Traffic Data Collection and its Standardization
Traffic Data Collection and its Standardization
A nice night of October 2007, in Beijing, during the XV World Conference on ITS a number of colleagues met informally for a dinner party that spontaneously became a vivid discussion on the importance of traffic data for all types of purposes. Researchers can hardly do any progress in modeling, developing, and testing theories without suitable data, and what practitioners can do in real life is limited not only by technology but also by the availability of the required data. Quite frequently, the data and not the technologies are what determine how far we can go.

Any discussion about traffic data leads in a natural way to a discussion on the variety of traffic data sources, formats, levels of aggregation, accuracies, and so on. Consequently, we moved to talk on the initiative that Kuwahara had undertaken in his traffic laboratory at the University of Tokyo, known as the International Traffic Data Base, and thus smoothly but inexorably we came to agree that it would be convenient to organize a workshop to continue our discussion at a more formal level, share our points of view with other colleagues, listen what they had to say and, if possible, disseminate the findings in our professional and academic communities.

This was the way in which we came to organize an International Workshop on Traffic Data Collection and its Standardization that was held on September 8–9, 2008 in Barcelona, whose objectives were rooted on a reflection about: What do we need traffic data for, the dependencies on the quality, other relevant properties of the data, and their uses.

Thinking of traffic data usage, one primarily thinks of a variety of applications ranging from the most classical traffic control ones to the most advanced real-time control and management implementing modern ITS applications. All these applications are primarily based on the availability of traffic data supplied by a Data Collection system, which, equipped with more or less sophisticated technologies, provides measurements on the fundamental traffic variables – supposedly with the required level of temporal aggregation – and perhaps, when the technology allows it, additional measurement on other variables of interest depending on the type of application in which they will be used.

But, applications are supported by models and in fact the primary use of the data is to provide the input to traffic models, whose qualities strongly depend on the quality, consistency, robustness, completion and other characteristics of the data.
This book contains a selected collection of papers presented at the Workshop dealing with

- Which kind of data are available and under what conditions
- What kind of data are needed for
  - Online applications
  - Model calibration/validation
  - Safety analysis
- How reliable the sources of traffic data are
- How a standardization can accelerate developments in the field, and
- How traffic data can be accessed easily

We sincerely hope that our colleagues, professionals as well as academicians, find this book a useful contribution. If this is the case we will feel rewarded with the service provided to our community.

Summer 2009 Jaume Barceló
Barcelona, Spain Masao Kuwahara
## Contents

1 **Traffic Data Collection and Its Standardization** ........................................... 1  
   Jaume Barceló, Masao Kuwahara, and Marc Miska

2 **Data Collection, Use and Provision at the Transport**  
   Data Centre, New South Wales, Australia .................................................. 11  
   Peter Hidas

3 **Data Collection for Measuring Performance**  
   of Integrated Transportation Systems ....................................................... 25  
   Wei-Bin Zhang, Alex Skabardonis, Meng Li, Jingquan Li,  
   Kun Zhou, and Liping Zhang

4 **International Traffic Database: Gathering Traffic**  
   Data Fast and Intuitive ............................................................................. 47  
   Marc Miska, Hiroshi Warita, Alexandre Torday, and Masao Kuwahara

5 **Data Mining for Traffic Flow Analysis: Visualization Approach** ........ 57  
   Takahiko Kusakabe, Takamasa Iryo, and Yasuo Asakura

6 **The Influence of Spatial Factors on the Commuting**  
   Trip Distribution in the Netherlands ....................................................... 73  
   Tom Thomas and Bas Tutert

7 **Dynamic Origin–Destination Matrix Estimation**  
   Using Probe Vehicle Data as A Priori Information .................................. 89  
   Rúna Ásmundsdóttir, Yusen Chen, and Henk J. van Zuylen

8 **Using Probe Vehicle Data for Traffic State**  
   Estimation in Signalized Urban Networks ............................................... 109  
   Henk J. van Zuylen, Fangfang Zheng, and Yusen Chen
9 Floating Car Data Based Analysis of Urban Travel Times for the Provision of Traffic Quality ....................................................... 129
Jan Fabian Ehmke, Stephan Meisel, and Dirk Christian Mattfeld

10 A Cost-Effective Method for the Detection of Queue Lengths at Traffic Lights........................................................ 151
Thorsten Neumann

11 Extended Floating Car Data in Co-operative Traffic Management................................................................................ 161
Thomas Scheider and Martin Böhm

12 Microscopic Data for Analyzing Driving Behavior at Traffic Signals.............................................................................. 171
Francesco Viti, Serge P. Hoogendoorn, Henk J. van Zuylen,
Isabel R. Wilmink, and Bart van Arem

Index.................................................................................................................................................................................. 193
Contributors

Yasuo Asakura
Graduate School of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan

Rúna Ásmundsdóttir
Mannvit Engineering, Iceland

Jaume Barceló
Universitat Politècnica de Catalunya, CENIT (Center for Innovation in Transport), Campus Nord, Edifici C5, Jordi Girona Salgado 1-3, 08034 Barcelona, Spain

