a way, more detailed models are developed. On the other hand, data–driven models may not provide as much insight into traffic flow theory as the traditional models. Correlations between data are identified but only significant ones should be included. Otherwise over-fitted models may be produced. In addition, hidden biases in data may lead to biased models.

Moving to opportunities, machine learning techniques contribute into the deployment of ITS and the effective analysis of large datasets analyzing modern traffic data from multiple sources with different time resolution and spatial coverage. They could also provide robust policies for simulation and autonomous driving. However, there might be hidden threats in opportunities, such as data compatibility and protection of personal data.

Table 2.2: SWOT analysis of using data–driven transportation models

Strengths	**Weaknesses**
• easy incorporation of additional variables • calibration of few hyperparameters instead of complex and time–consuming parameter calibration of conventional models • identification of correlations among data instead of based on fixed formulas • more detailed models	• not theoretically verified • hidden biases in data collection and analysis
Opportunities	**Threats**
• Exploitation of available data • Deployment of ITS	• Data compatibility • Protection of personal data

2.7 Limitations of existing models and research directions

From the foregoing review of the literature, the main limitations of existing models are identified and presented below.

- The majority of models could not represent driving characteristics, for instance reaction time, as vary with traffic conditions (Al-Obaedi et al., 2009). Vehicle–dependent models need to be developed.

- Driving characteristics vary not only for different traffic conditions, but also for different drivers. Driver heterogeneity influences drivers' behavior, perception, aggressiveness, risk awareness and safety constraints.

- The effect of the vehicle size in driving behavior of the follower vehicle is not considered as a factor influencing distance from the leading vehicle (Al-Obaedi et al., 2009) or lane–changing patterns (Moridpour et al., 2010).

- Nowadays, models tend to be more complicated without understanding the physical meaning and the qualitative effect of their parameters.

- Incorporation of additional explanatory variables in conventional models may lead to cumbersome reformulations of formulas.

- Hoogendoorn et al. (2011) have concluded that the assumption of drivers accelerating smoothly may not be valid. Drivers may not pay attention to car–following situation all the time and do not adapt their acceleration respectively (Oud, 2016). Distraction needs to be taken into consideration in future models.

- Stop–and–go waves constitute a common driving experience but most models could not explain wave features and replicate this traffic behavior (Wilson and Ward, 2011).

- Most lane–changing models focus on lane–changing decision and ignore lane changing execution. A lane change is treated as an instantaneous event. However, a driver needs several seconds to execute a lane change (Moridpour et al., 2010). Furthermore, lane changing behavior should depend not only on lead and lag vehicles but also the conditions of the broader traffic range, such as traffic density around lead and lag vehicles (Rahman et al., 2013).

- The strict separation of lane changes into mandatory and discretionary is not realistic, as except for very special cases, such as on-ramp merging traffic, the emergence of a mandatory lane change is unobservable (Toledo et al., 2003).

- Some lane–changing models assume that a lane change is executed when the available space gap is satisfactory. However, these models may not be appropriate for modeling traffic in congested conditions that gaps are created with courtesy of surrounding drivers or smaller headways are accepted.

- Most lane–changing approaches rely on predefined logic–based rules, which explain driving behavior to some extent. However, simplicity and inflexibility of such rules may lead to unrealistic lane–changing simulation.

- Most existing models ignore interactions between car–following and lane-changing decisions and model them separately.

- Traffic simulation researches in mixed traffic conditions is limited and existing models may not perform well in mixed traffic conditions under non-lane discipline. The current models do not consider the wider range of situations that drivers in mixed traffic may face compared to drivers in homogeneous lane-based traffic, such as staggered following, following between two vehicles, and passing and lateral shifts.

- Data collected from multiple sources will play a key role in ITS. However, existing models could neither exploit the information generated from the available data nor incorporate ITS in their functions.

Application of these models in micro-simulation softwares may result in unrealistic traffic flow simulation. The literature review has highlighted a number of areas where further research needs to focus on in order to overcome gaps in existing knowledge. Future models should capture driver heterogeneity through time in various conditions. Furthermore, research should be directed towards integrated behavior models, vehicle–dependent models and traffic simulation in mixed traffic conditions. Data–driven models offer flexibility in incorporating more explanatory variables influencing driving behavior without using complex functions.

Chapter 3

Methodology

The objective of this research is to provide an alternative modeling approach for microscopic traffic simulation models. This modeling approach can take advantage of a wide range of available data, and is therefore suitable to implementation in the context of ITS systems.

3.1 Modeling framework for microscopic data–driven models

The overall process for data–driven model development is outlined in Figure 3.1. The approach includes two parts: training and application. First the required explanatory variables of the model are determined and the appropriate surveillance data are collected. In the training step traffic models are estimated according to the available surveillance data, while in the application step these traffic models are applied to provide predictions using new observations.

The training process is initialized with the identification of clusters based on underlying patterns and in the available data, corresponding to traffic states with similar characteristics. A flexible regression technique is applied to each cluster separately and representative models are formed for each group of the data (calibration). The fitted models are stored into a knowledge database.

In the application step follows, when new measurements become available, the new data are classified to the appropriate classes based on their characteristics. The model that has been estimated for that class is then retrieved from the knowledge base and applied to the new data for the estimation of the response variable.The predicted values are evaluated and the next iteration improves the model.

A methodology, separated in two approaches is presented in Figure 3.2. The first one employs a flexible regression technique, while the second one comprises

[0]The sub–chapter 3.1 is based on Papathanasopoulou and Antoniou (2015a)

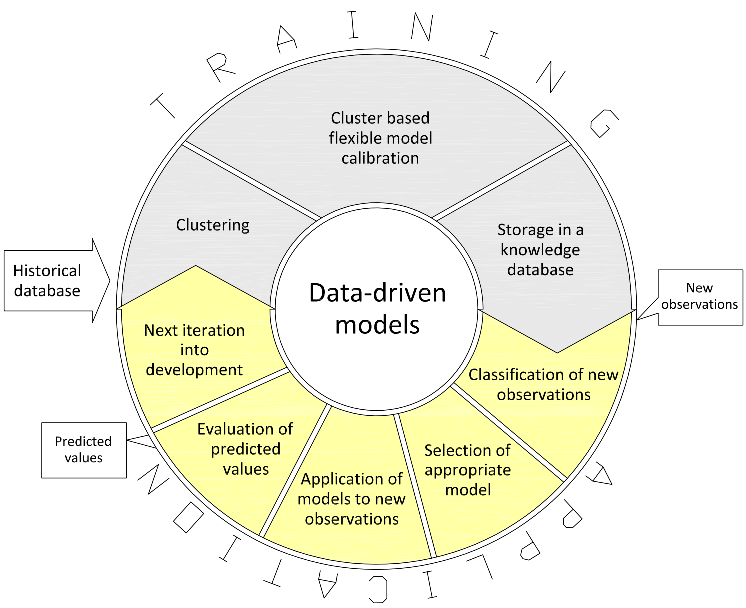


Figure 3.1: Process diagram for data–driven model development

a combination of computational methods, such as flexible regression techniques, model-based clustering and classification algorithms.

Both methodological approaches include two parts: training and application. First the required explanatory variables of the microscopic model are determined and respectively the appropriate data are collected. In the training step traffic models are estimated according to the available surveillance data, while in the application step these traffic models are applied to provide speed predictions for the following vehicle and the next time instant using new observations.

The second methodological approach includes a clustering step to identify portions of the available data that correspond to traffic states with similar characteristics. Then, a flexible regression technique is applied to each cluster separately and representative models are formed for each group of the data (fitting). The application step follows, when new measurements arise. New data are classified to the appropriate classes based on their characteristics. The flexible model that has been estimated for that class is then retrieved from the knowledge base and applied to the new data for the estimation of the response variable, for instance speeds of the following vehicle.

In this research the first methodological approach has been used as the second more elaborate one was not necessary for the available data. T he explanatory variables per each time instant t have been considered as independent predictor variables for the estimation of the response variable (for instance speed) for the next time instant ($t+\tau$), where τ is the apparent reaction time. Estimation is achieved without assuming any predefined functional form; instead a flexible regression method can be used.

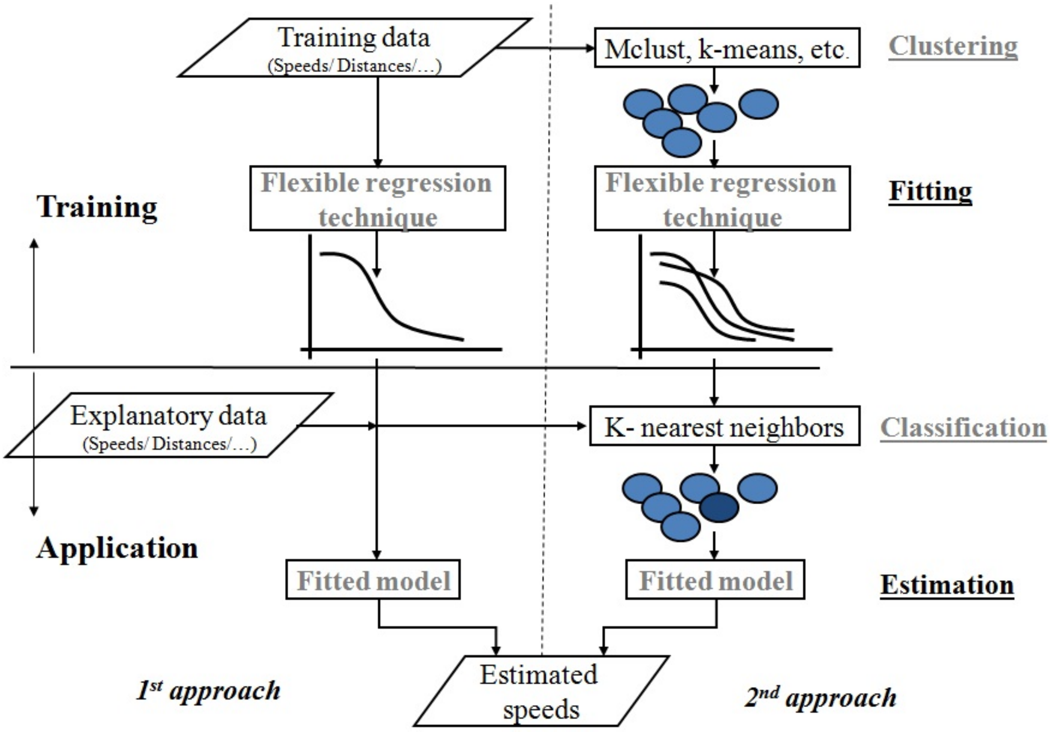


Figure 3.2: Methodology for estimation of flexible microscopic models

The type of driving situation is divided into free flow, car–following and lane–changing according to the Figure 3.3. Longitudinal and lateral positions are recorded per time instant and saved in a database. Then significant lateral changes are identified using appropriate algorithms that allow monitoring structural changes in linear regression models. If no significant lateral change is identified then lateral information is used for determination of lane identification and then a car-following model or a free flow model is applied if at least one preceding vehicle is identified or not respectively. It is common that multiple leader vehicles are identified in heterogeneous traffic conditions and thus the critical leader vehicle should be identified. The probability of a given front vehicle to be the governing leader depends on the type of the lead vehicle and the extent of lateral overlap with the following vehicle (Choudhury and Islam, 2016). On the other hand, if a breakpoint is observed in data sequence, namely if significant lateral changes are identified, then a lane-changing situation is indicated and the lane needs to be modified. A lane–changing model should be applied for time t_L, time of lane-changing duration. Then the process is iterated for the following time instants.

In order to explore car–following and lane–changing behavior, data–driven models could be applied, as described in Figure 3.2. Details on identification of lane–changing behavior and estimation of lane–changing duration is provided in the following chapters. Furthermore, the methodology needs to be modified to adapt to mixed traffic conditions.

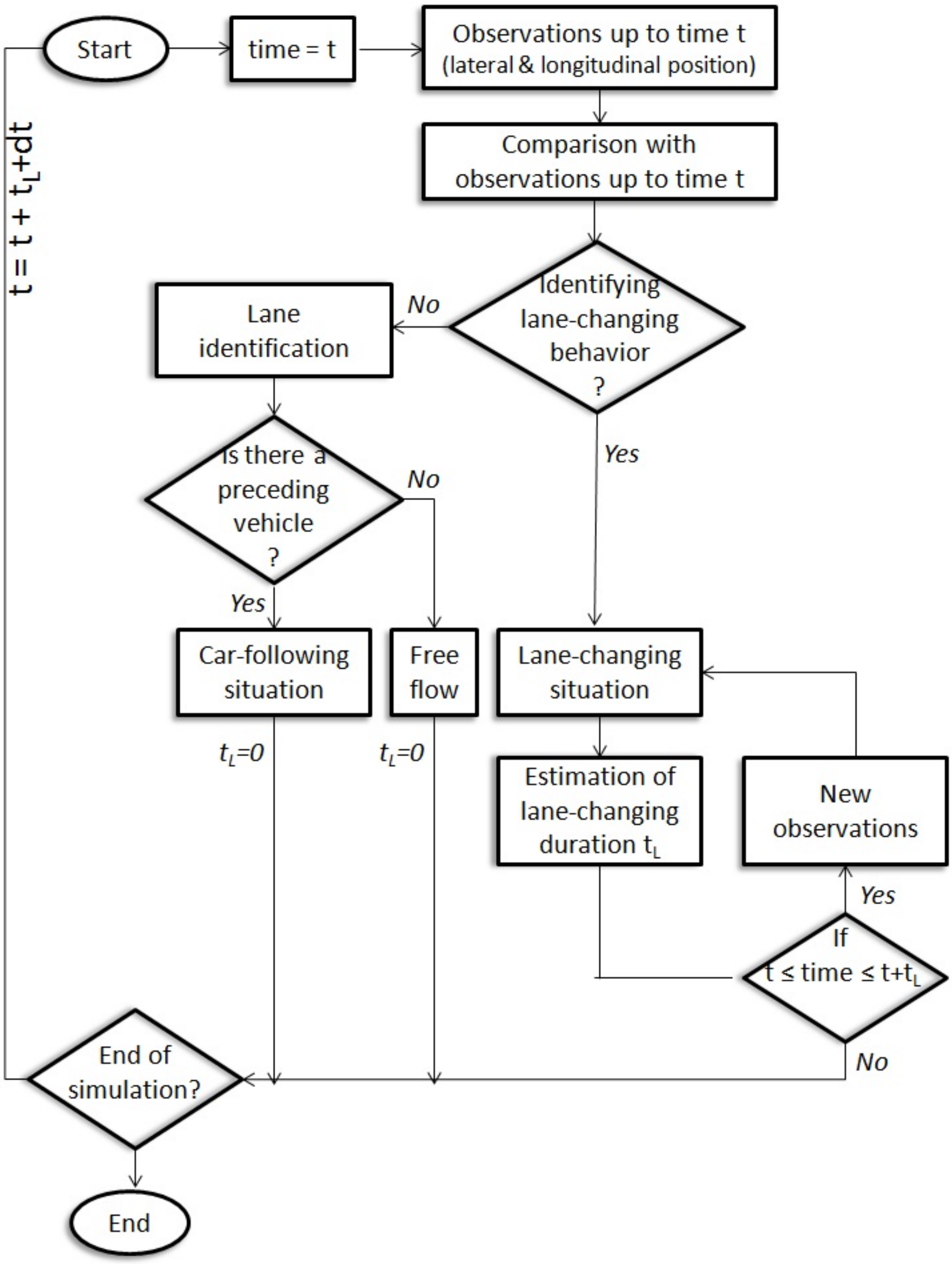


Figure 3.3: Operationalization process

3.2 Data

In order to implement the proposed methodology, trajectory data are required. The data were selected from available multiple sources so as to cover different aspects of three factors: the data collection technology, the environment and the driver (Figure 3.4). The feasibility of the proposed methodology should be checked using data collected from different technologies (cameras or GPS). Regarding the environment, traffic network and traffic rules may be differentiated between different continents or even between different countries (Oud, 2016). Furthermore, different conditions prevail in terms of traffic, such as congested conditions or mixed conditions. In order to capture heterogeneity in driving behavior, data should definitely capture different driving behaviors. A mixture of heterogeneous data allows the validation of the proposed methodology from different perspectives. Last but not least, the selected data should be appropriate for study of car–following, lane–changing models, as well as models for mixed traffic. Data selection plays a key role for data–driven models, as the models learn from the data. The data selected are briefly described in Figure 3.5 and are analyzed in detail in the following chapters.

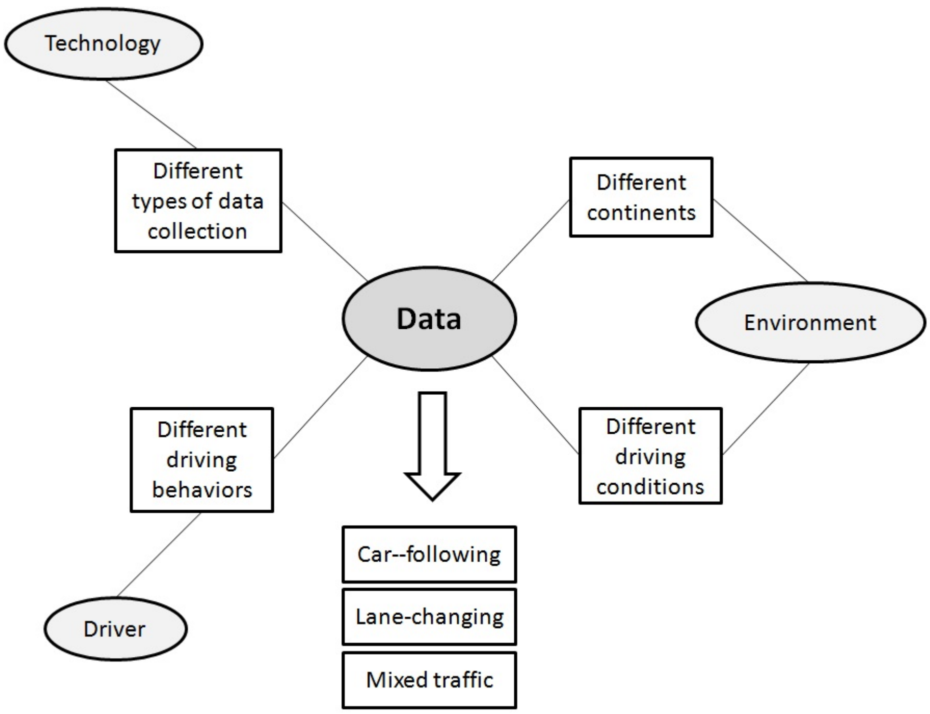


Figure 3.4: Data selection criteria

3.2.1 Naples data

A series of data–collection experiments were carried out on roads surrounding the city of Naples, in Italy (Punzo et al., 2005). All data were collected from the same platoon under real traffic conditions in October 2002. The same four drivers were moving by the same vehicles (vehicles 1, 2, 3, 4) in the same sequence

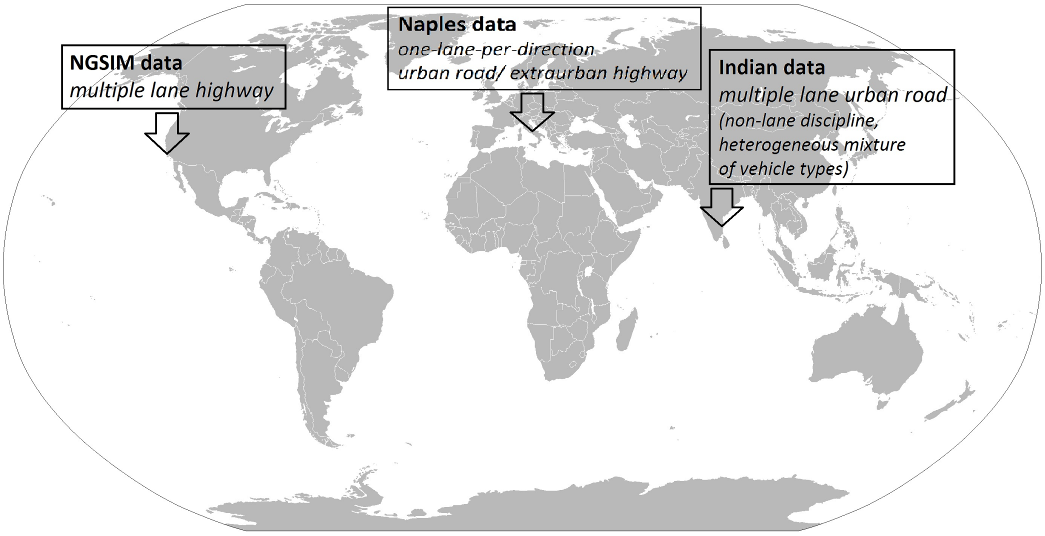


Figure 3.5: Data selection for implementation of the proposed methodology

(first vehicle 1 as the leader, followed by vehicle 2, which was in turn followed by vehicle 3, while the last vehicle was vehicle 4), but from different driving sessions. The driving routes and traffic conditions were differentiated among the datasets. Datasets with index A and C correspond to one–lane urban road, while datasets with index B to a two–lane extraurban highway (Figure 3.8). However, all selected roads have one lane per direction in order to avoid effects on driving behavior by lane changing. GPS receivers located on the vehicles were recording the coordinates X, Y, Z of each vehicle per 0.1s (i.e. in 10Hz). Thus, the speed of each vehicle ($v_1(t)$, $v_2(t)$, $v_3(t)$, $v_4(t)$) and the traveled distances for each vehicle could be calculated at each moment ($x_1(t)$, $x_2(t)$, $x_3(t)$, $x_4(t)$). In this research, data used are readily available observations from the field. No corrections and no interpolation have been performed. Therefore, only segments with consecutive measurements have been considered. The data series include location records of each vehicle (coordinates x, y, z and time) per 0.1 s for all the vehicles. Using the above information, the vehicle headways, the distance traveled per 0.1 s for each vehicle, and their respective speeds were calculated. The size and speed ranges of each data series are shown in Table 3.1 and Figures 3.6 and 3.7. A detailed description of the data could be found in Punzo et al. (2005), who kindly provided the data for this research. Trajectory and speeds are plotted indicatively for data series B1695 in Figures 3.9 and 3.10.

3.2.2 NGSIM data

The "Next Generation SIMulation (NGSIM)" program (http://ngsim.fhwa.dot.gov.) includes vehicle trajectories in real traffic conditions, which -–along with other output of the project– have become available to the scientific community for re-

Table 3.1: Characteristics of Naples data

a/a	Dataset	No. Observations	Duration (s)
1	B1695	1695	169.4
2	C621	621	62.0
3	A358	358	35.7
4	A172	172	17.1
5	C168	168	16.7
6	C171	171	17.0

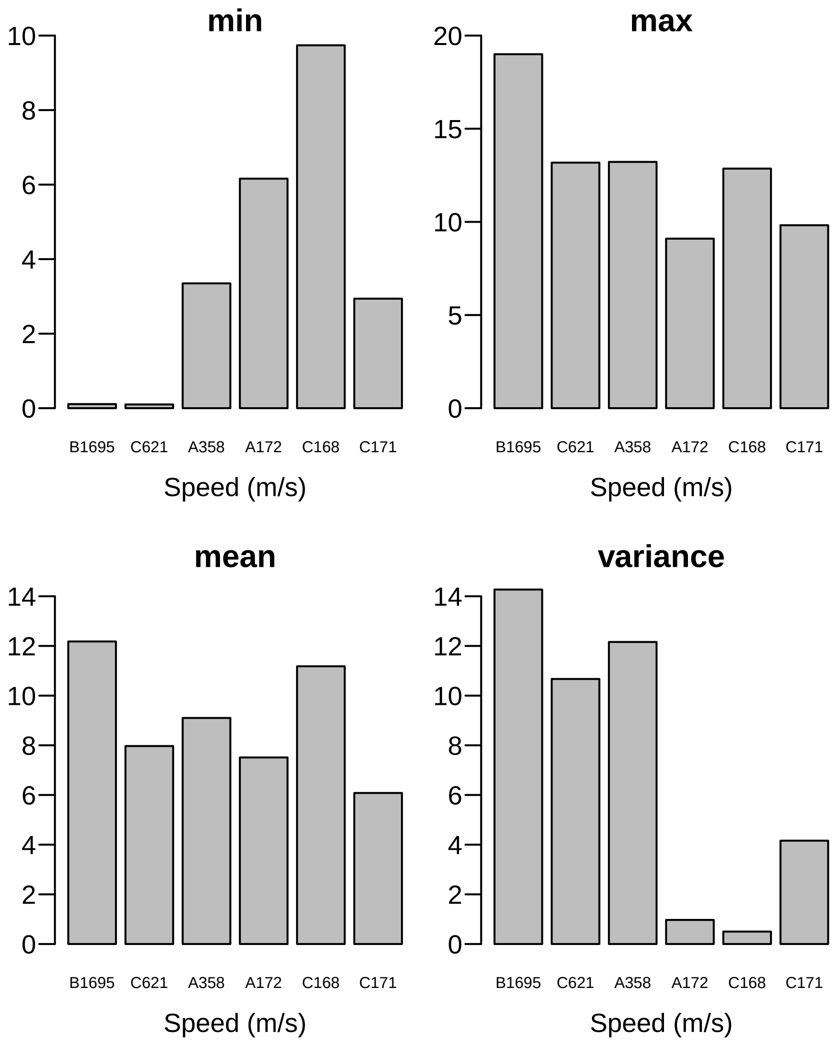


Figure 3.6: Summary statistics of speed for Naples data

search of microscopic driving behavior. The considered NGSIM data were collected on eastbound I-80 in the San Francisco Bay area in Emeryville on April

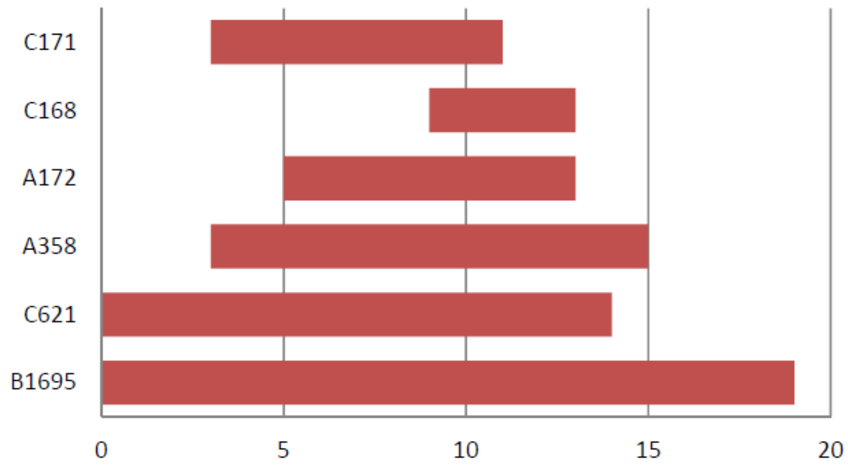


Figure 3.7: Speed ranges for Naples data

13, 2005 (US Department of Transportation 2012). The study area extends approximately 500m in length and consists of six freeway lanes (Figure 3.11). Seven modern digital cameras were mounted on the top of a 30-story-building adjacent to the freeway and were recording passing vehicles. The custom NG–VIDEO software application transformed video to vehicle trajectories data (at 10Hz). These data were recorded mainly in congested conditions. 45 minutes of data are available in a data set divided into three periods of 15 minutes and particularly in accordance with the register time, 4:00 p.m. to 4:15 p.m., 5:00 p.m. to 5:15 p.m., and 5:15 p.m. to 5:30 p.m.

NGSIM data have been used in many studies for calibration or validation of existing models (e.g. (Bevrani and Chung, 2011)). In the years 2007-2008 more than 30 studies used the NGSIM data (Punzo et al., 2011). However, only few studies have raised the issue of their accuracy (Hamdar and Mahmassani, 2008; Punzo et al., 2011; Thiemann et al., 2008). Although the way that the velocities and accelerations of vehicles were calculated and the errors were reduced is not known, studies suggest the existence of residual noise and errors in the data (Bevrani and Chung, 2011; Punzo et al., 2011). The complete set of NGSIM vehicle trajectory data from the I80–1 dataset (from 4.00 p.m. to 4.15 p.m.) was filtered with a multi–step procedure for vehicle trajectory reconstruction by Montanino and Punzo (2014). For each vehicle the available data which are taken into account are: Vehicle ID, Frame ID (Frame Identification number, ascending by start time), Lane ID, LocalY (Longitudinal Y coordinate of the front center of the vehicle with respect to the entry edge of the section in the direction of travel [m]), Mean Speed, Mean Acceleration, Vehicle length, Vehicle Class ID, Follower ID, Leader ID. More information about the Enhanced NGSIM data could be found in (M. Montanino and V. Punzo, 2015; Montanino and Punzo, 2013; Punzo et al., 2011).

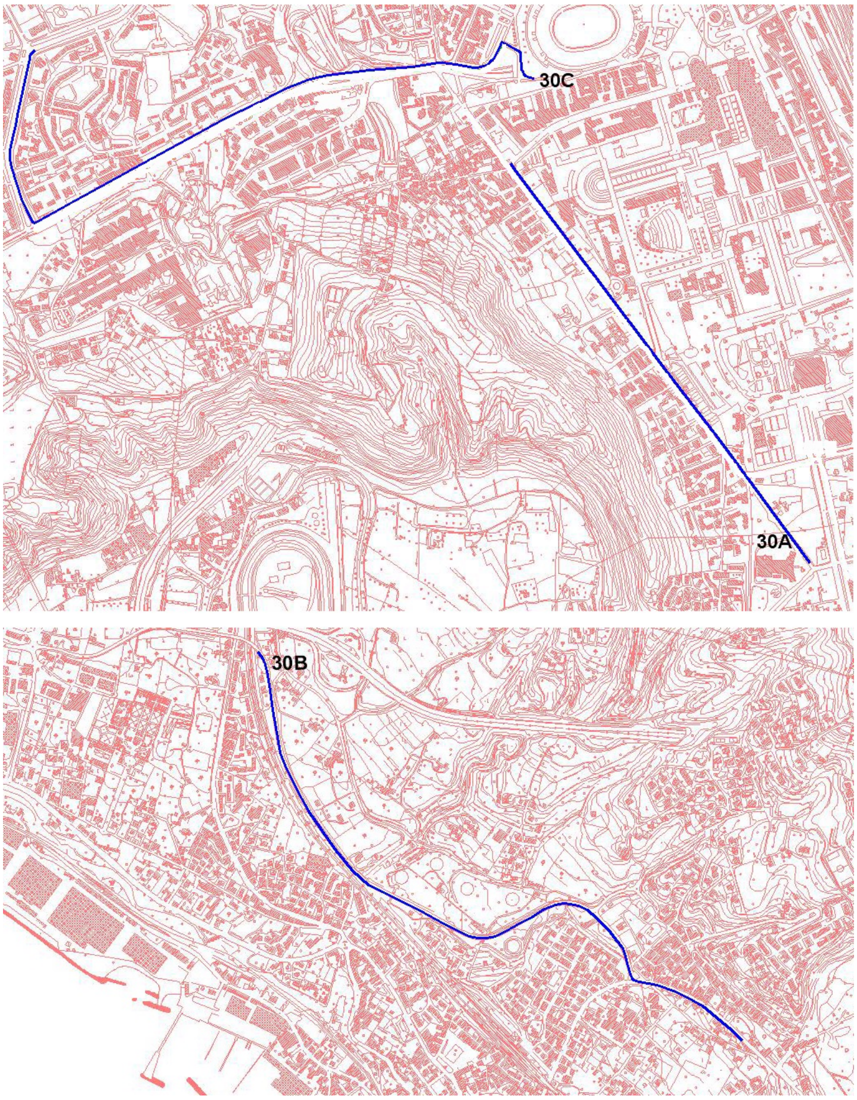


Figure 3.8: Routes from Naples data

The available enhanced NGSIM data include 1055800 observations. Due to frequent lane changing, 10 vehicles moving only in a car–following state were chosen for this analysis. Vehicles moving in the same lane and in sequence one after the other were easily recognized according to the lane identification and Follower/ Leader ID. The data selected are presented in Table 3.2. Speed densities of the selected data series are plotted in Figure 3.12.

3.2.3 Indian data

In order to evaluate the feasibility of the methodological framework, data collected in India were used (Kanagaraj et al., 2015). The video data were collected on a six-lane separated urban arterial road at the Maraimalai Adigalar Bridge in Saidapet, Chennai, India. Collection took place on the northbound approach, as shown in Table 3.3. The section was on a bridge, which ensured that the road geometry

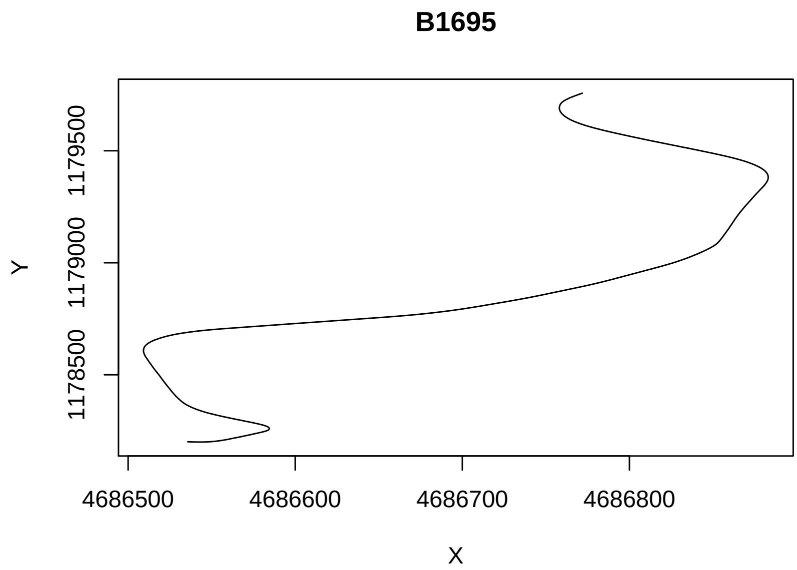


Figure 3.9: Trajectory for B1695 data series

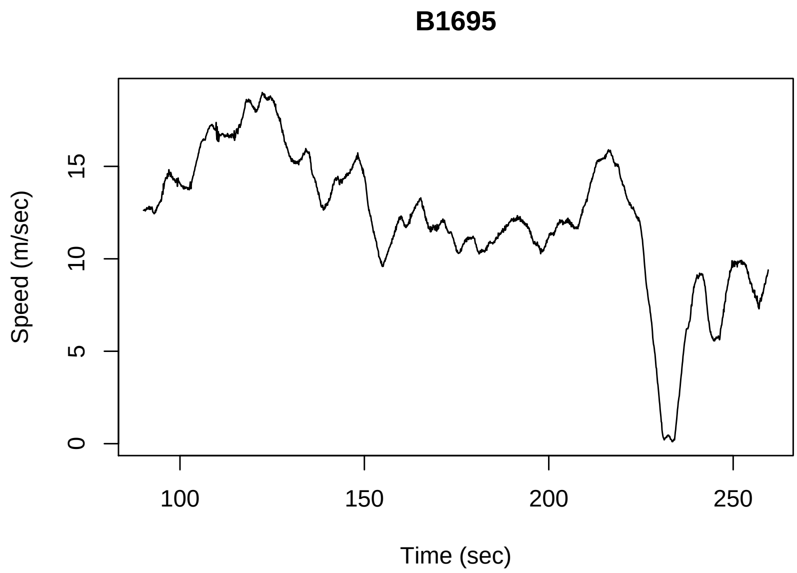
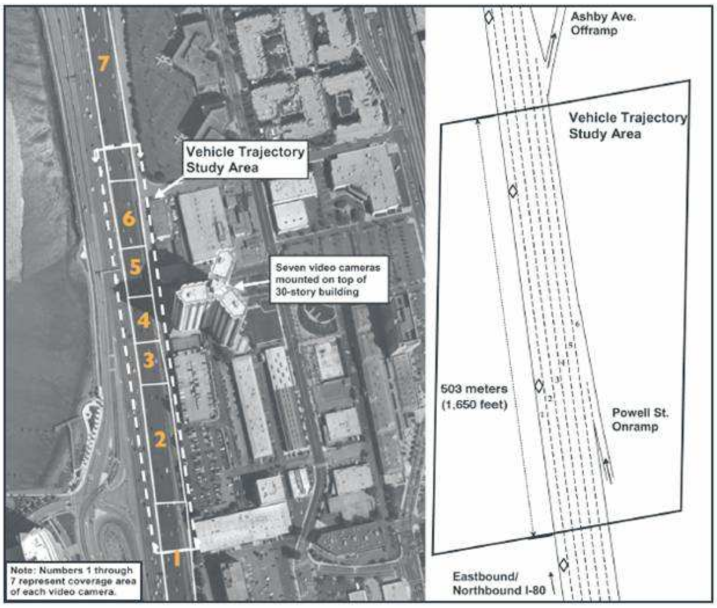


Figure 3.10: Speeds for B1695 data series

was uniform and that there were no nearby intersections, bus stops, parked vehicles, or other side factors that could affect drivers' behavior. Furthermore, there was no interaction between vehicle traffic and pedestrians, because the pedestrian walkway is segregated by a barrier. A detailed description of the data can be found in Kanagaraj et al. (2015). The data are presented in two parts – two excel files for the data collected in the periods 2:45–3:00 PM (data245) and

(a) A digital video camera recording vehicle trajectory data



(b) Aerial photograph and schematic drawing of the I–80 study area

Figure 3.11: NGSIM data collection

Table 3.2: Characteristics of Data Series

a/a	Observations	No. Observations	Duration (s)
1	405108:405521	414	41.4
2	422895:423105	211	21.1
3	733723:733967	245	24.5
4	954322:954709	388	38.8
5	1006645:1006788	144	14.4
6	1000238:1000463	226	22.6
7	84232:84592	361	36.1
8	971206:971558	353	35.3
9	972848:973219	372	37.2
10	972848:973219	330	33.0

3:00–3:15 PM (data300), respectively, on February 13, 2014. Each excel sheet contains 12 columns of data, as described in Table 3.3. The trajectory data is available publicly at the address http://toledo.net.technion.ac.il/downloads/.

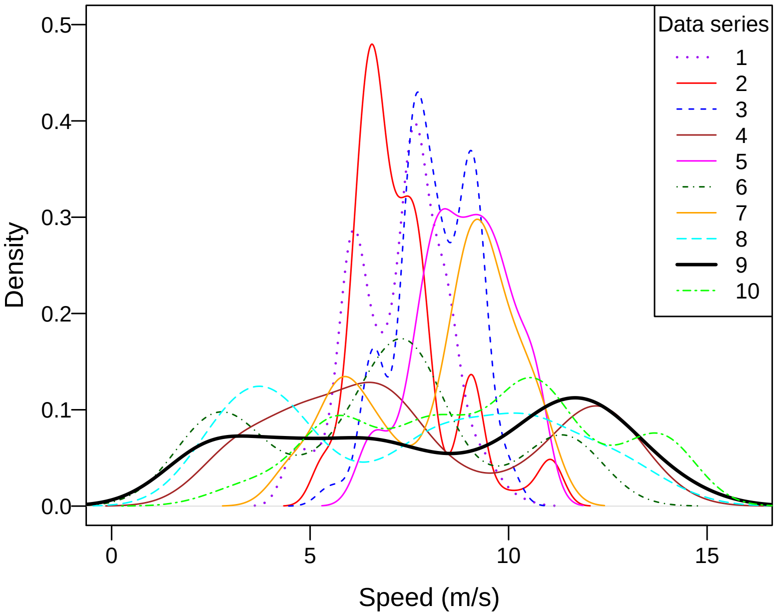


Figure 3.12: Speed densities in data series

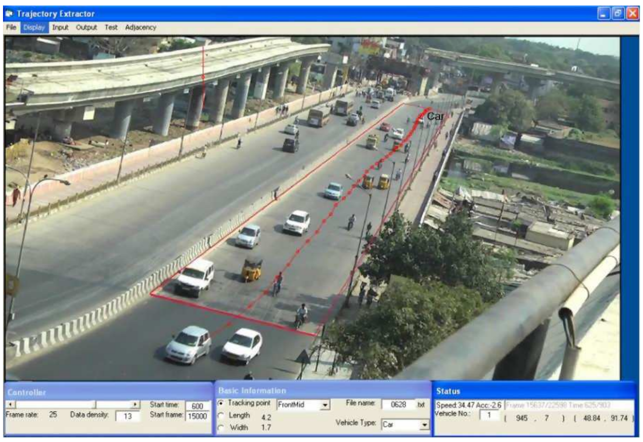


Figure 3.13: Trajectory extractor user interface showing road section and reference points in India

3.3 Methodological components

Data–driven models can greatly benefit from efficient machine learning algorithms. A number of machine learning techniques, including loess, splines, support vector machines, Gaussian processes and neural networks, are analyzed and used for implementation of the proposed methodology.

Table 3.3: Description of Indian data

Column	Item	Remarks
2	Vehicle type	1– motorcycle, 2– car, 3– bus, 4– truck, 5– light commercial vehicle, 6– auto–rickshaw
3	Time (s)	Time from beginning of recording. The time interval is 0.5 s.
4	Vehicle length (m)	–
5	Vehicle width (m)	–
6	Longitudinal position (m)	Position of the front of the vehicle, measured from the upstream end of the section.
7	Longitudinal speed (m/s)	Instantaneous speed of the vehicle
8	Longitudinal acceleration (m/s^2)	Instantaneous acceleration of the vehicle
9	Lateral position (m)	Position of the center of the vehicle, measured from the left–most side of the roadway
10	Lateral speed (m/s)	Instantaneous speed of the vehicle. Positive (negative) values represent movement to the right (left).
11	Lateral acceleration (m/s^2)	Instantaneous acceleration of the vehicle
12	Flag	Represents manual correction of data points where overlap among vehicles occurred. 0: No correction, 1 through 7: Data was manually modified to eliminate overlap with other vehicles. It is suggested that these observations will not be used for microscopic–level analysis.

3.3.1 Clustering and classification

Clustering

A simple form of clustering is **k-means** algorithm. As its name suggests, the k–means algorithm (MacQueen et al., 1967; Hartigan and Wong, 1979) minimizes the distance between each point and the center of its cluster for k given clusters. This is achieved by assigning each point to the nearest mean and reestimating or moving the mean to the center of its cluster. It is regarded as a maximum likelihood clustering. The objective function to be minimized is:

$$min_{(\mu_1,\ldots,\mu_k)} \sum_{h=1} \sum_{x \in X_h} \|X - \mu_h\|^2 \tag{3.1}$$

where μ_i is the mean of cluster i. A hypothesis $h_1 =< \mu_1, \ldots, \mu_k >$ with the means of the k different normal distributions is requested. A random hypothesis is assumed for the initialization of the procedure. Each instance could be written as $< x_i, z_{i1}, z_{i2}, \ldots, z_{ik} >$ where x_i is the observed variable and z_{ij} is equal to 1 if it was obtained by the j_{th} normal distribution or 0 otherwise. A maximum–likelihood hypothesis is sought after iterative re-estimations of the expected values of z_{ij}. Then, a new maximum likelihood hypothesis h_2 is calculated using the expected values in the previous step. Finally, the new hypothesis replaces the earlier one and iterations are going on until the algorithm converges to a value for the hypothesis.

Fraley and Raftery (2002, 2003) proposed a model based clustering which combines hierarchical clustering, expectation–maximization algorithm (EM algorithm) for mixture models and Bayesian information Criterion (BIC) for selection of models and number of classes (Schwarz et al., 1978). Hierarchical clustering, used for model–based hierarchical agglomeration, is initialized by default with each observation of the data in a cluster by itself and finished when all observations have been merged into a cluster. A classification maximum likelihood approach is required to determine which two groups are merged at each stage (Banfield and Raftery, 1993; McLachlan and Krishnan, 1997; Fraley, 1998). EM algorithm is included in the R **Mclust** package and is applied for maximum likelihood clustering with parameterized Gaussian mixture models (Dempster et al., 1977; McLachlan and Krishnan, 1997). The EM algorithm is implemented in two steps: E–step which calculates a matrix z_{ik}, which corresponds to the likelihood of an observation i to be merged into a cluster κ given the current parameter estimates, and M-step, which calculates maximum likelihood parameter estimates given z. Each cluster is represented by a Gaussian model $\phi K(x|\mu_\kappa, \Sigma_\kappa)$, where x are the data, κ an integer indicating a cluster centered at means μK and covariances Σ_κ. Then the maximum likelihood values for the Gaussian mixture model is given by Equation 3.2 (Fraley and Raftery, 2002), where τ_κ are the mixing proportions.

$$f(z) = arg \min_{y \in} d(z, y) \tag{3.2}$$

Banfield and Raftery (1993) suggested a clustering strategy based on a maximization algorithm and Bayes factors. This strategy was upgraded by Fraley

(1998) and later by Fraley and Raftery (2002, 2003) and could be carried out with the following steps:

- A maximum number of clusters and a subset of covariance structures are considered
- A hierarchical agglomeration that maximizes the classification likelihood for each model is performed and the appropriate classifications are illustrated up to M groups.
- The EM algorithm is applied for each model and each number of clusters 2,..., M. The procedure is initialized from the classification result of hierarchical agglomeration.
- The Bayesian information Criterion BIC is calculated for the one-cluster case for each model and for the mixture model with the optimal parameters from EM for 2,..., M clusters. Each combination corresponds to a unique probability model.
- The model with the highest BIC is selected and the best classification is recovered. Although in such a way the optimal number of classes is determined, a lower number of classes could be chosen, aiming at the development of more parsimonious models.

Classification

One of the most common methods of classification is k–nearest neighbors (Mitchell et al., 1997). According to this method, all observations correspond to points in n–dimensional space. Future data points are registered in the class of nearest neighbors of the already grouped data. Especially, the point of the nearest neighbor classification is the calculation of the correlation map:

$$d(x_i, x_j) = \sqrt{(\sum_{r=1}^{n}[a_r(x_i) - a_r(x_j)]^2)} \tag{3.3}$$

In a pattern space P, where $M \subseteq P$, $z \in P$ and d(z,y) is a metric in P–dimensional space. The evaluation of Equation 3.3 could be easily achieved on a computer following three steps: computation of an array with distances from z to each $y \in M$, finding the minimum distance after comparisons and exporting the final result $y^* \in M$ (Muezzinoglu and Zuracla, 2005). The nearest neighbors could be defined according to the Euclidean distance (Roughan et al., 2004), if a point x is described as $< a_1(x), a_2(x), \ldots, a_n(x) >$ where $a_r(x)$ corresponds to the value of the r–th attribute of x. Attributes of x could include density, traffic flow, and time. The distance between two points is defined by Equation 3.4 (Mitchell et al., 1997). Thus the class of a new observation x_i is the same as the class of point x_j, which minimizes the distance $\|x_i - x_j\|$.

$$f(z) = arg \min_{y \in M} d(z, y) \tag{3.4}$$

3.3.2 Flexible fitting models

Loess

Locally weighted regression (loess) could be considered as a generalization of the k-nearest neighbor method (Mitchell et al., 1997). It was firstly introduced by Cleveland (1979) and the following analysis is based on Cleveland and Devlin (1988).

Locally weighted regression y_i = $g(x_i)$ + ϵ_i, where i=1,..., n index of observations, g is the regression function and ϵ_i are residual errors, provides an estimate g(x) of each regression surface at any value x in the d-dimensional space of the independent variables. Correlations between observations of the response variable y_i and the vector with the observations d-tuples x_i of d predictor variables are identified. Local regression provides an estimation of function g(x) near x = x_0 according to its value in a particular parametric class. This estimation could be achieved by adapting a regression surface to the data points within a neighborhood of the point x_0, which is bounded by a smoothing parameter: span. The span determines the percentage of data that are considered for each local fit and hence the smoothness of the estimated surface is influenced (Cohen, 1999). The span ranges from 0 (wavy curve) to 1 (smooth curve). Each local regression uses either a first or a second degree polynomial that it is specified by the value of the "degree" parameter of the method (degree=1 or degree=2).

The data are weighted according to their distance from the center of neighborhood x, therefore a distance and a weight function are required. As a distance function p, Euclidean distance is used for a single independent variable; otherwise, for the multiple regression case, any variable should be evaluated on a scale before applying a standard distance function (Cleveland et al., 1988). A weight function defines the size of influence on fit for each data point taking for granted that nearby points have higher influence than the most distant. Therefore the weight function calculates the distances between each point and the estimation point and higher values in a scale from 0 to 1 are set for the nearest observations. A weight function should meet the requirements determined by Cleveland (1979) and the most common one is the tri–cube function:

$$W(u) = \begin{cases} (1 - u^3)^3, 0 \leq u \leq 1 \\ \\ 0, otherwise \end{cases} \tag{3.5}$$

The weight of each observation (y_i, x_i) is defined as following:

$$w_i(x) = W[p(x, x_i)/d(x)] = (1 - (\frac{x_i - x}{d(x)})^3)^3 \tag{3.6}$$

where d(x) is the distance of the most distant predictor value within the area of influence. In the loess method, weighted least squares are used so as linear or quadratic functions of the independent variables could be fitted at the centers of neighborhoods (Cleveland, 1979). The objective function that should be minimized is:

$$\sum_{n=1}^{n} w_i \cdot \epsilon_i^2 \tag{3.7}$$

Multivariate Adaptive Regression Splines (MARS)

Multivariate adaptive regression splines (MARS) have been introduced by Friedman (1991). It is a non-parametric method for flexible regression modeling of high dimensional data that identifies nonlinearities and interactions between variables. In this research, this method is implemented using package 'earth' (Milborrow, 2017) in R (R Core Team, 2017). MARS builds a model of the form:

$$f(x) = \sum_{i=1}^{k} c_i \cdot B_i(x) \tag{3.8}$$

The model is a weighted sum of basis functions $B_i(x)$ where c_i are coefficients estimated by minimizing the residual sum of squares (Happe et al., 2010). The model strategy is similar to stepwise linear regression, except that the basis functions are taken into account instead of the observations. An independent variable translates into a series of linear segments joint together at points called knots (Courtois and Woodside, 2000). Each segment uses a piecewise linear basis function which is constructed around a knot. MARS selects dynamically the knot locations. It is a forward pass– backward pass process in order to decrease the training error. Optimal number of terms in the model is estimated using generalized cross validation (Happe et al., 2010).

Kernel support vector machines (KSVM)

Support vector machines are based on the Structural Risk Minimization principle (Cortes and Vapnik, 1995). An SVM model is a representation of training data as points in space. Training a support vector machine (SVM) leads to the following quadratic optimization problem with bound constraints and one linear equality constraint (Cortes and Vapnik, 1995).

$$W(a_1 \ldots a_n) = -\sum_{i=1}^{n} a_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i \cdot y_j \cdot a_i \cdot a_j \cdot K(x_i, x_j) \tag{3.9}$$

subject to

$$\sum_{n=1}^{n} y_i \cdot a_i, 0 < a_i < C \tag{3.10}$$

Where n is the dimensionality of α_i, each component α_i corresponds to a training example (x_i, y_i), K(x_i,x_j) is the kernel function which is used as a similarity measure between objects x_i and x_j and C is an upper bound on α_i.

Gaussian processes

Gaussian processes are based on the idea that adjacent observations convey information about each other (Williams and Rasmussen, 1996). Observations are considered to be normal and the relationship between them is represented by a covariance matrix of a normal distribution. Kernel matrix is used as the covariance matrix in order to extend Bayesian modeling to non-linear situations. The following analysis is based on (Quiñonero-Candela and Rasmussen, 2005). For regression estimation it is assumed that observations $f(x)$ could be written as $y(x)$=$f(x)$+ϵ, where ϵ is Gaussian distribution noise with zero mean, $\epsilon\sim$ N(0, σ_n^2). The number of training data is n. A Gaussian distribution is fully described by the mean μ and covariance Σ of the distribution in terms of hyperparameters θ. The log marginal likelihood is given by Equation 3.11.

$$L = logp(y|x,\theta) = -\frac{1}{2}log|\Sigma| - \frac{1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu) - \frac{n}{2}log(2\pi) \tag{3.11}$$

Bayesian regularized neural networks (BRNN)

In the Bayesian framework, model parameters are treated as probabilistic variables. The posterior probability of the weights is given according to Bayes' rule by Equation 3.12.

$$p(w|D) = \frac{p(w|D)p(w)}{p(D)} \tag{3.12}$$

where D is a set of observations, $p(w|D)$ is the probability of observations given a choice of weights w, p(w) is a prior distribution of weights and p(D) is a normalization factor.

BRNN address one of the difficulties in building a neural network, i.e. determining the number of hidden neurons. To overcome this difficulty, the BRNN algorithm incorporates the Bayes' theorem into the regularization scheme. Foresee and Hagan (1997) and MacKay (1992) provide a detailed description of Bayesian regularized neural networks. It uses the Nguyen and Widrow algorithm (Nguyen, 1990) to assign initial weights and the Gauss–Newton algorithm to perform the optimization.

The model is given by:

$$y_i = g(x_i) + e_i = \sum_{k=1}^{s} w_k g_k(b_k + \sum_{j=1}^{p} x_{ij}\beta_j^{[k]}) + e_i, i = 1, \ldots, n \tag{3.13}$$

where
e_i ~N(0,σ_e^2), s is the number of neurons,
w_k is the weight of the k–th neuron, k=1,. . . , s
b_k is a bias of the k–th neuron, k=1,. . . , s
$\beta_j^{[k]}$ is the weight of the j–th input to the net, j=1,. . . , p
g_k is the activation function in this implementation

$$g_k(x) = \frac{exp(2x) - 1}{exp(2x) + 1} \tag{3.14}$$

The software will minimize

$$F = \beta E_D + \alpha E_W \tag{3.15}$$

where

$$E_D = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \tag{3.16}$$

i.e. the error sum of squares,
E_W is the sum of squares of network parameters (weights and biases)

$$\beta = \frac{1}{2\sigma_e^2} \tag{3.17}$$

$$\alpha = \frac{1}{2\sigma_\theta^2} \tag{3.18}$$

σ_θ^2 is a dispersion parameter for weights and biases.

3.4 Goodness–of–fit measures

3.4.1 Single aggregate measures

The performance of the models presented in this paper is evaluated using several goodness–of–fit measures: RMSN, RMSPE, MPE and Theil's U, U_m and U_s coefficients (for details and a discussion of these metrics, see e.g. Antoniou et al. (2013)). Different measures are used so that the properties of the calibration and validation results could be quantified from different views. For example, the normalized root mean square error (RMSN) assesses the overall error and performance of each method estimating the difference between the observed values Y_n^{obs} and their simulated counterparts Y_n^{sim}. The root mean square percentage error (RMSPE) penalizes large errors more heavily than small errors and the mean prediction error (MPE) indicates the existence of systematic under– or over–estimation in the simulated values. The measure of Theil's inequality coefficient U has been applied in transport model validation and includes three error proportions: the bias (U_m), the variance (U_s) and the covariance (U_c), whose sum is one. Values close to zero for U_m and U_s measures indicate an ideal fit, while values close to 1 suggest the worst fit. The goodness–of–fit measures are calculated from the following equations:

$$RMSN = \frac{\sqrt{N \cdot \sum_{n=1}^{N} (Y_n^{sim} - {Y_n}^{obs})^2}}{\sum_{n=1}^{N} Y_n^{obs}} \tag{3.19}$$

$$RMSPE = \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} \left(\frac{Y_n^{sim} - {Y_n}^{obs}}{{Y_n}^{obs}} \right)^2} \tag{3.20}$$

$$MPE = \frac{1}{N} \cdot \sum_{n=1}^{N} \left(\frac{Y_n^{sim} - Y_n^{obs}}{Y_n^{obs}} \right) \tag{3.21}$$

$$U = \frac{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{sim} - {Y_n}^{obs})^2}}{\sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{sim})^2} + \sqrt{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{obs})^2}} \tag{3.22}$$

$$U_m = \frac{(\bar{Y_n}^{sim} - \bar{Y_n}^{obs})^2}{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{sim} - {Y_n}^{obs})^2} \tag{3.23}$$

$$U_s = \frac{(\sigma^{sim} - \sigma^{obs})^2}{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{sim} - {Y_n}^{obs})^2} \tag{3.24}$$

$$U_c = \frac{2 \cdot (1 - p) \cdot \sigma^{sim} \cdot \sigma^{obs}}{\frac{1}{N} \cdot \sum_{n=1}^{N} (Y_n^{sim} - {Y_n}^{obs})^2} \tag{3.25}$$

3.4.2 Distribution–based evaluation

Unlike dealing with absolute numbers as the measure of effectiveness, which is the commonly used evaluation approach, in this research we are proposing the use of distributions. Besides the advantages that derive from the richness of these data, there are also significant challenges regarding the statistics that will be used for distributions' comparison. In particular, while the statistics that are needed to compare numerical values have been widely used and there is experience in using them, working with distributions imposes novel challenges.

The distribution–based evaluation approach assumes as input a set of measured distributions of the appropriate measures of effectiveness (e.g. speeds or accelerations). The data need to be appropriately pre–processed, to ensure they are not susceptible to measurement or equipment error, and that they comply with the experiment requirements. These distributions are then compared with the simulated distributions, based on a series of goodness–of–fit measures, depending on the requirements of the application. If, according to the set convergence criteria, the simulated distributions accurately capture their observed counterparts, then the process concludes that models are considered suitable for the application at hand. Otherwise, the models need to be fitted again using more data.

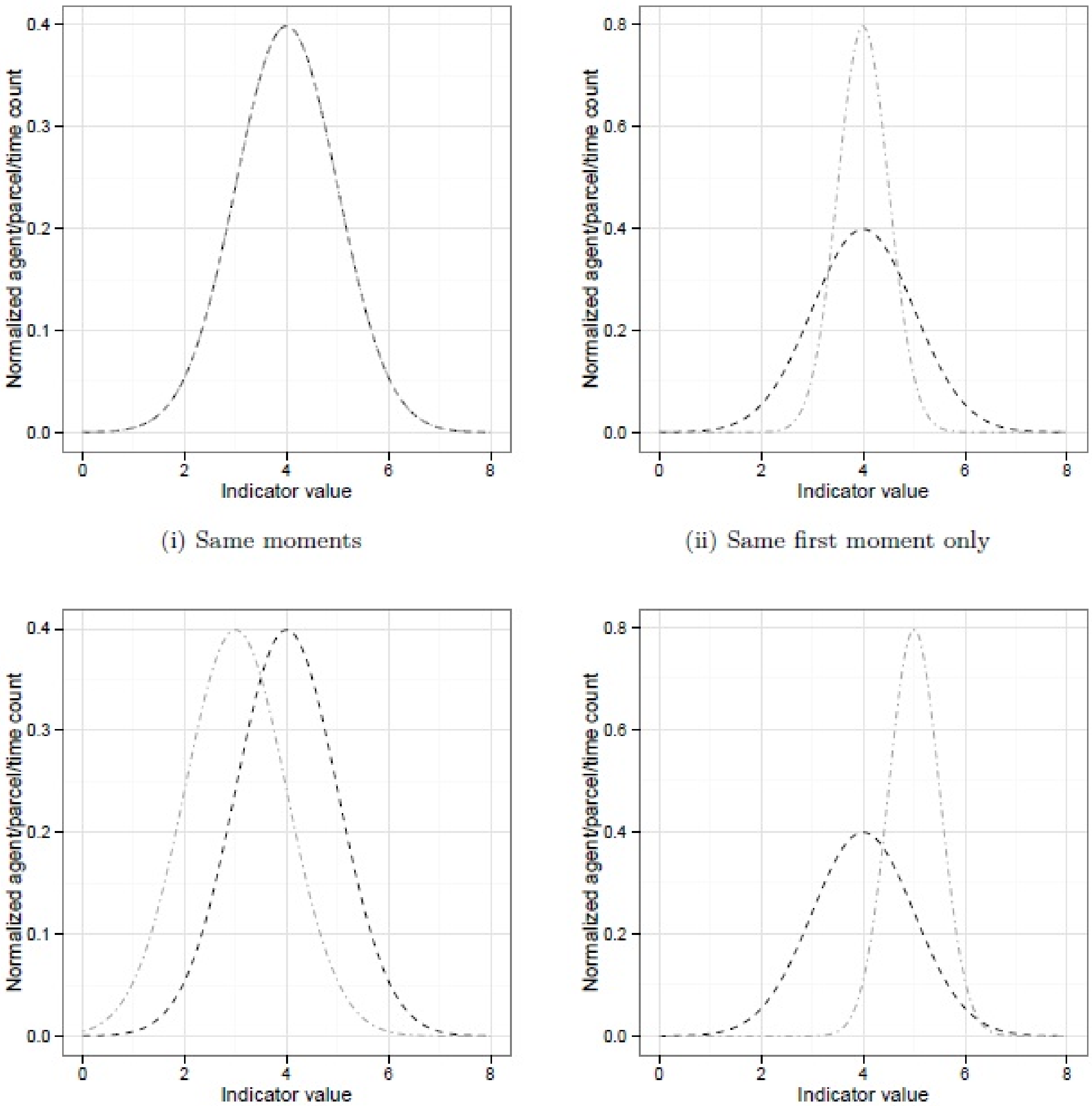
The simplest approach would probably be to compare the first few moments of the distributions (Figure 5.10). According to Ramsey et al. (2002) almost all the information about the shape of histograms (location, spread, symmetry, and peakedness) could be described by four numbers. These numbers are the averages of powers of the variable values, named moments. The mean of a distribution, namely the first moment, is a measure of location. In order to find out if two distributions have means that are significantly different, a statistical t–test may be applied (Press et al., 1992). The second moment is the average squared deviation of the variable's values from the mean and it is used as a measure of spread. For measuring how significantly different the variances of two distributions are, an appropriate statistical test is the F–test (Press et al., 1992). The F–test suggests that two variances are significantly different by rejecting the null hypothesis that they are both consistent. The third and fourth moments imply measures of skewness and kurtosis accordingly. Higher moments could also been taken into consideration. However, many distributions are not uniquely determined by their moments (Lindsay and Basak, 2000; McCullagh, 1994; Stoyanov, 2006).

Another approach could be to create a metamodel (Santos and Santos, 2007; Antoniou et al., 2014a; Ciuffo et al., 2013; Pereira et al., 2014) of the available data (and their simulated counterparts), and then work with the parameters of the metamodel. For example, instead of comparing the distributions themselves, one could perform F-tests with the null hypothesis that the parameters of the metamodels are jointly equal. Santos and Santos (2007) claimed that an improved comprehension of the system could be obtained by using the normal distribution mean and variance functions. The parameters of the distribution metamodel could be estimated using the least squares method. Metamodels are not accurate representations of the original model and may include a degree of uncertainty. If two or more metamodels ensure sufficiently accurate fit capture, then it is preferable to choose the simplest one. These approaches are computationally and conceptually attractive, but might lead to some loss of information. Therefore, this tradeoff between loss of information and gain in simplicity should be carefully evaluated.

The selection of the most appropriate distribution could be assisted by distribution fitting software packages. Such packages calculate goodness of fit statistics and can support decisions regarding the distribution with the best fit on the data. However, there is the dilemma if the chosen distribution and its parameters capture well the original output or not. The cumulative of a continuous distribution is equal to the integral $F(x) = \int_{xmin}^{x} f(y)dy$. Therefore, the cumulative of a sample could be calculated as $\sum_i H(x - X_i)$. The comparison of two distributions could be easily done visually by looking at the cumulative of these distributions. The smaller the difference between the cumulative of two distributions is, the most two distributions approach each other. The Kolmogorov–Smirnov statistic test (K–S test) is based on this concept and is an approach to obtain a quantitative assessment. The K–S test is defined as the maximum value of the absolute difference between two distribution functions (Press et al., 1992). Therefore, the maximum vertical difference between the measured and the simulated cumulative distribution functions could be evaluated. The K–S test is sensitive and captures effectively changes especially in the median value. On the other hand, it is less sensitive to the tails of the distribution. In this case, other suitable statistics can be used, such as Kuiper's variant (Tygert, 2010). There are many available metrics to evaluate the distance between probability distributions. Although it is difficult enough to choose which of them to prefer, Gibbs and Su (2002) have presented an overview of the most important probability metrics/distances. They offer guidance for a better choice of metrics and suggest that complementary insights could be provided using several metrics for an analysis instead of only one. Some extra important statistical distances found in the literature are the following: discrepancy, Hellinger distance, relative entropy (or Kullback-Leibler divergence), Levy metric, Prokhorov metric, separation distance, total variation distance, Wasserstein (or Kantorovich) metric, x2 distance, maximum mean discrepancy (Gibbs and Su, 2002; Gretton et al., 2012). The maximum mean discrepancy is the only method applicable to structured data, such as graphs, and is used in Kernel approaches for comparison between distributions on graphs (Gretton et al., 2007).

3.5 Resume

Machine learning techniques are innovatively integrated into a methodological framework for estimation of data–driven microscopic models. These techniques are easily employed and could help us to overcome limitations of conventional models and develop more robust and flexible models. Different goodness–of–fit measures allow for a more comprehensive evaluation of models performance.



(i) Same moments

(ii) Same first moment only

(iii) Different first, but same second moment

(iv) Different first and second moments

Figure 3.14: Comparison of distributions

Chapter 4

Data–driven car–following models

Car following models have been studied with many diverse approaches for decades. Nowadays, technological advances have significantly improved our traffic data collection capabilities. Conventional car following models rely on mathematical formulas and are derived from traffic flow theory; a property that often makes them more restrictive. On the other hand, data–driven approaches are more flexible and allow the incorporation of additional information to the model; however, they may not provide as much insight into traffic flow theory as the traditional models. In this research, an innovative methodological framework based on a data–driven approach is proposed for the estimation of car–following models, suitable for incorporation into microscopic traffic simulation models. The focus is given on simulation optimization of car-following models, mainly the error between simulation and real traffic to be minimized, using a flexible method. An existing technique, i.e. locally weighted regression (loess), is defined through an optimization problem and is employed in a novel way. The proposed methodology is demonstrated using data collected from a sequence of instrumented vehicles in Naples, Italy. Gipps' model, one of the most extensively used car–following models, is calibrated against the same data and used as a reference benchmark. Optimization issues are raised in both cases. The obtained results suggest that data-driven car–following models could be a promising research direction.

4.1 Reference benchmark: conventional car–following models

In this research, Gipps' model is used as a reference model for the comparison with the proposed methodology on the same data. Punzo et al. (2012) suggested a

[0]The sub–chapters 4.1–4.3.1 are based on Papathanasopoulou and Antoniou (2015b) and 4.5 on Papathanasopoulou and Antoniou (2016).

calibration process of Gipps' model using three different algorithms. However, in this research in order to solve the optimization problem (finding optimal values for Gipps' parameters), two ways are used: a thorough sensitivity analysis of the ranges of all model parameters and a constrained global optimization. The optimal values of model parameters are defined using as objection function to be minimized the RMSN. In such way it is ensured that the calibrated Gipps' model for this data–set was as good as possible, thus providing a fair reference model for this comparison.

4.1.1 Sensitivity analysis

First, a sensitivity analysis of the ranges of all model parameters has been performed. In particular, changing one-factor-at-a-time, an one–at–a–time (OAT) sensitivity analysis is implemented. Parameter ranges for the sensitivity analysis are defined by the suggested ranges by Gipps (1981) and Ranjitkar et al. (2005) as presented in Figure 4.1 for some parameters. The range for the reaction time was defined taking into consideration a larger number of available references (Johansson and Rumar, 1971; Gipps, 1981; Fambro et al., 1998; Ahmed, 1999; Green, 2000; Summala, 2000; Taieb-Maimon and Shinar, 2001; Brunson et al., 2002; Yang et al., 2004; Ranjitkar et al., 2005; Magister et al., 2005; Bilban et al., 2009) (Figure 4.2). The selected ranges for the parameters are: reaction time τ: 0.4s – 3 s, maximum desired speed V_n: 10.4 m/s – 29.6 m/s, distance s_{n-1}: 5.6 m – 7.5 m, maximum desired acceleration a_n: 0.8m/s^2 – 2.6 m/s^2, maximum desired deceleration b_n: 1.6 m/s^2 – 5.2 m/s^2, maximum estimated deceleration $\hat{b}$: 3.0 m/s^2 – 4.5 m/s^2.

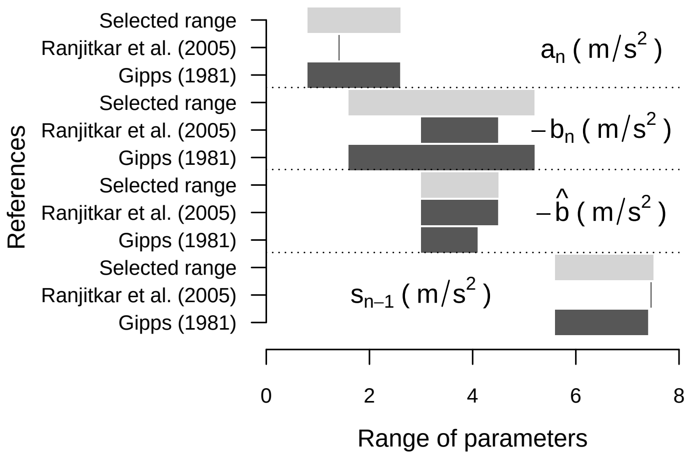


Figure 4.1: Range of Gipps' parameters values according to references

The most extensive data series (B1695, comprising 1695 triplets) is selected for the sensitivity analysis, as it includes more traffic states and more variable

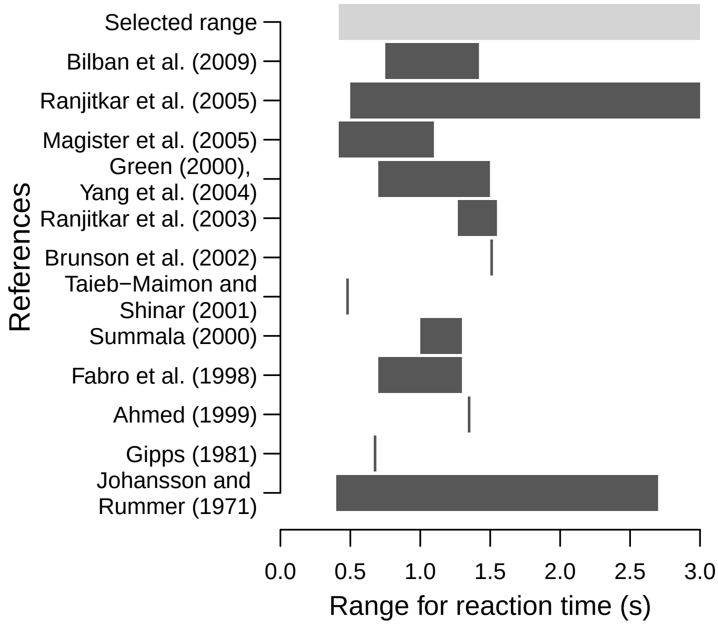


Figure 4.2: Range for drivers' reaction time according to references

speed profiles than others. Hence it may lead to a more representative model, thus potentially avoiding overfitting. For this data series, the best combination of parameter values is requested for the entire trajectory. The performance of each combination of parameter values is evaluated with the normalized root mean square error (RMSN). The combination of parameter values, which provides outputs with the least overall error, is chosen. The effect of each parameter on the performance of the model (namely how the RMSN measure increases or decreases) was examined separately with respect to how its value affects the entire trajectories of the given data series B1695. The values of the other parameters were set at the middle of their identified range. The value of each parameter, which results in the most limited simulation error, is the solution of the optimization problem and could thus be determined. The sensitivity analysis is illustrated in Figure 4.3. The influence of speed V_n and maximum acceleration a_n is not clear from Figure 4.3, as some ranges of these parameters do not seem to affect the RMSN at all. This is due to the observation that both parameters are found only in the first equation of Gipps' model and for certain parameter values the speed estimation may be provided by the second equation of the model. Consequently, the behavior of these parameters was examined again after setting the other parameters to their "optimal" values, as determined by the sensitivity analysis. Figure 4.4 shows the sensitivity analysis that resulted in this case. The best performance of Gipps' model (RMSN=2.7%) in the calibration data–set was achieved with the following combination of parameters: τ=0.4 s, V_n=14 m/s, a_n=0.8 m/s^2, s_{n-1}=5.6 m, b_n=-5.2 m/s^2 and $\hat{b}$=-3.0 m/s^2. According to Brackstone and McDonald (2003), time steps between 0.1 and 1 second are commonly used in micro–simulation models. Brackstone and McDonald (2003) also suggest that small time steps do not allow for driver error. In addition, Simonelli et al. (2009), applying Gipps' model using the same data, have also tested values for this apparent reaction time in the range of 0.4- 1 s. Therefore, a second model was also calibrated, in which a value for the reaction time of τ=1.0 s was considered and accordingly a sensitivity analysis was revised. The values of the final parameters are: τ=1.0 s, V_n=16 m/s, a_n=1.6 m/s^2, s_{n-1}=5.6 m, b_n=-5.2 m/s^2 and $\hat{b}$=-3.0 m/s^2, and the minimum RMSN that was achieved was 4.9%.

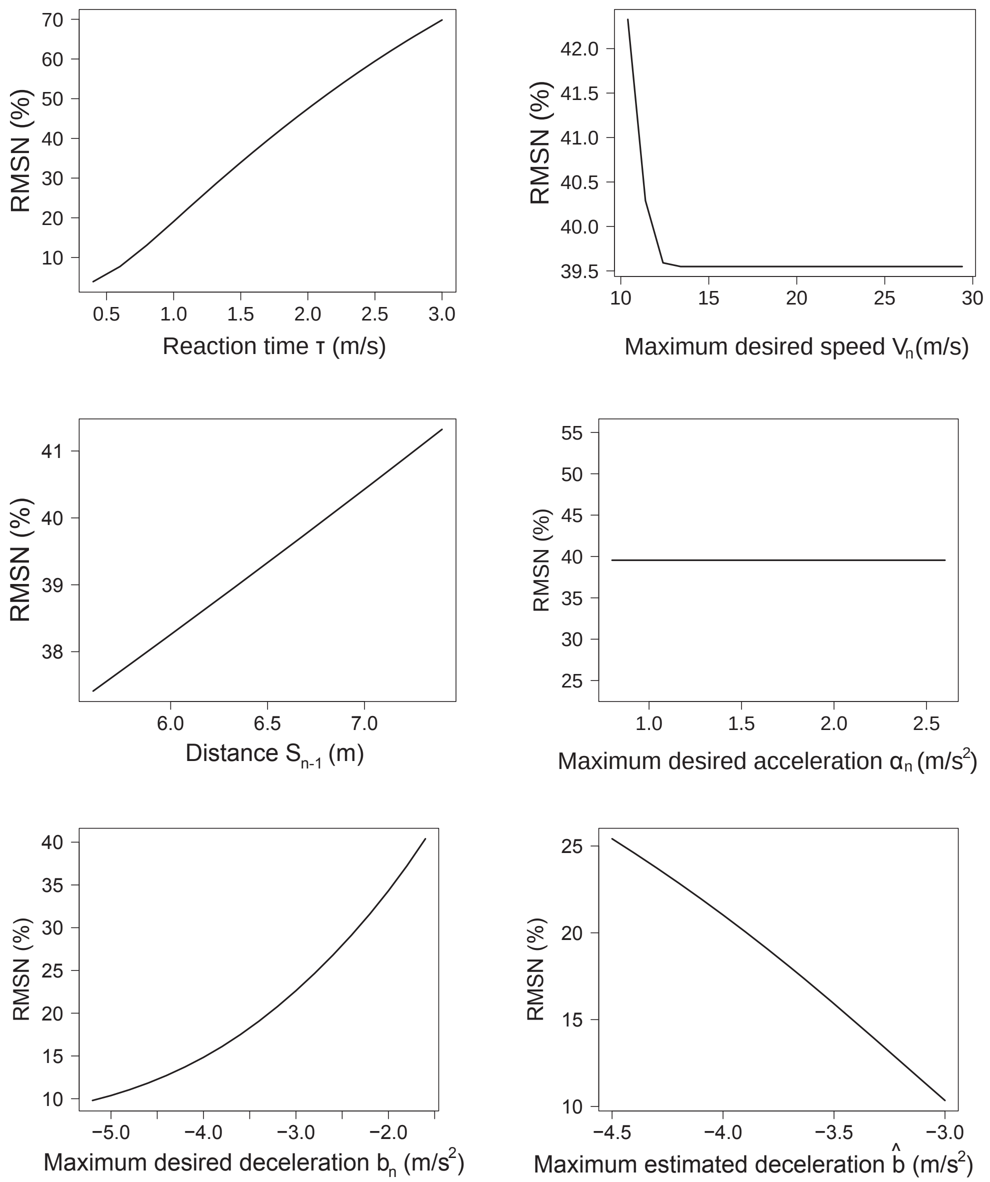


Figure 4.3: Sensitivity analysis of Gipps' parameters

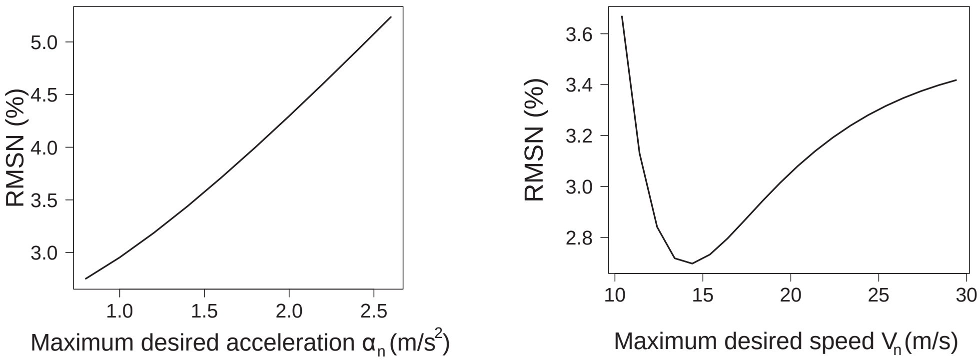


Figure 4.4: Calibration of Gipps' model

4.1.2 Parameters Optimization

To confirm the robustness of the sensitivity analysis results and for a more evidence–based approach, Gipps' model was calibrated using the same data and an optimization algorithm. The longest data series (B1695, longer than 3 minutes) was used again for model calibration. It is worth noting that –besides being the longest– this time series includes the most extensive range of speed values. The calibration process was performed within the R software for statistical computing (R Core Team, 2016). In particular, the Improved Stochastic Ranking Evolution Strategy (ISRES) algorithm was used, which is included in the package “nloptr" (Runarsson and Yao, 2005) and is appropriate for nonlinearly constrained global optimization. This method is implemented in a simple way and supports arbitrary nonlinear inequality and equality constraints in addition to the bound constraints. Furthermore, it incorporates heuristics to escape local optima. On the other hand, although a lot of research has been performed on determining which algorithm is best suited for a given problem, there has not been a satisfactory answer to this question. Thus, various algorithms should be tested in future research.

The objective function that was minimized is: RMSN(v_3^{obs},v_3^{sim}). The range of model parameters, shown in Table 4.1, has been defined in Figures 4.1 and 4.2. In addition, as initial values for the optimization process, optimal values, defined through the sensitivity analysis of Gipps' model, were used. However, it is noted that interactions among the parameters had not been taken into account in the sensitivity analysis. A global optimization is performed, considering all combinations of these model parameters concurrently. For the whole dataset B1695 the optimization process has converged to the optimal set of parameters after approximately 10000 iterations. All parameter combinations that were tested are presented in Figures 4.5, 4.6 and 4.7. In Figure 4.5 optimization results for parameters, included in the first equation of the Gipps model, are presented. These figures confirm the results of sensitivity analysis and indicate that the apparent reaction time is the most critical parameter. The optimal values are presented in Table 4.1, where "initial values" refers to the model parameter values obtained by the sensitivity analysis and "optimal values" refers to the parameters obtained from the optimization using the ISRES algorithm within this research. The minimum value of the objective function, namely the RMSN, that was achieved with these optimal values of parameters was 2.2%, which is slightly smaller than this obtained from the sensitivity analysis.