Martin Böhm
AustriaTech, Donau-City-Straße 1; 1220 Vienna, Austria

Yusen Chen
CYStone, The Netherlands

Jan Fabian Ehmke
Carl-Friedrich Gauß Department, Business Information Systems, Decision Support Group, Universität Braunschweig, Mühlenpförtstraße 23; 38106 Braunschweig, Germany

Peter Hidas
Manager, Transport Modelling, Transport Data Centre, NSW Ministry of Transport, GPO Box 1620, Sydney NSW 2001, Level 7, 227 Elizabeth Street, Sydney 2000, Australia

Serge P. Hoogendoorn
Delft University of Technology, Stevinweg 1, 2628CN Delft, The Netherlands

Takamasa Iryo
Graduate School of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan

Takahiko Kusakabe
JSPS Research Fellow, Graduate School of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada, Kobe 657-8501, Japan
Contributors

Masao Kuwahara
Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8505 Tokyo, Japan

Jingquan Li
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA

Meng Li
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA

Dirk Christian Mattfeld
Carl-Friedrich Gauß Department, Business Information Systems, Decision Support Group, Universität Braunschweig, Mühlenpfordtstraße 23, 38106 Braunschweig, Germany

Stephan Meisel
Carl-Friedrich Gauß Department, Business Information Systems, Decision Support Group, Universität Braunschweig, Mühlenpfordtstraße 23, 38106 Braunschweig, Germany

Marc Miska
Institute of Industrial Science, University of Tokyo, 4-6-1 Komaba, Meguro-ku, 153-8505 Tokyo, Japan

Thorsten Neumann
German Aerospace Center, Institute of Transportation Systems, Rutherfordstr. 2, 12489 Berlin, Germany

Thomas Scheider
AustriaTech, Donau-City-Straße 1, 1220 Vienna, Austria

Alex Skabardonis
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA

Tom Thomas
University of Twente, Enschede, The Netherlands

Alexandre Torday
TSS - Transport Simulation Systems, Pg. de Gracia 12, 08007 Barcelona, Spain

Bas Tutert
University of Twente, Enschede, The Netherlands

Bart van Arem
TNO, business unit Mobility & Logistics, Van Mourik Broekmanweg 6, 2628XE Delft, The Netherlands
Contributors

**Henk J. van Zuylen**
Delft University of Technology, The Netherlands

**Francesco Viti**
Delft University of Technology, Stevinweg 1, 2628CN Delft, The Netherlands

**Hiroshi Warita**
Planning and Environmental Department, Tokyo Metropolitan Expressway Company Limited, 1-4-1 Kasumikaseki, Choyoda-ku, 100-8930 Tokyo, Japan

**Isabel R. Wilmink**
TNO, Business unit Mobility & Logistics, Van Mourik Broekmanweg 6, 2628XE Delft, The Netherlands

**Wei-Bin Zhang**
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA

**Liping Zhang**
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA

**Fangfang Zheng**
Delft University of Technology, Delft, The Netherlands and TRAIL Research School, Delft, The Netherlands

**Kun Zhou**
California PATH, University of California at Berkeley, 1357 S. 46th Street, Richmond, CA 94804, USA
Chapter 1
Traffic Data Collection and Its Standardization

Jaume Barceló, Masao Kuwahara, and Marc Miska

1.1 Introduction

Traffic engineers are involved in transport modeling, traffic simulation, operation optimization, and the development of methods to control and analyze traffic itself. New developments of individual traffic, public transport as well as pedestrian movements are the hope to ensure mobility and accessibility in urban areas to secure mobility in the less profitable countryside, to increase safety, and to limit the effects on the environment caused by transportation. Around the globe, governments declare goals in each of the mentioned fields, mostly under the umbrella of intelligent transport systems (ITS) to develop a sustainable transportation for everyone.

Since all models, methods, and solutions developed have to represent the real world, and the studied effects are supposed to be found after implementation, all developments are highly dependent on data for calibration and validation. With data being such a necessity, it is hard to fathom that the availability and standardization of this data is not in a stage as it should be. Research is delayed with a long and costly procedure of finding and gathering data, and the data found are non-standardized, which adds the task of data formatting to the list of things to do before research can begin. So the choice is either to focus more on the theoretical research using synthetic as well as error free data or to gather, format, and process the real world data limiting the available time for the research.

Our aim is to raise the awareness for the important topic of traffic data collection and its standardization, by introducing approaches and projects dealing with data collection, processing and distribution, by exploring the needs of traffic data given research examples and to trigger the development of standardization. As a mean to this end, in September 2008, in Barcelona, Spain, we organized an International
Workshop of Traffic Data Collection and its Standardization. In the following lines of this Chapter, we give a summary of the topics discussed in this Workshop as an introduction to the selected contributions contained in this book.