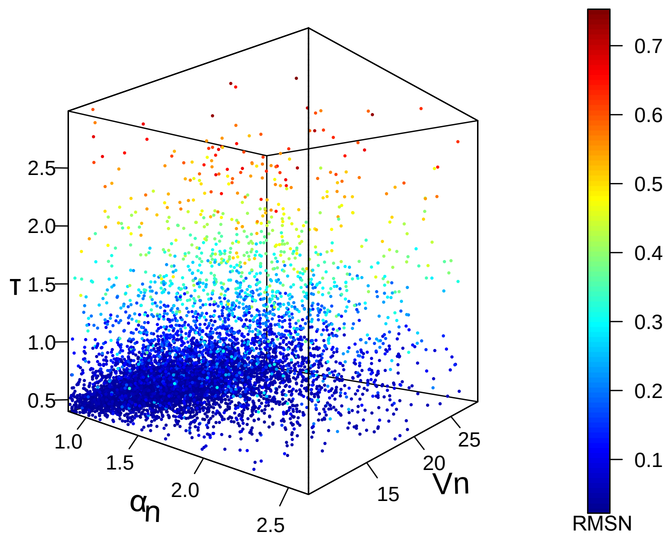


Figure 4.5: Optimization results for Gipps' parameters a_n, V_n, τ

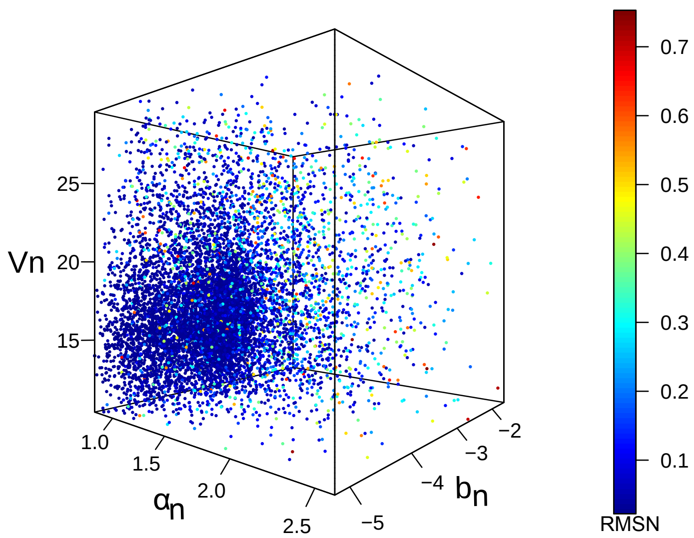


Figure 4.6: Optimization results for Gipps' parameters a_n, V_n, b_n

As mentioned in the sensitivity analysis section, a second model was also calibrated allowing for a more relaxed apparent reaction time. A value for the re-

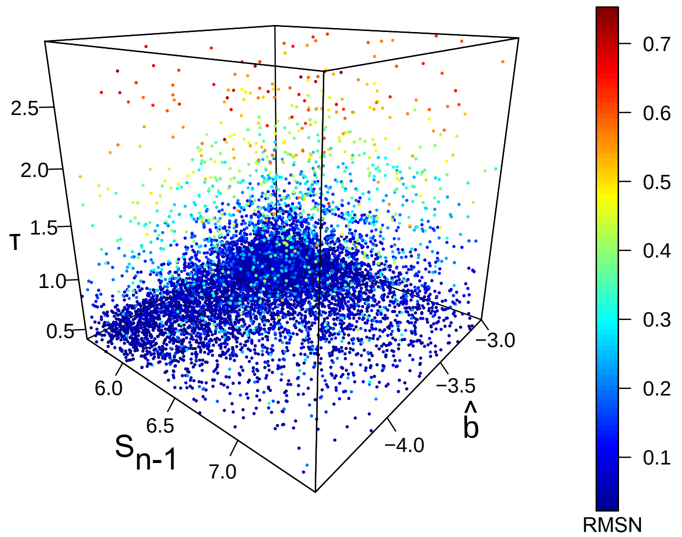


Figure 4.7: Optimization results for Gipps' parameters S_{n-1}, $\hat{b}$, τ

Table 4.1: Optimization of model parameters using ISRES algorithm

Parameters of Gipps' model	Parameters range	Initial values	Optimal values
a_n (m/s^2)	[0.8, 2.6]	0.8	0.8
b_n (m/s^2)	[-5.2, -1.6]	-5.2	-3.2
V_n (m/s)	[10.4, 29.6]	14.0	14.4
S_{n-1} (m)	[5.6, 7.5]	5.6	5.9
$\hat{b}$ (m/s^2)	[-4.5, -3.0]	-3.0	-3.1
τ (s)	[0.4, 3.0]	0.4	0.4

action time of τ=1.0 s was considered and accordingly the optimization process was revised for the rest of parameters. In the optimization process initial values obtained from the sensitivity analysis, as indicated in Table 4.2. The results are presented in Figures 4.8 and 4.9. The isres algorithm converged after 10000 iterations to the optimal set of parameters: τ=1.0 s, V_n=15.8 m/s, a_n=0.8 m/s^2, s_{n-1}=5.6 m, b_n=-5.0 m/s^2 and $\hat{b}$=-3.0 m/s^2. The minimum value of the objective function, namely the RMSN, that was achieved with these optimal values of parameters was 4.2%.

The two calibrated models that ensure the best performance for data series B1695 and that will be used for a fair comparison with the proposed methodology are summarized in Table 4.3.

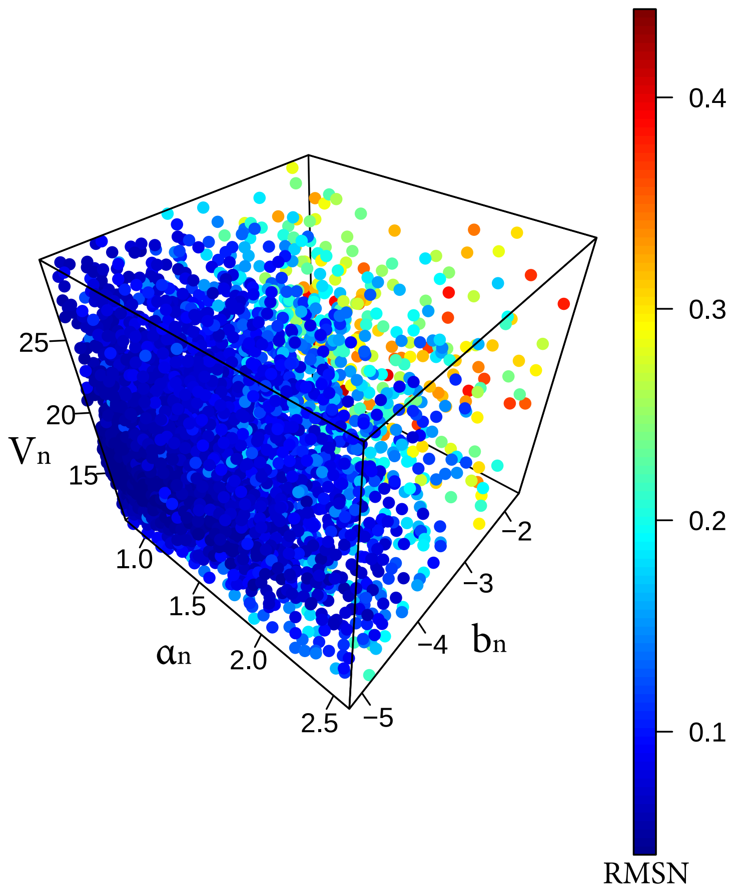


Figure 4.8: Optimization results for Gipps' parameters a_n, V_n, b_n considering τ=1.0 s

Table 4.2: Optimization of 5 model parameters using ISRES algorithm, considering τ=1.0 s

Parameters of Gipps' model	Parameters range	Initial values	Optimal values
a_n (m/s^2)	[0.8, 2.6]	1.6	0.8
b_n (m/s^2)	[-5.2, -1.6]	-5.2	-5.0
V_n (m/s)	[10.4, 29.6]	16	15.8
S_{n-1} (m)	[5.6, 7.5]	5.6	5.6
$\hat{b}$ (m/s^2)	[-4.5, -3.0]	-3.0	-3.0

4.1.3 Gipps' model application – a benchmark case

The two calibrated models with the fixed parameters values could be validated to the rest of data series. The validation results are presented in Figure 4.14.

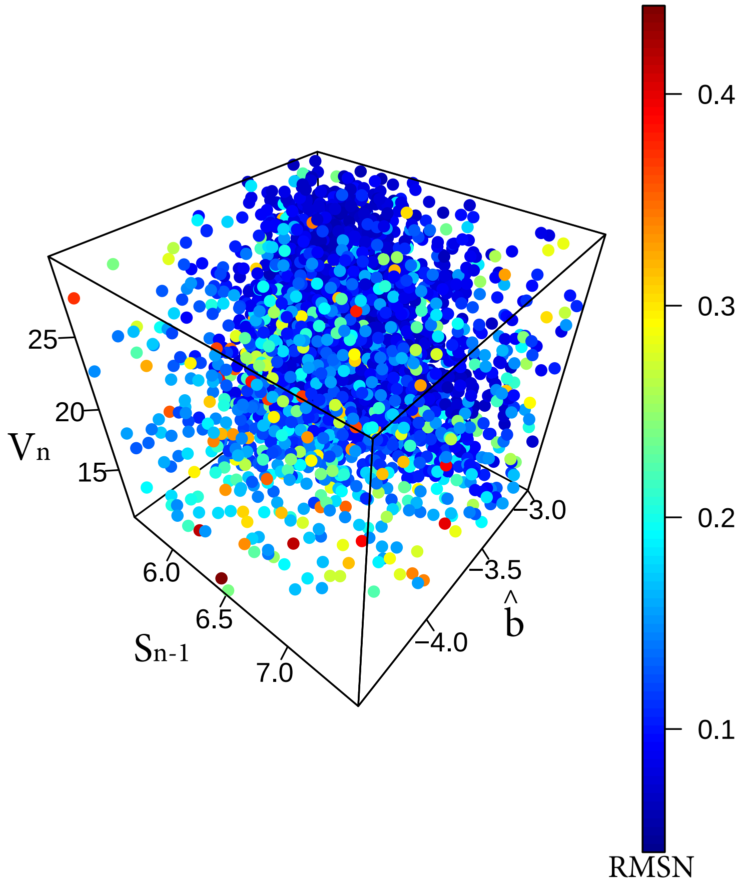


Figure 4.9: Optimization results for Gipps' parameters S_{n-1}, $\hat{b}$, V_n, considering τ=1.0 s

Table 4.3: Calibrated Gipps' model

Parameters of Gipps' model	Model (τ=0.4 s)	Model (τ=1.0 s)
a_n (m/s^2)	0.8	0.8
b_n (m/s^2)	-3.2	-5.0
V_n (m/s)	14.4	15.8
S_{n-1} (m)	5.9	5.6
$\hat{b}$ (m/s^2)	-3.1	-3.0
τ (s)	0.4	1.0

4.2 Estimation of data–driven car–following models

The proposed methodology is composed of two parts: training and application, outlined in Figure 4.11. In the training step traffic models are estimated according to the available surveillance data, while in the application step these traffic

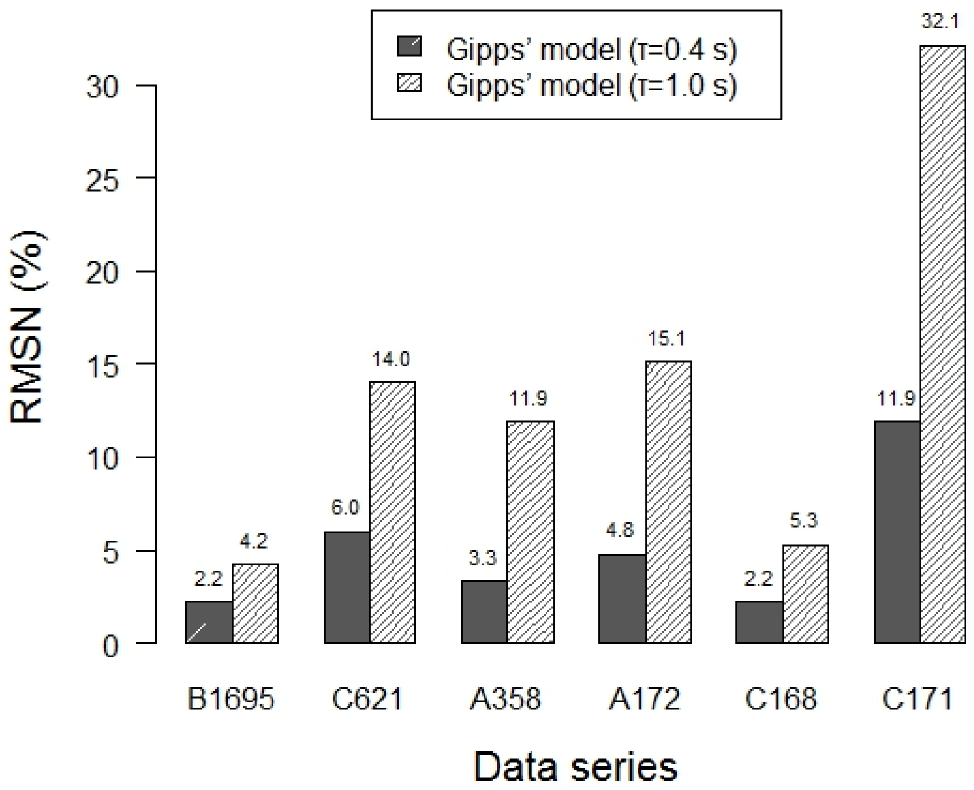


Figure 4.10: Validation results for Gipps' model using Naples data

models are applied to provide speed predictions for the following vehicle and the next time instant using new observations. In particular, the required explanatory variables of the car-following process are determined and respectively the appropriate data are collected. In this research the triples $v_i(t)$, $v_{i-1}(t)$, $d_{i,i-1}(t)$ (leader and follower speed and their distance) per each time instant t have been considered as independent predictor variables for the estimation of the response variable $v_{i-1}(t+\tau)$, i.e. the follower speed, for the next time instant (t+τ), where τ is the apparent reaction time. Estimation has been achieved without assuming any predefined functional form; instead a flexible regression method can be used. Portions of the available data are identified and correspondingly various representative models are formed (fitting). The application step follows, when new measurements arise. The flexible model that has been estimated for each traffic state is then retrieved from the knowledge base and applied to the new data for the estimation of the speeds of the following vehicle.

4.2.1 Application of loess model

The models presented in this research were all implemented using the R Software for Statistical Computing (R Core Team, 2017). Application of loess (locally weighted regression) requires the 'stats' package and the determination of its pa-

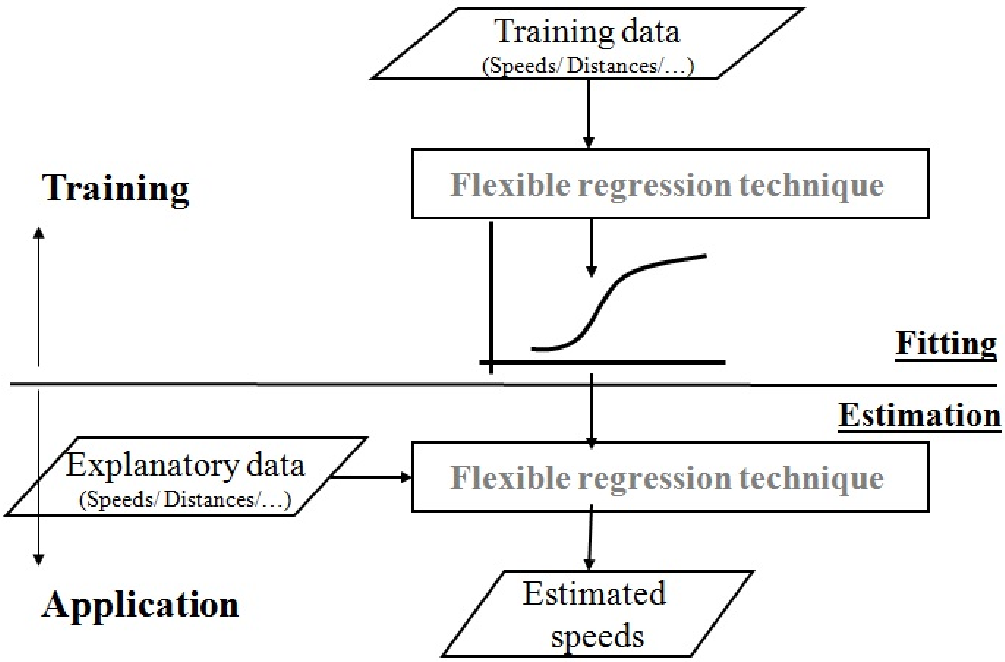


Figure 4.11: Estimation of data–driven car–following models

rameter values, i.e. span (a) and degree (presented in Section 3.2), to ensure a good fit to the data. The span determines how smooth the curve is and it ranges from 0 (wavy curve) to 1 (smooth curve). The degree determines the degree of local polynomials, which are used in each local regression. In the used implementation, a value of 1 refers to a linear function, while 2 in quadratic function. Optimal values of the loess model parameters can be estimated through an optimization approach. A sensitivity analysis was preferred here. The performance of the proposed method for different values of span and degree is presented in Figures 4.12 and 4.13 for all available data series in order for appropriate values to be selected. The optimal values are these for which RMSN is minimized.

It is noted that the data that are taken into account for loess are the same with those used in Gipps' model [speed $v_2(t)$ and $v_3(t)$ of vehicles 2 and 3 and distance $D_{23}(t)$ between vehicles 2 and 3, as they were estimated by their coordinates], so that a direct and fair comparison between them is possible. It should be mentioned that different combinations of data (v_1, v_2, v_3, v_4, D_{23}, D_{34}) have also been tested. However, the best performing loess model was this taking into account the same data with Gipps' model, mainly speed $v_2(t)$ and $v_3(t)$ and distance $D_{23}(t)$. In addition, for all data series the speed estimation for speed $v_3(t+\tau)$ with the proposed method relies on the pattern resulting from the entire leader–follower trajectory of data series B1695, as well as the calibration of Gipps' model. In more detail, the proposed method firstly recognizes the relationships between observations ($v_2(t)$ and $v_3(t)$ and distance $D_{23}(t)$) and the response data

v_3(t+τ) of the B1695 data series. After the relevant pattern from the B1695 data series has been identified, the proposed method is applied to the remainder of the data series. It requires the input data (here speed $v_2(t)$ and $v_3(t)$ and distance $D_{23}(t)$) and exports the estimated output v_3(t+τ). It should be clarified that reaction time is not a parameter of the loess method. However, it plays a significant role in loess method application, as for different values of reaction time τ different data, mainly data of different time instants, are selected for prediction. For instance, if prediction for time instant t is required, then data for time instant (t-τ) are used. In this research, the same values of reaction time as those used for Gipps' calibration are used, ensuring a fair comparison. In Figure 4.12, the solid lines illustrate the RMSN of speed v_3(t+τ) estimation with proposed method considering degree = 1 for each data series and for each value of span among its range, while the dashed lines illustrate the corresponding results for degree = 2. The solid lines (degree = 1) are smoother and represent lower RMSN than dashed lines (degree = 2), and therefore the preferred degree in this case is selected equal to 1. Regarding the span, the solid lines are almost flat for values of span between 0.4 and 1.0 for all data sets and for both reaction times (0.4 s or 1.0 s). Consequently, excluding low values, the span does not appear to affect significantly the results. Furthermore, the ranges of the span, for which the lowest RMSN was observed for all data series, are presented in Figure 4.13. The value 0.75 is chosen as average and more representative of the data.

4.2.2 Application of other machine learning techniques

The proposed methodology for estimation of data–driven car–following models has been applied using different state–of–the–art machine learning techniques, which are currently finding a lot of researchers' attention and have described in methodology chapter, such as: locally weighted regression, splines, Gaussian process, kernel support vector machine and neural network. Computational intelligence in general has proven its applicability to traffic simulation models. However, the question of which machine learning technique could be the most appropriate one for traffic simulation models has not been answered conclusively. This research aims to provide some more input into this ongoing active research field.

In this case study traffic models are trained using as input data the most representative data series, B1695. Relationships among predictor variables $v(t)$, $v_{front}(t)$, $D_{front}(t)$) and the response variable v(t+τ) are identified using observations of data series B1695. After the model fitting, the proposed methods are applied to the remainder of the data series for validation. The proposed methodology is further implemented using MARS, KSVM, GP and BRNN as regression techniques. All the models have been applied in the R statistical software and specifically MARS using 'earth' package (Milborrow, 2017), KSVM and GP using the 'kernlab' package (Zeileis et al., 2004) and finally BRNN using the 'brnn' pack-

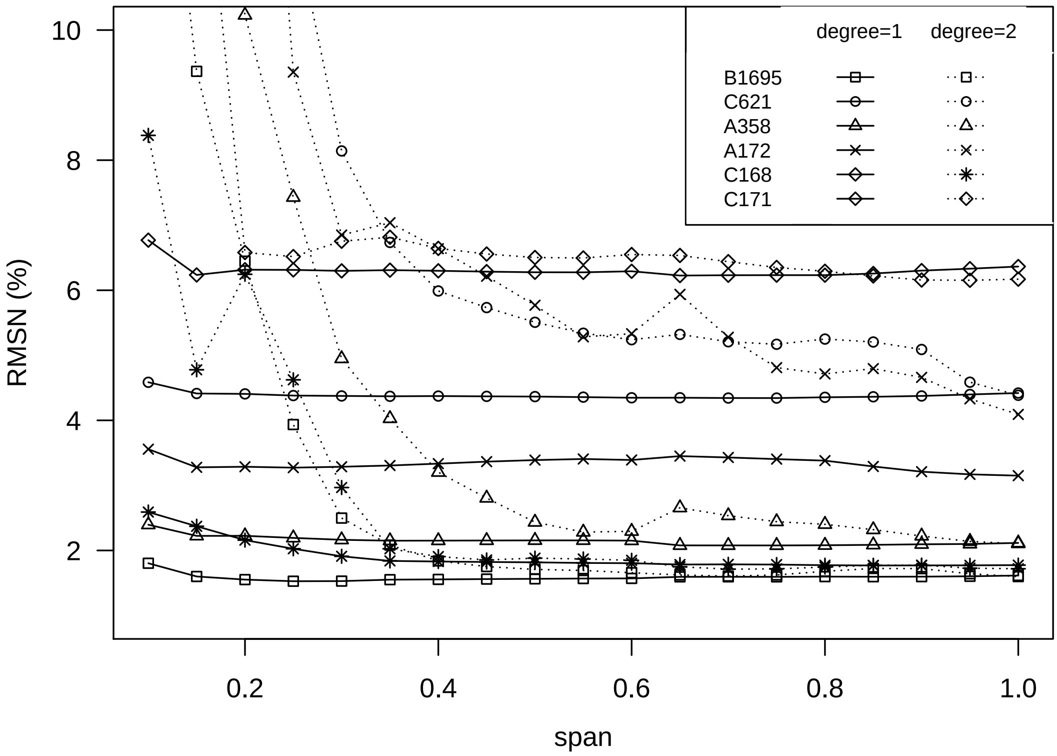


Figure 4.12: RMSN for different values of span and degree, by applying the method loess for a reaction time τ = 0.4 s

age (Pérez-Rodriguez and Gianola, 2013). For each model hyperparameters need to be calibrated.

Calibration of MARS models includes tuning parameters, such as nprune, maximum number of terms (including intercept) in the pruned model, and degree, which defines the maximum degree of interaction. The degree is set to one, but more complicated response curves may be necessary in certain instances. The value for nprune is semi-automatically calculated from the number of predictors. The optimal number of terms in the model is estimated using generalized cross validation (Happe et al., 2010). The additive model includes 4 terms at the degree of interaction. The distance between the two vehicles was not used because it was not considered as an important predictor variable by the algorithm.

A KSVM is trained using 110 Support Vectors. Between the Gaussian and polynomial kernels, Ben-Hur and Weston (2010) claim that the Gaussian kernel usually outperforms the polynomial kernel in both accuracy and convergence time. The hyperparameter sigma, the inverse kernel width for the Radial Basis kernel function "Gaussian" (rbfdot), is estimated using automatic sigma estimation for the regression by the kernlab package. For the available data the estimated value is sigma = 2.799. The cost of constraints violation is the C-constant

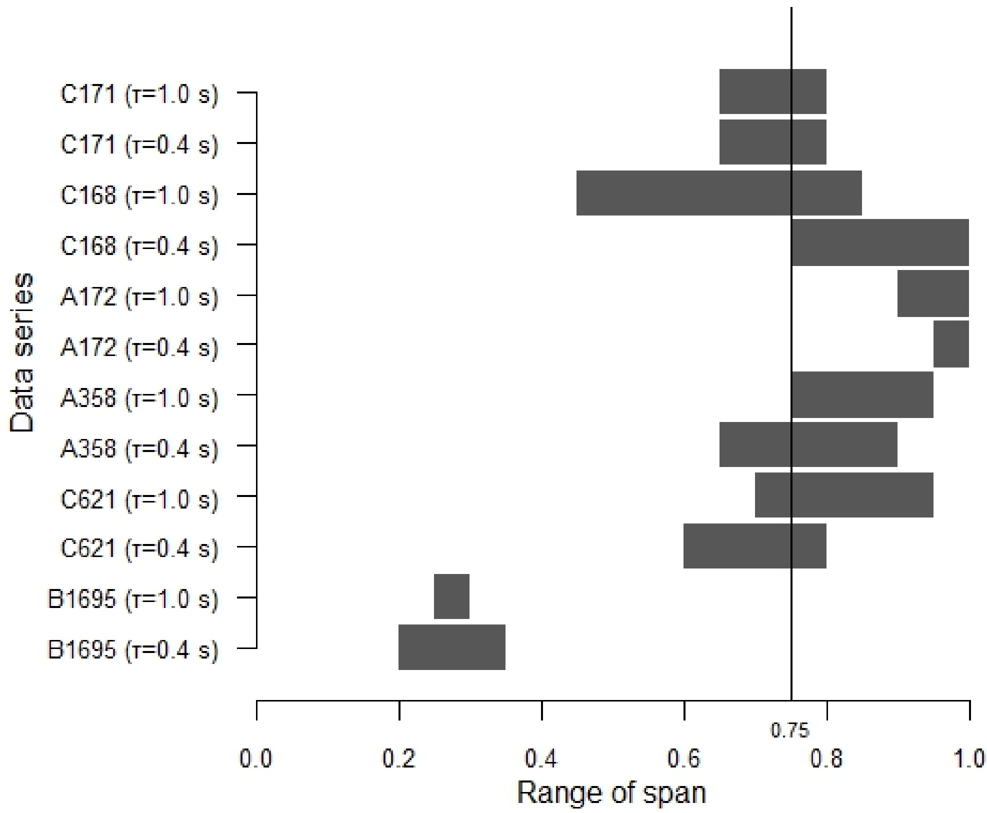


Figure 4.13: Ranges of span, which minimize the RMSN for each data series

of the regularization term in the Lagrange formulation. For a large value of C a large penalty is assigned to errors/ margin errors (Ben-Hur and Weston, 2010). The value C=1 has been used.

The GP is a generalisation of a LOESS, where the span ("bandwidth") parameter varies in the dataset. Data are scaled internally to zero mean and unit variance. The center and scale values are returned and used for later predictions. The list of hyper-parameters (kernel parameters) contains the parameters to be used with the kernel function.The Radial Basis kernel function "Gaussian" (rbfdot) has been used. The hyperparameter sigma, the inverse kernel width for the Radial Basis kernel function "rbfdot", is estimated using automatic sigma estimation for the regression by the kernlab package. For the available data the estimated value is sigma = 3.059.

Before applying a BRNN, the optimal number of neurons should be determined. If the number of neurons is too small, the network cannot learn correctly. If it is too large, it will increase complexity and training time and may lead to overfitting, thus the network will model random noise in the data. In order to determine the optimal number of neurons, 2 to 10 neurons have been tested. The RMSN for the training data B1695 is 1.6 using 2 neurons, while using from 3 up to 10 neurons the RMSN remains 1.5. The number of neurons seems that does not impact the model efficiency for the available data. Therefore, the simplest BRNN with 2 neurons is selected for application. A Bayesian regularized neural network with 2 neurons and 10 weights, biases and connection strengths has been applied. Inputs and output were normalized (scaling factor: 0.700287). Nguyen and Widrow algorithm (Nguyen, 1990) has been used to assign initial weights. The training process finished because changes in Equation 3.15 in last 3 iterations were less than 0.001.

In this research locally weighted regression has been used for further analysis, as it comprises much of the simplicity of linear least squares regression with the flexibility of nonlinear regression.

4.3 Validation results- accuracy comparison

4.3.1 Comparison of Gipps' model and loess model

The accuracy of estimation of the speed v_3(t+τ) of the third vehicle was estimated with both approaches and their performance in terms of RMSN is presented in Figure 4.14. The loess method provides more reliable results (smaller RMSN errors) for all data sets than Gipps' model. Figure 4.15 present the same comparison (for reaction time τ=0.4 s), but considering more measures of goodness of fit, used so that both approaches could be evaluated from different points of view, as described in the methodology section. Figure 4.15 confirms that the proposed method outperforms Gipps' model. This result confirms the claim that the proposed method comprising locally linear regressions could provide satisfactory results and that data–driven methods could outperform the performance of conventional models. For reaction time equal to 1 s, these measures of goodness of fit were also calculated and it was found that the comparative advantage of the loess method was even larger. Furthermore, we notice that the performance of both models is significantly better for lower values of the reaction time variable t. This could be explained by the fact that a driver with a smaller reaction time could react faster and respond more abruptly to the changes in traffic conditions. Therefore, a model with shorter reaction time would also be able to replicate this driving behavior better.

Besides the aggregate analysis of the model fit, an analysis of the produced residuals is also undertaken, in order to check whether the estimation of speed is biased or not. This could be achieved by testing if the assumptions of normality, linearity and homoscedacity are met or violated. Linearity and homoscedacity could be detected in a plot of residuals versus predicted values. The linearity assumption is supported to the extent that the amount of points scattered above and below the line is equal. The homoscedacity refers to the homogeneity of variance, which is sufficient to the extent that the vertical scatter is the same across

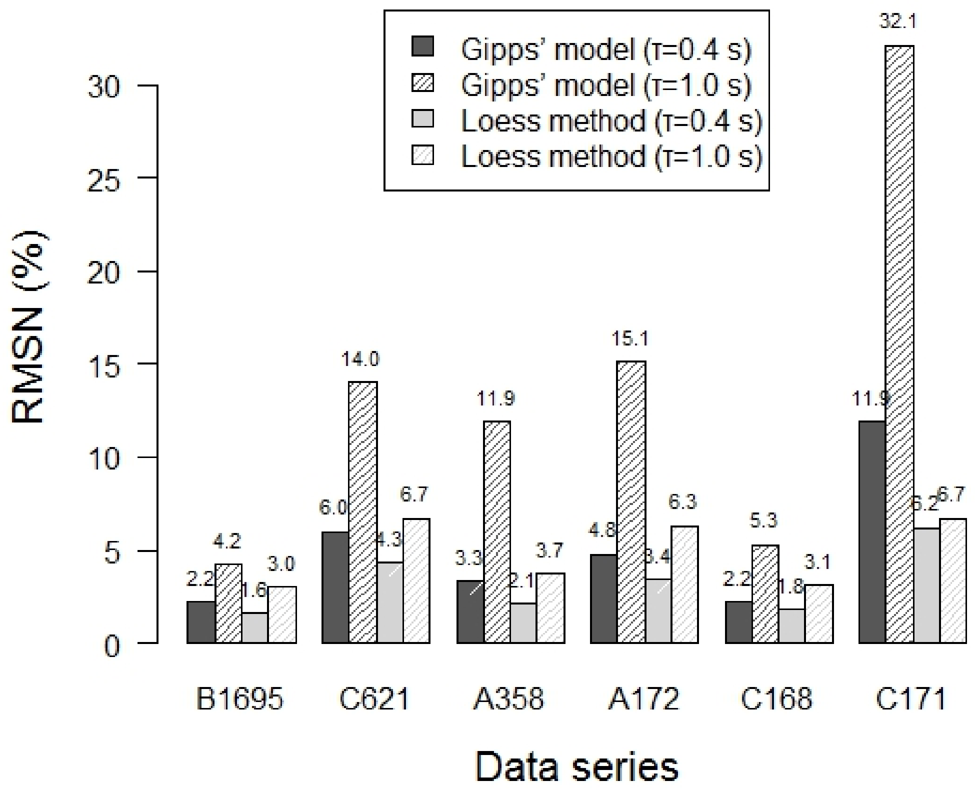


Figure 4.14: Comparison of RMSN by applying Gipps' model and loess method

all x values. The normality assumption could be tested using normal quantile (Q–Q) plots or normal probability (P–P) plots. Normality is achieved when the points on such a plot fall close to the diagonal reference line. The analysis is outlined in Figures 4.16, 4.17 and 4.18. Standardized residuals have been used. Residuals of all the data series are presented together in each plot. The assumption of normality seems to be probably sufficient looking at the Q–Q plot (Figure 4.16). The deviations from the diagonal line in the center of the plot are minimal. The pattern is slightly differentiated at both ends, which may indicate a light tail on both sides. The P-P plot (Figure 4.17) also shows that the distribution of the residuals tends to be normal. However, it may suggest some skew, though not so sensitive. The plot of standardized residuals versus standardized predicted values (Figure 4.18) suggests that points are around the horizontal line and therefore the assumptions of homoscedacity and linearity are met. Based on this analysis, the assumptions of normality, linearity and homoscedacity seem to be supported. There is no evidence for a biased estimation of speeds.

4.3.2 Comparison of Gipps' model and other flexible models

The goodness–of–fit measures have been estimated in order to compare predicted and observed speed values and to evaluate the performance of other models. The

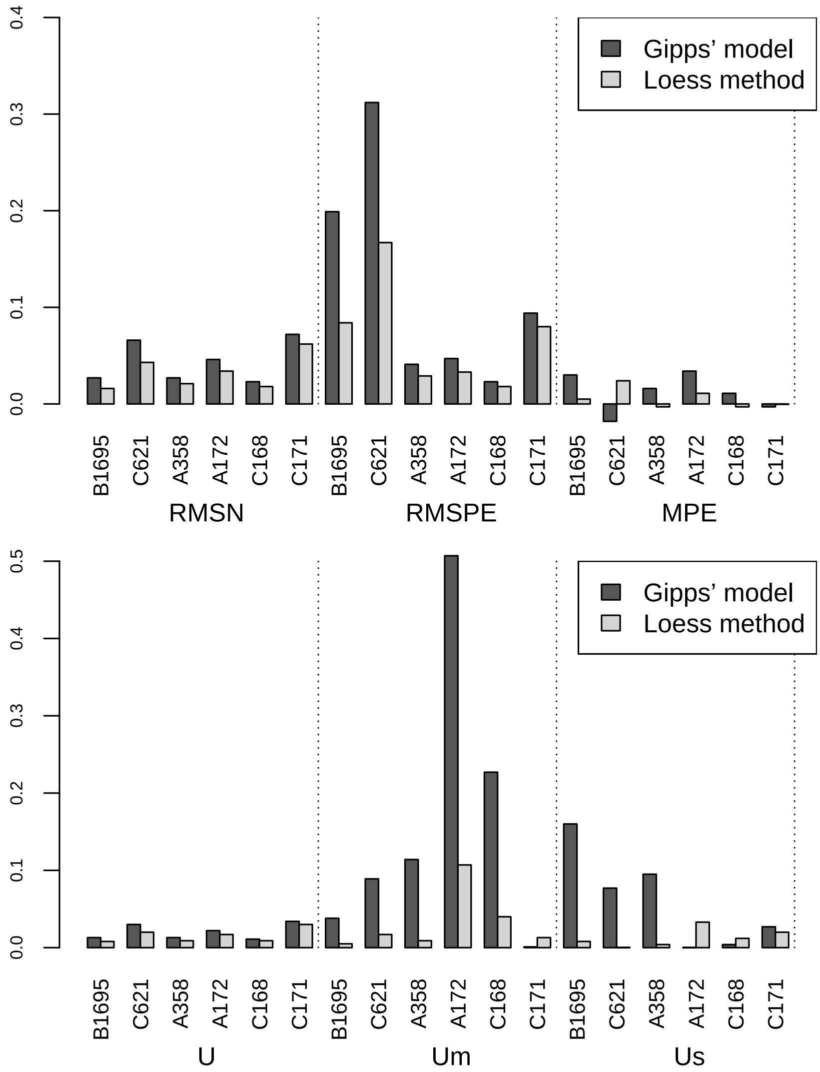


Figure 4.15: Comparison of Gipps' model and loess method for reaction time τ = 0.4 s with different measures of goodness of fit for the available data series

results are presented in Figure 4.19 and indicate that for the considered problem the most stable performance is achieved by loess method and Gaussian processes for the majority of the data series. Loess method, which combines the benefit of being very simple to implement, seems to be the best choice for this case study, as speed estimation with the lowest error is consistently achieved. Further analysis is presented in the computational cost section. Similar behavior is observed using other machine learning techniques, such as KSVM, MARS and BRNN, and all of them provide good alternatives for estimation of data–driven models using the available data.

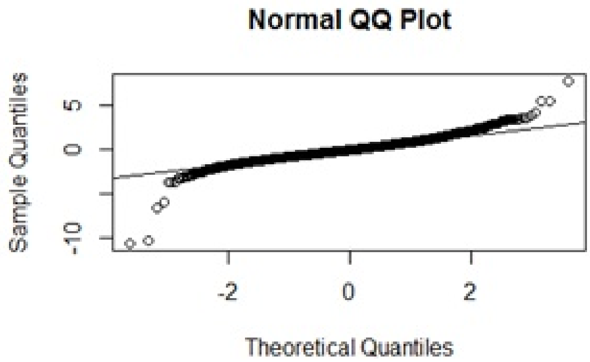


Figure 4.16: Q–Q plot

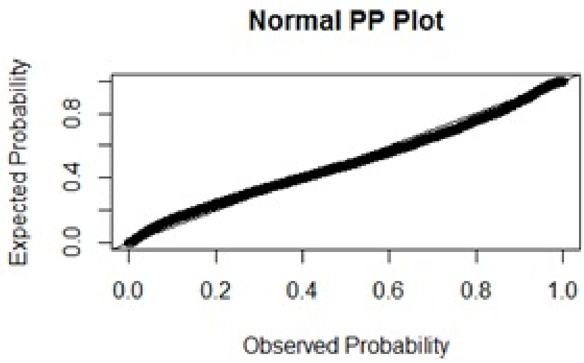


Figure 4.17: P–P plot

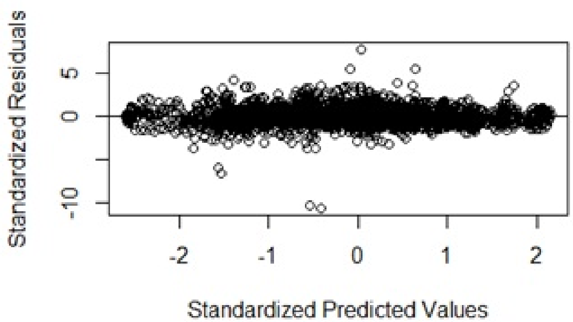


Figure 4.18: Linearity and homoscedacity plot

4.4 Further exploration of the models

4.4.1 Computational cost

As far as computational effort is concerned, the execution time, including training and application time, has been estimated for each model. All models have

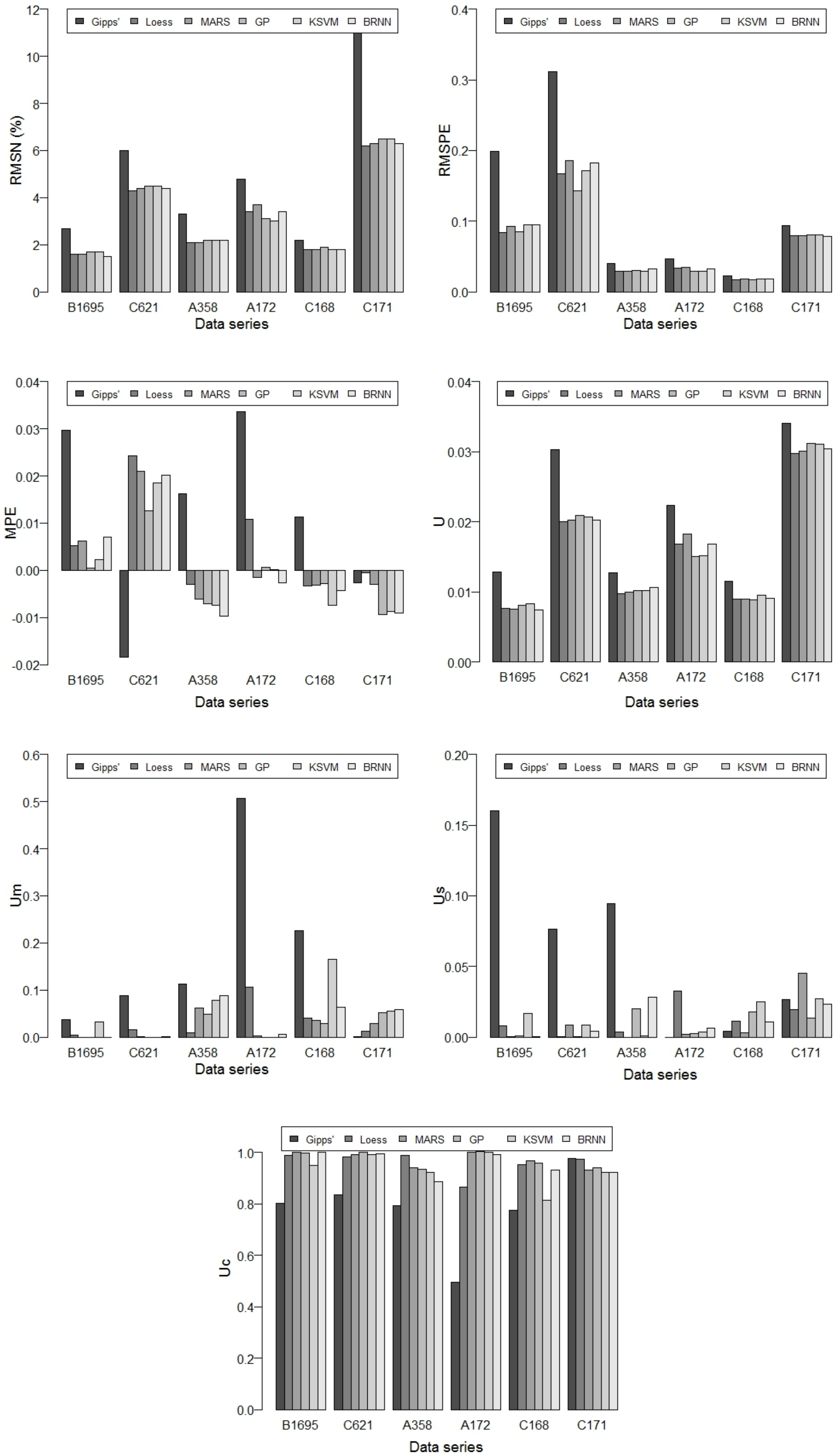


Figure 4.19: Results for Gipps' model and loess, MARS, GP, KSVM, BRNN models

been performed in R using the same computer workstation. In this respect, some considerations can be drawn for the computational cost of each method. Although Gaussian Processes allow for a reliable estimation of car–following models, they seem to require more execution time than the other methods, as outlined in Figure 4.20. Gaussian Processes seem to be slow to learn but fast to use. Training MARS and KSVM models seem to be faster processes. The calibration time for Gipps' model has not been estimated, as it is a time consuming process. For calibration or training of all methods B1695 dataset was used. In Figure 4.21 it is observed that after Gipps' model, GP tend to require more time for model application on the validation data, while MARS and KSVM are less time consuming models. Computational cost of models plays an important role and should be taken into account in order to choose the appropriate model for each application, especially if it is an on–line application which requires speed and accuracy.

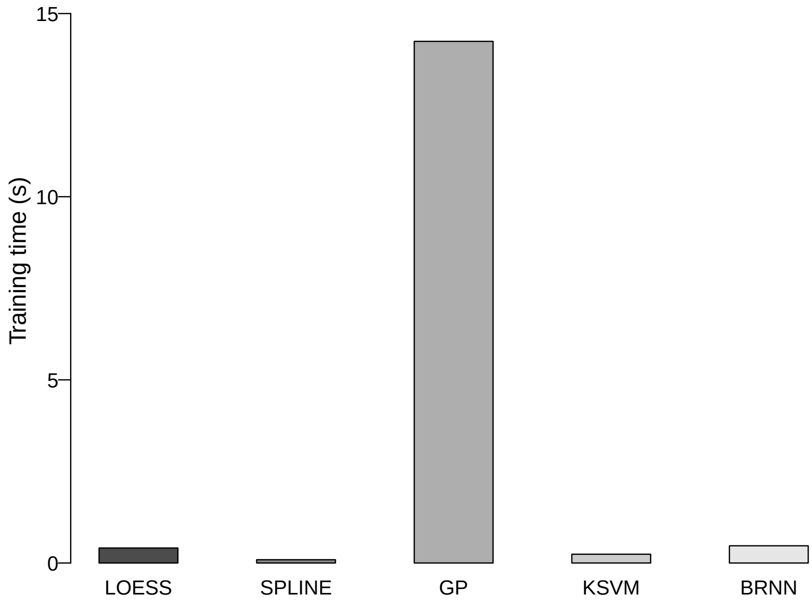


Figure 4.20: Training time for loess, MARS, GP, KSVM and BRNN models

4.4.2 Exploration of noise interference in the models

Data used in this research are readily available observations from the field without corrections. The proposed methodology is applied to measured trajectory data without filtering measurement noise. Although Gipps' model has a fixed equation structure, the proposed flexible models are not based on a specific relationship. Therefore, one could claim that the noise affects the effectiveness of each model in a different way and thus a fair comparison is not feasible. In order

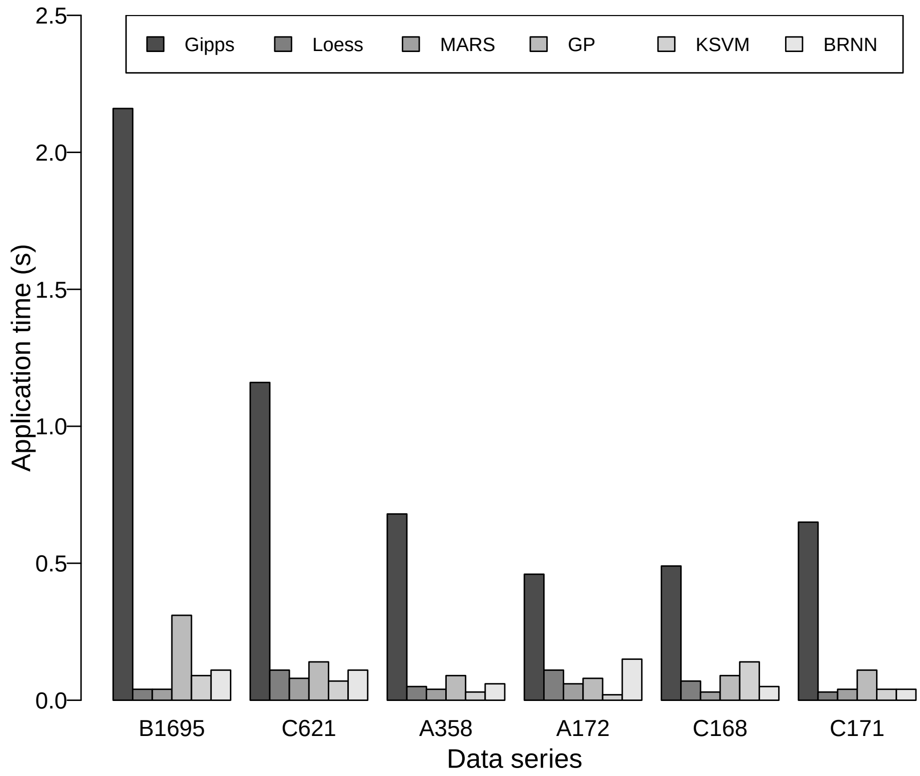


Figure 4.21: Application time for loess, MARS, GP, KSVM and BRNN models

to address this issue, it is proposed to create extra noise on the trajectory data and then to explore the noise interference in the after effect prediction process. Jitter function, written by Werner Stahel and Martin Maechler (ETH Zurich), is used to apply noise on X, Y, Z coordinates of the vehicles in R Statistical Software. This function is appropriate to add a small amount of noise to a numeric vector. The result ϵ is given as per Equation 4.1.

$$\epsilon = \mathbf{x} + runif(n, -a, a) \tag{4.1}$$

where x: numeric vector to which jitter should be added. n = length(x) α: the amount argument, α <- factor * d/5 where d is the smallest difference between adjacent unique (apart from fuzz) x values. Factor=1 was set.

The coordinates X_i, Y_i and Z_i after the application of jitter function are described by Equation 4.2.

$$\begin{aligned} X_i &= X_{real_i} + \delta_{x_i} + \epsilon_{x_i} \\ Y_i &= Y_{real_i} + \delta_{y_i} + \epsilon_{y_i} \\ Z_i &= z_{real_i} + \delta_{z_i} + \epsilon_{z_i} \end{aligned} \tag{4.2}$$

where i: time instant X_{real},Y_{real}, Z_{real}: the real coordinates of the vehicles δ_x, δ_y, δ_z: the unspecified noise included in measured observation ϵ_{x_i}, ϵ_{y_i}, ϵ_{z_i}: the noise added to the coordinates using jitter function

Then speeds and distances are estimated using the noisy coordinates. In Figure 4.22 the measured speed (black line) and the noisy speed (red line) of the third vehicle is plotted against time for dataset B1695.

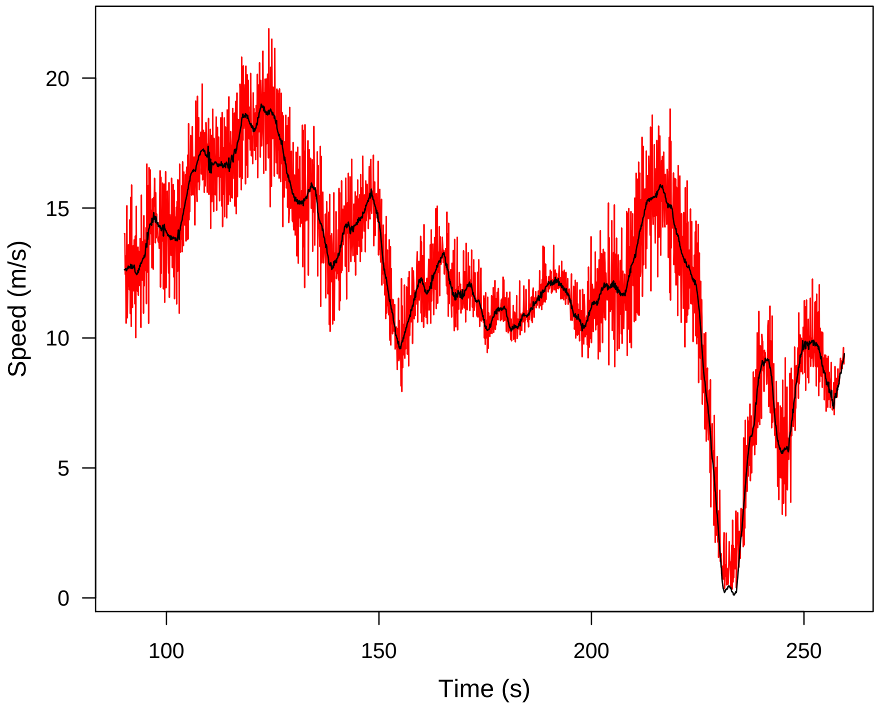


Figure 4.22: Speed with noise applied by jitter function

The speed estimation for the next time instants is revised using all the applicable models. The concept is to explore the effect of noise ϵ_{x_i}, ϵ_{y_i}, ϵ_{z_i} on the prediction process for all the methods. The results are indicated in Figure 4.23. The various models seem to react in a similar way to the noise and no significant changes in their between comparison were observed. It is noted that B1695 dataset is used both for calibration and validation. Gaussian processes allow also to define the initial noise variance, which can improve significantly the model efficiency.

4.4.3 Input data: more traffic observations but irrelevant or less traffic observations but relevant?

The updated version of the methodology is based on an on–line fitting of data–driven models (Figure 4.24). The proposed methodology may benefit from a system that allows a fleet of connected vehicles to exchange information, such as (X, Y, Z) coordinates, using a central data system. However, even a single instrumented vehicle, with the ability to geo–locate itself, and obtain (e.g. via suitable instruments and cameras) estimates of the speed and distance of surrounding vehicles, has access to all information required to apply this methodology. Speeds,

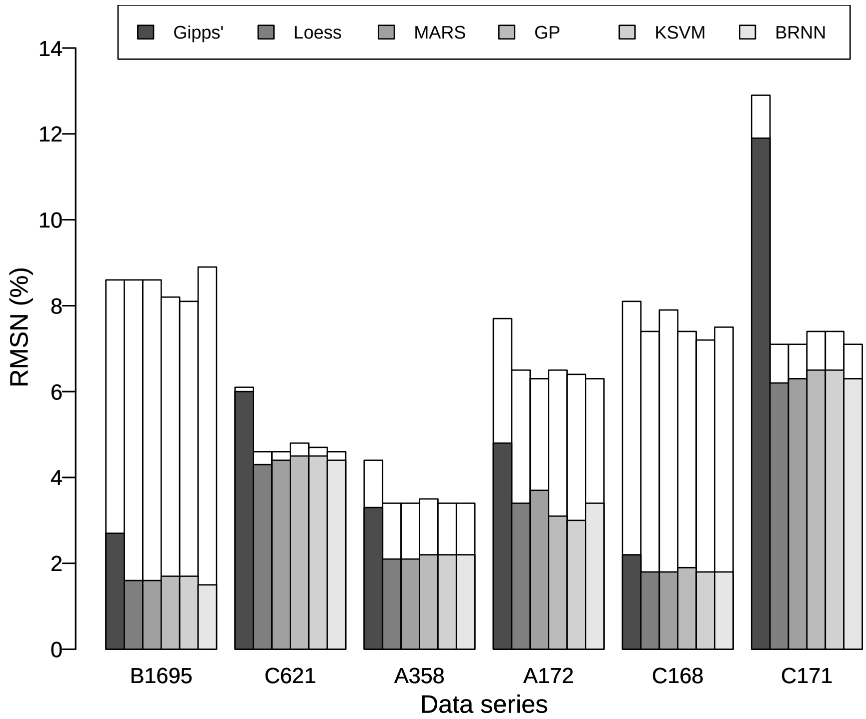


Figure 4.23: Speed with noise applied by jitter function

accelerations or gaps could be calculated per time instant and used to dynamically calibrate or fit flexible car–following models. Therefore, fitting of the flexible models is achieved using as input data the most recent and thus more relevant observations obtained from the same driver, the same vehicle, the same network and the same traffic conditions. Specifically, observations up to time t are used as input in a flexible regression technique and a pattern of speed prediction for the following vehicle and the next time instant is identified. In the next time instant when a new observation arises, the calibrated model from the previous time instant is used and estimated speeds from time t onwards are produced. In the meantime, the new observation has been stored to a database with the previous observations and the whole process is iterated per time instant t. In each iteration a certain amount of the most recent and relevant observations is used. The question that arises at this point is what is the amount of the most recent observations that is required for the suggested methodological approach.

In the second column of the Table 4.4 off–line fitting has been already demonstrated using Naples data as presented in the subsection 5.2.1. In this case traffic models are calibrated using as input data the most representative data series, namely this with the largest speed range, B1695. Relationships among predictor variables $v(t)$, $v_{front}(t)$, $D_{front}(t)$) and the response variable v(t+τ) are identified using observations of B1695 data series (1695 observations). After the relevant pattern from the B1695 data series has been formed, the suggested method is

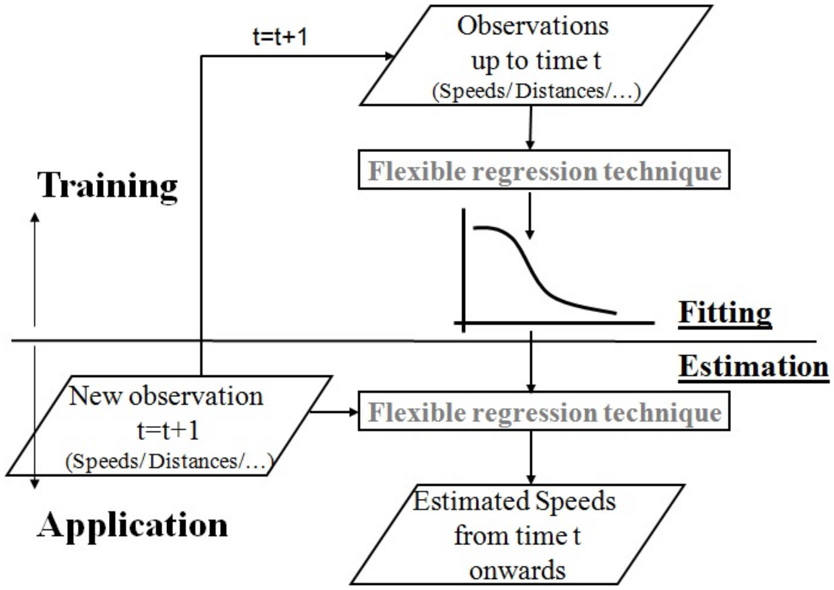


Figure 4.24: Recent traffic observations as input in data–driven car–following models

applied to the remainder of the data series for validation. The RMSN values have been estimated in order to compare predicted and observed speed values and estimate the performance of this methodological approach. The results are presented in the second column of Table 4.4. As regards the loess parameters and the reaction time, degree=1, span=0.75 and τ=0.4 s have been considered.

In order to start with the implementation of method presented in Figure 4.24, a certain amount of few observations from each data series is required. This depends on the number of most recent observations required to calibrate traffic models. In this case, traffic models are fitted using as input data the observations of each data series up to time t. Relationships among predictor variables $v(t)$, $v_{front}(t)$, $D_{front}(t)$ and the response variable v(t+τ) are identified using the most recent n observations of each data series. Therefore, traffic patterns are formed using less but more relevant data in comparison with the first methodological approach. When a new observation in time instant $t+1$ arises, the calibrated traffic models are used for speed estimation on time $t+1+\tau$. Then the whole process is revised and models are fitted again using as input data the n most recent observations up to time $t+1$ and the process goes on. Regarding the amount of the most recent observations that is required, values from 10 up to 100 observations have been tested, in order to define the most appropriate value of n. The RMSN values have been estimated in all occasions and a sensitivity analysis is illustrated in Table 4.4. For the majority of data series there is a sharp decrease when using only the 20 most recent observations, namely a period of 2 seconds. However, a second level of improvement is obtained for model fitting in data series C621 and A358. When 60 or more of the most recent observations are taken into account in each step of model fitting, the RMSN seems to be stabilized for all data series.

In Table 4.4 a comparison of their performance is attempted from 10 up to 100 observations. The RMSN values are indicative for the overall error of each approach. The second column refers to an off–line fitting of data–driven models, while the rest of the columns refer to an on–line approach taking into consideration from 10 up to 100 most recent observations in each step of model fitting. B1695 data series should be omitted and not be used for a fair comparison as the whole B1695 data series is used for both calibration and validation. For the majority of the remaining data series, the loess method based on on–line calibration and taking into account only the 100 most recent observations outperforms the loess method based on off–line calibration (considering the entire data–set). For data series A358 and C621 it seems that more observations are required for a better performance. This could be attributed to different traffic conditions. Specifically, if 150 of the most recent observations are considered in each step of model calibration for A358 data series, the RMSN for the second methodological approach is reduced to 1.8% and outperforms the first methodological approach. Respectively, if 250 of the most recent observations are considered in each step of model calibration for C621 data series, the RMSN for the second methodological approach is reduced to 4.1%. The results are presented in Figure 4.25. The RMSN for the off–line process is indicated with red dashed line, while the RMSN against the number of the considered observations for the on–line process is plotted with black line. Consequently, the results indicate that if fewer observations, but more relevant, are used as input for data–driven car–following models, a better performance of the flexible models could be achieved.

Table 4.4: Modeling improvement versus number of observations

RMSN (%)											
Data series	Loess (off–line calibration)	Loess on–line calibration/ Number of observations									
		10	20	30	40	50	60	70	80	90	100
B1695	1.6	8.6	3.2	2.9	2.8	2.8	2.7	2.7	2.7	2.7	2.6
A358	2.1	10.6	4.9	3.5	2.8	2.8	2.6	2.5	2.4	2.5	2.5
C621	4.3	26.6	29.6	9.3	8.4	7.3	7.2	6.1	5.9	6.0	5.8
C171	6.2	37.7	9.2	8.2	9.3	8.8	5.2	4.4	4.5	4.2	4.2
A172	3.4	18.7	6.4	5.3	4.5	3.7	3.0	2.9	3.0	3.0	2.9
C168	1.8	6.4	5.0	3.3	2.4	2.2	2.0	1.9	1.6	1.5	1.4

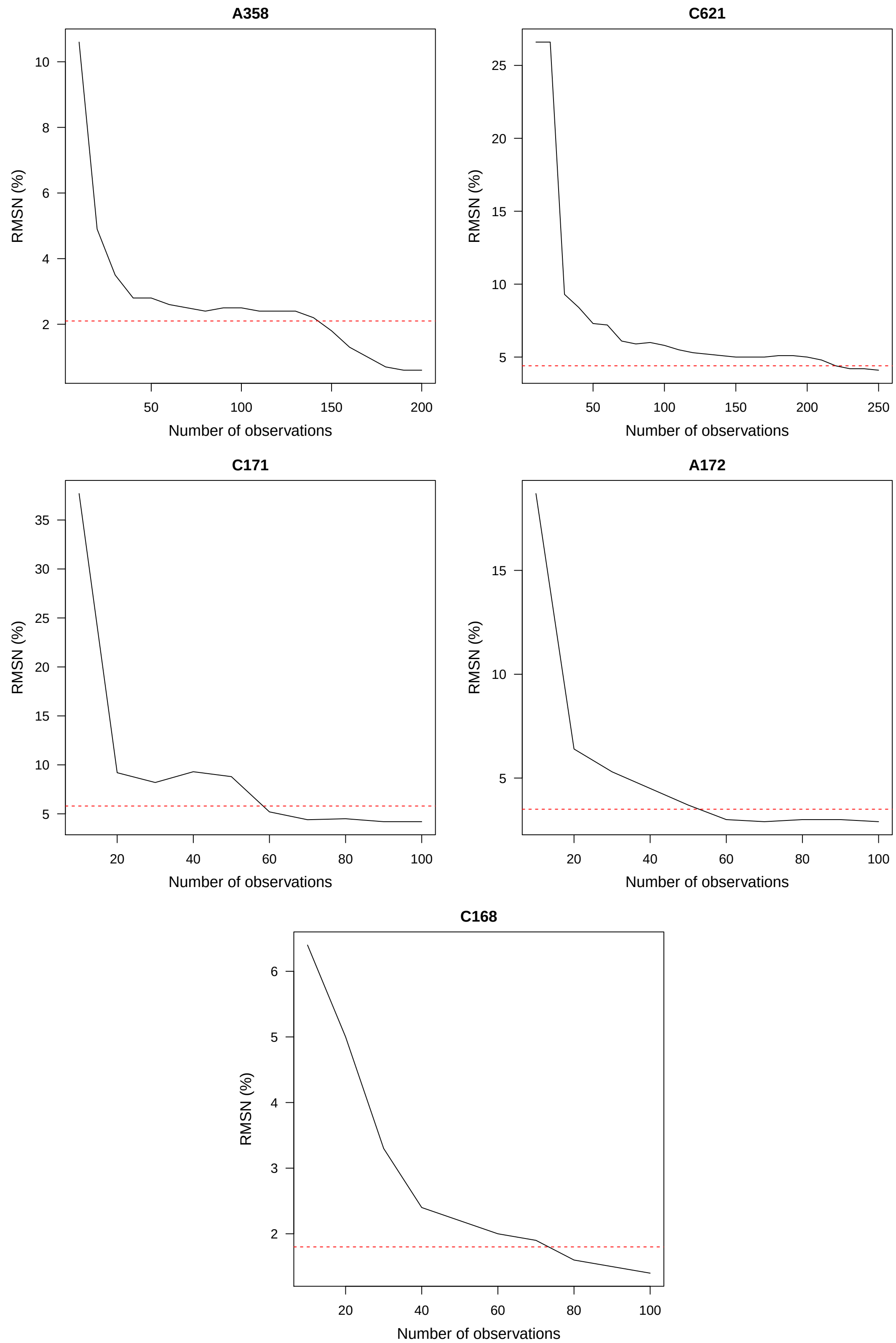


Figure 4.25: On–line application of loess method

4.5 Incorporation of additional variables

4.5.1 Motivation

In recent years, technological advances have significantly improved Driver Assistance Systems and there has been an increasing interest in autonomous vehicles. Aiming at safety, reliability and convenience, autonomous vehicles require detailed car–following models that could model driving behavior in an efficient way. In this section, the proposed methodology model is enriched by incorporating additional information about density of two adjacent lanes. It is explored if the additional information on density of adjacent lanes could improve the accuracy of a car–following model. More realistic detailed models could provide a robust solution to autonomous driving. The updated model is applied to reconstructed NGSIM data using a flexible regression technique, loess method. For a more in depth analysis, a meta–model is developed to evaluate the magnitude of the effect of the considered predictor variables on the proposed model.

Car–following models and driving behavior have been studied with many diverse approaches for decades. In recent years, technological advances have significantly improved Driver Assistance Systems and Intelligent Transportation Systems. Moreover, increasing volumes of potentially useful data are readily available in low-cost opportunistic sensors. Nowadays, there has been an increasing interest in self–driving or autonomous vehicles. Aiming at safety, reliability and convenience, an autonomous vehicle should adapt to user preferences and simulate human driving reactions naturally, preventing abrupt acceleration and jerk (Kuderer et al., 2015). Undoubtedly, in this direction, machine learning techniques have played a key role in learning driving styles and realizing Autonomous Driving. Wachenfeld and Winner (2016) have paid attention to collective learning in the context of autonomous driving, as directly exchanging with and copying from the learned is one of the particular advantages machine learning has over the human version. Machine learning methods can capture driving behavior in an efficient way taking into account various variables. In contrast, traditional car–following models based on a mathematical formula may not allow the incorporation of all these variables because of the high number of parameters (Papathanasopoulou and Antoniou, 2015b; Antoniou and Koutsopoulos, 2006a).

There are several successful demonstrations of machine learning algorithms in the field of intelligent autonomous vehicles (Ding et al., 2015; Xu et al., 2015). Riedmiller et al. (2007) used reinforcement learning to learn a steering controller from scratch. Their approach learns a controller that is able to navigate the vehicle on the track within 25 min of driving a real car. Kuderer et al. (2015) presented an inverse reinforcement learning method to learn individual driving styles for self–driving cars from demonstration. In order to capture the relevant properties of highway driving, they proposed a set of features that captures distances to other vehicles, the distance to the desired lane as well as higher order properties such as velocities and accelerations. Zhou and Qu (2016) have developed a microscopic car–following model for autonomous vehicles using Reinforcement Learning. Huval et al. (2015) showed how existing convolutional neural networks (CNNs) can be used to perform lane and vehicle detection while running at frame rates required for a real-time system. They rely only on the robustness of a neural network to make reasonable predictions. In addition, some researchers have introduced new insights for car–following models by exploring how traffic flow in the adjacent lanes could affect car-following behavior. Relative speed from one or two adjacent lanes has been taken into consideration (Ponnu and Coifman, 2015; Yu et al., 2015). However, features such as lane density of the adjacent lanes has not been included in the aforementioned studies. In this research it is attempted to improve further the existing model developed by (Papathanasopoulou and Antoniou, 2015b). This could be achieved by incorporating additional information to the model aiming at more detailed models. This research aims to explore if a car–following model depends on new features such as the density of adjacent lanes and if this additional information could improve the accuracy of speed prediction. Furthermore, the significant contribution of machine learning methods into autonomous driving is recognized. More realistic detailed models could provide a robust solution to autonomous driving. Attention is also given to flexibility of these methods and a metamodel for evaluation of model parameters is suggested.

4.5.2 Methodology extension

The proposed methodology for estimation of data–driven models is flexible enough to allow the incorporation of additional information to microscopic models. This is the main advantage of machine learning methods against traditional mathematical models. It is assumed that the speed of a vehicle in the next time instant is a function of various features.

$$\mathbf{v}(t+\tau) = f(\mathbf{x}_1(t), \mathbf{x}_2(t), \ldots, \mathbf{x}_n(t)) \tag{4.3}$$

where:
t: time instant
τ: the apparent reaction time
$\mathbf{v}$: the speed of a vehicle in a car–following state
$\mathbf{x_1}$, $\mathbf{x_2}$, ...,$\mathbf{x_n}$: predictor variables (such as speed of the preceding vehicle, distance from the front vehicle, lane density, weather, information from adjacent lanes etc.) that affect the driving behavior.

The accurate formula of Equation 5.4 is unknown as machine learning techniques are used. While they may not provide as much insight into traffic flow theory as the traditional models, new predictor variables could be easily added to the process. The opportunity to incorporate new kind of data is explored in this subsection (Figure 4.26). The existing model is enriched by adding the information about the density of adjacent lanes. The purpose of this choice was to examine if a car–following model depends on traffic of the adjacent traffic lanes or not.

The proposed methodology may benefit from a system that allows a fleet of connected vehicles to exchange information, such as (X, Y, Z) coordinates, using a central data system. However, even a single instrumented vehicle, with the ability to geo–locate itself and obtain (e.g. via suitable instruments and cameras) estimates of the speed and distance of surrounding vehicles, has access to all information required for this methodology.

	Gipps' model	Flexible regression technique
Application using variables including v, v_front, D_front	✓	✓
Incorporation of further variables	NA	✓

Figure 4.26: Modeling improvement including further explanatory variables

The performance of the models is evaluated using goodness–of–fit measures described in Section 3. In addition, a metamodel is used in order to evaluate the significance of predictor variables.

4.5.3 Application and Results

Four models are applied to the available data using loess method. The models presented in this research were all implemented using the R Software for Statistical Computing R Core Team (2017).

As training data for the calibration of traffic models, the longest data set is selected, the first one in Table 3.2. The rest of available data series are used as explanatory data for validation of the models. For the application of the loess method, parameters degree=1 and span =0.75 are set, as they have been chosen in the previous work Papathanasopoulou and Antoniou (2015b). In addition, the apparent reaction time is considered as 0.4 s, as the previous research, too. The predictor variables considered in each model are outlined in Table 4.5. The Model I is the existing one from the previous research Papathanasopoulou and Antoniou (2015b). Specifically, predictor variables are defined as following.

v: the current speed of the vehicle
v_{front}: the current speed of the preceding vehicle
D_{front}: the distance from the front vehicle
N_{left}: the number of existing vehicles in a distance of 100 m (50 m forwards and backwards, see Figure 4.27) in relation to the moving vehicle in the left adjacent lane
N_{right}: the number of existing vehicles in a distance of 100 m (50 m forwards and backwards, see Figure 4.27) in relation to the moving vehicle in the adjacent lane on the right
N_{all}: the number of existing vehicles in a distance of 100 m (50 m forwards and backwards, see Figure 4.27) in relation to the moving vehicle in both adjacent lanes, given by Equation 4.4. It is assumed that the density of both adjacent lanes would contribute equally to the model.

$$N_{all} = 0.50 \cdot N_{left} + 0.50 \cdot N_{right} \tag{4.4}$$

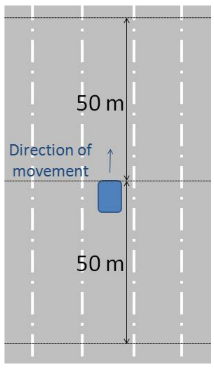


Figure 4.27: Area considered for lane density

The predictor variables, which are used in each model (Table 4.5), are referred to time instant t, while the response variable in all cases is the speed v of the vehicle at time instant t+τ.

The application results of four models are presented in Table 4.6 and Figure 4.28. It is observed that for some data series Model II or Model III outperform Model I. However, for all data series Model IV outperforms the other three models. Therefore, the accuracy of the method has increased by adding information about the density of the two adjacent lanes.

4.5.4 Effect of the predictor variables on the model

Visual inspection

A visual inspection is attempted for dependent and independent variables used in Model IV. In Figure 4.29, variables and their correlations are outlined. In the

Table 4.5: Predictor Variables used in Models

Models	v	v_{front}	D_{front}	N_{left}	N_{right}	N_{all}
Model I	✓	✓	✓			
Model II	✓	✓	✓	✓		
Model III	✓	✓	✓		✓	
Model IV	✓	✓	✓			✓

Table 4.6: Results

Data series	RMSN (%) Model I	Model II	Model III	Model IV
1	3.7	3.7	**3.6**	**3.6**
2	7.7	6.0	7.1	**5.4**
3	7.6	13.7	**5.7**	5.9
4	9.5	**9.4**	9.9	9.5
5	4.5	**4.0**	4.5	4.3
6	11.7	15.5	13.9	**10.5**
7	7.4	9.1	6.0	**5.9**
8	11.0	12.9	11.4	**10.4**
9	10.5	12.4	10.8	**9.7**
10	7.7	7.9	7.7	**7.1**

diagonal line, histograms of the variables are presented. In the upper triangle, values of correlations have been estimated for each potential pair of variables, while scatterplots of variables are displayed in the lower triangle. Looking at the lower triangle, two trends are evident, one for speeds lower than 6 m/s and one for speeds higher than 6 m/s. As concerns as speeds greater than 6m/s, RMSN decreases, as speeds v approach the value of 6 m/s. On the other hand, for speeds higher than 6 m/s, a balance with small variations of RMSN is observed.

Qualitative analysis

For a more in-depth and quantitative analysis of the proposed model, a meta-model is used. This is a way to test the predictors, but of course there are other ways as well in the machine learning literature.

The effect of the predictor variables on RMSN is estimated. In order to evalu-

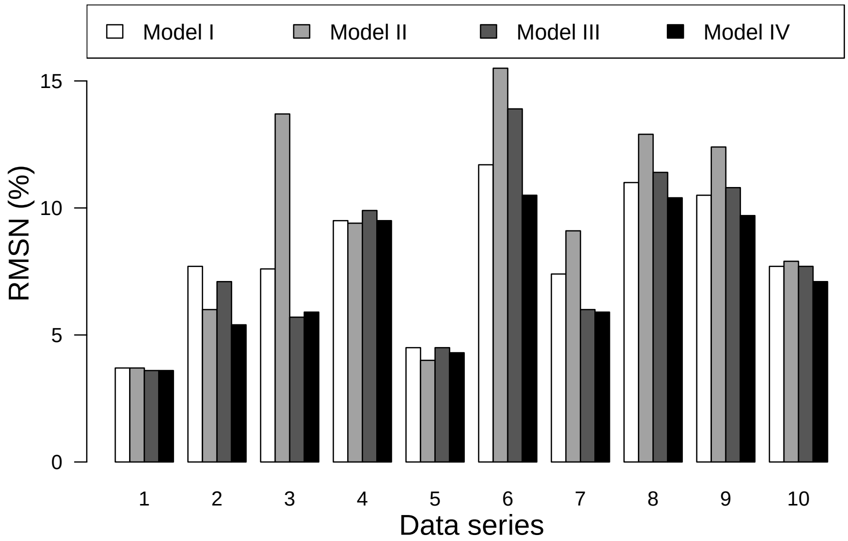


Figure 4.28: RMSN (%) for all data series and all models

ate if the considered predictor variables are highly significant, a multiple linear regression is performed. It is assumed that RMSN is the dependent variable and that the predictor variables v, v_{front}, D_{front} and N_{all} are independent variables that might affect RMSN. The general format of the metamodel is:

$$\mathbf{RMSN} = \beta_0 + \beta_1 \cdot \mathbf{v} + \beta_2 \cdot \mathbf{v_{front}} + \beta_3 \cdot \mathbf{D_{front}} + \beta_4 \cdot \mathbf{N_{all}} \tag{4.5}$$

A multiple linear model is fitted to the observed data and the RMSN resulting from the application of the proposed model. Due to two trends observed, it is suggested that two metamodels are developed. Data are divided into two categories, observations with speeds v higher than 6 m/s or lower than 6 m/s. Speeds v and v_{front} are highly correlated. Thus the division is implemented taking into account only speeds v.

The model estimation results for the first group (observations with v<6 m/s) are presented in Table 4.7. As it is indicated by t values, all the variables are highly significant. Specifically, there is an overall positive relationship between the variable N_{all} and RMSN, which could be physically interpreted as higher values of lane density combined with low speeds lead to more conservative car–following behavior (moving to a more congested environment) and therefore to higher values of RMSN. On the other hand, there is a negative correlation between the other variables (speeds and distances) suggesting that higher speeds and higher distances lead to lower RMSN (and therefore a better fitting model). Higher speeds and reasonable distances lead to a more uniform flow, and therefore it is expected that the fit of the model would be better (i.e. lower RMSN).

Regarding the second group (observations with v>6 m/s), a regression was attempted and all the variables were included. However, variable D_{front} was not

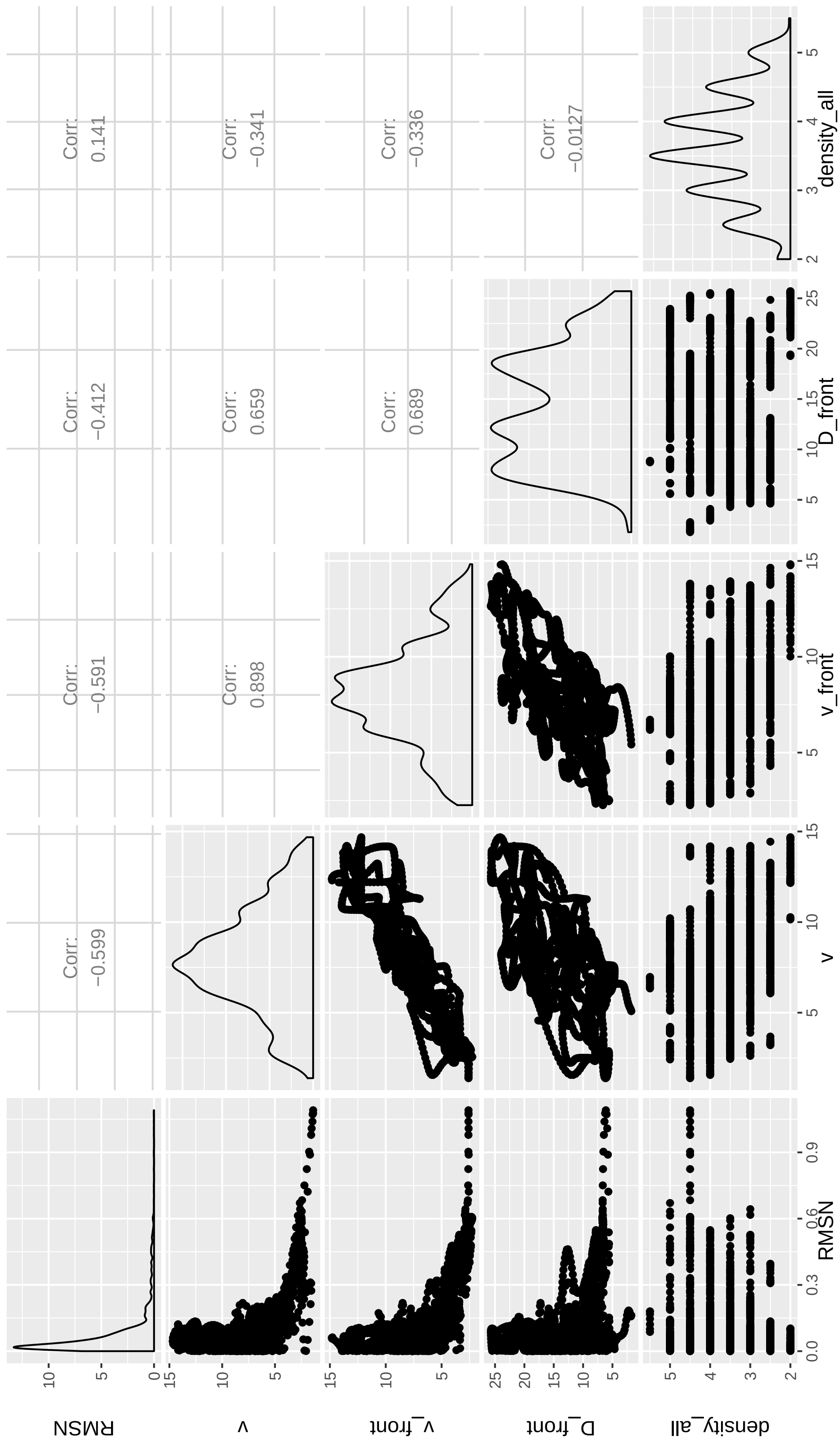


Figure 4.29: Visual inspection of model variables including scatterplots, histograms and correlation values

significant and it was excluded. This could be attributed to the fact that higher speeds often lead to higher distances moving from car–following to independent movement. Then, the regression was revised and the results are presented in Table 4.8. As it is indicated by t values, all the remaining variables are significant. Specifically, there is an overall negative relationship between all the variables and RMSN, suggesting that higher speeds and densities lead to lower RMSN (and therefore to a better fitting model).