1.2 Data Collection

With the ever growing traffic demand and rising challenge of controlling it, our networks are equipped with different sensors to measure the actual traffic state. This data is essential for evaluating the performance of transportation systems and for supporting the development of new approaches and technologies that address traffic problems.

1.2.1 Data Sources and Measurement Stations

The range of traffic detection devices is wide, and the major sources of traffic and transport data are listed below.

- Infrared detectors, which detect passing vehicles when a beam of light is interrupted. Active infrared detectors are additionally able to recognize temperature differences (engine heat, body warmth). Usually gathered information: aggregated flows and aggregated speeds in a few minute intervals.
- Radar detectors, which measure the presence and the speed of vehicles using the Doppler Effect. Usually gathered information: aggregated flows and aggregated speeds in a few minute intervals. Further, they may measure the height of the passing vehicle.
- Induction loop detectors, which detect vehicles entering a created electro-magnetic field by the induction of Foucault currents. With two induction loops placed closely together (commonly 1 m apart) not only the vehicle, but also its speed can be detected. Usually gathered information: aggregated flows and aggregated speeds in a few minute intervals. Vehicle types can be obtained by induction patterns and are experimentally in use.
- Ultrasonic detectors, which transmit ultrasonic sound waves instead of electromagnetic radar waves. Usually gathered information: aggregated flows in a few minute intervals plus a record of vehicle types, distinguished by their length and/or height. As for induction loops, two ultrasonic detectors placed closely together could measure vehicle speed, too.
- Video cameras, which detect vehicles when entering and exiting a road stretch. Usually gathered information: aggregated flows and aggregated speeds in a few minute intervals plus individual travel time data if license plate recognition is used.
- Probe vehicle transmitting traffic messages containing location, speed, and others at regular time or space intervals.
- Mobile data, which are recently used to gather travel time data, dynamic origin-destination relationships, and mode choice of the transport network. Bluetooth detection as a special case is emerging as a valuable technology, independent of mobile operators, providing high quality measurements – namely speeds and travel times – and with the additional advantage of Bluetooth having no privacy concerns.
- ETC (electronic toll collection) data that reveals location and passing times of the ETC infrastructures. From the data, the origin and destination of each individual vehicle and the travel time can accordingly be observed.
- Control programs of signalized intersections that enable reproduction of the actual signal timings for instance, in a simulation model.
- Variable message sign logs that give insight about the en route information drivers had, while other measurements were performed.
- Weather conditions, which have an influence on the traffic flow.
- Individual person trips have been observed to understand the whole picture of the city-wide travel pattern of people. The survey is normally conducted by the home interview and/or mailing survey which is obviously quite labor intensive, but still its sampling rate is limited.
- Origin-destination volumes of vehicle traffic are estimated based on the OD survey because OD cannot be measured directly from conventional sensing data. Normally, interviews of vehicle owners and business offices are carried out to identify trip origins and destinations as well as trip time, vehicle types, etc. The survey is also quite labor intensive and conducted only every few years in most cities. Meanwhile, the original target matrices resulting from the survey are adjusted based on link flow counts, a methodology which is widely used in practice raises a complementary problem, that of determining the optimal layout of detectors for such purposes.

The list is certainly not complete, and a growing variety of sources that process raw data to determine congestion levels, travel times, and other measures keep adding to the list. Maybe it is worthwhile to mention that a fast change of perspective is expected due to the evolution of technologies. Mobile detection is opening a completely new perspective and devices fitted with, for example, Bluetooth are making new results in a short term possible.

A more recent concern is the environmental impact of traffic. Dense urban areas are getting increasingly polluted and research aims to limit the pollution to a minimum. Key pollutants from traffic are particulate matters (fine dust and soot particles – PM), carbon monoxide (CO), nitrogen oxides (NOx), benzene and hydrocarbons (HCs) which can lead to serious health effects as follows:

- **Particulate matters**: High levels of particles have also been linked with increased hospital admissions and asthma attacks. Smaller particles can carry carcinogenic particles into the lungs.
- **Nitrogen dioxide**: May aggravate asthma symptoms. Can cause a tightening of the chest and reduced lung function. Can make airways more sensitive to allergens such as house dust mite. By disrupting the body’s natural cleansing mechanisms, nitrogen dioxide may increase the body’s susceptibility to viral infections.
- **Carbon monoxide**: Slows reflexes, impairs thinking, and causes drowsiness by reducing the oxygen-carrying capacity of the blood. Can increase the likelihood of exercise-related pain in people with coronary heart disease.
- **Benzene**: A known carcinogen which can cause leukemia.
- **Ozone**: Irritates the mucous membrane of the respiratory system causing coughing, choking, and impaired lung function, particularly in people who exercise. Other symptoms include headaches, eye/nose/throat irritation, and chest pain on deep breathing. Can make airways more sensitive to allergens such as pollen. Can also impair defenses against bacteria and viruses.

Up to now we do not see many continuous measurement stations for these kinds of pollutants. However, many countries have employed driving restrictions during high ozone concentrations and some countries measure the fine dust concentration in locations of major cities and ban cars without particle filters from using the road during measurements of high concentrations. Since measurements are mostly difficult and effects of buildings and winds cannot be taken into account in general, the measurements are mostly experimental, which results in a different basis of that kind of traffic data.

Next to pollution, governments and society are also concerned about the amount of carbon dioxide (CO₂) which is not always classed as a “pollutant,” but is connected with global warming. An instantaneous emission model could be used with traffic simulation models to evaluate the impacts of policies.

### 1.2.2 Data Storage, Provision, and Needs

Data storage, the provision of the data, and the needs of data users are major issues. In this book, the reader will find different perspectives on this subject from across the world.

In Australia, the Transport Data Centre (TDC) within the New South Wales (NSW) Ministry of Transport is the premier source of transport data for NSW. The TDC’s role is to collect, analyze, process, and provide reliable and up-to-date information on current and future travel patterns and employment and population trends. The contribution by Peter Hidas provides an overview of the data-sets collected and maintained by TDC. For each data-set, a brief summary of the purpose, the methodology, and the available outputs is presented, and some issues related to specific problems, such as privacy, data accuracy, funding, and inappropriate use of data are highlighted.

In California, significant efforts have been made to collect freeway data and to develop freeway performance measures over the past 10 years. Transit agencies have developed the capabilities for collecting transit operations data. However, detailed data for arterial highways and transit systems are still significantly lacking. Additionally, data collection and evaluation have been focusing on individual networks. California PATH Program at the Institute of Transportation Studies, University of California has devoted significant efforts to collect traffic and transit data to develop performance measures for the evaluation of integrated multimodal transportation systems and to support research of advanced technologies for traffic
and transit applications. The paper by Zhang describes the data needs, efforts by PATH to collect high quality data, and the applications of such data for the evaluation and research of ITS technologies.

The paper by Miska et al. describes International traffic database (ITDb) project. The project, funded by the Japanese Government, aims to make the access to traffic data fast and intuitive, so that research and practice can benefit from the time they gain and money they safe during the process of data acquisition. ITDb collects and provides traffic data world-wide. In the long run, ITDb tries to define a future standard for traffic data storage, and provision.

All the collected data can be quite overwhelming without the right approaches to analyze them. Kusakabe et al. propose a visualization approach for analyzing traffic flow data, later on in the book. With the improvement in information technology systems, many detectors are set up and they generate large amounts of traffic flow data. These data provide useful information, and visualization is one of the methodologies for discovering certain characteristics from large amounts of data. They developed a visualization system for long term traffic detector data (several years) and showed their results of an empirical analysis.

Of course, the data analysis in the end leads to insights and applications to improve the transportation network development and operation. In the following section, we introduce examples of data usage described in this book.

### 1.3 Data Usage

As mentioned before, collecting data is vital for traffic engineers because the data contain the information that we try to extract, use for testing existing models, and to create improved or new ones. With more and more data coming from traditional and new sources, we can have another look and test our state of the art beliefs.

Thomas et al. have estimated multimodal distribution functions for commuting trips in the Netherlands. Contrary to most studies, their function form follows directly from aggregated survey data. They have shown that a negative exponential to the power law with power equal to 0.4 gives a good description of distribution as a function of distance. The function has only two free parameters, i.e., a scaling factor and a slope. The slope of the distribution function had proven to be strongly dependent on the size of the city. They parameterized this dependence, creating a model with only a few parameters and estimated commuting trip frequencies between and within cities. By comparing these with the observed frequencies, they uncovered other spatial factors that have a strong influence on trip distribution, but that are ignored in most traditional models.

While survey data of that kind have been collected for a long time now, and still surprisingly contains new information, probe vehicle data (PVD) are rather new and widely used as shown in the following.

For most dynamic Origin-Destination (OD) matrix estimation methods, an a-priori matrix is necessary as an initial guess. The more this a-priori matrix matches the real traffic situation, the better the final outcome of the estimation
will be. Until this date, a-priori matrices have been acquired from non-current data that are potentially outdated. In order to increase the reliability of the estimated OD matrix, Chen et al. used PVD for the estimation of a-priori matrices. In their contribution, methods to fuse the PVD information with traditional methods of a-priori matrix estimation and mapping are suggested and tested. The PVD used in the paper come from taxis that are driven in the city center of Chengdu in southwest China.

With data sets being rare, the same data set has been used by van Zuylen et al. for delay estimations at signalized intersections. By matching the experienced delay with delay patterns for different traffic states, they developed a method for dynamic state estimation by applying a Markov chain model that describes a traffic flow at signalized intersections. The approach adopted in their paper is interesting. As mentioned by many previous studies, including authors’, the phenomena at signalized intersections are definitely affected by stochastic processes unless they are not heavily oversaturated. Therefore, estimating the distribution of travel time instead of travel time itself seems to be a rational approach.