Two parts of the dataset imply a different error distribution. A model that assumes varying variance could be also implemented (e.g. Heteroskedastic GP, heteroskedastic time series ARCH). Incorporation of information from adjacent lanes contributes into creating more detailed models. Machine learning methods are promising for modeling driving behavior, as well as for self–driving vehicles. The updated model was applied to the available data and outperforms the existing model. Higher distances and speeds could imply a driving behavior that is more consistent, thus leading to better model fit (i.e. lower RMSN). Machine learning techniques could provide robust policies for autonomous vehicles.

The proposed methodology should be implemented not only for speed estimation but also for estimation of spacing, as suggested by Punzo and Montanino (2016). In addition, in this case study it is considered that both adjacent lanes contribute equally to modeling driving behavior. Different weights of density on the right or left lane should be also tested. Further information on various predictor variables (such as weather, lighting, road geometry, etc.) could be also added to the model and tested if it is significant.

Table 4.7: Multiple linear Regression Results for Observations with v<6 m/s

	Estimate	Std. Error	t value
v (m/s)	-0.271	0.035	-7.745
$v_{front}(m/s)$	-0.284	0.035	-8.205
D_{front}	-0.057	0.012	-4.588
N_{all}	0.272	0.029	9.275

Table 4.8: Multiple linear Regression Results for Observations with v>6 m/s

	Estimate	Std. Error	t value
v (m/s)	-0.050	0.021	-2.427
$v_{front}(m/s)$	-0.150	0.021	-7.206
N_{all}	-0.518	0.020	-26.398

4.6 Conclusions

Data driven estimation of car following models appears to be a promising tool that could offer considerable benefits if integrated into traffic simulation models, resulting in higher accuracy and reliability of model outputs. In this research, an alternative methodology for estimation of car following models has been presented. A simulation optimization of car–following models could be achieved using the proposed method. The proposed method outperforms the reference (Gipps') model for all available data. This research corroborates results from other studies (Simonelli et al., 2009; Bifulco et al., 2013b), which imply that data–driven methods could provide better estimation than conventional models. The additional contribution of this work is to suggest an easy and quick methodology for estimation of car–following models, especially for applications that individual and personalized models are not so critically necessary. Machine learning methods present great flexibility and speed in managing data, avoiding the time consuming process of parameter calibration, which is essential for traditional car following models. The most important advantage of the proposed method is that it does not require the specification of a function to fit a model simultaneously to all the data in a sample. There is only the need to define the smoothing parameter and the degree of the local polynomial, avoiding time–consuming calculations for traffic model calibration. Moreover, the loess method could be helpful and flexible in specific and complex traffic situations (for instance emergency cases), for which no theoretical models may exist or it is complicated to be specified. Therefore the proposed method can be used in calibration issues, where computational ability of classical methods is limited. Furthermore, it provides the opportunity for incorporation of additional parameters without the need to resort to cumbersome reformulations of the model functional form. The loess procedure is suitable for data with outliers (not extreme), when a robust fitting method is essential. On the other hand, the loess method cannot be represented by a mathematical formula and it is thus difficult to transfer the results to other cases. It should be also referred that too large data series may require too much memory for loess application and that extreme outliers may overcome the method. While this research focused on the longitudinal interaction among vehicles moving in sequence, there are many other perspectives of the issue that have not been considered, such as the lateral interaction of a vehicle with vehicles in near–by lanes, as well as interactions between vehicles and road infrastructure (road curvature, existence of stop or traffic light at intersections, etc.). However, it is noted that Brackstone et al. (2009) detected little correlation between road type and driving behavior. Other factors, which could influence driving behavior and may be incorporated into the process, include drivers' characteristics, such as age, reaction time and experience. Moreover, heterogeneity of data could be addressed by more specific car-following models moving towards clustered models. This could be an interesting topic for future work. This research suggests that data–driven methods could provide reliable results and potentially more accurate than traditional car– following models. However, traditional car–following models have the advantage of basing their output on traffic flow theory. In contrast, despite their flexibility, computational approaches do not contribute as much in the advancement of traffic flow theory, may not be necessarily comprehensible by the human mind. Integration of data-driven methods in the simulation process requires additional studies that will further confirm their validity. Technological advancements will contribute significantly into the collection of data, which could in turn result to more robust and reliable models. Although many theoretical models have been developed so far, there is lack of a robust model that could generally represent car–following behaviors under all conditions. Data driven methods can be a plausible substitute for theory–based models and this research provides a contribution towards this direction. Naturally, a lot of further research is needed to elucidate further aspects of its applications and verify its robustness.

Chapter 5

Dynamic calibration and multiple time step prediction

Simulating driving behavior in high accuracy allows short–term prediction of traffic parameters, such as speeds and travel times, which are basic components of Advanced Traveler Information Systems (ATIS). Models with static parameters are often unable to respond to varying traffic conditions and simulate effectively the corresponding driving behavior. It has therefore been widely accepted that the model parameters vary in multiple dimensions, including across individual drivers, but also spatially across the network and temporally. While typically on–line, predictive models are macroscopic or mesoscopic, due to computational and data considerations, nowadays microscopic models are becoming increasingly practical for dynamic applications. In this research, we develop a methodology for online calibration of microscopic traffic simulation models for dynamic multi–step prediction of traffic measures, and apply it to car–following models, one of the key models in microscopic traffic simulation models. Online calibration of model parameters is data–driven in the sense that the data from previous time instants indicate parameter values for the next time instants. The methodology is illustrated using real trajectory data available from an experiment conducted in Naples, using a well–established car–following model. The performance of the application with the dynamic model parameters consistently outperforms the corresponding static calibrated model in all cases, and leads to less than 10% error in speed prediction even for ten steps into the future, in all considered data–sets.

5.1 Problem statement

While Advanced Traveler Information Systems have been around for decades, current developments, such as the increasing interest in Active Traffic Management, make them more relevant (Kurzhanskiy and Varaiya, 2010). Indeed, ATIS can be

[0]The chapter 5 is based on Papathanasopoulou et al. (2016).

effective in supporting active traffic management policies by Real-Time Decision Support Systems, whose core engine is a real-time traffic simulation model. The real–time requirements bring to the forefront the limitations of static calibration, and accelerate the need for procedures like the ones discussed in this research. An example of such applications is the Integrated Corridor Management initiative in the US (Miller and Skabardonis, 2010).

Reliable representation of driving behavior is a crucial issue for traffic simulation. Appropriate simulation models are chosen according to the requirements of each application; when considering the modeling detail, traffic simulation models can be divided into microscopic, mesoscopic and macroscopic. Microscopic models provide the highest level of detail for advanced transport applications (Antoniou and Koutsopoulos, 2006a). However, the traditional static calibration approach may not allow the incorporation of driving heterogeneity in the simulation.

In recent years there has been a tendency towards more flexible and dynamic methods than static car–following models. It is widely accepted that driving behavior in general (and therefore car–following parameters) vary in multiple dimensions, i.e. exhibit inter–personal, temporal and spatial heterogeneity. Higgs and Abbas (2014) compare car–following models at different levels of analysis: driver, car–following period and cluster of drivers. For example, Ossen and Hoogendoorn (2005) have identified considerable differences between the car–following behavior of individual drivers. Ma (2006) has developed a neural fuzzy framework for modeling car–following behavior. It illustrates human knowledge of car–following in a more understandable manner and can be rather flexible as the regime parameters and model forms may vary according to the application context. According to Hoogendoorn et al. (2006) and Ellison et al. (2013), real driving behavior is variable in time and space. Some researchers have attempted to capture heterogeneity across drivers spatially (e.g. Papathanasopoulou and Antoniou (2015b)) or temporally, which is one of the aspects investigated in this research.

Many car–following models predict a stable car–following behavior with a very small fluctuation around an equilibrium value. However, in reality these fluctuations are much larger than these models predict. Wagner (2012) has attributed them not due to driver heterogeneity, but to an internal stochasticity of the driver itself. Randomness is thus incorporated in traffic flow and model calibration requires the flexibility to adapt to it. On the other hand, several empirical analyses performed by Ossen and Hoogendoorn (2007) showed a high degree of driver heterogeneity in car–following. Inter–driver differences could be described not only by different parameter values, but also different model specifications may be needed. All above researchers conclude that different optimal parameter values, as well as different optimal car–following models, should be applied to overcome this problem.

Static calibration requires a database with historical data. It could feed a simulation model with initial parameter values, which allow a good representation of a general traffic state (Balakrishna, 2006). However, dynamic calibration could take advantage of real–time data and adapt model parameters to the current traffic state. Dynamic estimation of model parameters and especially reaction time have been attempted by e.g. Hoogendoorn et al. (2006) and Ma and Jansson (2007). Hoogendoorn et al. (2006) and Lorkowski and Wagner (2005) use the Unscented Kalman Filter (UKF), while Ma and Jansson (2007) have proposed a dynamic model estimation method based on iterative usage of the Extended Kalman Filter (IEKF) algorithm. Ma and Andréasson (2005) have suggested a dynamic car following data collection and noise cancellation based on the Kalman smoothing. However, according to Treiber and Kesting (2013), smoothing the data had no significant influence on the calibration quality. Naturally, calibration and sampling issues in estimation car–following parameters have been studied extensively in the literature (Monteil et al., 2014; Schultz and Rilett, 2004)

Rahman et al. (2014) develop a calibration approach based on the Markov Chain Monte Carlo (MCMC) simulation that uses the Bayesian estimation theory. The authors use a linear model (Helly) with a different number of vehicle trajectories on a highway network. Ossen and Hoogendoorn (2011) investigate the level of heterogeneity in car–following behavior, using a large number of trajectory observations collected via helicopter. The authors observe different vehicle type interactions and develop eight models with different car–following assumptions.

Simulation models do not always adequately reflect field conditions outside of the time period for which they have been calibrated (Balakrishna et al., 2007b; Daamen et al., 2014; Henclewood et al., 2012). Microscopic models often comprise different detailed models, including car–following, lane–changing and gap–acceptance models. In most cases, the parameters of these models are assumed to be stable, both across space and time, and also across drivers. The online calibration of car–following models is a promising approach to capture the heterogeneity of driver behavior and traffic conditions. By continuously supplying a car–following model with surveillance data, an online calibration process could be applied in order to adapt model parameters to the current traffic state. In this view, the use of richer data, such as real–time Floating Car Data (FCD), based on traces of Global Navigation Satellite Systems (GNSS), could be leveraged as a reliable and cost–effective way to gather accurate traffic data (De Fabritiis et al., 2008; Antoniou et al., 2011).

Calibration of car–following models (Brackstone and McDonald, 1999) has been an issue for a long time (Aycin and Benekohal, 1999), but nowadays it has received a new boost (Hoogendoorn and Hoogendoorn, 2010; Monteil et al., 2014), in light of new data–collection techniques, mostly related to the increasing availability of trajectory data (Kesting and Treiber, 2008; Punzo et al., 2005; Papathanasopoulou and Antoniou, 2015b), which of course introduce other challenges (Punzo et al., 2012).

Online calibration has been used in many macroscopic and mesoscopic modeling approaches (Papageorgiou et al., 1989; Kim, 2002; Antoniou et al., 2005; Fei et al., 2011). The use of the Kalman Filter (and its extensions) for online parameter calibration has shown encouraging results (Antoniou et al., 2007). However, in recent years there has been an increasing interest in online applications of microscopic traffic models. Moreover, Henclewood et al. (2012) suggest that a real-time calibration algorithm should be included in online, data–driven microscopic traffic simulation tools.

The objective of this chapter is to motivate, develop and demonstrate with real data a practical approach for the online calibration of microscopic traffic simulation models, which considers dynamic parameters for individual drivers, in time and space. At each time instance, the dynamically obtained model parameters are being used for short–term prediction (up to ten steps into the future), and the performance of this prediction is compared with the reference case of static model parameters.

5.2 On–line calibration process

In this research, a methodological framework for the dynamic calibration of car–following models using real–time data is proposed. The approach has two main steps, an estimation phase and a prediction phase. The estimation phase relies on a constrained global optimization algorithm. Once an optimal set of parameters is identified for each individual time–instance (and each individual driver), multi–step prediction is performed. In each time step, prediction is achieved using the estimated values from the previous time step, i.e. the best available estimates of the drivers' behavior. The evolving patterns of the driver behavior during the past few intervals is used to forecast the expected behavior during the next few time–steps. Naturally, this logic could be adapted to data–availability or computational concerns, e.g. it could only be invoked when there is an indication that traffic conditions may be changing (e.g. automatically detected anomaly in the data).

The dynamic calibration problem can be mathematically stated as

$$\min f(\mathbf{x_t}), \quad \mathbf{x_t} = (x_{1t}, \ldots, x_{nt}) \in \mathcal{R}^n \tag{5.1}$$

where $f(\mathbf{x_t})$ is the objective function (e.g. error estimating the difference between observed and simulated values of a traffic measure), t is the current time interval, $\mathbf{x_t}$ stands for time dependent parameters, and the feasible region $\mathcal{F}$ is defined by

$$\mathcal{F} = \{\mathbf{x_t} \in \mathcal{R}^n | g_j(\mathbf{x_t}) \leq 0 \quad \forall j\} \tag{5.2}$$

where $g_j(\mathbf{x_t}), \quad j = 1, \ldots, m$ are inequality constraints. For instance, these constraints could ensure that a traffic measure, such as space gaps, will not take unacceptable values. The way that the constraints are handled may vary according to the type of optimization algorithm that is adopted, including the penalty function method, special representations and operators, and the co–evolutionary method, the repair method or the multiobjective method (Runarsson and Yao, 2005). The dynamic calibration problem could also be stated as a multiobjective optimization problem including various measures of effectiveness (such as speeds, space gaps, and accelerations) and various measures of goodness–of–fit (such as normalized root mean square error and mean prediction error, as described in Chapter 3).

$$\min(f_1(\mathbf{x_t}), f_2(\mathbf{x_t}), \ldots, y_1(\mathbf{x_t}), y_2, (\mathbf{x_t}) \ldots z_1(\mathbf{x_t}), z_2, (\mathbf{x_t}), \ldots), \quad \mathbf{x_t} = (x_{1t}, \ldots, x_{nt}), \in \mathcal{R}^n \tag{5.3}$$

where $f_1(\mathbf{x_t}), f_2(\mathbf{x_t}), \ldots$ correspond to the same measure of effectiveness (e.g. comparison of observed and simulated speeds), but different measures of goodness–of–fit, while $f_i(\mathbf{x_t}), y_i(\mathbf{x_t}), z_i(\mathbf{x_t}) \ldots$ correspond to different measures of effectiveness, but for the same measure of goodness–of–fit.

The estimated model parameters can then be easily propagated into the future, thus providing plausible predictions of these values for the next intervals, as shown in Figure 5.1 for a hypothetical parameter. A general way to express this evolution could be

$$\hat{\mathbf{x}}_{t+1} = h(\mathbf{x}_t, \mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \ldots) \tag{5.4}$$

where h is a function relating an estimate of the car–following model parameter values in the next time interval to the best available estimates of the same model parameters in previous time steps. However, in this way all observations from time t=t_0 up to time t are considered as equally important and included to define the objective function. This approach could lead to errors since drivers may drive in a different way at the current moment in comparison with previous time instants. The apparent solution could be based on weighting the data:

$$\hat{\mathbf{x}}_{t+1} = h(w_0 \cdot \mathbf{x}_t, w_1 \cdot \mathbf{x}_{t-1}, w_2 \cdot \mathbf{x}_{t-2}, \ldots) \tag{5.5}$$

where w_i are the weights of observations. Presumably largest weights would be applied to the most recent data and therefore $w_1 > w_2 > w_3 > \cdots$. The decaying weights thus imply that in practice only a finite number of past observations affects the process. A more explicit example of an operationalization of this process is e.g. an autoregressive process, such as:

$$\hat{\mathbf{x}}_{t+1} = f_0 \cdot \mathbf{x}_t + f_1 \cdot \mathbf{x}_{t-1} + f_2 \cdot \mathbf{x}_{t-2} + \epsilon \tag{5.6}$$

where f_i are the autoregressive factors, and ϵ a random error term. A process for determining these factors is presented e.g. in (Antoniou, 2004, p. 79).

Through repeating this process, multi–step predictions can be obtained in the following manner:

$$\hat{\mathbf{x}}_{t+2} = h(w_0 \cdot \hat{\mathbf{x}}_{t+1}, w_1 \cdot \mathbf{x}_t, w_0 \cdot \mathbf{x}_{t-1}, \dots) \quad (5.7)$$

$$\hat{\mathbf{x}}_{t+3} = h(w_0 \cdot \hat{\mathbf{x}}_{t+2}, w_1 \cdot \hat{\mathbf{x}}_{t+1}, w_2 \cdot \mathbf{x}_t, \dots) \quad (5.8)$$

Naturally, as the number of time–intervals (for which the prediction is performed) increases, the accuracy of these predictions is expected to deteriorate, as evidenced schematically by the widening bounds of the predicted values in Figure 5.1. A practical way to determine these bounds is presented in Pereira et al. (2014). As the process is repeated, though, when new data are available, these forecasts can be improved, as shown by the lower subfigure in Figure 5.1. This concept of "rolling horizon" is being used in software for real–time traffic predictions (Ben-Akiva et al., 2010). In that case, at time $t+1$ the previously predicted values of the model parameters for time $t+1$ can now be estimated using available data. Furthermore, the previous two–step predicted model parameters for time interval $t+2$ can now be replaced by one–step predicted model parameters, since estimates of the model parameters are available up to time interval $t+1$. Similarly, e.g. at time $t+1$ for time interval $t+7$, the previous seven–step prediction (using estimated values up to time interval t and predicted parameter values for subsequent intervals) can be replaced by an —arguably more accurate— six–step prediction (using estimated values up to time interval $t+1$ and predicted parameter values for subsequent intervals).

The proposed methodological framework may benefit from a system that allows a fleet of connected vehicles to exchange information, such as (X,Y,Z) coordinates, using a central data system. However, even a single instrumented vehicle, with the ability to geo–locate itself, and obtain (e.g. via suitable instruments and cameras) estimates of the speed and distance of surrounding vehicles, has access to all information required to apply this methodology. Therefore, speeds, accelerations or gaps could be calculated per time instant and used to dynamically calibrate driver–specific, time–varying parameters for the considered model. The overall methodology is illustrated in Figure 5.2. A calibration process is demonstrated for each time instant t, using observations available up to that time and applying a suitable optimization algorithm, solving Equation 5.1 or Equation 5.3. The estimated optimal model parameters for time intervals up to t are being used to predict dynamic parameter values for the subsequent intervals, using Equations 5.4 or Equations 5.5 and so on. These predicted parameters can then be used as time–dependent inputs into the microscopic traffic models to provide predicted outputs for the horizon from t onwards. In the next time–interval, this process is repeated, but this time using also measured information

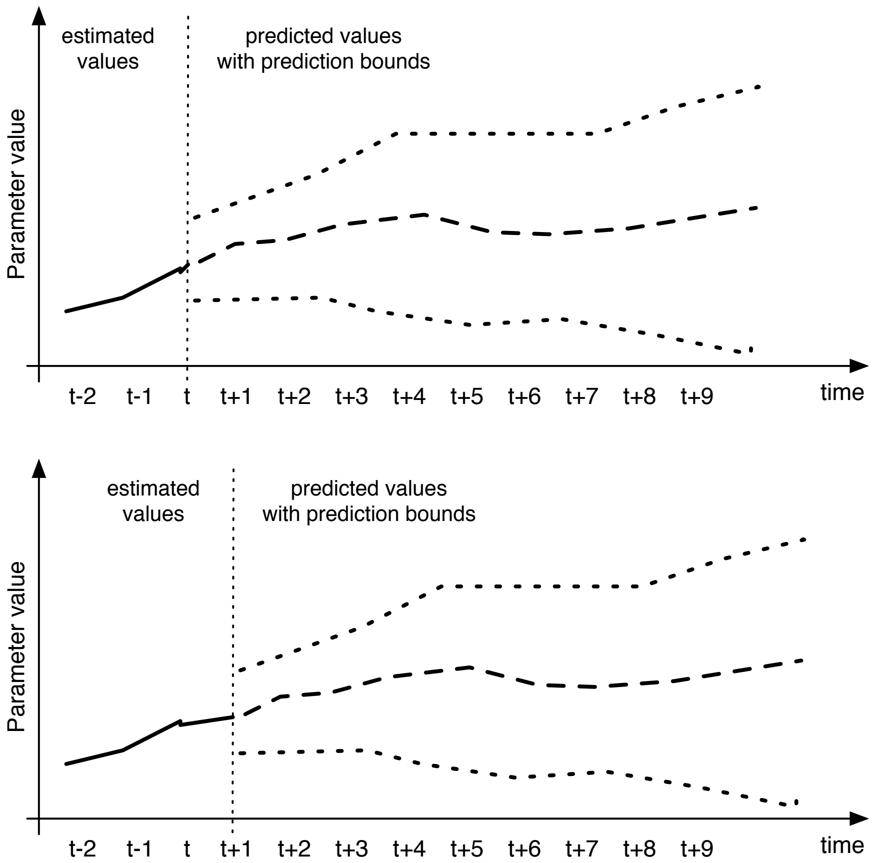


Figure 5.1: A rolling horizon for prediction with improving bounds

from time–interval $t+1$, thus obtaining an estimated set of model parameters for time interval $t+1$, updated predicted values for the subsequent intervals, plus a new predicted value for a new time–interval (for example, if performing n–step prediction, during time t we obtained predicted parameter values for intervals $t+1$ through $t+n$, but when we move to time interval $t+1$ then we obtain predictions for intervals $t+2$ through $t+n+1$, thus revising the previous estimates for intervals $t+2$ through $t+n$ and extending our prediction horizon to $t+n+1$).

The methodology is applied to a car–following model, which is arguably the most critical component of microscopic traffic simulation models. In particular, Gipps' model (used e.g. in the widely used Aimsun traffic simulation model and described in detail in chapter 2) is calibrated using available data from an experiment conducted in Naples (Punzo et al., 2005). A static calibration is also performed in order to be used as a reference benchmark. The main difference from dynamic calibration is that the parameter values are constant for all the time instants and all prediction steps. For both methods, the same optimization algorithm was used. Therefore, for the static calibration Equation 5.1 is simplified through the removal of subscript t, while of course the forecasting Equations 5.4

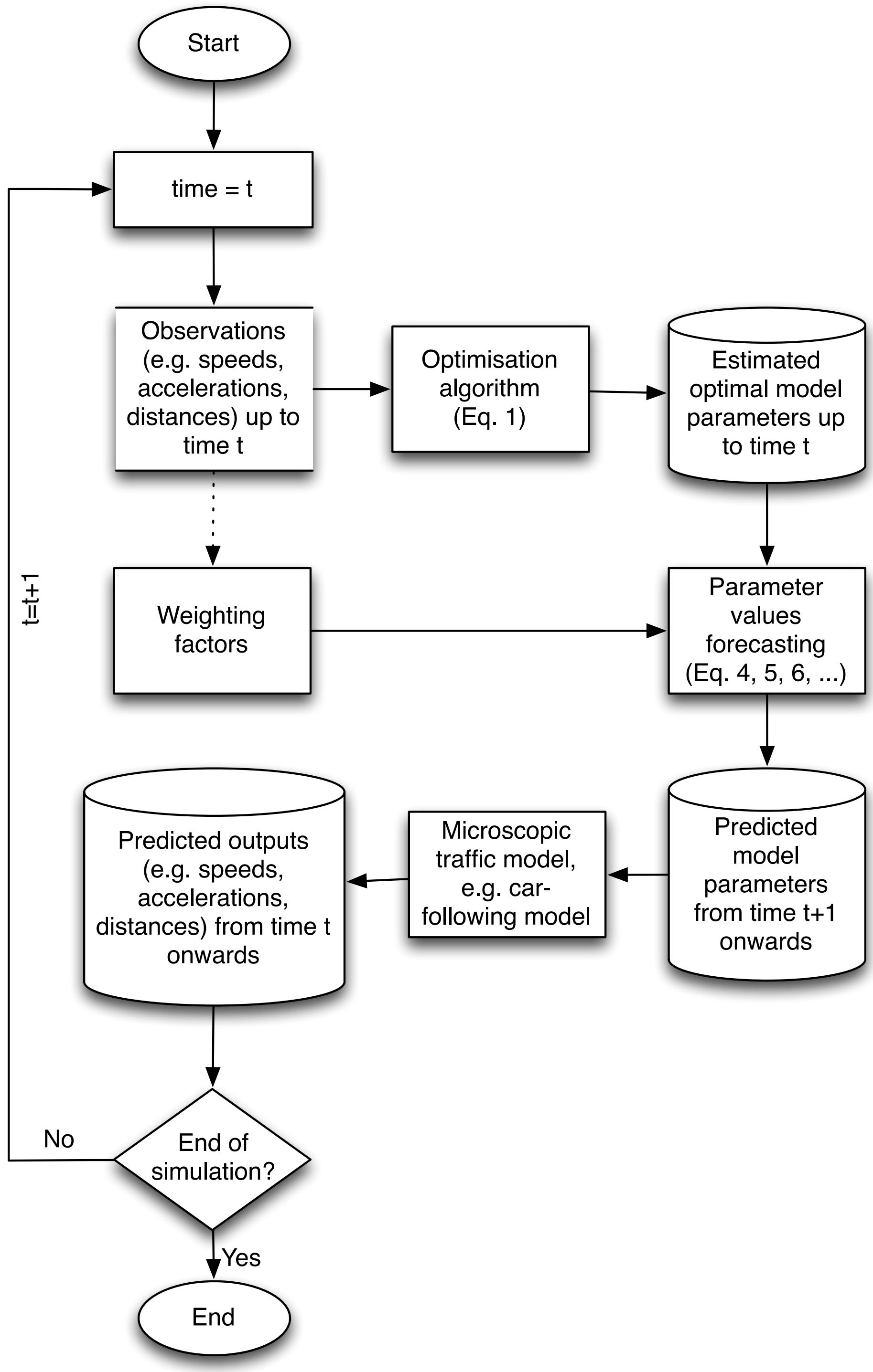


Figure 5.2: Methodological overview: dynamic calibration and multiple steps prediction

and so on are not relevant (since the parameter values are constant).

5.3 Static calibration

For a more evidence–based approach, the calibration process was performed using an optimization algorithm within the R software for statistical computing (R Core Team, 2016). In particular, the Improved Stochastic Ranking Evolution Strategy (ISRES) algorithm was used, which is included in the package "nloptr" (Runarsson and Yao, 2005) and is appropriate for nonlinearly constrained global optimization. This method is implemented in a simple way and supports arbitrary nonlinear inequality and equality constraints in addition to the bound constraints. In addition, it incorporates heuristics to escape local optima. On the other hand, although a lots of research has been performed on determining which algorithm is best suited for a given problem, there has not been a satisfactory answer to this question. Thus, various algorithms should be tested in future research.

The objective function that was minimized is: RMSN(v_3^{obs},v_3^{sim}). The range of model parameters are shown in Figures 4.1 and 4.2 and these constraints are defined by Equation 5.9:

$$\underline{x}_{it} \leq x_{it} \leq \bar{x}_{it} \tag{5.9}$$

which is a special case of Equation 5.2. In addition, as initial values for the optimization process, optimal values defined through the sensitivity analysis were used. It is noted that interactions among the parameters had not been taken into account in the sensitivity analysis. However, a global optimization is performed, considering all combinations of these model parameters concurrently. For the whole dataset B1695 the optimization process has converged to the optimal set of parameters after approximately 10000 iterations. The optimal values are presented in Table 4.1, where "initial values" refers to the model parameter values obtained by the sensitivity analysis and "optimal values" refers to the parameters obtained from the static calibration using the ISRES algorithm within this research. The minimum value of the objective function, namely the RMSN, that was achieved with these optimal values of parameters was 2.2%.

5.4 Dynamic calibration and results analysis

Gipps' model is calibrated dynamically in order to simulate the speed of the third vehicle (v_3(t+τ)). Gipps' model requires as input data $v_2(t)$, $v_3(t)$, $x_2(t)$, $x_3(t)$ and the appropriate parameter values. The superiority of this calibration over the static calibration presented before is demonstrated both for estimation and also for multiple step prediction of traffic speeds.

5.4.1 Dynamic calibration

In online calibration an optimization process with the same characteristics (parameters range, initial values, objective function) is iterated per time instant, namely per observation ($v_2(t)$, $v_3(t)$, $x_2(t)$ and $x_3(t)$), and not for the whole data series such as in static calibration. Therefore, a different optimal set of parameters is determined per time instant t in order to be characteristic of the current traffic conditions. In order to simplify the optimization problem, the apparent reaction time is considered equal to τ=0.4 sec (for a discussion and motivation of this choice, see e.g. Papathanasopoulou and Antoniou (2015b)). Five parameters are optimized per iteration (the remaining five parameters shown in Table 4.1).

This implies that —with the addition of subscript t to the time–dependent parameters— in effect Equation 2.3 becomes

$$v_{nt}[t+\tau] = min \begin{cases} v_{nt}[t] + 2.5 \cdot a_{nt} \cdot \tau \cdot (1 - \frac{v_{nt}[t]}{V_{nt}} \cdot \sqrt{(0.025 + \frac{v_{nt}[t]}{V_{nt}})} \\ b_{nt} \cdot \tau + \sqrt{(b_{nt} \cdot \tau)^2 - b_{nt} \cdot [2 \cdot (x_{n-1}[t] - s_{n-1,t} - x_n[t]) - v_{nt}[t] \cdot \tau - \frac{v_{n-1}^2[t]}{\hat{b_t}}]} \end{cases} \tag{5.10}$$

where:

a_{nt} : the maximum acceleration that the driver of vehicle n wishes to acquire (m/s^2).
b_{nt} : the maximum braking that the driver of vehicle n wishes to apply in order to avoid a crash, b_n<0 (m/s^2).
$\hat{b}_t$: the estimated maximum braking that the driver of the preceding vehicle (n-1) wishes to apply (m/s^2).
$s_{n-1,t}$ = L_{n-1} + Safety, namely the size of the preceding vehicle (n-1) including its length and the safety distance at which vehicle n is unwilling to compromise even when at rest (m).
V_{nt} : the speed at which the driver of vehicle n wishes to travel (m/s).
x_n[t], x_{n-1}[t] : the location of the front side of the respective vehicle (n or n-1) at time t (m)
v_{n-1}[t] : the speed of the preceding vehicle (n-1) at time t (m/s)
v_n[t] : the speed of the following vehicle (n) at time t (m/s)
τ : the apparent reaction time (a constant for all vehicles) (s)

In setting up the case study, a single objective optimization problem was formulated and the objective function f that was set to be minimized is:

$$\min f(\mathbf{x_t}) = \min(RMSN(v_3^{obs}, v_3^{sim}(\mathbf{x_t}))) \tag{5.11}$$

where

$$\mathbf{x_t} = (a_{nt}, V_{nt}, b_{nt}, s_{n-1,t}, \hat{b_t}) \tag{5.12}$$

In the optimization process for the case study, observations of only the previous time instant have been taken into consideration and therefore the formulation of the weights was straightforward, i.e. the weight of the last observation was 1 and that of all previous observations was 0. In Figure 5.3, the RMSN goodness–of–fit measure assesses the overall error and performance of the static and dynamic calibration procedures estimating the difference between the observed and the simulated values of speed v_3 per time instant.

RMSN(static) was calculated considering fixed parameters values for all observations. It becomes evident that lower RMSN values are consistently achieved through dynamic calibration, as expected. While the difference is not so large most of the time, there is a specific period, in which the static model seems to perform particularly poorly. In order to investigate the conditions that led to this performance deterioration, we have looked at the prevailing traffic conditions, which indeed provide some insight. In particular, in the bottom subfigure of Figure 5.3 observed speeds are also plotted over time, in order to clarify when the static model performance deteriorates. During the time period of interest (the sharp increase in RMSN), very low speeds are observed, and it is therefore reasonable to assume that the static calibration does not allow the model to adapt to these extreme traffic conditions. Therefore, it is not expected to provide a good speed prediction, when significant changes in speed take place. On the other hand, dynamic calibration is more flexible and adaptable to the current traffic conditions, reacting considerably better, even in such extreme situations.

In this case study, we use a single measure of effectiveness (speed). The failure of the dynamic calibration in the low–speed situation in Figure 5.3 could be overcome by considering additional measures of effectiveness (e.g. vehicle gap).

5.4.2 Multiple time step prediction

Multiple steps prediction is achieved using both methods of calibration, static and dynamic. In each prediction step, predicted values from the previous step and optimal values used in the initial step are imported to the car–following model. However, after the first prediction step, only the speed of the third vehicle v_{3t} is predicted and could be used to the next step. Therefore, the speed v_2 of the lead vehicle, which is also required as input to the model, is considered as constant through the prediction steps, since there is no way to know what the evolution of that speed would be. Extensions to this model could include predicting of other traffic parameters, such as speeds of surrounding vehicles. Values for vehicle positions x_2 and x_3 are calculated according to the distance that was traveled in the corresponding time (i.e. assuming the appropriate speed values).

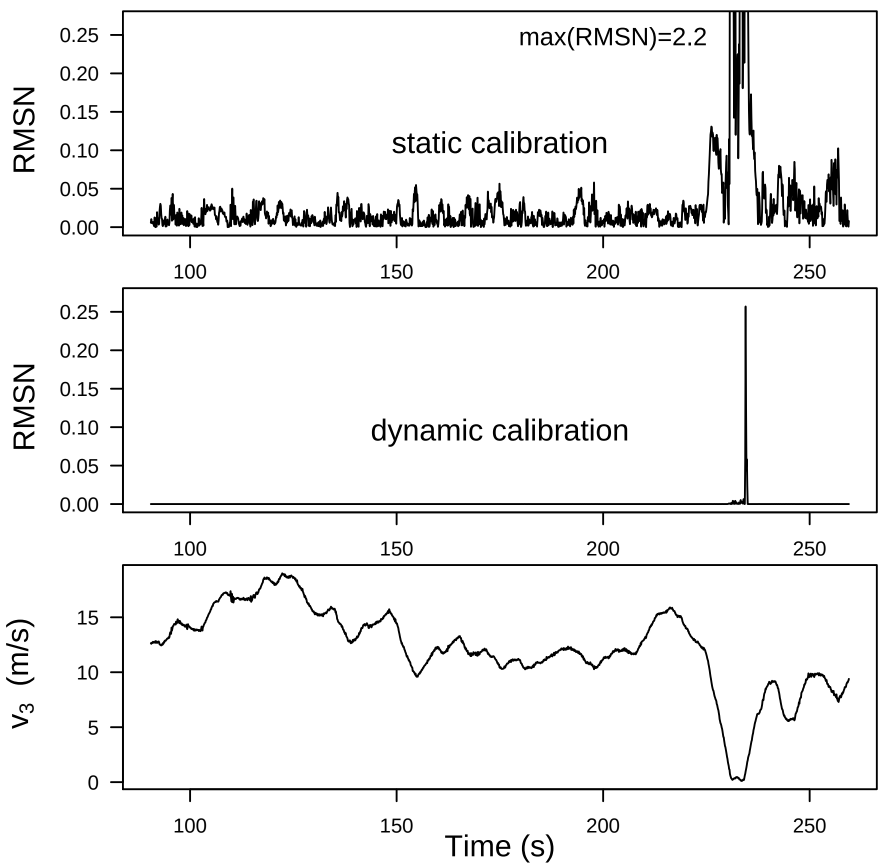


Figure 5.3: Dynamic calibration versus static calibration and speed profile (observed values) for B1695 data series

The proposed methodology has been applied for all data series. In the presentation of the results in the following subsections, we make a distinction between measures of effectiveness and measures of goodness–of–fit that have been used in the objective function, versus those that have not. In particular, the objective function has been formulated in terms of RMSN (as the goodness–of–fit measure) and following car speed (as the measure of effectiveness). Looking only at these metrics could leave us susceptible to over–fitting. However, to demonstrate that this is not the case, we also present results for other measures of effectiveness (in particular distance between the lead and following vehicle), as well as a number of other goodness–of–fitness measures (as described in Chapter 3).

Measures of effectiveness included in the objective function

Figure 5.4 presents the aggregate goodness–of–fit measures for the estimated speeds. The left subfigure of Figure 5.4 presents the calibration results based on the RMSN metric, which has been included in the objective function. Indeed, as expected, dynamic calibration outperforms static calibration in all cases for one–step prediction. This is evident also from all other goodness–of–fit measures (right subfigure of Figure 5.4). The RMSPE measure indicates that using static calibration results in large errors. Values of Theil's inequality U_m and U_s are close to zero for all data series and indicate an ideal fit. It is noted that for the B1695 data series, prediction with the static calibration is using additional information, as the information for the entire data–set has been considered for the estimation of the static parameter values. Therefore, this value is "better" than it would otherwise be. Put differently, the improvement obtained by the dynamic calibration is arguably underestimated. In Figure 5.5, the Empirical Cumulative Distribution Functions (ECDFs) of observed and simulated speeds are illustrated. The ECDF of simulated speeds using dynamic calibration is almost identical with the ECDF of observed speeds in the majority of data series. While in many cases the static calibration also performs reasonably well, in some cases, such as in the first half of data series A172 and C171, the poorer performance of the static calibration (compared to dynamic calibration) is clearly evident.