Taxis seem to be a reasonable accessible source for probe vehicle data, and so Ehmke et al. describe in their article how to use the huge amount of data collected from taxis’ GPS devices in the city of Stuttgart, Germany. Their data consist of localization ID, date stamp, and speed. From this, the method presented proposes to compute mean characteristics for 24×7 periods for a typical week, with a mean speed (or speed index) for each link of the Stuttgart’s network. The paper summarizes the main necessary steps when one wants to use individual speed data.

As Neumann points out in his contribution, floating car technology is mostly used to get information about travel times which, however, are more attractive to navigation and route guidance. Concerning traffic management applications such as traffic light control, it is more likely to know about exact delay times induced by traffic lights or the queue lengths. So, he developed a new method based on existing floating car systems, which in its first version provides estimations of queue lengths at traffic lights. This would allow for an area-wide urban traffic monitoring without any additional infrastructure which is capable for typical traffic management applications.

An overview of the extended floating car (xFCD) data system that could be utilized for traffic management is given by Scheider et al. Their paper first outlines the system and then discusses the usefulness of the xFCD for not only efficiency, but also safety improvement. The paper also raises several issues for the system deployment, such as economic and legal aspects of the data, and includes the necessity of the data management so as to combine other data from different sources.

But probe vehicle data are not the only method to acquire just recently available individual vehicle data, needed for modeling realistic driving behavior. Viti et al. describe in full detail the image processing technology and the design of the data collection process to collect empirical evidence on traffic behavior at signalized intersections, using a camera installed on a high-rise building. Of particular interest in this contribution are the specifications on the data requirements and the choice of the study area. Requirements should influence standards for traffic data collection and storage, but are there standards?
1.4 Data Standards

Transport authorities are storing their data in the most convenient format for their operations, and that is mostly dictated by the systems they are using, or better, the collection of systems they are using and not driven by any standards.

Most standards in the field are concerned with the communication of different systems, like sending information to variable message signs for example, but not for storing data or even handling data inside an application. This confusion of standards triggered the government of the United Kingdom to finance a project that created a catalog of all variables used in the governmental systems. The outcome was the ITS Registry that now enables developers to integrate new systems into the existent instead of having an additional software package incompatible with the rest of the systems. This was the first step in organizing the system structures used into something manageable that opens the chance to create commonly used storage to avoid redundancy and to limit data transfer.

Other standards for mostly data exchange between systems are:

- The Traffic Management Data Dictionary (TMDD),
- The P1512 Incident Management Data Dictionary (P1512-IEEE),
- Traffic Model Markup Language (TMML),
- Geographic Markup Language (GML),
- Universal Traffic Data Format (UTDF),
- Digital Geospatial Metadata (FGDC-STD-001-1998),
- European traffic information exchange standard (DATEX2), and others.

This, however, is not enough to ensure a standardized way of data storage and access that would be so necessary for effective research and development. Therefore, one will find, with every new data set from a different provider, a new format that has to be studied in advance to be able to use the information. Several data platforms existing around the world and are all born out of in-house solutions which have the disadvantage that they have not taken the general needs into account, but only the specific needs of the developers that might not be enough for others in the community. One could argue that the data could be standardized if research is performed within a small limited space. However, in many cases, the authority of even a small scale network is separated in, for example, national roads, provincial road, and local roads. Most of the time, we therefore have to contact several different authorities to acquire the necessary data.

1.5 Data Availability

As stated before, there is a huge variety of data sources and modern networks are equipped with them. In other words, all the inputs for testing new developments in traffic engineering are given. But are they? Answering this question appears not to be as straightforward as it sounds. Yes, the availability of data is much more limited
than it could be, but what are the reasons? In the following, we are trying to investigate the problems of data provision from the perspective of the transport authorities in different categories.

### 1.5.1 Data Storage and Usage

Depending on the network size and the amount of measurement points, the daily amount of data can be huge and transport authorities have to think about the cost benefit relationship to decide what data to store. It is obvious that, for instance, raw pulse data from loop/ultrasonic detectors contain a lot of useful information, but the data amount is equivalent to the number of vehicles passing the measurement point and that it is much more efficient to store aggregated values, such as volume and average speed instead. This means that the data storage can be easily calculated and the most beneficial aggregation period can be determined. Using these aggregated values to determine congestion levels, to identify bottlenecks, and to observe the overall performance of the network is beneficial for the transport authority and therefore worth the costs of storage. Control systems, forecasting tools, incident detection systems, and decision support application in use now are designed to use these values as input. So from the side of the transport authority, it is questionable when researchers require them to collect and store more data.

From the researcher’s point of view however, the systems built and used today came after the detection systems. The data were available and methods have been developed to extract the most useful information from these data sources. That implies, that if systems should become more intelligent, better performing and more reliable, we have to test and experiment with new data sources, lower aggregation periods and fuse data from different sources to get a more accurate picture of what is happening in the network. This, however, can only be done if the data are collected and provided to the researchers and developers. Looking at this scenario once again, one will recognize that the initiative has to come from the transport authority by investing money for data collection and storage without a guarantee of a benefit. This is why mostly such kind of data will only be provided temporarily. But, one does not have to go that far and it can be even difficult to receive status quo data that is available, and we will pick this up in the category of provision costs.

### 1.5.2 Privacy Concerns

For data that relates to individuals, there is an additional hurdle to take to get access to traffic data: privacy concerns. Video camera detection, number plate recognition, probe vehicles, ETC payment systems, and so on are all collecting information of the individual. This information is among the most valuable information, but the concerns of its misuse are very high. Transport authorities have to protect the users
as long as they have not declared their personal data to be used, which is a common practice for cell phone movements, but not for vehicle movements.

The first thought of allowing the usage of such data is to anonymize the individual with a random number that is not traceable in any way. However, certain parties claim that such anonymous data from different sources could open possibilities to track individuals even without their privacy information. Consider the timestamp of an ETC entry gate: just a few physical observations could be enough to link an individual to the random number and therewith would allow a continuous tracking.

### 1.5.3 Provision Costs

A further important issue is the cost of data provision. Transport authorities are collecting data to feed their systems and integrated tools, but the extracting of data and its provision by any kind of media is not their core business. This means that providing data involves man power to gather the requested data, possibly from different systems, and, without having designated personnel to do so, the disruption of operations. However, the request might not come in frequently enough to create a position for this task, especially, since normally, this work will remain unpaid.

A few thoughts about this will reveal an easy solution to this problem. If the traffic control systems used would have been designed to commonly use a single data storage facility, it would be easy to extend the system with a data provision tool. Such a tool could break down the workload for the data providing task at most to several minutes without disrupting the operation. However, usually, the systems in traffic control centers are extended over during the time without re-engineering and the basis for such an easy way out is not given.

### 1.5.4 Other Factors

Several other factors were discovered during talks with transport authorities and we would like to introduce some of those briefly. There seems to be an issue about the value of traffic data. While traffic data is mostly collected using tax money, the transport authorities are not selling the data, and they do not want someone else to make a profit out of it. This goes so far, as to make the data available to the public and mirror the data to a sponsored website, so that the owner of the website would indirectly make a profit from the data that is already paid for.

### 1.6 Final Thoughts

After discussions and sharing the experience of difficulties in data collection and gathering, it is clear that both sides, data provider and data users, have issues that need to be resolved to make research and development in traffic engineering more efficient.
Data providers have to consider the benefits of disclosing their data and to evaluate the potential of feedback they will receive from research being done with their input. Certainly, transport authorities will always have to fear that research unveils problems in the operation, but this also includes that the problems, if any, will be detected faster and studies can show and evaluate solutions for these problems. Research activities using the data have not only pointed out deficiencies in the system, but they have also often pointed out low cost adoptions to make controls work as they have been intended to. New projects by the authority have now got an additional decision support and effects of changes are well investigated. Transparency means that users will understand the problems and issues and the acceptance for certain measures might be higher as a result, since the users have a chance to understand the background.

That however solves just a part of the problem and technical help from developers and researchers should be given in return. Disclosing data has to become an easy task if it takes minimum time and does not cause any disruption of operations. Standards have to be established, so that various sensing systems store data in compatible ways and with easy access. Further, these standards have to be flexible enough to cope with future developments in sensing techniques and new data sources to avoid running into the same problem again and again.

The benefits of such changes are obvious. Research can be more effective, control strategies can be evaluated faster, and so on with all the necessary inputs given. Thinking more globally, it also opens possibilities for internationally comparable efforts to tackle our transportation problems. New methods and algorithms could be tested with transport data from all over the world – a task that is nowadays too time consuming. The internet gives us the possibility to access knowledge from around the globe in no time. Real time traffic information helps us to navigate and find best possible routes and is accessible in all developed countries. It is available because it is seen as a service, but the service of disclosing historical data to improve our status quo in research and development seems mistakenly recognized as not being beneficial enough.
Chapter 2
Data Collection, Use and Provision at the Transport Data Centre, New South Wales, Australia

Peter Hidas

2.1 Introduction

The Transport Data Centre (TDC) within the New South Wales Ministry of Transport is the premier source of transport data for NSW. TDC’s role is to assist those involved in transport and land use planning to make informed decisions by collecting, analysing, processing, and providing reliable and up-to-date information on current and future travel patterns and employment and population trends. This information is used by Government and private sector clients for the evaluation of all major transport infrastructure developments, and strategic and service planning in NSW.

TDC staff have expertise in data collection, analysis and reporting, project management, transport modelling, employment and travel forecasting, population projections, and geographic information systems applications.

Information available from TDC covers the travel patterns of the resident population, transport infrastructure and service networks, commercial vehicle travel, employment and population data, forecasts, and geographic information. It is available in a range of formats to suit users’ needs. These are standard “off the shelf” products ranging from printed summary publications and pre-defined electronic data tables available through the internet. Alternatively, TDC Client Services can provide a customised data request tailored to users’ specific data needs.