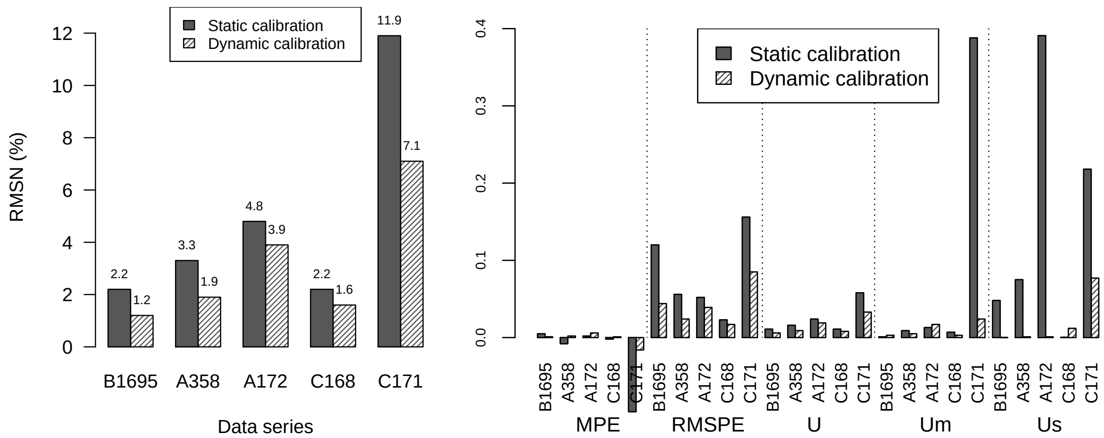


Figure 5.4: Comparison of estimated speeds between static and dynamic calibration for one–step prediction

Measures of effectiveness not included in the objective function

In order to check if the methodology over–fits to speed prediction due to the objective function defined in Equation 5.11, other measures of effectiveness such as space, acceleration etc. not included in the objective function may be checked. For instance, space prediction is not included in the objective function (Equation 5.11) and this could make the parameter estimation rather insensitive to space–gap related properties.

Space headways for dynamic calibration have been estimated using the assumption that the the speed v_2 is constant throughout the prediction steps. Figure 5.6 is similar to Figure 5.4, but this time presents the same results for space

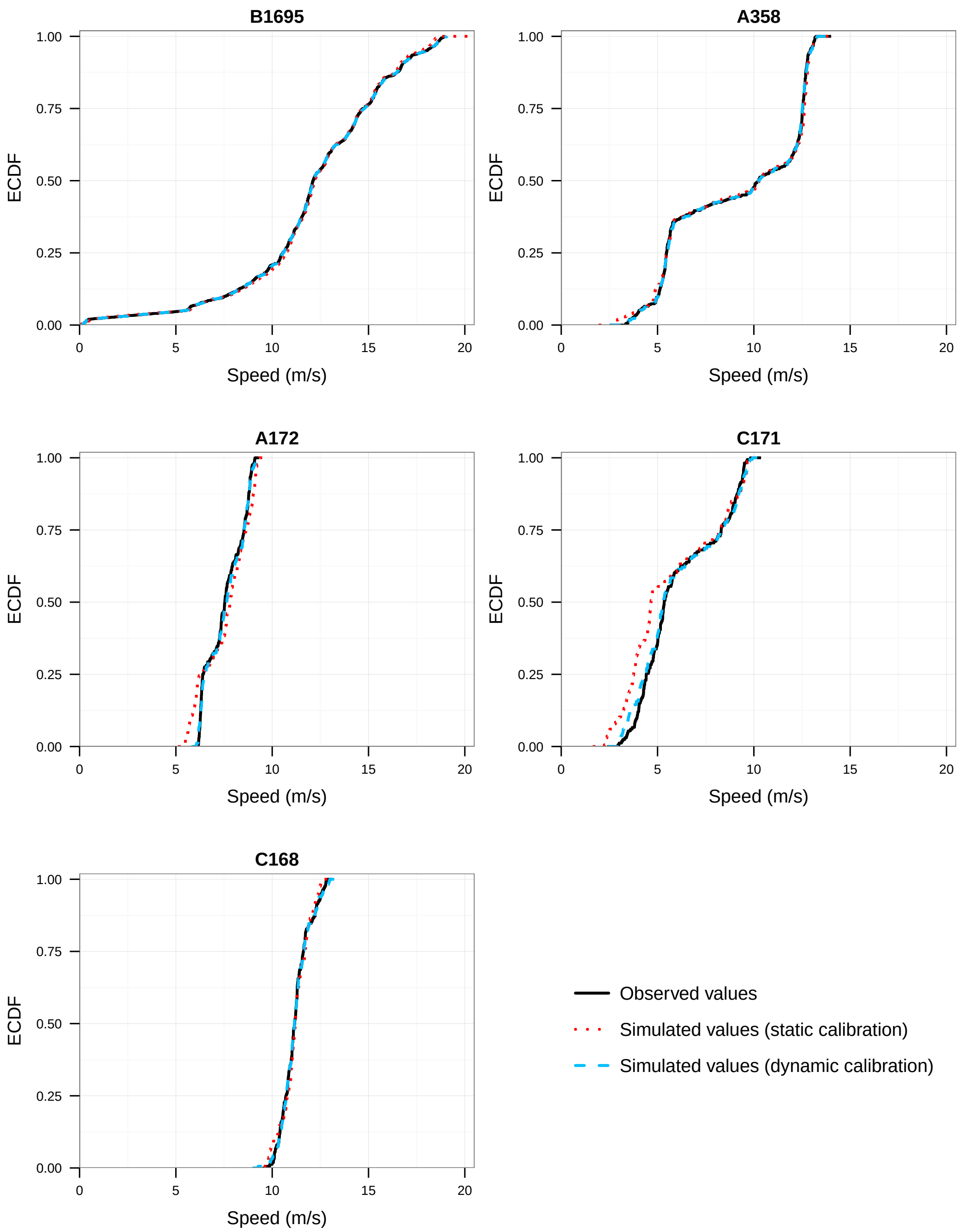


Figure 5.5: ECDFs of observed and simulated speeds for one–step prediction

headways. Indeed, the dynamic calibration once again performs considerably better than the static calibration, in most cases. Furthermore, Figure 5.7 illustrates the Empirical Cumulative Distribution Functions (ECDFs) of the observed and simulated space gaps. Although space gaps were not included as a traffic measure in the objective function, the estimated space gaps from the dynamic calibration track the observed values much better than those obtained from the static calibration.

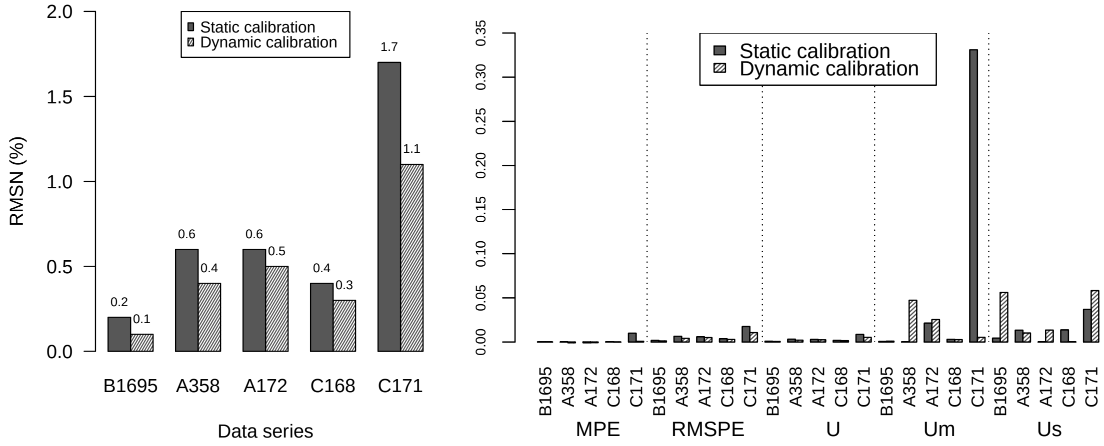


Figure 5.6: Comparison of estimated space gaps between static and dynamic calibration for one–step prediction

5.4.3 Multiple time step prediction results for the entire prediction horizon

The RMSN results for all five data series and for the entire prediction horizon (one– to ten–step prediction) are presented in Figure 5.8. The superior performance of the prediction based on the dynamic calibration is evident in all data series. While prediction error consistently increases as the prediction horizon increases, the results obtained from the dynamic calibration never exceed 10% error, even for ten–step prediction into the future.

Figure 5.9 presents a more detailed overview of the predicted speeds by the static and dynamic calibration, compared with the observed speed for one of the data series (C171, the one for which the biggest improvement has been achieved through the use of dynamic vs. static calibration).

5.4.4 Exploration of calibrated parameter values

The overall results presented in the previous subsection eloquently illustrate the superior performance of the dynamic calibration over the static case. In this subsection, we explore the values of the parameters obtained from each approach, in order to gain some further insight. Figure 5.10 presents the densities of the

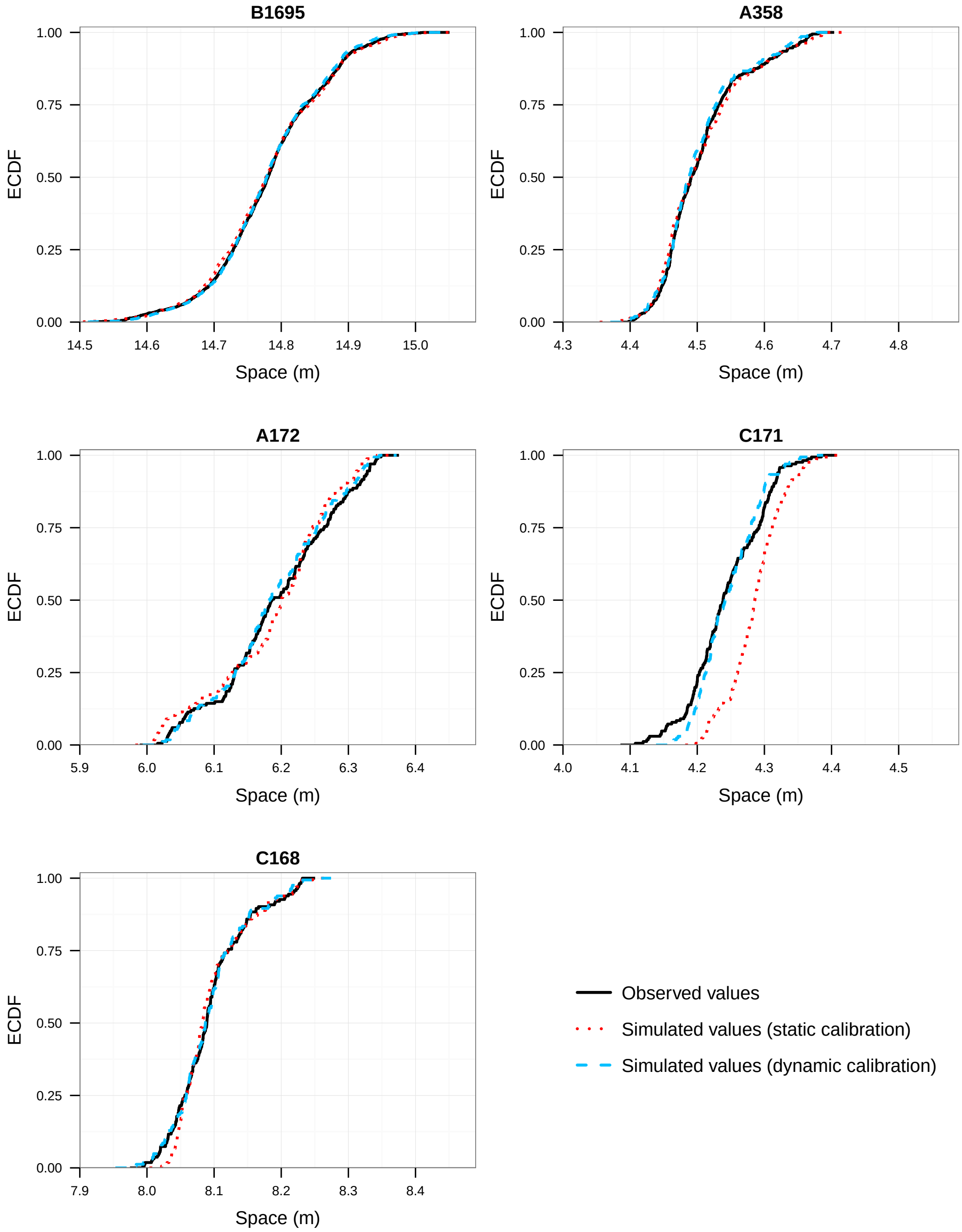


Figure 5.7: ECDFs of observed and simulated space gaps for one–step prediction
Note: Different ranges, but same scale, have been used for the x axis in all sub-figures

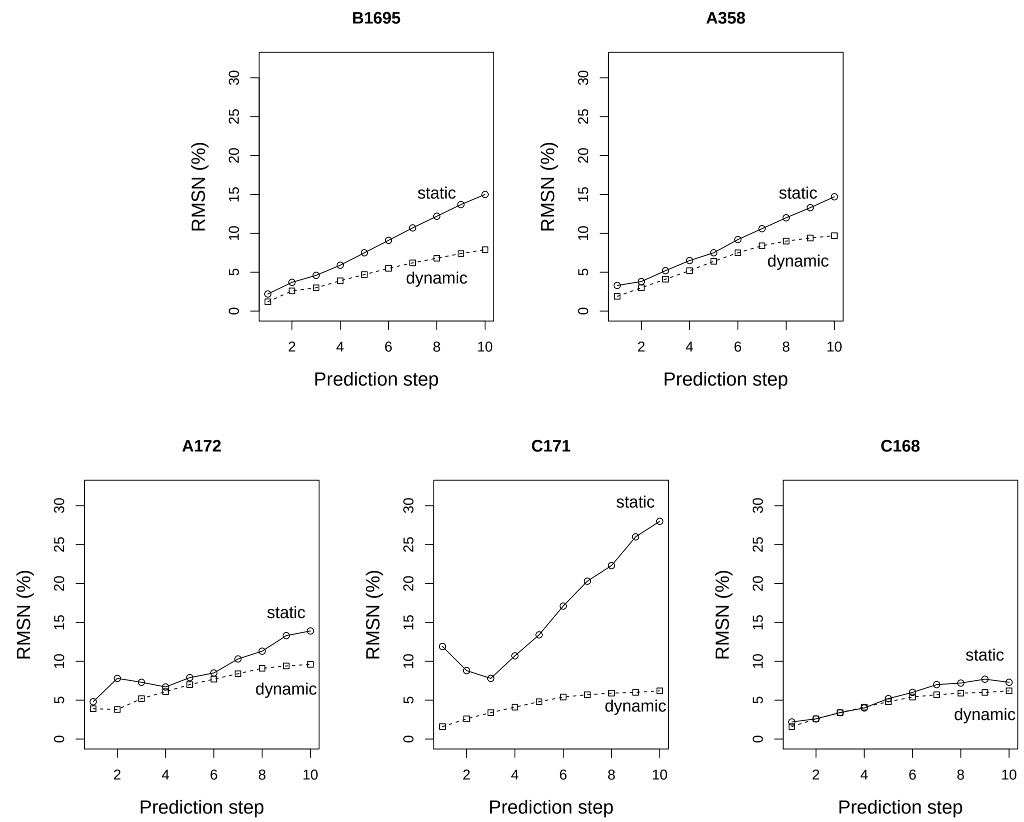
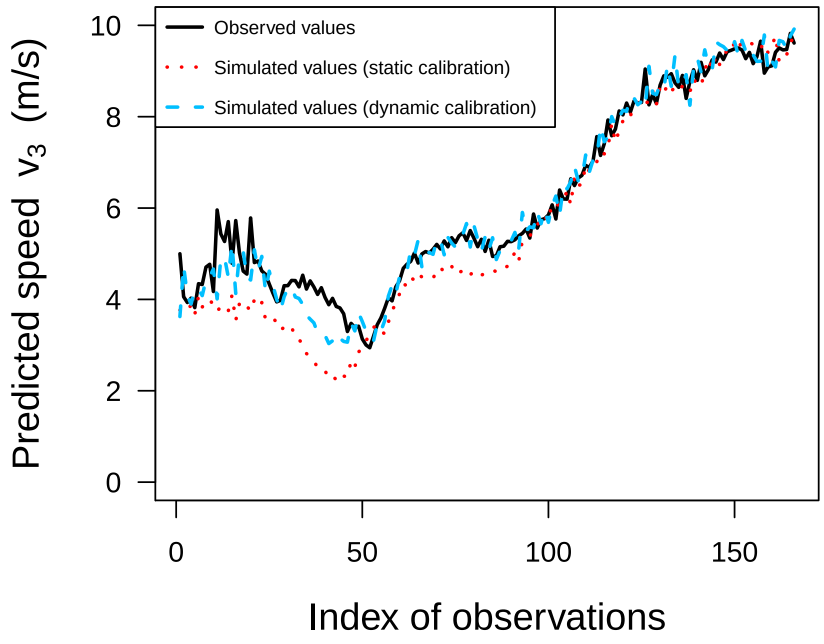


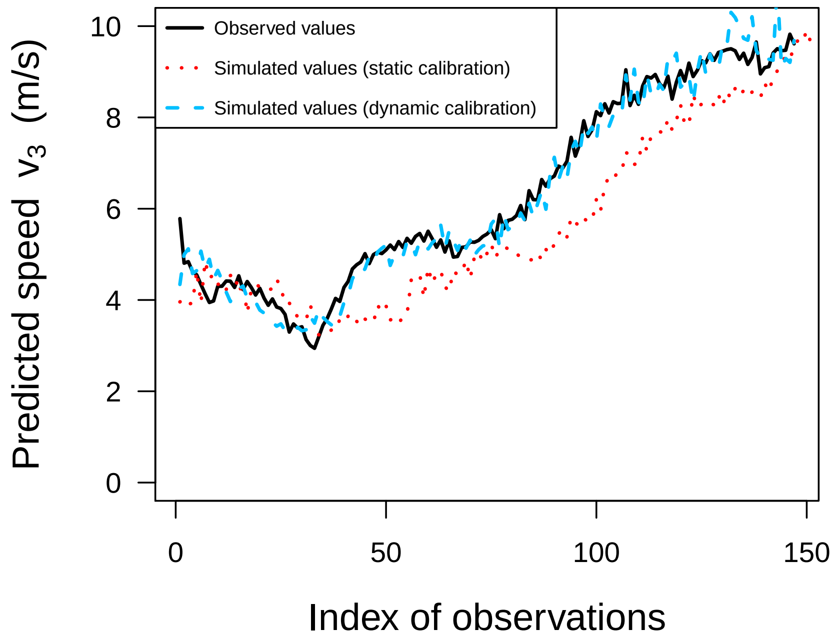
Figure 5.8: Multiple time steps prediction using static and dynamic calibration

obtained parameters for the considered model and data–set. It is noted that, since for each time instant only one of the two equations is critical, the parameter values for that equation are considered at each time point. For example, if the top equation of Equation 2.3 is active at the particular point, then the parameters a_{nt} and V_{nt} are used (and therefore these values contribute to the densities in Figure 5.10), while if the bottom equation is active then the values for the parameters b_{nt}, s_{nt} and $\hat{b}_t$ are considered.

In each figure, the value of the static calibration is also indicated with a vertical dashed line. It becomes apparent that the dynamic values are not distributed symmetrically around the statically obtained value. This could have several implications. One question could be whether the static calibration is not really optimal. To check for this, we repeated the estimation and prediction using constant parameter values; however, this time, instead of using the value obtained from the static calibration, we used the median from the densities obtained from the dynamic calibration (i.e. the distributions shown in Figure 5.10). In that case, the estimation RMSN ended up actually being inferior to that obtained from the static calibration results (with an RMSN of 3.4% instead of 2.2%). Therefore, it seems that there is something different going on, and that indeed the median



(a) Step 1



(b) Step 5

Figure 5.9: Predicted speed v_3 for data series C171 for steps 1 and 5

values from the distributions cannot be used as best values for the determination of constant/average values. The explanation for this may come from the nature of the Gipps model, i.e. the fact that there are two different equations, and at each given time the parameters of a single one are in effect considered.

Therefore, while during the dynamic calibration the model steers only these parameters towards their desired values, using the available information, in the case of a static calibration one needs to determine single values that are relevant for all observations.

Another concern could, of course, be that the dynamic model is actually overfitting. However, the fact that it outperforms the static model even in ten–step predictions (as shown in Figure 5.8) provides compelling evidence that this is not the case.

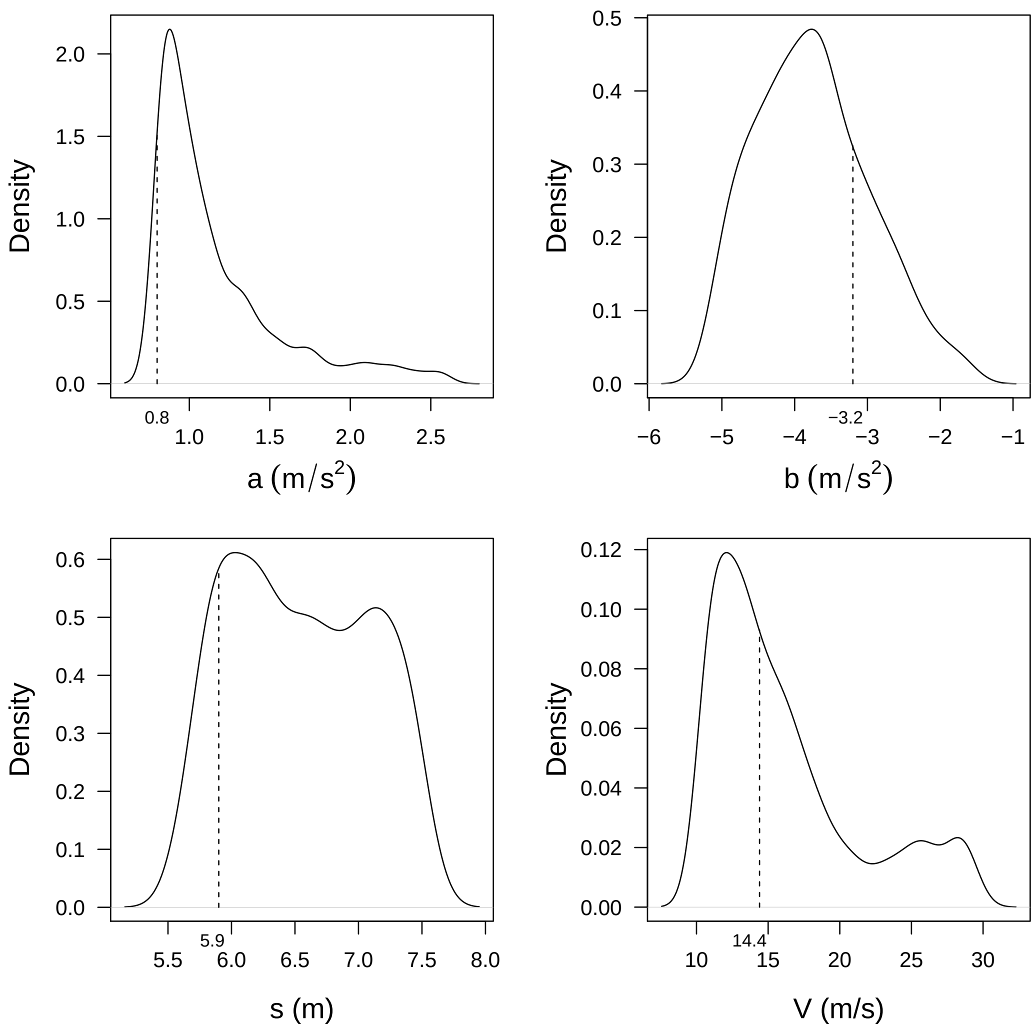


Figure 5.10: Distributions of dynamically calibrated parameters
Notes: (i) dashed line: static calibration value, (ii) for each subfigure, only values from instances, in which the corresponding equation is critical, are used

5.5 Conclusions

The findings of this research suggest that dynamic calibration for microscopic traffic models could be promising and should be further studied. In this research, the prediction of the dynamic parameters was simple, in the sense that the dynamically calibrated parameters were assumed as the best available estimate for the short–term values of these parameters. Further research could consider secondary models that would actually aim at predicting the evolution of these parameters, as well, e.g. via autoregressive, polynomial or other statistical forecasting specifications.

Furthermore, the procedure could be applied to other car–following models. Regarding the optimization process, different optimization algorithms could be tested, besides the ISRES optimization algorithm, which has been selected for the case study presented in this research. In addition, the results of this research should be compared with those obtained by a multiobjective optimization process in a future research. While in this research the optimization of the car–following model parameters has been performed using an optimization algorithm across all parameters concurrently, further exploration of the correlation of the model parameters (Kim and Mahmassani, 2011) could provide deeper insight into the problem. The heterogeneity of the drivers and their behavior (Ossen and Hoogendoorn, 2011) could also be further explored, using a larger number of characteristics trajectories from a larger sample of drivers.

While in this case study we only use a small number of vehicles, for which data are available, it is practical to apply this methodology to all individual vehicles in a network, in real time. The computational and data requirements are such, that allow the application of the methodology to each individual vehicle (as the required data could be obtained e.g. from a GPS trace of the vehicle, and e.g. radars/cameras providing information about the vehicles around it). Furthermore, the computational overhead to apply the methodology is minimal and it can be dealt by in-vehicle processing facilities, in a decentralized way. This would allow to not only obtain different parameters estimated and predicted per each vehicle class, but indeed for each individual vehicle, in real time.

In the realistic situation that not all vehicles have the ability to collect/receive the required data, an extrapolation could be used to generalize the obtained results. In this case, it could be practical to identify classes of vehicles, estimate and predict these parameters for the vehicles comprising the sample for each class, and then extrapolate this information to the entire population of vehicles of this class in the studied area. Besides temporal variability of these parameters, by class, one can of course foresee a spatial distribution, as traffic conditions, road characteristics, fleet mix, and other parameters could influence their value.

In the case study, presented in this research, we use a single measure of effectiveness (speed). The failure of the dynamic calibration in the low–speed situation in Figure 5.3 could be overcome by considering additional measures of effectiveness (e.g. vehicle gap). Furthermore, the data set used in this research has intentionally considered the movement of vehicles in a single lane, avoiding the complications offered by lane–changing opportunities. Richer data–sets, offering also lane-changing opportunities, would allow the extension of this research to more elaborate models, such as joint car–following/lane–changing models.

Car–following models are also used as input to Adaptive Cruise Control systems, which also need to be adaptive and consider the heterogeneity of the driver behavior (Bifulco et al., 2013a,b). The work presented in this research could be leveraged for the improvement of such models.

In conclusion, as we move towards more detailed models, even for online applications, it is expected that dynamic calibration will play an increasing role for this type of models. While a lot of experience exists in the online calibration of macroscopic and mesoscopic models, it is likely that this expertise will not transfer directly to the more detailed microscopic models. Therefore, novel research is needed, to validate the existing techniques and develop new ones, specifically suited to leverage the benefits and unique characteristics of microscopic traffic simulation models.

Chapter 6

Flexible microscopic models in mixed traffic conditions

6.1 Problem statement

Intelligent transportation systems require detailed car-following models that could represent driving behavior in an efficient way. Moving towards autonomous vehicles, models should be able to reflect heterogeneity in driving behaviour and traffic networks. While many driving behavior models have been developed over the years, there are still aspects that remain unsolved. Most existing studies focus on driving behavior using trajectory data of vehicles moving in lanes. However, modeling driving behavior in mixed traffic streams is still a challenge.

A heterogeneous mixture of vehicle types and a violation of lane discipline are common characteristics in cities in developing countries. These conditions lead to driving manoeuvres that combine both longitudinal and lateral movements. Modeling this driving behavior tends to be complex and cumbersome, as various phenomena, such as multiple–leader following, should be addressed. Traffic flow in the developing countries, as well as in urban road networks mainly in South European countries, is very complex in nature and safety issues arise. This research attempts to simplify mixed traffic modeling by developing a methodology which is based on data–driven models. The focus is given on recognizing leader–follower pairs and identifying significant lateral changes that may be indicative of the traffic situation of a vehicle (car–following, lane–changing or free flow). The objective of this research is to develop a dynamic methodology that could describe mixed traffic in a more realistic way using different virtual lanes for each driver which are constantly modified according to the traffic conditions. The proposed methodology is validated on mixed traffic trajectory data, which have been collected in India.

[0]The sub–chapters 6.1, 6.2.1 and 6.3 are based on Papathanasopoulou and Antoniou (2017), while the sub–chapter 6.2.2 is based on Papathanasopoulou and Antoniou (2018).

6.2 Non–lane based approach

It is assumed that all vehicles are moving without lane discipline. In order to simplify this traffic situation, temporary virtual lanes for each vehicle are defined. The methodology is based on the idea that each driver follows his own temporary virtual traffic lane until his lane overlaps with the virtual lane of another driver and thus he is forced to modify it. The proposed methodological approach is outlined in Figure 6.3. Longitudinal and lateral positions are recorded per time instant and saved in a database. Then significant lateral changes are identified using appropriate algorithms that allow monitoring structural changes in linear regression models. If no significant lateral change is identified then lateral information is used for determination of a temporary virtual lane and then a car-following model or a free flow model is applied if at least one preceding vehicle is identified or not respectively. For identification of the front vehicle more details are provided in the next subsection. On the other hand, if a breakpoint is observed in data sequence, namely if significant lateral changes are identified, then a lane-changing situation is indicated and the virtual lane needs to be modified. A lane–changing model should be applied for time t_L, time of lane-changing duration. Then the process is iterated for the following time instants.

6.2.1 Identification of lead and lag vehicle

Since multiple leader vehicles may be present in heterogeneous traffic conditions and thus the critical leader vehicle should be identified. The probability of a given front vehicle to be the governing leader depends on the type of the lead vehicle and the extent of lateral overlap with the following vehicle Choudhury and Islam (2016).

In order to apply a microscopic model, it should be determined whether there is a vehicle pair of follower–leader. The main characteristic of mixed traffic is that the size of overlap between the leader and the follower varies. Assuming that the lateral and longitudinal coordinates of the front center of each vehicle (x_{c_i}, x_{c_i}) are known, it could be defined which vehicle follows the other. The coordinates for the left and the right lateral bound of each vehicle are estimated per time instant t by Equations 6.1 and 6.2 (as shown in Figure 6.1).

$$x_{l_i}(t) = x_{c_i}(t) - \frac{w_i}{2} - s_i(t) \tag{6.1}$$

$$x_{r_i}(t) = x_{c_i}(t) + \frac{w_i}{2} - s_i(t) \tag{6.2}$$

where i: 0,1,2,...n vehicle index x_{c_i}: lateral coordinate of the front center of vehicle i, x_{l_i}: lateral coordinate of the front left bound of vehicle i, x_{r_i}: lateral coordinate of the front right bound of vehicle i, w_i: width of vehicle i s_i: a lateral safety distance for vehicle i

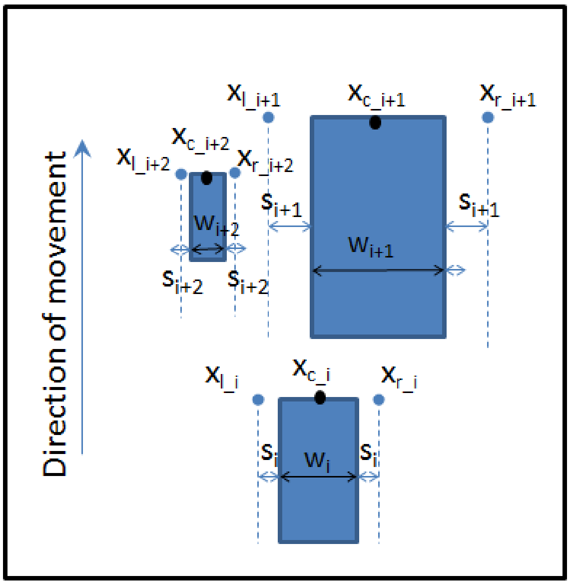


Figure 6.1: Estimation of lateral coordinates

In order to define the car–following vehicle pairs, the longitudinal position of the leader should be in front of the following vehicle and in a distance L that could influence the movement of the following vehicle (Equation 6.3). In addition, a part of the front side of a vehicle should overlap a part of the front side of another vehicle (Equation 6.4). This overlap is evident in Figure 6.2 with light blue color. Each vehicle i is considered as follower and then a leader vehicle is required to fulfill the conditions, described by Equations 6.3 and 6.4, at the same instant t:

$$y_{follower}(t) \leq y_{leader}(t) \leq y_{follower}(t) + L \tag{6.3}$$

$$\begin{aligned} x_{l_{follower}}(t) \leq x_{r_{leader}}(t) \\ x_{l_{leader}}(t) \leq x_{r_{follower}}(t) \end{aligned} \tag{6.4}$$

Four cases of vehicle pair leader– follower have been identified, as shown in Figure 6.2. Furthermore, a scenario with two leaders and one follower case is also possible. For instance, a bus could be the follower and a part of its front side may overlap with two leaders such as two motorcycles or a small vehicle and a motorcycle. In this case the closest vehicle according to the direction of movement is chosen as the most critical leader. If no vehicles are identified as leaders, then the driving situation of the vehicle is free flow.

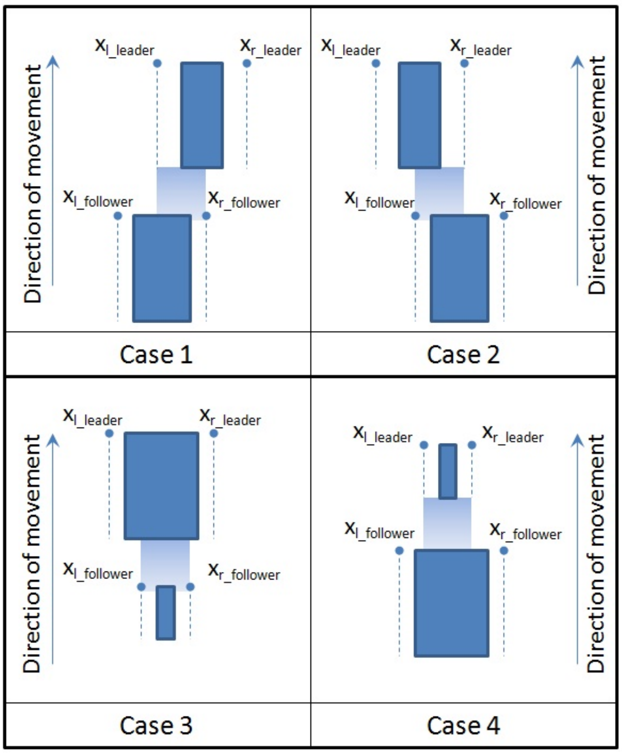


Figure 6.2: Identification of pair leader- follower

6.2.2 Determination of virtual lane changes

A typical example of modification of virtual lane change is illustrated in Figure 6.4. In this figure, there are two vehicles. The first vehicle follows the virtual lane i. While there are small lateral movements, it is considered that it does not change lane. However, when its movement is constrained by the hatched vehicle at the breakpoint, it is considered that it changes lane and then follows virtual lane i+1. The challenge is that vehicles are moving constantly laterally. This could be addressed in two distinct ways. The first one is to estimate the threshold that indicates a lane change. The second one is using change detection algorithms. In this research the focus is given on the second approach, namely on identifying significant changes in lateral positions, so as the appropriate microscopic model to be applied. Algorithms that are capable of finding major changes in data sequence could be used.

Heterogeneity in vehicle types implies various widths of vehicles and thus various widths of virtual lanes. The width of a temporary virtual lane W could be estimated by Equation 6.5, if no significant lateral changes and breakpoints are identified. The estimation of temporary virtual lanes is also illustrated in 6.5.

$$\begin{aligned} W =& max(x_t, x_{t+1} + \cdots + x_{t+n}) - \\ & min(x_t, x_{t+1} + \cdots + x_{t+n}) + w_v \end{aligned} \tag{6.5}$$

where x_t is the position of the center of the vehicle, measured from the left–most side of the roadway for each time instant t+i and w_v is the width of the vehicle.

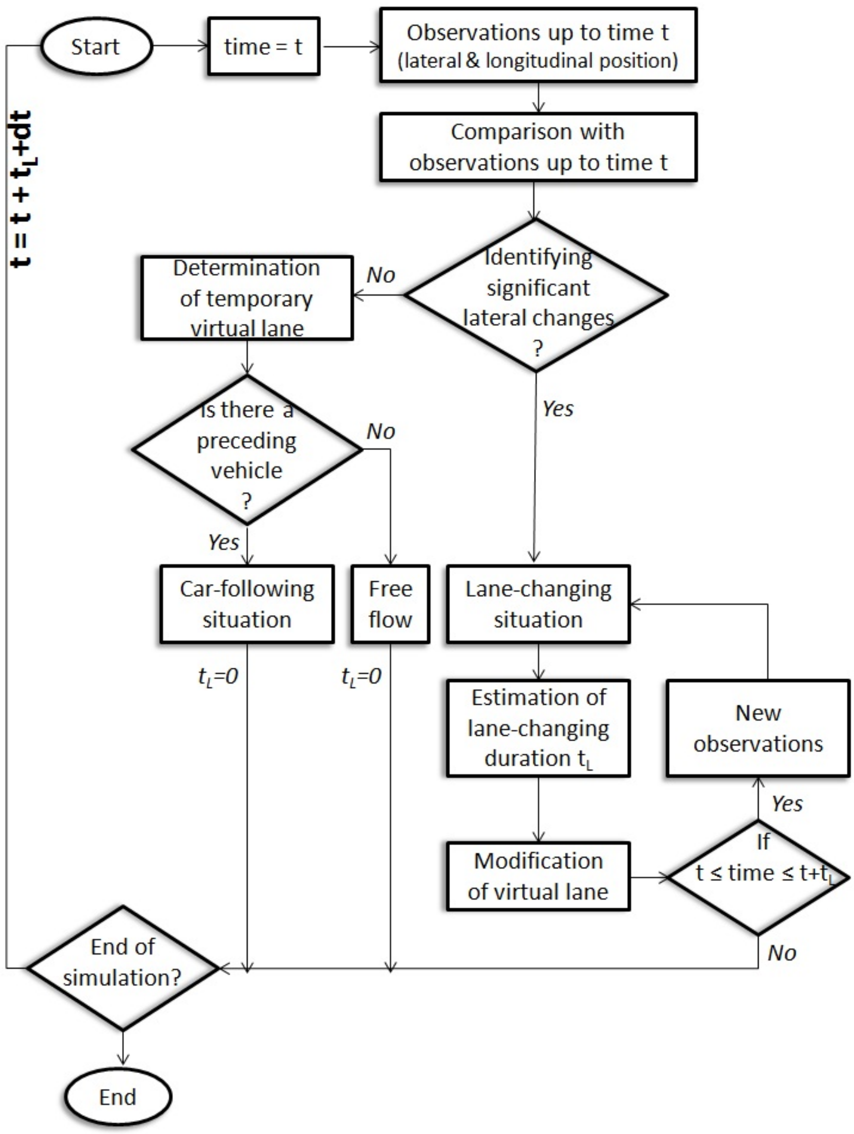


Figure 6.3: Methodological approach for non–lane based traffic

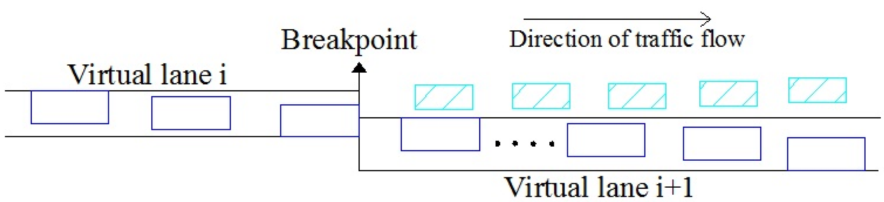


Figure 6.4: Virtual lanes

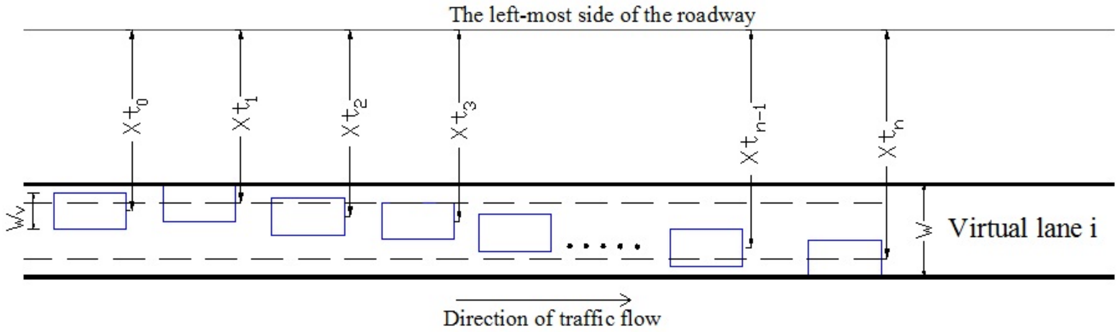


Figure 6.5: Estimation of virtual lane width

6.3 Car–following application

The next step is the fitting of the proposed methodology for car–following situations using data–driven models. The problem to be addressed is the speed estimation of each vehicle, when the available data include its speed, the speed of the preceding vehicle and the distance between the two vehicles (in the previous time instant). Locally weighted regression could be used for the application. In the training step the flexible car-following model is fitted or calibrated on the surveillance data and validated on the other dataset. If no vehicles are identified as leaders, then these observations are omitted, as they do not correspond to car-following state.

6.3.1 Data processing

First, data were organized in ascending order of vehicle ID, so as the trajectory of each vehicle to be continuous and observations of other vehicles not to interfere. Then, only observations appropriate for microscopic analysis are selected (flag=0). As coordinates of the front center of each vehicle longitudinal and lateral positions are used. Regarding the considered speed for each vehicle, the resultant speed is estimated by Equation 6.6.

$$v_i(t) = \sqrt{{v_{long_i}}^2 + {v_{lat_i}}^2} \tag{6.6}$$

where v_i: resultant speed of vehicle i, v_{long_i}: longitudinal speed of vehicle i and v_{lat_i}: lateral speed of vehicle i.

In addition, a new column is added which includes the observed speed for the next time instant, namely the speed that should be predicted for each observation. Actually this is the speed that corresponds to time t + 0.5 s and to the same vehicle ID. If there is no observation for this vehicle and for the next time instant, NA is given. Afterwards, rows with NA in this column are omitted, as there is no observed speed to compare with the estimated one by the proposed methodology.

Due to the mature of mixed traffic data the next step was to define the car-following sequence, namely which vehicle is in front of the other. Kanagaraj et al. (2015) have identified that in 45% of the observations the overlap between the leader and the follower is less than half the follower width. The methodology described in section was adopted for the identification of the front vehicle. Observations that correspond to vehicles with no leading vehicle were excluded. As lateral safety distance, s=0.20m is considered for each vehicle on both sides. As distance L in Equation 6.3, L=200m is considered. The same procedure was also used with the validation on dataset data300. Finally, dataset "data245" includes 47036 observations corresponding to 1511 vehicle pairs and dataset "data300" 45982 observations corresponding to 1488 vehicle pairs.

6.3.2 Estimation of conventional models

A conventional car–following model, the Gipps model (Gipps, 1981), is used as reference in order to monitor and evaluate the effectiveness of the proposed method. It is a well–known model that is used in AIMSUN traffic simulator. This model requires as input the same data as the proposed method and thus a direct comparison would be feasible. First, a calibration of model parameters is required. There are six parameters in this model that have to be calibrated. The apparent reaction time is considered as 0.5 s and for calibration of the rest of parameters an optimization process is implemented. Dataset "data245" was used for calibration and "data300" for validation. The calibration process was performed within the R software for statistical computing (R Core Team, 2017). In particular, the Improved Stochastic Ranking Evolution Strategy (ISRES) algorithm was used, which is included in the package "nloptr" (Runarsson and Yao, 2005) and is appropriate for nonlinearly constrained global optimization. This method is implemented in a simple way and supports arbitrary nonlinear inequality and equality constraints in addition to the bound constraints. In addition, it incorporates heuristics to escape local optima. The objective function that was minimized is the RMSN between the observed and simulated values of speeds:

$$RMSN(v^{obs}_{follower}, v^{sim}_{follower}) \tag{6.7}$$

Bounds and initial values for model parameters have been defined in Figures 4.1 and 4.2. These initial values have been defined as optimal values for data with lane discipline by algorithm ISRES in that research. Thus, it is expected that there will be a differentiation in optimal values due to different nature of data. Three samples of 5000 observations were selected randomly from dataset "data245". The amount of observations used in each sample are summarized in Table 6.1 per vehicle type. A representative amount for each vehicle type is included in each sample. The optimization process was implemented for each

Table 6.1: Observations per vehicle type used for calibration of each data sample

a/a	Vehicle type	Sample 1	Sample 2	Sample 3
1	motorcycle	2665	2701	2626
2	car	1347	1292	1347
3	bus	145	156	156
4	truck	41	29	15
5	light commercial vehicle	56	59	78
6	auto–rickshaw	746	763	778

Table 6.2: Optimized parameters values for dataset "data245" using ISRES algorithm

Parameters	Initial values	Constraints min	Constraints max	Optimized values for sample 1	Optimized values for sample 2	Optimized values for sample 3	Mean parameter values
$a(m/s^2)$	0.8	0.8	-2.6	0.81	0.82	0.82	0.82
$b(m/s^2)$	-3.2	-5.2	-1.6	-5.18	-5.17	-5.08	-5.14
$V(m/s)$	14.4	10.4	29.6	10.45	10.44	10.44	10.44
$s(m)$	5.9	5.6	7.5	5.62	5.60	5.60	5.61
$\hat{b}(m/s^2)$	-3.1	-4.5	-3.0	-3.01	-3.01	-3.00	-3.01
RSMN	-	-	-	0.21	0.22	0.21	-

sample separately and the results are presented in Table 6.2. For these samples the optimization process has converged to the optimal set of parameters after approximately 10000 iterations. For all samples similar parameter values have been produced and thus the optimization process for the whole dataset is considered unnecessary. Instead, the mean of the three optimized sets of parameters is selected and is presented in the last column of Table 6.2. Furthermore, the authors explored the impact of different initial values and the algorithm converged to the same solution, suggesting robustness of the optimization process. Looking into initial values that were appropriate for traffic under normal conditions and optimized values optimized for mixed traffic conditions, the main difference is observed in maximum braking b that the driver of vehicle wishes to apply in order to avoid a crash. This could be attributed to the fact that more abrupt driving is observed in a mixed traffic environment. The minimum value of the objective function, namely the RMSN that was achieved with these optimal values of parameters was 21%. Then, the calibrated model is validated on dataset "data300" and RMSN is estimated between observed and predicted speed per time instant. The results are shown in Figure 6.6 and a comparison with the proposed method is feasible.

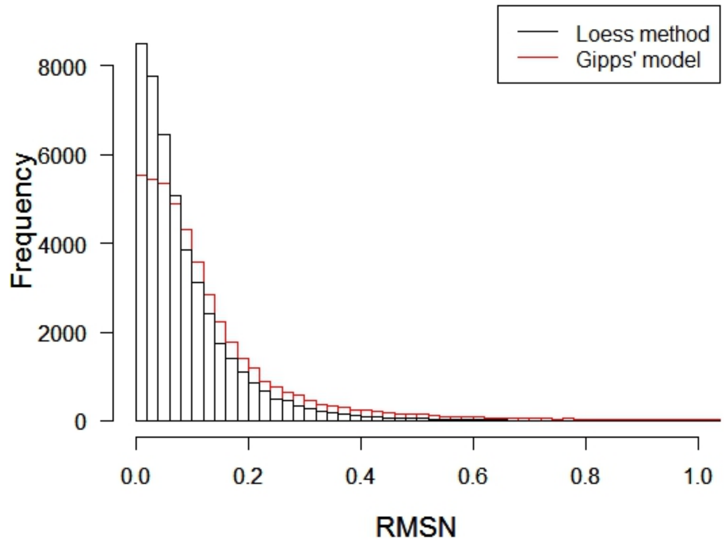


Figure 6.6: Histograms of RMSN using loess method and Gipps' model for dataset "data300"

6.3.3 Exploration of data–driven models

The proposed method identifies the relationships between predictor variables $v_{leader}(t)$, $v_{follower}(t)$, the distance $D(t)$ between the two vehicles and the response data $v_{follower(t+\tau)}$, where τ=0.5 s. After the relevant pattern from "data245" data series has been identified, the proposed method is applied to "data300" data series. It requires the input data ($v_{leader}(t)$, $v_{follower}(t)$ and distance $D(t)$) and exports the estimated $v_{follower}(t + 0.5)$. The RMSN values have been estimated per time instant t in order to compare predicted and observed speed values and estimate the performance of this methodological approach. The validation results are presented in Figures 6.7, 6.8 and 6.9.

In Figure 6.6, the proposed method outperforms Gipps' model and produces a more reliable speed prediction. In Figure 6.7, distance between the follower and the leader is plotted versus residuals produced by loess method. Residuals are estimated as the difference between observed and predicted speeds. Higher values of residuals are observed for smaller distances. Then, in Figures 6.8a and 6.8b, an analysis of the results per vehicle type is attempted. Figure 6.8a shows the density of RMSN per vehicle type. The best performance of loess method is achieved for cars and light commercial vehicles, while higher RMSN are observed for other vehicle types, especially for trucks and auto–rickshaws. In Figure 6.8b densities of RMSN are outlined per vehicle type of the leader when the follower is a car. Vehicles pairs car– car and motorcycle– car (leader– follower) have a peak of density curve lower than RMSN=0.1. The vehicle type plays a significant role in accuracy of the model. This could be attributed to different driving behavior in

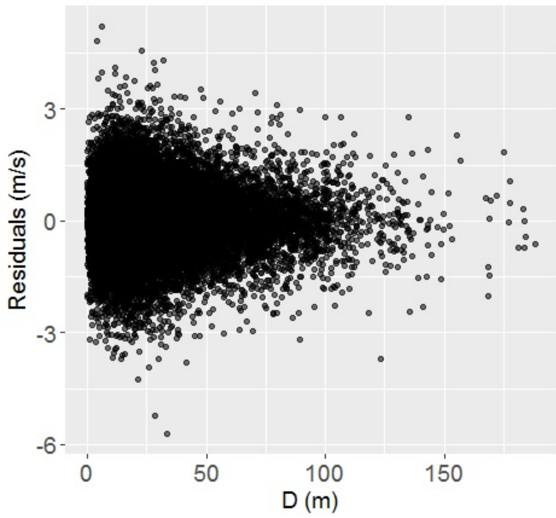
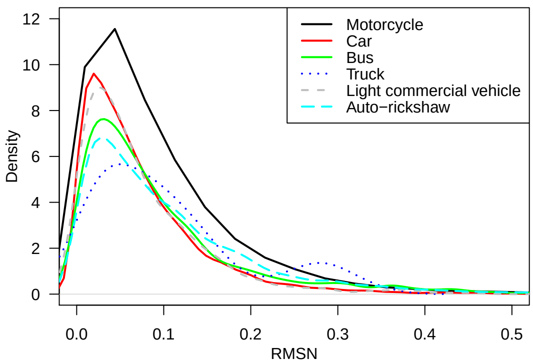
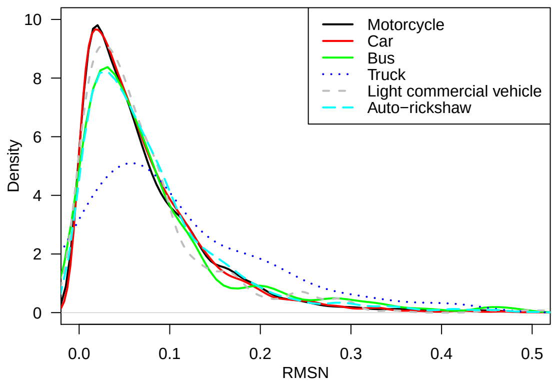


Figure 6.7: Distance between the follower and the leader versus residuals produced by loess method for dataset “data300”



(a)



(b)

Figure 6.8: (a) Density of RMSN per vehicle type for dataset “data300”, (b) Density plot of RMSN per vehicle type of the leading vehicle when the follower is a car

different vehicle types, as well as varying vehicle kinematics. The density curve of vehicle pair truck– car corresponds to higher RMSN than the other vehicle pairs. Finally, in Figure 6.9 observed speeds are plotted versus predicted speeds per vehicle type. Linearity is evident for all vehicle types.

Finally, a visual inspection is attempted for dependent and independent variables used in the proposed method. In Figure 6.10, variables and their correlations are outlined. In the diagonal line, histograms of the variables are presented. In the upper triangle, values of correlations have been estimated for each potential pair of variables, while scatterplots of variables are displayed in the lower triangle. Looking at the first column of Figure 6.10, the lower speeds are, the higher RMSN is. In addition, a negative relationship is observed for distances, too.

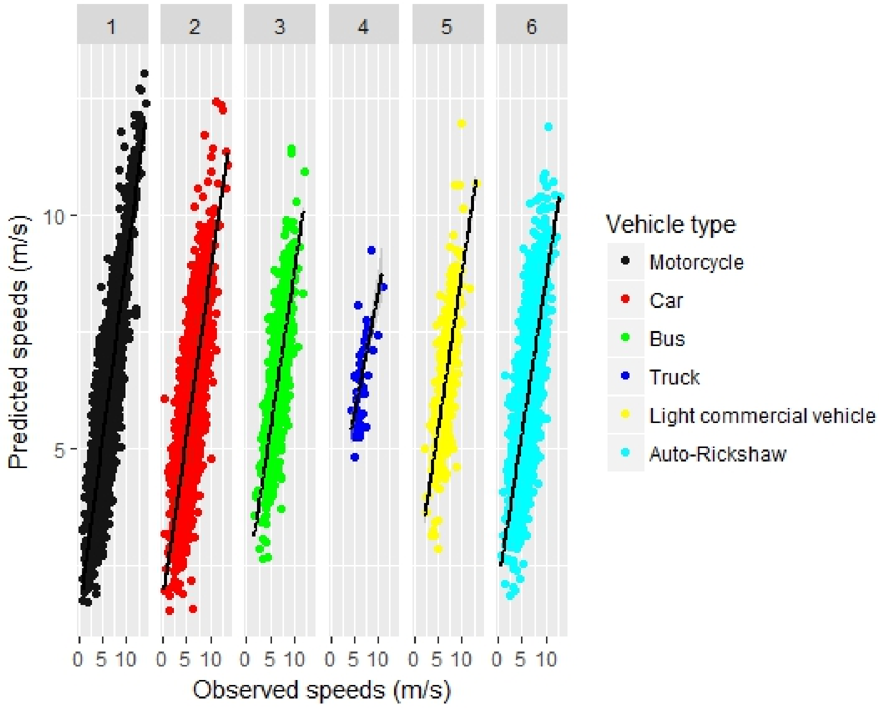


Figure 6.9: Observed speeds versus predicted speeds by loess method

6.4 Conclusions

Data driven approaches could be a promising tool for modeling mixed traffic. They lead to flexible car-following models and thus to more robust and reliable representation of driving behavior. In this research, an existing methodology for estimation of car following models has been validated on mixed traffic trajectory data. This simple methodological approach outperforms the reference (Gipps') model for the available data. Data-driven estimation techniques are designed to address cases in which the traditional approaches do not perform well or cannot be effectively applied without including undue labor. The data-driven approach presented in this research is based on a non-parametric method, locally weighted regression (loess), which might be considered as a generalization of multi-regime approaches (Antoniou and Koutsopoulos, 2006b; Antoniou et al., 2013). There are also other data-driven methods such as neural networks (Vlahogianni et al., 2005b; van Lint et al., 2005) and kernel methods offering similar capabilities. Karlaftis and Vlahogianni (2011) provide an interesting discussion of such methods against statistical methods. However, locally weighted regression comprises much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. Deep learning networks and Gaussian processes could be also explored as future prospects.

Models developed for lane–based traffic conditions may not be appropriate to

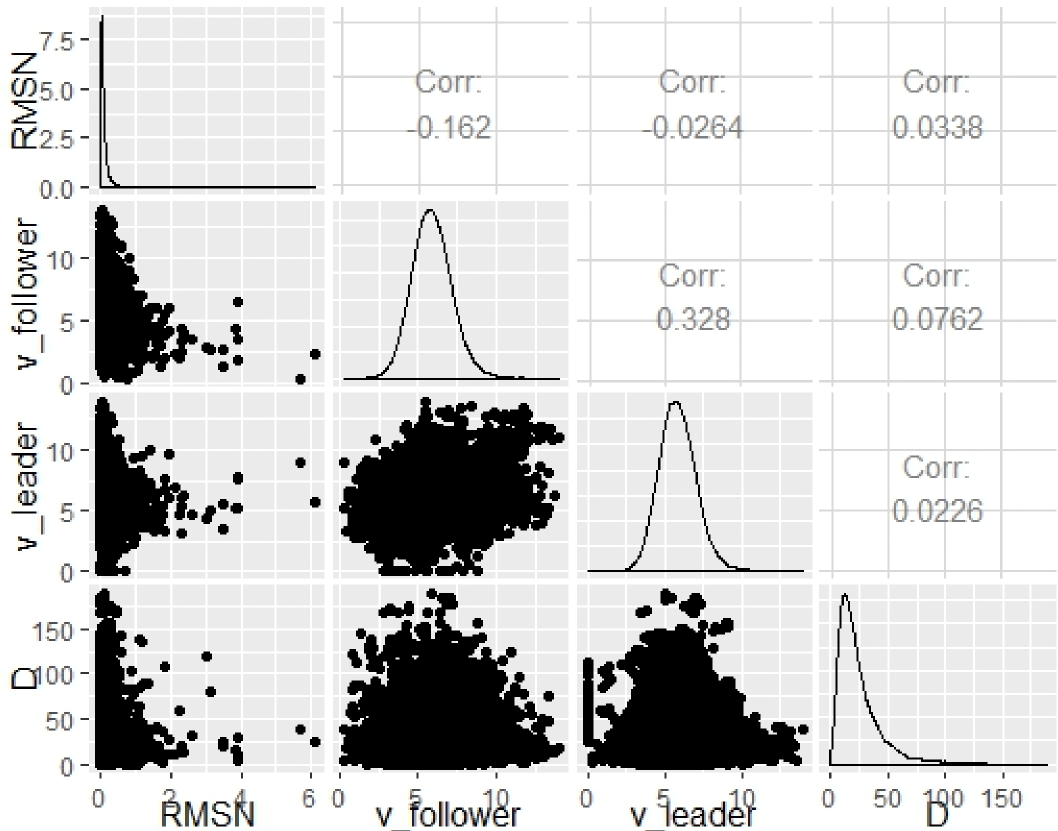


Figure 6.10: Visual inspections of model variables including scatterplots, histograms and correlation values

simulate traffic situations in developing countries, where weak lane discipline is often observed. Traffic in developing countries is so heterogeneous that often lane–based models cannot be realistic. To overcome some of the associated limitations, in this research a methodology is proposed using temporary virtual lanes. An algorithm for the identification of significant lateral changes has been applied and the feasibility of the method has been explored. As future prospects, there are some components in the proposed methodology that require further analysis. The estimation of lane–change duration is one of these. In addition, the proposed methodology could be employed with flexible data–driven models Papathanasopoulou and Antoniou (2017), which allow incorporation of further variables moving towards an integrated solution for the simulation of mixed traffic. For instance, vehicle–dependent models need to be developed in case of heterogeneous traffic, as the drivers of vehicles with unequal dimensions tend to have different driving behaviors; furthermore, different vehicle types are characterized by varying vehicle kinematics. Thus, it is expected that further exploration of data–driven approaches could open up opportunities to understand and simulate driving behavior in non–lane discipline conditions with heterogeneity of vehicle types. Additional research should be performed to determine the factors that determine its performance, too. The proposed methodological framework is flexible, less time-consuming and allows the incorporation of additional parameters that may influence driving behavior (such as density of the surrounding area, vehicle type, drivers' age, road infrastructure etc.). Vehicle type seems to play a significant role in speed estimation and should be incorporated as a categorical variable in the process. Resorting cumbersome reformulations of a fixed model form could be impractical. However, conventional models such as Gipps' model may provide better insight into traffic flow theory. The integration of data-driven methods in advanced driver assistance systems under mixed traffic conditions could be very interesting, though additional research should be conducted.

Chapter 7

Discussion and conclusion

7.1 Conclusions

In this research a data–driven methodological approach has been developed and successfully demonstrated on experimental data, in order to overcome some of the associated limitations of conventional microscopic models. The proposed approach is suitable for incorporation into microscopic traffic simulation models. Data driven approaches could be a promising tool for optimization of microscopic models, as it may lead to more robust and reliable representation of driving behavior. The proposed approach outperforms the reference (Gipps') model and could be an innovative perspective for estimation of microscopic data–driven models. Flexible models are derived from causal inference and allow the incorporation of additional predictor variables, while cumbersome reformulations of a fixed model form could be impractical. However, conventional models such as Gipps' model may provide better insight into traffic flow theory.

On the other hand, model parameters vary in multiple dimensions, including across individual drivers, spatially and temporally. The computational and data requirements in the proposed methodology are such, that allow the application of the methodology to each individual vehicle (as the required data could be obtained e.g. from a GPS trace of the vehicle, and e.g. radars/cameras providing information about the vehicles around it). Furthermore, the computational overhead to apply the methodology is minimal and it can be dealt by in-vehicle processing facilities, in a decentralized way; This is very important in the context of autonomous vehicles. The proposed methodology allows to have not only different parameters estimated and predicted for each vehicle class, but indeed for each individual vehicle, in real time. The extrapolation will be needed in the case of only having a sample of vehicles with the ability to collect/receive the required data. In this case, it could be practical to identify classes of vehicles, estimate and predict these parameters for the vehicles comprising the sample for each class, and then extrapolate this information to the entire population of vehicles of this class in the studied area. Besides temporal variability of these parameters, by class, one can of course foresee a spatial distribution, as traffic conditions, road characteristics, fleet mix, and other parameters could influence their value. Microscopic models are also used as input to Adaptive Cruise Control Systems, which also need to be adaptive and consider the heterogeneity of the driver behavior. The work presented in this research could be leveraged for the improvement of such models.

On–line calibration of microscopic traffic simulation models for dynamic multi–step prediction outperforms the static calibrated models and less than 10% error in speed prediction even for ten steps into the future. As we move towards more detailed models, even for online applications, it is expected that dynamic calibration will play an increasing role for this type of models.

Furthermore, models developed for lane-based traffic conditions may not be appropriate to simulate traffic situations in developing countries, where weak lane discipline is often observed. Traffic in the developing world is so heterogeneous that often lane-based models cannot be realistic. In this research an alternative methodology based on temporary virtual lanes is proposed. The integration of data-driven methods in advanced driver assistance systems under mixed traffic conditions could be very helpful, though additional research should be conducted.

Data driven estimation of microscopic models appears to be a promising tool that could offer considerable benefits if integrated into traffic simulation models, resulting in higher accuracy and reliability of model outputs.

7.2 Research Contributions

Modeling driving behavior plays a fundamental role in traffic management, safety research and the development of Intelligent Transportation Systems. This research makes several contributions to the state–of–the–art of microscopic traffic simulation:

- A methodological framework based on non–parametric approaches has been developed for simulation of driving behavior. Microscopic conventional models represent individual vehicles and their interactions and are capable of simulating traffic to a high level of detail, but they do require a long execution time, as their successful application depends on the effectiveness of calibration process. On the other hand, the proposed methodology offers great flexibility and there is no need for time consuming calibration process. Data–driven models result in better fit, comparing to the traditional models, and thus could be a plausible substitute for theory–based models.

- Computational approaches allow the easy incorporation of additional parameters. Conventional models do not allow the easy incorporation of additional variables without labor undue, since cumbersome reformulations of functional form should be performed. Data–driven microscopic models have been proposed in this research as a way to overcome these limitations and capture driving behavior in an efficient way taking into account various variables. Additional variables, such as traffic density, have been incorporated into the proposed models and more detailed models have been developed.

- The use of various machine learning techniques for estimation of microscopic models is explored. The question of which machine learning technique could be the most appropriate one for traffic simulation models has not been answered conclusively. This research provides some more input into this ongoing active research field.
- The impact of various predictor variables on the models is investigated. A meta–model is developed to evaluate the magnitude of the effect of the considered predictor variables on driving behavior.
- Data–driven models are validated on non-disciplinary trajectory data with heterogeneous mixture of vehicle types and are proved to be a promising perspective for microscopic traffic simulation in the developing world, where these conditions occur in a common basis.
- Data–driven models and on–line calibration of microscopic models provide a robust solution to autonomous driving. Aiming at safety, reliability and convenience, an autonomous vehicle should adapt to user preferences and simulate human driving reactions naturally, preventing abrupt acceleration and jerk. Undoubtedly, in this context, this research contributes significantly into learning driving styles and realizing autonomous driving.
- Data–driven estimation of microscopic models appears to be a promising tool that could offer considerable benefits if integrated into microscopic traffic simulators, resulting in higher accuracy and reliability of model outputs.

Conclusively, better representation of driving behavior contributes substantially into the development of Intelligent Transportation Systems, which are directly related to the concept of smart cities. This contribution could be translated to sustainable transportation solutions, reduced costs in terms of safety, time, money, energy and environmental impact, and by extension to benefits of the humanity.

7.3 Directions for future research

Directions for further research are outlined in this section.

- **Data–driven simulators**

  Data–driven microscopic models could be integrated in a traffic microsimulator to be used for real–time applications. The results presented in the research provide clear evidence that data–driven traffic approaches have the potential to contribute to improved modeling capabilities, in light of new data and emerging simulation needs. A network–wide validation using a microscopic traffic simulator would offer a flexible environment. Implementation aspects should be carefully considered.

- **Clustering of sub–models**

  In Section 3.1 two methodological approaches are proposed for estimation of data–driven models. In this research the focus is given on the first methodological approach which is applied directly on the data. In the second methodological approach, more elaborate approach, data are divided into clusters before the model fitting. In such a way more detailed models could be developed and testing on data should be performed. Guidelines for the selection of one or the other approach and the best way of clustering should be given.

- **Space gaps in the optimization algorithm**

  In this research speeds have been used in the optimization algorithm so that the model minimizes the difference between observed and simulated speeds. It is proposed as a future prospect that space gaps instead of speeds can be used for the model optimization.

- **Incorporation of additional variables in the models**

  The proposed methodological framework is flexible, less time-consuming and allows the incorporation of additional variables that may influence driving behavior (such as density of the surrounding area, vehicle type, drivers' age, road infrastructure etc.). In addition, the proposed methodology could be employed with flexible data-driven models which allow incorporation of further variables moving towards an integrated solution for the simulation of mixed traffic. In Section 4.5 the incorporation of additional information is feasible and density of the surrounding area is explored as an extra variable. However, there are many variables, the influence of which on driving behavior has not been explored, yet. Further information on various predictor variables (such as weather, lighting, road geometry, percent of autonomous vehicles etc.) could be also added to the model and tested if it is significant.

- **Parameters evolution**

In Chapter 5, the prediction of the dynamic parameters was simple, in the sense that the dynamically calibrated parameters were assumed as the best available estimate for the short–term values of these parameters. Further research could consider secondary models that would actually aim at predicting the evolution of these parameters, as well, e.g. via autoregressive, polynomial or other statistical forecasting specifications.

- **Model calibration separately for each vehicle type**

  In case study for mixed traffic conditions (chapter 6) the calibration for the Gipps model as well as for the data–driven model is implemented using a representative sample from each vehicle type category. For a more in depth analysis, different models for each vehicle type category could be calibrated in order to develop more detailed models.

- **Vehicle–dependent models**

  Vehicle-dependent models need to be developed, as the drivers of vehicles with unequal dimensions tend to have different driving behaviors; furthermore, different vehicle types are characterized by varying vehicle kinematics. Especially, in case of heterogeneous traffic vehicle type plays a significant role as it is indicated in Section 6.3.3. The performance of a model seems to be differentiated as per the vehicle type. The best performance is achieved for cars and light commercial vehicles, while higher RMSN are observed for other vehicle types, especially for trucks and auto-rickshaws. Vehicle type should be incorporated as a categorical variable in the process. Thus, it is foreseen that further exploration into this could open up opportunities to understand and simulate driving behavior in non–lane discipline conditions with heterogeneity of vehicle types.

- **Exploration of longitudinal and lateral movement separately for mixed traffic conditions**

  In this research speed for mixed conditions is explored as a sum of longitudinal and lateral speed. However, longitudinal and lateral speed could be explored separately in order to investigate the model efficiency for each direction.

- **Deep learning and tree–based modeling**

  In this research various machine learning techniques have been used. However, other methods, such as deep learning and tree–based algorithms, should also be applied in order to offer an overall comparison of machine learning techniques for the estimation of microscopictraffic simulation models.

- **Mixed traffic**

  Crowd simulation and swarm–like models could be also used for modeling mixed traffic conditions due to weak–lane discipline characteristics.

- **Integrated behavior models**

  Car–following and lane–changing behaviors should be incorporated into one data–driven model, as there is interaction between these two behaviors.

- **Further experimental analysis**

  The estimation of data–driven models as well as the dynamic calibration and multiple time step prediction have been successfully demonstrated using actual data from a variety of facilities. However, additional testing on richer data and further applications in different networks need to be performed.

Appendix A

List of publications

*JA: journal article, BC: book chapter, CP: conference proceedings

Chapter	Title	Type*
3.1	Papathanasopoulou, V. and Antoniou (2015). Simulation optimization of car–following models using flexible techniques. In N. D. Lagaros and M. Papadrakakis (eds), Engineering and Applied Sciences Optimization. Dedicated to the memory of Professor M.G. Karlaftis, Springer, pp. 87-106.	BC
3.3.2	Antoniou, C., V. Gikas, V. Papathanasopoulou, C. Danezis, A. Panagopoulos, I. Markou, H. Perakis, D. Efthymiou and G. Yannis (2015). Localization and driving behavior classification using smartphone sensors in the direct absence of GNSS. Transportation Research Record: Journal of the Transportation Research Board, 2489, pp. 66-76.	JA
3.4.2	Antoniou, C., V. Gikas, V. Papathanasopoulou, T. Mpimis, I. Markou and H. Perakis (2014). Towards distribution-based calibration for traffic simulation. 17th International IEEE Conference on Intelligent Transportation Systems, Qingdao, China, October 8-11, 2014.	CP
	Markou, I., V. Papathanasopoulou and C. Antoniou (2015). A demonstration of distribution-based calibration, Proceedings of Models and Technologies for Intelligent Transportation Systems (MT-ITS), Budapest, 3-5, June 2015.	CP
4.1–4.3	Papathanasopoulou, V. and Antoniou, C. (2015). Towards data-driven car-following models. Transportation Research Part C: Emerging Technologies, 55, pp. 496-509.	JA
	Papathanasopoulou, V. and C. Antoniou (2014). Simulation optimization of car-following models using flexible models. Opt-i: 1st International Conference on Engineering and Applied Sciences Optimization, Kos, Greece, 4-6 June 2014.	CP
4.4.3	Papathanasopoulou V. and C. Antoniou (2016). On–line calibration for data–driven car–following models: can fewer, but more relevant, traffic observations suffice? Proceedings of the 5th Symposium of the European Association for Research in Transportation, September 14-16, Delft, Netherlands.	CP
	Papathanasopoulou, V. and C. Antoniou (2017). A comparison of machine learning techniques for data-driven car-following models. mobil.TUM 2017 – International Scientific Conference on Mobility and Transport, 4-5 July, Munich, Germany.	CP

Chapter	Title	Type*
4.5	Papathanasopoulou, V. and Antoniou, C. (2016). Flexible car–following models incorporating information from adjacent lanes. Proceedings of the 19th IEEE International Conference on Intelligent Transportation Systems, November 1-4, Rio de Janeiro, Brazil, pp. 701-706.	CP
5	Papathanasopoulou, V., Markou, I. and Antoniou, C. (2016). Online calibration for microscopic traffic simulation and dynamic multi-step prediction of traffic speed. Transportation Research Part C: Emerging technologies, 68, 144-159.	JA
	Papathanasopoulou V. and C. Antoniou (2015). On-line calibration of car–following models exploiting input parameters correlation. Proceedings of the 4th Symposium of the European Association for Research in Transportation, September 9-11, Copenhagen, Denmark.	CP
6	Papathanasopoulou, V. and C. Antoniou (2017). Data-driven models for identification of lane-changing characteristics and duration using NGSIM data, Proceedings of 6th Symposium of the European Association for Research in Transportation (hEART), 12-14 September 2017, Haifa, Israel.	CP
6	Papathanasopoulou, V. and C. Antoniou (2017). Flexible car-following models for mixed traffic conditions, Proceedings of 8th International Congress on Transportation Research (ICTR 2017), September 27-29, 2017, Thessaloniki, Greece.	CP
6.2	Papathanasopoulou, V. and Antoniou, C. (2018). Identification of lane changes manoeuvres on mixed traffic trajectory data. Proceedings of the 97th Annual Meeting of the Transportation Research Board, January 7-11, Washington, D.C., USA.	CP
6.3	Papathanasopoulou, V. and Antoniou, C. (2017). Flexible car–following models on mixed traffic trajectory data. Proceedings of the 96th Annual Meeting of the Transportation Research Board, January 8-12, Washington, D.C., USA. (No. 17-06671).	CP

Chapter	Title	Type*
–	Antoniou, C., V. Gikas, V. Papathanasopoulou, T. Mpimis, H. Perakis, C. Kyriazis (2017). A framework for risk reduction for indoor parking facilities under constraints using positioning technologies, International Journal of Disaster Risk Reduction (in press)	JA
	Antoniou, C., V. Papathanasopoulou, V. Gikas, C. Danezis and H. Perakis (2014). Classification of driving characteristics using smartphone sensor data, hEART 2014: 3rd Symposium of the European Association for Research in Transportation, Leeds, UK, September 10-12, 2014.	CP
	Sarlas, G., V. Papathanasopoulou and Antoniou, C. (2013). A simulation-based analysis of road pricing prospects for Athens, Greece. ASCE Journal of Urban Planning and Development, 139(3), pp. 206–215. (doi: 10.1061/(ASCE)UP.1943-5444.0000145).	JA
	Antoniou, C., V. Gikas, V. Papathanasopoulou, T. Mpimis, H. Perakis and C. Kyriazis (2017). Data collection and traffic modelling of indoor parking facilities under constraints, Proceedings of 8th International Congress on Transportation Research (ICTR 2017), September 27-29, 2017, Thessaloniki, Greece.	CP
	Antoniou, C., V. Gikas, V. Papathanasopoulou, T. Mpimis, H. Perakis and C. Kyriazis (2017). A framework for efficient data collection and modeling of indoor parking facilities under constraints. Proceedings of the 96th Annual Meeting of the Transportation Research Board, January 2017, Washington, D.C.	CP
	Papathanasopoulou V., I. Markou, C. Antoniou, V. Gikas, A. Mpimis, H. Perakis and G. Yannis (2015). "Efficient management of parking under constraints", Proceedings of the 7th International Congress on Transport Research in Greece, Hellenic Institute of Transportation Engineers, Hellenic Institute of Transport, Athens, November 2015. .	CP
	Antoniou, C., V. Gikas, V. Papathanasopoulou, C. Danezis, A. Panagopoulos, I. Markou, H. Perakis, D. Efthymiou and G. Yannis (2015). Localization and driving behavior classification using smartphone sensors in the direct absence of GNSS. Proceedings of the 94th Annual Meeting of the Transportation Research Board, January 2015, Washington, D.C.	CP

Chapter	Title	Type*
–	Antoniou, C., V. Papathanasopoulou, V. Gikas, T. Mpimis, I. Markou and H. Perakis (2014). Monitoring indoor driver behavior using opportunistic smartphone sensor data. ITS2014 - ITS and Smart Cities, Patras, November 19-21, 2014.	CP
	Papathanasopoulou, V. and C. Antoniou (2011). Assessment of congestion pricing prospects for Athens, Greece. Proceedings of the 90th Annual Meeting of the Transportation Research Board, January 2011, Washington, D.C.	CP

Bibliography

Ahmed, K., Ben-Akiva, M., Koutsopoulos, H., and Mishalani, R. (1996). Models of freeway lane changing and gap acceptance behavior. *Transportation and traffic theory*, 13:501–515.

Ahmed, K. I. (1999). *Modeling drivers' acceleration and lane changing behavior*. PhD thesis, Massachusetts Institute of Technology.

Al-Obaedi, J., Yousif, S., et al. (2009). The use of visual angle in car following traffic micro-simulation models.

Antoniou, C. (2004). *On–line calibration for dynamic traffic assignment*. PhD thesis, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology.

Antoniou, C., Balakrishna, R., and Koutsopoulos, H. N. (2011). A synthesis of emerging data collection technologies and their impact on traffic management applications. *European Transport Research Review*, 3(3):139–148.

Antoniou, C., Barcelò, J., Brackstone, M., Celikoglu, H., Ciuffo, B., Punzo, V., Sykes, P., Toledo, T., Vortisch, P., and Wagner, P. (2014a). Traffic simulation: case for guidelines.

Antoniou, C., Ben-Akiva, M., and Koutsopoulos, H. N. (2005). Online calibration of traffic prediction models. *Transportation Research Record: Journal of the Transportation Research Board*, 1934(1):235–245.

Antoniou, C., Ben-Akiva, M., and Koutsopoulos, H. N. (2007). Nonlinear Kalman filtering algorithms for on–line calibration of dynamic traffic assignment models. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):661–670.

Antoniou, C., Gikas, V., Papathanasopoulou, V., Mpimis, T., Markou, I., and Perakis, H. (2014b). Towards distribution-based calibration for traffic simulation. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 786–791. IEEE.

Antoniou, C. and Koutsopoulos, H. (2006a). Estimation of traffic dynamics models with machine-learning methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1965):103–111.

Antoniou, C. and Koutsopoulos, H. N. (2006b). A Comparison of Machine Learning Methods for Speed Estimation. In *Proceedings of the 11th IFAC Symposium on Control in Transportation Systems*.

Antoniou, C. and Koutsopoulos, H. N. (2006c). Estimation of Traffic Dynamics Models with Machine Learning Methods. *Transportation Research Record: Journal of the Transportation Research Board*, (1965):103–111.

Antoniou, C., Koutsopoulos, H. N., and Yannis, G. (2013). Dynamic data–driven local traffic state estimation and prediction. *Transportation Research Part C: Emerging Technologies*, 34:89–107.

Asaithambi, G., Kanagaraj, V., and Toledo, T. (2016). Driving behaviors: Models and challenges for non-lane based mixed traffic. *Transportation in Developing Economies*, 2(2):19.

Aycin, M. and Benekohal, R. (1999). Comparison of car–following models for simulation. *Transportation Research Record: Journal of the Transportation Research Board*, 1678:116–127.

Balakrishna, R. (2006). *Off–line calibration of dynamic traffic assignment models*. PhD thesis, Massachusetts Institute of Technology.

Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H., and Wen, Y. (2007a). Calibration of microscopic traffic simulation models: Methods and application. *Transportation Research Record: Journal of the Transportation Research Board*, (1999):198–207.

Balakrishna, R., Antoniou, C., Ben-Akiva, M., Koutsopoulos, H., and Wen, Y. (2007b). Calibration of microscopic traffic simulation models: Methods and application. *Transportation Research Record: Journal of the Transportation Research Board*, 1999:198–207.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A., and Sugiyama, Y. (1994). Structure stability of congestion in traffic dynamics. *Japan Journal of Industrial and Applied Mathematics*, 11(2):203–223.

Bando, M., Hasebe, K., Nakayama, A., Shibata, A., and Sugiyama, Y. (1995). Dynamical model of traffic congestion and numerical simulation. *Physical review E*, 51(2):1035.

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pages 803–821.

Barceló, J. and Casas, J. (2005). Dynamic network simulation with aimsun. *Simulation approaches in transportation analysis*, pages 57–98.

Barceló, J., Codina, E., Casas, J., Ferrer, J., and Garcia, D. (2005). Microscopic traffic simulation: A tool for the design, analysis and evaluation of intelligent transport systems. *Journal of Intelligent and Robotic Systems*, 41(2-3):173–203.

Barceló, J. et al. (2010). *Fundamentals of traffic simulation*, volume 145. Springer.

Barth, M., Todd, M., and Xue, L. (2004). User-based vehicle relocation techniques for multiple-station shared-use vehicle systems.

Bellemans, T., De Schutter, B., and De Moor, B. (2002). Models for traffic control. *JOURNAL A*, 43(3/4):13–22.