This paper provides an overview of the datasets collected and maintained by TDC. For each dataset, a brief summary of the purpose, the methodology, and the available outputs are presented; we then highlight some issues related to specific problems, such as privacy, data accuracy, funding, and inappropriate use of data.

P. Hidas (✉)
Manager, Transport Modelling, Transport Data Centre, NSW Ministry of Transport,
GPO Box 1620, Sydney, NSW 2001, Australia
Level 7, 227 Elizabeth Street, Sydney, 2000, Australia
e-mail: peter.hidas@transport.nsw.gov.au
2.2 TDC Datasets and Methodology

The Transport Data Centre maintains a number of core datasets related to travel, demographic, and employment trends in NSW. This section lists these datasets, the type of information held in each, and outlines the main methodological issues which will be of interest to data users.

2.2.1 Household Travel Survey (HTS)

The Household Travel Survey (HTS) is a continuous survey of the personal travel patterns of a large sample of residents of the Sydney Greater Metropolitan Area (GMA), which covers Newcastle Statistical Sub-division, Sydney Statistical Division, and the Illawarra Statistical Division (Fig. 2.1). The survey is financed by the NSW Ministry of Transport with contributions from the NSW Roads and Traffic Authority (RTA). The HTS provides annual travel and demographic data. The data are collected by face-to-face interviews with every resident of selected households in occupied private dwellings within the GMA (TDC 2008a).

The HTS sample is selected for three annual waves, based on a stratified, multi-stage cluster sampling method with sample sizes for Statistical Local Areas (SLA) determined using the optimal allocation method. The sample for the HTS was allocated to achieve a relative standard error (RSE) of total trips of approximately 10% at the 95% confidence interval for each SLA.

- Every SLA is sampled each year
- All Travel Zones (TZ) are sampled over the 3 years
- Census Collector Districts (CDs) within TZs are selected using probability proportional to size
- Each CD is divided into blocks of approximately 50 dwellings each
- One block is randomly selected, then a random start point is generated, and every seventh dwelling is selected from the start point until seven dwellings have been selected
- The seven clustered dwellings are randomly allocated a different day of the week so that each of the day of the week is represented
- Household members then record their travel over a 24 h period for this allocated day

With 3,000–4,000 fully responding households per annum, this method provides a degree of data reliability at the SLA level acceptable after three waves, plus an even temporal and geographic spread.

Each household is allocated a single travel day, Monday to Friday. At the end of a year, each day of the year will be represented in the data. Approximately 8,500 people in 3,500 households are surveyed every year. The interviews are conducted by the same contractor since the beginning of the survey in 1997. The interview questions relate to the following key variables:

- Household: household and family type, dwelling structure, number of vehicles and bicycles
Data Collection, Use and Provision at the Transport Data Centre

Person: age, sex, employment status, income
Vehicle: make, model, fuel type, ownership
Trip: origin, destination, purpose, mode, time, costs

This makes it possible to examine travel patterns for different population groups, for example, to compare the travel patterns of people who live in units versus a house or those who do not have a license or car with those who do.

The sample data collected by the HTS are expanded (weighted) to represent the travel patterns of the whole population in the survey area, using information on households and individuals from the Australian Bureau of Statistics (ABS), including the Census of Population and Housing and annual ERP (Estimated Resident Population). Annual estimates are produced based on three waves to represent the population as on June 30 of each year. The data are maintained in a large relational database with separate tables for households, vehicles, persons, and trips. These tables can be

Fig. 2.1 The greater metropolitan area in the state of NSW
linked by unique ID variables in each record. No identifying information is held in the dataset, and unit record data is not made available to users. All geographic data are geocoded and allocated to a geographic unit and no address information is held in the dataset.

From this database, TDC provides standard output tables by geographic distribution on counts of trips by mode, by purpose, persons, households, and vehicles, weighted to the total population for an average day (weekday or weekend day). The geographic unit may be Travel Zone or Statistical Local Area (SLA), or aggregation of these. A summary of key transport indicators is presented in Table 2.1. Alternatively, TDC can provide detailed cross-tabulations and/or subselections of the above variables on request for a fee. As the HTS is a sample survey, data broken down by a number of variable categories, or at a very fine geographic level can be subject to high standard errors.

HTS data are used widely across the transport portfolio, as well as by other NSW and Federal Government Departments, Local Councils, Universities, and the private sector for monitoring travel trends, estimating patronage demand for infrastructure projects, strategic planning, policy development, and local transport studies.

2.2.2  Journey to Work (JTW)

Information on where a person usually works, and how they travel to work on census day, is collected every 5 years as part of the Australian Bureau of Statistics (ABS) Census of Population and Housing covering all states of Australia. The data set that results from these work-travel related questions is referred to as the Journey to Work (JTW) data set. The census is a complete enumeration survey via a self completion census form. The ABS provides information on the JTW between trip origin (home location) and destination (work location) by SLA of LGA (Local Government Area), method of travel, plus industry and occupation category of the person (TDC 2008c).