Ben-Akiva, M., Choudhury, C., and Toledo, T. (2009). Integrated lane-changing models. *Transport Simulation*.

Ben-Akiva, M., Koutsopoulos, H. N., Antoniou, C., and Balakrishna, R. (2010). Traffic simulation with DynaMIT. In Barceló, J., editor, *Fundamentals of traffic simulation*, pages 363–398. Springer.

Ben-Hur, A. and Weston, J. (2010). A user's guide to support vector machines. In *Data mining techniques for the life sciences*, pages 223–239. Springer.

Benekohal, R. and Treiterer, J. (1988). Carsim: Car-following model for simulation of traffic in normal and stop-and-go conditions. *Transportation research record*, (1194).

Bevrani, K. and Chung, E. (2011). Car following model improvement for traffic safety metrics reproduction. In *Proceedings of the Australasian Transport Research Forum 2011*, pages 1–14. PATREC.

Bham, G. H. and Benekohal, R. F. (2004). A high fidelity traffic simulation model based on cellular automata and car-following concepts. *Transportation Research Part C: Emerging Technologies*, 12(1):1–32.

Bi, H., Mao, T., Wang, Z., and Deng, Z. (2016). A data-driven model for lane-changing in traffic simulation. In *Symposium on Computer Animation*, pages 149–158.

Bifulco, G. N., Pariota, L., Brackstione, M., and Mcdonald, M. (2013a). Driving behaviour models enabling the simulation of advanced driving assistance systems: revisiting the action point paradigm. *Transportation Research Part C: Emerging Technologies*, 36:352–366.

Bifulco, G. N., Pariota, L., Simonelli, F., and Di Pace, R. (2013b). Development and testing of a fully adaptive cruise control system. *Transportation Research Part C: Emerging Technologies*, 29:156–170.

Bilban, M., Vojvoda, A., and Jerman, J. (2009). Age affects drivers' response times. *Collegium antropologicum*, 33(2):467–471.

Blue, V., Bonetto, F., and Embrechts, M. (1996). A cellular automata model of vehicular self-organization and nonlinear speed transitions. In *Proceedings of the Transportation Research Board 75th Annual Meeting*, number 961336.

Bonsall, P., Liu, R., and Young, W. (2005). Modelling safety-related driving behaviour—impact of parameter values. *Transportation Research Part A: Policy and Practice*, 39(5):425–444.

Bonte, L., Espié, S., and Mathieu, P. (2006). Modélisation et simulation des usagers deux-roues motorisés dans archisim. *JFSMA*, 6:17.

Box, G. E. (1979). All models are wrong, but some are useful. *Robustness in Statistics*, 202.

Boxill, S. A. and Yu, L. (2000). An evaluation of traffic simulation models for supporting its. *Houston, TX: Development Centre for Transportation Training and Research, Texas Southern University.*

Brackstone, M. and McDonald, M. (1999). Car–following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2(4):181–196.

Brackstone, M. and McDonald, M. (2003). Driver behaviour and traffic modelling. are we looking at the right issues? In *Intelligent Vehicles Symposium, 2003. Proceedings. IEEE*, pages 517–521. IEEE.

Brackstone, M., Waterson, B., and McDonald, M. (2009). Determinants of following headway in congested traffic. *Transportation Research Part F: Traffic Psychology and Behaviour*, 12(2):131–142.

Braskstone, M. and McDonald, M. (2000). Car following: a historical review, transportation research part f. 2.

Brunson, S., Kyle, E., Phamdo, N., and Preziotti, G. (2002). Alert algorithm development program: Nhtsa rear-end collision alert algorithm. Technical report.

Chandler, R. E., Herman, R., and Montroll, E. W. (1958). Traffic dynamics: studies in car following. *Operations research*, 6(2):165–184.

Chaniotakis, E., Antoniou, C., and Pereira, F. (2016). Mapping social media for transportation studies. *IEEE Intelligent Systems*, 31(6):64–70.

Chen, X.-Y., Pao, H.-K., and Lee, Y.-J. (2014). Efficient traffic speed forecasting based on massive heterogenous historical data. In *Big Data (Big Data), 2014 IEEE International Conference on*, pages 10–17. IEEE.

Chong, L., Abbas, M. M., Flintsch, A. M., and Higgs, B. (2013). A rule-based neural network approach to model driver naturalistic behavior in traffic. *Transportation Research Part C: Emerging Technologies*, 32:207–223.

Choudhury, C. F. and Islam, M. M. (2016). Modelling acceleration decisions in traffic streams with weak lane discipline: A latent leader approach. *Transportation research part C: emerging technologies*, 67:214–226.

Ciuffo, B., Casas, J., Montanino, M., Perarnau, J., and Punzo, V. (2013). From theory to practice: Gaussian process meta-models for sensitivity analysis of traffic simulation models: Case study of aimsun mesoscopic model. In *Proc. 92nd Annu. Meet. Transp. Res. Board*.

Ciuffo, B., Punzo, V., and Montanino, M. (2012). Thirty years of gipps' car-following model: Applications, developments, and new features. *Transportation Research Record: Journal of the Transportation Research Board*, (2315):89–99.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association*, 83(403):596–610.

Cleveland, W. S., Devlin, S. J., and Grosse, E. (1988). Regression by local fitting: methods, properties, and computational algorithms. *Journal of econometrics*, 37(1):87–114.

Cohen, R. A. (1999). An introduction to proc loess for local regression. In *Proceedings of the 24th SAS users group international conference, Paper*, volume 273. Citeseer.

Colombaroni, C. and Fusco, G. (2014). Artificial neural network models for car following: experimental analysis and calibration issues. *Journal of Intelligent Transportation Systems*, 18(1):5–16.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Courtois, M. and Woodside, M. (2000). Using regression splines for software performance analysis. In *Proceedings of the 2nd international workshop on Software and performance*, pages 105–114. ACM.

Daamen, W., Buisson, C., and Hoogendoorn, S. P. (2014). *Traffic Simulation and Data: Validation Methods and Applications*. CRC Press.

De Fabritiis, C., Ragona, R., and Valenti, G. (2008). Traffic estimation and prediction based on real time floating car data. In *11th International IEEE Conference on Intelligent Transportation Systems (ITSC 2008)*, pages 197–203. IEEE.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Ding, C., Wang, W., Wang, X., and Baumann, M. (2013). A neural network model for driver's lane-changing trajectory prediction in urban traffic flow. *Mathematical Problems in Engineering*, 2013.

Ding, D., Yoo, J., Jekyo, J., Jin, S., and Kwon, S. (2015). Efficient road-sign detection based on machine learning. *Bulletin of Networking, Computing, Systems, and Software*, 4(1):15–17.

El Faouzi, N.-E., Leung, H., and Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges–a survey. *Information Fusion*, 12(1):4–10.

Ellison, A., Greaves, S., and Bliemer, M. (2013). Examining heterogeneity of driver behavior with temporal and spatial factors. *Transportation Research Record: Journal of the Transportation Research Board*, (2386):158–167.

Fambro, D., Koppa, R., Picha, D., and Fitzpatrick, K. (1998). Driver perception-brake response in stopping sight distance situations. *Transportation Research Record: Journal of the Transportation Research Board*, (1628):1–7.

Fei, X., Lu, C.-C., and Liu, K. (2011). A bayesian dynamic linear model approach for real–time short–term freeway travel time prediction. *Transportation Research Part C: Emerging Technologies*, 19(6):1306–1318.

Foresee, F. D. and Hagan, M. T. (1997). Gauss-newton approximation to bayesian learning. In *Neural Networks, 1997., International Conference on*, volume 3, pages 1930–1935. IEEE.

Fraley, C. (1998). Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20(1):270–281.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458):611–631.

Fraley, C. and Raftery, A. E. (2003). Enhanced model-based clustering, density estimation, and discriminant analysis software: Mclust. *Journal of Classification*, 20(2):263–286.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Fritzsche, H.-T. (1994). A model for traffic simulation. *Traffic Engineering+ Control*, 35(5):317–21.

Gazis, D. C., Herman, R., and Rothery, R. W. (1961). Nonlinear follow-the-leader models of traffic flow. *Operations research*, 9(4):545–567.

Gibbs, A. L. and Su, F. E. (2002). On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435.

Gipps, P. G. (1981). A behavioural car–following model for computer simulation. *Transportation Research Part B: Methodological*, 15(2):105–111.

Gipps, P. G. (1986). A model for the structure of lane-changing decisions. *Transportation Research Part B: Methodological*, 20(5):403–414.

Green, M. (2000). " how long does it take to stop?" methodological analysis of driver perception-brake times. *Transportation human factors*, 2(3):195–216.

Gretton, A., Borgwardt, K. M., Rasch, M., Schölkopf, B., and Smola, A. J. (2007). A kernel approach to comparing distributions. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1637. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773.

Guido, G., Gallelli, V., Rogano, D., and Vitale, A. (2016). Evaluating the accuracy of vehicle tracking data obtained from unmanned aerial vehicles. *International journal of transportation science and technology*, 5(3):136–151.

Gundaliya, P., Mathew, T. V., and Dhingra, S. L. (2008). Heterogeneous traffic flow modelling for an arterial using grid based approach. *Journal of Advanced Transportation*, 42(4):467–491.

Hamdar, S. H. and Mahmassani, H. S. (2008). Driver car-following behavior: From discrete event process to continuous set of episodes. In *Transportation Research Board 87th Annual Meeting*, number 08-3134.

Happe, J., Westermann, D., Sachs, K., and Kapová, L. (2010). Statistical inference of software performance models for parametric performance completions. In *International Conference on the Quality of Software Architectures*, pages 20–35. Springer.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Helly, W. (1959). *Dynamics of Single Lane Vehicular Traffic Flow*. Center for Operations Research, Massachusetts Institute of Technology.

Henclewood, D., Suh, W., Rodgers, M., Hunter, M., and Fujimoto, R. (2012). A case for real-time calibration of data-driven microscopic traffic simulation tools. In *Proceedings of the Winter Simulation Conference*, page 148. Winter Simulation Conference.

Hidas, P. (1998). A car-following model for urban traffic simulation. *Traffic engineering & control*, 39(5):300–305.

Hidas, P. (2002). Modelling lane changing and merging in microscopic traffic simulation. *Transportation Research Part C: Emerging Technologies*, 10(5):351–371.

Hidas, P. (2005). Modelling vehicle interactions in microscopic simulation of merging and weaving. *Transportation Research Part C: Emerging Technologies*, 13(1):37–62.

Higgs, B. and Abbas, M. M. (2014). Multi–resolution comparison of car–following models using naturalistic data. In *Transportation Research Board 93rd Annual Meeting*, number 14-4528.

Hoogendoorn, S., Hoogendoorn, R., and Daamen, W. (2011). Wiedemann revisited: new trajectory filtering technique and its implications for car-following modeling. *Transportation Research Record: Journal of the Transportation Research Board*, (2260):152–162.

Hoogendoorn, S., Ossen, S., Schreuder, M., and Gorte, B. (2006). Unscented particle filter for delayed car–following models estimation. In *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*, pages 1598–1603. IEEE.

Hoogendoorn, S. P. and Hoogendoorn, R. (2010). Generic calibration framework for joint estimation of car–following models by using microscopic data. *Transportation Research Record: Journal of the Transportation Research Board*, 2188:37–45.

Hou, Y., Edara, P., and Sun, C. (2014). Modeling mandatory lane changing using bayes classifier and decision trees. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):647–655.

Huang, W., Song, G., Hong, H., and Xie, K. (2014). Deep architecture for traffic flow prediction: deep belief networks with multitask learning. *IEEE Transactions on Intelligent Transportation Systems*, 15(5):2191–2201.

Husch, D. and Albeck, J. (2000). Simtraffic 5.0 user guide for windows.

Huval, B., Wang, T., Tandon, S., Kiske, J., Song, W., Pazhayampallil, J., Andriluka, M., Cheng-Yue, R., Mujica, F., Coates, A., Rajpurkar, P., Migimatsu, T., and Y. Ng, A. (2015). An empirical evaluation of deep learning on highway driving. *arXiv preprint arXiv:1504.01716*.

Jenelius, E. and Koutsopoulos, H. N. (2013). Travel time estimation for urban road networks using low frequency probe vehicle data. *Transportation Research Part B: Methodological*, 53:64–81.

Jenelius, E. and Koutsopoulos, H. N. (2018). Urban network travel time prediction based on a probabilistic principal component analysis model of probe data. *IEEE Transactions on Intelligent Transportation Systems*, 19(2):436–445.

Johansson, G. and Rumar, K. (1971). Drivers' brake reaction times. *Human factors*, 13(1):23–27.

Kaisler, S., Armour, F., Espinosa, J. A., and Money, W. (2013). Big data: Issues and challenges moving forward. In *System sciences (HICSS), 2013 46th Hawaii international conference on*, pages 995–1004. IEEE.

Kanagaraj, V., Asaithambi, G., Kumar, C. N., Srinivasan, K. K., and Sivanandan, R. (2013). Evaluation of different vehicle following models under mixed traffic conditions. *Procedia-Social and Behavioral Sciences*, 104:390–401.

Kanagaraj, V., Asaithambi, G., Toledo, T., and Lee, T.-C. (2015). Trajectory data and flow characteristics of mixed traffic. *Transportation Research Record: Journal of the Transportation Research Board*, (2491):1–11.

Karlaftis, M. G. and Vlahogianni, E. I. (2011). Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399.

Kesting, A. and Treiber, M. (2008). Calibrating car–following models by using trajectory data: Methodological study. *Transportation Research Record: Journal of the Transportation Research Board*, 2088:148–156.

Kesting, A., Treiber, M., and Helbing, D. (2007). General lane-changing model mobil for car-following models. *Transportation Research Record: Journal of the Transportation Research Board*, (1999):86–94.

Kesting, A., Treiber, M., and Helbing, D. (2010). Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1928):4585–4605.

Kikuchi, S. and Chakroborty, P. (1992). Car-following model based on fuzzy inference system. *Transportation Research Record*, pages 82–82.

Kim, J. and Mahmassani, H. (2011). Correlated parameters in driving behavior models. car–following example and implications for traffic microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, 2249:62–77.

Kim, Y. (2002). Online traffic flow model applying dynamic flow–density relation. Technical report, Int. At. Energy Agency.

Kleyko, D., Hostettler, R., Birk, W., and Osipov, E. (2015). Comparison of machine learning techniques for vehicle classification using road side sensors. In *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, pages 572–577. IEEE.

Kometani, E. and Sasaki, T. (1958). On the stability of traffic flow (report-i). *J. Oper. Res. Soc. Japan*, 2(1):11–26.

Koutsopoulos, H. N. and Farah, H. (2012). Latent class model for car following behavior. *Transportation research part B: methodological*, 46(5):563–578.

Krajzewicz, D. (2009). Kombination von taktischen und strategischen einflüssen in einer mikroskopischen verkehrsflusssimulation.

Krauß, S. (1998). *Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics*. PhD thesis.

Kuderer, M., Gulati, S., and Burgard, W. (2015). Learning driving styles for autonomous vehicles from demonstration. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 2641–2646. IEEE.

Kumar, P., Perrollaz, M., Lefevre, S., and Laugier, C. (2013). Learning-based approach for online lane change intention prediction. In *Intelligent Vehicles Symposium (IV), 2013 IEEE*, pages 797–802. IEEE.

Kurzhanskiy, A. A. and Varaiya, P. (2010). Active traffic management on road networks: a macroscopic approach. *Philosophical Transactions of the*

Royal Society of London A: Mathematical, Physical and Engineering Sciences, 368(1928):4607–4626.

Laval, J. A. and Leclercq, L. (2008). Microscopic modeling of the relaxation phenomenon using a macroscopic lane-changing model. *Transportation Research Part B: Methodological*, 42(6):511–522.

Leclercq, L., Chiabaut, N., Laval, J., and Buisson, C. (2007). Relaxation phenomenon after lane changing: Experimental validation with ngsim data set. *Transportation Research Record: Journal of the Transportation Research Board*, (1999):79–85.

Leutzbach, W. and Wiedemann, R. (1986). Development and applications of traffic simulation models at the karlsruhe institut für verkehrswesen. *Traffic engineering & control*, 27(5):270–278.

Li, Y., Zhang, L., Peeta, S., Pan, H., Zheng, T., Li, Y., and He, X. (2015). Non-lane-discipline-based car-following model considering the effects of two-sided lateral gaps. *Nonlinear Dynamics*, 80(1-2):227–238.

Lindsay, B. G. and Basak, P. (2000). Moments determine the tail of a distribution (but not much else). *The American Statistician*, 54(4):248–251.

Lorkowski, S. and Wagner, P. (2005). Parameter calibration of traffic models in microscopic online simulations. In *84th Annual Meeting of the Transportation Research Board*.

Lv, Y., Duan, Y., Kang, W., Li, Z., and Wang, F.-Y. (2015). Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873.

M. Montanino and V. Punzo (2015). Trajectory data reconstruction and simulation-based validation against macroscopic traffic patterns. *Transportation Research Part B: Methodological*, 80:82–106.

Ma, X. (2006). A neural–fuzzy framework for modeling car–following behavior. In *IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC'06*, volume 2, pages 1178–1183. IEEE.

Ma, X. and Andréasson, I. (2005). Dynamic car following data collection and noise cancellation based on the kalman smoothing. In *IEEE International Conference on Vehicular Electronics and Safety, 2005*, pages 35–41. IEEE.

Ma, X. and Jansson, M. (2007). Model estimation for car-following dynamics based on adaptive filtering approach. In *IEEE Intelligent Transportation Systems Conference (ITSC 2007).*, pages 824–829. IEEE.

Maciejewski, M. (2010). A comparison of microscopic traffic flow simulation systems for an urban area. *Transport Problems*, 5(4):27–38.

MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, 4(3):415–447.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Magister, T., Krulec, R., Batista, M., and Bogdanović, L. (2005). The driver reaction time measurement experiences. *Innovative Automotive Technology–IAT*, 5.

Mathew, T. V. (2014). Lane changing models. *Transportation systems engineering anonymous*, pages 15–1.

Mathew, T. V., Munigety, C. R., and Bajpai, A. (2013). Strip-based approach for the simulation of mixed traffic conditions. *Journal of Computing in Civil Engineering*, 29(5):04014069.

McCullagh, P. (1994). Does the moment–generating function characterize a distribution? *The American Statistician*, 48(3):208–208.

McDonald, M., Wu, J., and Brackstone, M. (1997). Development of a fuzzy logic based microscopic motorway simulation model. In *Intelligent Transportation System, 1997. ITSC'97., IEEE Conference on*, pages 82–87. IEEE.

McLachlan, G. J. and Krishnan, T. (1997). Wiley series in probability and statistics. *The EM Algorithm and Extensions, Second Edition*, pages 361–369.

Metkari, M., Budhkar, A., and Maurya, A. K. (2013). Development of simulation model for heterogeneous traffic with no lane discipline. *Procedia-Social and Behavioral Sciences*, 104:360–369.

Michaels, R. (1963). Perceptual factors in car following. In *Proceedings of the 2nd International Symposium on the Theory of Road Traffic Flow (London, England), OECD*.

Milborrow, S. (2017). Package 'earth' 4.4.9.1. Multivariate adaptive regression spline models.

Miller, M. and Skabardonis, A. (2010). San diego i-15 integrated corridor management (icm) 7 system: Stage ii (analysis, modeling, and simulation. california path program, institute of 8 transportation studies. *University of California at Berkeley*, 9.

Mitchell, T. M. et al. (1997). Machine learning. wcb.

Montanino, M. and Punzo, V. (2013). Making NGSIM data usable for studies on traffic flow theory: Multistep method for vehicle trajectory reconstruction. *Transportation Research Record: Journal of the Transportation Research Board*, (2390):99–111.

Montanino, M. and Punzo, V. (2014). Reconstructed NGSIM I80-1. *Cost Action TU0903—Multitude, Mar.*

Monteil, J., Billot, R., Sau, J., Buisson, C., and El Faouzi, N.-E. (2014). Calibration, estimation, and sampling issues of car–following parameters. *Transportation Research Record: Journal of the Transportation Research Board*, 2422:131–140.

Moridpour, S., Sarvi, M., and Rose, G. (2010). Lane changing models: a critical review. *Transportation letters*, 2(3):157–173.

Muezzinoglu, M. K. and Zuracla, J. (2005). A recurrent rbf network model for nearest neighbor classification. In *Neural Networks, 2005. IJCNN'05. Proceedings. IEEE International Joint Conference on*, volume 1, pages 343–348. IEEE.

Munigety, C. R. and Mathew, T. V. (2016). Towards behavioral modeling of drivers in mixed traffic conditions. *Transportation in Developing Economies*, 2(1):6.

Nagel, K. and Schreckenberg, M. (1992). A cellular automaton model for freeway traffic. *Journal de physique I*, 2(12):2221–2229.

Nguyen, D. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Proc. International Joint Conference on Neural Networks*, volume 3, pages 21–26.

Olstam, J. J. and Tapani, A. (2004). Comparison of car–following models. *Swedish National Road and Transport Research Institute, Project VTI meddelande*, 960 A.

Ossen, S. and Hoogendoorn, S. (2005). Car–following behavior analysis from microscopic trajectory data. *Transportation Research Record: Journal of the Transportation Research Board*, (1934):13–21.

Ossen, S. and Hoogendoorn, S. (2007). Driver heterogeneity in car following and its impact on modeling traffic dynamics. *Transportation Research Record: Journal of the Transportation Research Board*, pages 95–103.

Ossen, S. and Hoogendoorn, S. P. (2011). Heterogeneity in car–following behavior: Theory and empirics. *Transportation Research Part C: Emerging Technologies*, 19:182–195.

Oud, M. (2016). Performance of existing integrated car following and lane change models around motorway ramps.

Panwai, S. and Dia, H. (2005). Comparative evaluation of microscopic car-following behavior. *IEEE Transactions on Intelligent Transportation Systems*, 6(3):314–325.

Panwai, S. and Dia, H. (2007). Neural agent car-following models. *IEEE Transactions on Intelligent Transportation Systems*, 8(1):60–70.

Papageorgiou, M., Blosseville, J.-M., and Hadj-Salem, H. (1989). Macroscopic modelling of traffic flow on the Boulevard Périphérique in Paris. *Transportation Research Part B: Methodological*, 23(1):29–47.

Papathanasopoulou, V. and Antoniou, C. (2015a). Simulation optimization of car-following models using flexible techniques. In *Engineering and Applied Sciences Optimization*, pages 87–106. Springer.

Papathanasopoulou, V. and Antoniou, C. (2015b). Towards data-driven car-following models. *Transportation Research Part C: Emerging Technologies*, 55:496–509.

Papathanasopoulou, V. and Antoniou, C. (2016). Flexible car-following models incorporating information from adjacent lanes. In *Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pages 701–706. IEEE.

Papathanasopoulou, V. and Antoniou, C. (2017). Flexible car-following models on mixed traffic trajectory data. In *Transportation Research Board 96th Annual Meeting*, number 17-06671.

Papathanasopoulou, V. and Antoniou, C. (2018). Identification of lane changes manoeuvres on mixed traffic trajectory data. In *Transportation Research Board 97th Annual Meeting*.

Papathanasopoulou, V., Markou, I., and Antoniou, C. (2016). Online calibration for microscopic traffic simulation and dynamic multi–step prediction of traffic speed. *Transportation Research Part C: Emerging technologies*, 68:144–159.

Parker, M. (1996). The effect of heavy goods vehicles and following behaviour on capacity at motorway roadwork sites. *Traffic engineering & control*, 37(9):524–531.

Parsuvanathan, C. (2015). Proxy-lane algorithm for lane-based models to simulate mixed traffic flow conditions. *International Journal of Traffic and Transportation Engineering*, 4(5):131–136.

Pereira, F. C., Antoniou, C., Fargas, J. A., and Ben-Akiva, M. (2014). A metamodel for estimating error bounds in real–time traffic prediction systems. *IEEE Transactions on Intelligent Transportation Systems*, 15(3):1310–1322.

Pérez-Rodriguez, P. and Gianola, D. (2013). brnn: brnn (bayesian regularization for feed-forward neural networks). r package version 0.3.

Pipes, L. A. (1953). An operational analysis of traffic dynamics. *Journal of applied physics*, 24(3):274–281.

Ponnu, B. and Coifman, B. (2015). Speed-spacing dependency on relative speed from the adjacent lane: New insights for car following models. *Transportation Research Part B: Methodological*, 82:74–90.

Press, W. H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., et al. (1992). Numerical recipes (cambridge.

Punzo, V., Borzacchiello, M. T., and Ciuffo, B. (2011). On the assessment of vehicle trajectory data accuracy and application to the Next Generation SIMulation (NGSIM) program data. *Transportation Research Part C: Emerging Technologies*, 19(6):1243–1262.

Punzo, V., Ciuffo, B., and Montanino, M. (2012). Can results of car–following model calibration based on trajectory data be trusted? *Transportation Research Record: Journal of the Transportation Research Board*, 2315:11–24.

Punzo, V., Formisano, D., and Torrieri, V. (2005). Part 1: Traffic flow theory and car following: Nonstationary kalman filter for estimation of accurate and consistent car-following data. *Transportation Research Record: Journal of the Transportation Research Board*, (1934):1–12.

Punzo, V. and Montanino, M. (2016). Speed or spacing? cumulative variables, and convolution of model errors and time in traffic flow models validation and calibration. *Transportation Research Part B: Methodological*, 91:21–33.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, M., Chowdhury, M., Khan, T., and Bhavsar, P. (2014). A parameter estimation and calibration method for car–following models. In *Transportation Research Board 93rd Annual Meeting*, number 14-4634.

Rahman, M., Chowdhury, M., Xie, Y., and He, Y. (2013). Review of microscopic lane-changing models and future research opportunities. *IEEE transactions on intelligent transportation systems*, 14(4):1942–1956.

Ramsey, J. B., Newton, H. J., and Harvill, J. L. (2002). *The elements of statistics: With applications to economics and the social sciences*. Duxbury/Thomson Learning.

Ranjitkar, P., Suzuki, H., and Nakatsuji, T. (2005). Microscopic traffic data with real-time kinematic global positioning system. In *Proceedings of annual meeting of infrastructure planning and management, Japan Society of Civil Engineer, Miyazaki, Preprint CD*.

Reuschel, A. (1950). Fahrzeugbewegungen in der kolonne. *Osterreichisches Ingenieur Archiv*, 4:193–215.

Rickert, M., Nagel, K., Schreckenberg, M., and Latour, A. (1996). Two lane traffic simulations using cellular automata. *Physica A: Statistical Mechanics and its Applications*, 231(4):534–550.

Riedmiller, M., Montemerlo, M., and Dahlkamp, H. (2007). Learning to drive a real car in 20 minutes. In *Frontiers in the Convergence of Bioscience and Information Technologies, 2007. FBIT 2007*, pages 645–650. IEEE.

Roughan, M., Sen, S., Spatscheck, O., and Duffield, N. (2004). Class-of-service mapping for qos: a statistical signature-based approach to ip traffic classification. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 135–148. ACM.

Runarsson, T. P. and Yao, X. (2005). Search biases in constrained evolutionary optimization. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(2):233–243.

Saidallah, M., El Fergougui, A., and Elalaoui, A. E. (2016). A comparative study of urban road traffic simulators. In *MATEC Web of Conferences*, volume 81, page 05002. EDP Sciences.

Saifuzzaman, M. and Zheng, Z. (2014). Incorporating human-factors in car-following models: a review of recent developments and research needs. *Transportation research part C: emerging technologies*, 48:379–403.

Santos, I. R. and Santos, P. R. (2007). Simulation metamodels for modeling output distribution parameters. In *Simulation Conference, 2007 Winter*, pages 910–918. IEEE.

Schultz, G. and Rilett, L. (2004). Analysis of distribution and calibration of car–following sensitivity parameters in microscopic traffic simulation models. *Transportation Research Record: Journal of the Transportation Research Board*, 1876:41–51.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

Simonelli, F., Bifulco, G., De Martinis, V., and Punzo, V. (2009). Human-like adaptive cruise control systems through a learning machine approach. *Applications of Soft Computing*, pages 240–249.

Smith, L., Beckman, R., Anson, D., Nagel, K., and Williams, M. (1995). Transims: Transportation analysis and simulation system. Technical report, Los Alamos National Lab., NM (United States).

Stoyanov, J. (2006). Determinacy of distributions by their moments. In *Proc. International Conf. Mathematical and Statistical Modeling in Honor of Enrique Castillo*.

Summala, H. (2000). Brake reaction times and driver behavior analysis. *Transportation Human Factors*, 2(3):217–226.

Taieb-Maimon, M. and Shinar, D. (2001). Minimum and comfortable driving headways: Reality versus perception. *Human factors*, 43(1):159–172.

Thiemann, C., Treiber, M., and Kesting, A. (2008). Estimating acceleration and lane-changing dynamics from next generation simulation trajectory data. *Transportation Research Record: Journal of the Transportation Research Board*, (2088):90–101.

Toledo, T. (2003). *Integrating driving behavior modeling*. PhD thesis, Massachusetts Institute of Technology.

Toledo, T. (2007). Driving behaviour: models and challenges. *Transport Reviews*, 27(1):65–84.

Toledo, T., Koutsopoulos, H., and Ben-Akiva, M. (2003). Modeling integrated lane-changing behavior. *Transportation Research Record: Journal of the Transportation Research Board*, (1857):30–38.

Tordeux, A., Lassarre, S., and Roussignol, M. (2010). An adaptive time gap car-following model. *Transportation research part B: methodological*, 44(8):1115–1131.

Treiber, M., Hennecke, A., and Helbing, D. (2000). Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805.

Treiber, M. and Kesting, A. (2013). Microscopic calibration and validation of car–following models–a systematic approach. *Procedia-Social and Behavioral Sciences*, 80:922–939.

Tygert, M. (2010). Statistical tests for whether a given set of independent, identically distributed draws comes from a specified probability density. *Proceedings of the National Academy of Sciences*, 107(38):16471–16476.

Van Arem, B., Van Driel, C. J., and Visser, R. (2006). The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Transactions on Intelligent Transportation Systems*, 7(4):429–436.

van Lint, J., Hoogendoorn, S., and van Zuylen, H. (2005). Accurate freeway travel time prediction with state–space neural networks under missing data. *Transportation Research Part C: Emerging Technologies*, 13(5–6):347 – 369.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2005a). Optimized and meta–optimized neural networks for short–term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, 13(3):211 – 234.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2005b). Optimized and meta-optimized neural networks for short-term traffic flow prediction: a genetic approach. *Transportation Research Part C: Emerging Technologies*, 13(3):211–234.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2008). Temporal evolution of short–term urban traffic flow: A nonlinear dynamics approach. *Computer-Aided Civil and Infrastructure Engineering*, 23(7):536–548.

Wachenfeld, W. and Winner, H. (2016). Do autonomous vehicles learn? In *Autonomous Driving*, pages 451–471. Springer.

Wagner, P. (2012). Analyzing fluctuations in car–following. *Transportation Research Part B: Methodological*, 46(10):1384–1392.

Wang, E.-g., Sun, J., Jiang, S., and Li, F. (2017). Modeling the various merging behaviors at expressway on-ramp bottlenecks using support vector machine models. *Transportation Research Procedia*, 25:1327–1341.

Wang, L., Rong, J., and Liu, X. (2005). The classification of car-following behavior in urban expressway based on fuzzy clustering analysis. In *Proceedings of the 84th annual meeting of the transportation research board, Washington, DC.*

Wei, D. (2014). *Data-driven modeling and transportation data analytics*. PhD thesis.

Wei, H., Meyer, E., Lee, J., and Feng, C. (2000). Characterizing and modeling observed lane-changing behavior: lane-vehicle-based microscopic simulation on urban street network. *Transportation Research Record: Journal of the Transportation Research Board*, (1710):104–113.

Wiedemann, R. (1991). Modelling of rti-elements on multi–lane roads. In *Drive Conference (1991: Brussels, Belgium)*, volume 2.

Wiedemann, R. (1994). Simulation des straßenverkehrsflusses. schriftenreihe heft 8. *Institute for Transportation Science, University of Karlsruhe, Germany.*

Wiedemann, R. and Reiter, U. (1992). Microscopic traffic simulation: the simulation system mission, background and actual state. *Project ICARUS (V1052) Final Report*, 2:1–53.

Williams, C. K. and Rasmussen, C. E. (1996). Gaussian processes for regression. *Advances in neural information processing systems*, pages 514–520.

Wilson, R. E. and Ward, J. A. (2011). Car-following models: fifty years of linear stability analysis–a mathematical perspective. *Transportation Planning and Technology*, 34(1):3–18.

Xu, J.-X. and Lim, J. (2007). A new evolutionary neural network for forecasting net flow of a car sharing system. In *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*, pages 1670–1676. IEEE.

Xu, X., He, H., Zhao, D., Sun, S., Busoniu, L., and Yang, S. X. (2015). Machine learning with applications to autonomous systems. *Mathematical Problems in Engineering*, 2015.

Yang, Q. and Koutsopoulos, H. N. (1996). A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies*, 4(3):113–129.

Yang, X., Liu, L., Vaidya, N. H., and Zhao, F. (2004). A vehicle-to-vehicle communication protocol for cooperative collision warning. In *Mobile and Ubiquitous Systems: Networking and Services, 2004. MOBIQUITOUS 2004. The First Annual International Conference on*, pages 114–123. IEEE.

Yeo, H., Skabardonis, A., Halkias, J., Colyar, J., and Alexiadis, V. (2008). Oversaturated freeway flow algorithm for use in next generation simulation. *Transportation Research Record: Journal of the Transportation Research Board*, (2088):68–79.

Yu, G., Wang, P., Wu, X., and Wang, Y. (2015). Linear and nonlinear stability analysis of a car-following model considering velocity difference of two adjacent lanes. *Nonlinear Dynamics*, pages 1–11.

Zeileis, A., Hornik, K., Smola, A., and Karatzoglou, A. (2004). kernlab-an s4 package for kernel methods in r. *Journal of statistical software*, 11(9):1–20.

Zhang, H. M. and Kim, T. (2005). A car-following theory for multiphase vehicular traffic flow. *Transportation Research Part B: Methodological*, 39(5):385–399.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.

Zheng, J., Suzuki, K., and Fujita, M. (2013). Car-following behavior with instantaneous driver–vehicle reaction delay: A neural-network-based methodology. *Transportation research part C: emerging technologies*, 36:339–351.

Zheng, Z. (2014). Recent developments and research needs in modeling lane changing. *Transportation research part B: methodological*, 60:16–32.

Zhou, M. and Qu, X. (2016). Microscopic car-following model for autonomous vehicles using reinforcement learning. In *Symposium on Innovations in Traffic Flow Theory and Characteristics and TFT Midyear Meeting*, volume 2, page 3.

CURRICULUM VITAE

PERSONAL INFORMATION

Name	PAPATHANASOPOULOU VASILEIA
Address	102 Perikleous Str., Petroupoli, GR-13231, Athens, Greece
Telephone	+30 6974506098, +30 210 5063595
E-mail	**vasileia.papathanasopoulou@gmail.com**

Nationality	Hellenic
Date of birth	14 January 1987
Gender	Female

AREAS OF EXPERTISE

Transportation, Traffic Simulation Models, Machine Learning, Data-driven Models

EDUCATION AND TRAINING

2014- 2018 — **National Technical University of Athens**, School of Rural & Surveying, Ph.D. in the field of Transportation Engineering

Title of doctoral dissertation: “Towards data-driven microscopic traffic simulation models” (Supervisor C. Antoniou, Associate Professor NTUA)

2010- 2012 — **National Technical University of Athens**, School of Rural & Surveying, MS in Geoinformatics
Degree grade: 8.92 (Very Good)

Master thesis: “Data-driven estimation of car following models with machine learning methods” (Supervisor C. Antoniou, Assistant Professor of National Technical University of Athens)

2004- 2010 — **National Technical University of Athens**, School of Rural & Surveying, Diploma (5-year) in Surveying Engineering
Degree grade: 7.86 (Very Good)

Diploma thesis: “Assessment of congestion pricing prospects for Athens, Greece”
(Supervisor C. Antoniou, Assistant Professor of National Technical University of Athens)

WORK EXPERIENCE

2016- Present — Transportation Engineer at “APION KLEOS Construction Joint Venture”, which has been assigned with the design and construction of Athens-Korinthos-Patra Motorway in Greece

2014-2015 — Transportation Engineer in EMPARCO Project: Effective Management of PARking under Constraints. ARISTEIA II

2010- 2014 — Cooperation with the engineering consulting company “YPODOMI Consulting Engineers Ltd”, participation in transportation and drainage studies

7/2007 — Geodetic practice in Oinousses (Greece) organized by the National Technical University of Athens

PROFESSIONAL MEMBERSHIPS

2016- Present	Hellenic Institute of Transportation Engineers (HITE)
2010- Present	Technical Chamber of Greece (TEE)
2010- Present	Association of Surveying Engineers of Greece (PSDATM)

PERSONAL SKILLS AND COMPETENCES

MOTHER TONGUE	**Greek**
OTHER LANGUAGES	**English**- Certificate of Proficiency in English (CPE) of Cambridge University (2003)– Excellent knowledge
	German- Zentrale Mittelstufenprüfung, Goethe Institute (2002)- Very good knowledge
TECHNICAL SKILLS	Microsoft Windows, Microsoft Office (Word, Excel, Power Point, Access), Adobe Photoshop, Corel Draw, Internet, AutoCAD Land Development Desktop, AutoCAD Civil 3D, AutoCAD Map 2000i and new editions, ArcMap, ER Mapper, ODOS, *Programming languages*: Matlab, R, MySQL, Simple– Spatial databases, *Traffic simulation software:* TransModeler , AIMSUN

AWARDS/ SCHOLARSHIPS

2017	Thomaidio Award for Scientific Publication in International Journal:" Papathanasopoulou, V., Markou, I., & Antoniou, C. (2016). Online calibration for microscopic traffic simulation and dynamic multi-step prediction of traffic speed. Transportation research part C: emerging technologies, 68, 144-159."
2016	Thomaidio Award for Scientific Publication in International Journal:" Papathanasopoulou, V., and C. Antoniou (2015). Towards data-driven car-following models. Transportation Research Part C, Vol. 55, pp. 496-509."
2016-2017	Special Fund for Research Grants of NTUA for the PhD studies
2015-2016	Scholarship for the PhD studies by the Alexander S. Onassis Foundation
2010	Dimitriou, Konstantinou & Vasileiou Kontodimou Award for outstanding grades during undergraduate studies
2010	Thomaidio Award for outstanding grades during undergraduate studies
2004	Eurobank Award for outstanding grades of high school graduation

INFORMATION AND AWARENESS MEETINGS/ EDUCATION PROGRAMS

27/9/2017- 29/9/2017	8th International Congress on Transportation Research (ICTR 2017), Thessaloniki, Greece
4/7/2017- 5/7/2017	International Scientific Conference on Mobility and Transport (mobil.TUM), Munich, Germany
8/1/2017- 12/1/2017	96th TRB Annual Meeting, Transportation Research Board (TRB 2017), Washington, D.C.
14/9/2016- 16/9/2016	5th Symposium of the European Association for Research in Transportation (hEART 2016), Delft, The Netherlands
5/11/2015- 6/11/2015	7th International Congress on Transport Research (ICTR 2015), Athens, Greece
8/9/2015- 10/9/2015	4th Symposium of the European Association for Research in Transportation (hEART 2015), DTU, Denmark
19/5/2014- 23/5/2014	Course on "Advanced Modelling and Simulation of Transportation Networks" in Naples, Italy
3/9/2013- 6/9/2013	4th Multitude Summer School 2013 "Uncertainty in traffic simulation" at the University of Aegean and "Maria Tsakos" Foundation, organized by the European COST Action TU0903
8/7/2013- 19/7/2013	Professional education program "Business to Business- European programs" (40hrs), offered by the Educational Institute of Technical chamber in Greece (IEKEM TEE)
30/9/2013- 22/10/2013	Professional education program "Logistics management" (40hrs), offered by the Small Enterprises' Institute of the Hellenic Confederation of Professionals, Craftsmen and Merchants (IME/GSEVEE)