In New South Wales, TDC adds value to the JTW data by geocoding all origin and destination addresses and allocating them to Travel Zones within the GMA. From this dataset, TDC provides standard output tables on the following variables:

- Employment: industry, occupation, mode of travel to work, hours
- Person: age, sex, income
- Household: household/family type, dwelling structure, number of household vehicles

The geographic level of data distribution may be SLA, Travel Zone, or aggregation of these. Historical data are available for 1961, 1966, 1971, 1976 (SLA/LGA level O/D data), 1981, and every 5 years since 1991 (Travel Zone level O/D data). These standard tables and customised data requests are sold by TDC under license to ABS.
### Table 2.1  Key transport indicators, GMA 2005

<table>
<thead>
<tr>
<th></th>
<th>Sydney</th>
<th>Newcastle</th>
<th>Illawarra</th>
<th>GMA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Population</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persons (‘000)</td>
<td>4,191</td>
<td>503</td>
<td>409</td>
<td>5,103</td>
</tr>
<tr>
<td>Households (‘000)</td>
<td>1,545</td>
<td>196</td>
<td>156</td>
<td>1,897</td>
</tr>
<tr>
<td>Average household size</td>
<td>2.7</td>
<td>2.6</td>
<td>2.6</td>
<td>2.7</td>
</tr>
<tr>
<td><strong>Vehicles</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private vehicles (‘000)</td>
<td>2,312</td>
<td>308</td>
<td>246</td>
<td>2,876</td>
</tr>
<tr>
<td>Vehicles per household</td>
<td>1.50</td>
<td>1.58</td>
<td>1.58</td>
<td>2.876</td>
</tr>
<tr>
<td><strong>Total travel</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trips, weekday (‘000)</td>
<td>15,737</td>
<td>2,008</td>
<td>1,594</td>
<td>19,440</td>
</tr>
<tr>
<td>Trips, weekend (‘000)</td>
<td>13,703</td>
<td>1,675</td>
<td>1,266</td>
<td>16,739</td>
</tr>
<tr>
<td>Av trips per person, weekday</td>
<td>3.75</td>
<td>3.99</td>
<td>3.90</td>
<td>3.81</td>
</tr>
<tr>
<td><strong>Distance travelled</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total distance travelled (‘000 km)</td>
<td>147,636</td>
<td>20,725</td>
<td>17,603</td>
<td>185,957</td>
</tr>
<tr>
<td>Distance travelled per person</td>
<td>35.2</td>
<td>41.2</td>
<td>43.1</td>
<td>36.4</td>
</tr>
<tr>
<td>Av trip length (km)</td>
<td>9.4</td>
<td>10.3</td>
<td>11.0</td>
<td>9.6</td>
</tr>
<tr>
<td>Total VKT (‘000 kms)</td>
<td>82,729</td>
<td>13,166</td>
<td>10,842</td>
<td>106,807</td>
</tr>
<tr>
<td>VKT per person (kms)</td>
<td>19.7</td>
<td>26.2</td>
<td>26.5</td>
<td>20.9</td>
</tr>
<tr>
<td><strong>Trip duration</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Av. work trip duration (min)</td>
<td>33</td>
<td>24</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td>Av. non-work trip duration (min)</td>
<td>18</td>
<td>15</td>
<td>16</td>
<td>17</td>
</tr>
<tr>
<td><strong>Purpose of travel (% of total trips)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social/recreation</td>
<td>22.9</td>
<td>23.4</td>
<td>23.3</td>
<td>23.0</td>
</tr>
<tr>
<td>Serve passenger</td>
<td>18.2</td>
<td>17.6</td>
<td>19.1</td>
<td>18.2</td>
</tr>
<tr>
<td>Shopping</td>
<td>15.8</td>
<td>17.6</td>
<td>16.0</td>
<td>16.2</td>
</tr>
<tr>
<td>Commuting</td>
<td>15.1</td>
<td>11.6</td>
<td>12.1</td>
<td>14.4</td>
</tr>
<tr>
<td>Other work related travel</td>
<td>8.9</td>
<td>8.3</td>
<td>8.8</td>
<td>8.8</td>
</tr>
<tr>
<td>Education/ childcare</td>
<td>8.4</td>
<td>8.1</td>
<td>9.1</td>
<td>8.4</td>
</tr>
<tr>
<td>Personal business</td>
<td>7.8</td>
<td>11.3</td>
<td>9.4</td>
<td>8.3</td>
</tr>
<tr>
<td>Other</td>
<td>3.0</td>
<td>2.1</td>
<td>2.2</td>
<td>2.8</td>
</tr>
<tr>
<td><strong>Mode of travel (% of total trips)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vehicle driver</td>
<td>48.3</td>
<td>56.6</td>
<td>54.7</td>
<td>49.7</td>
</tr>
<tr>
<td>Vehicle passenger</td>
<td>21.1</td>
<td>24.2</td>
<td>23.6</td>
<td>21.7</td>
</tr>
<tr>
<td>Train</td>
<td>4.8</td>
<td>0.7</td>
<td>1.7</td>
<td>4.1</td>
</tr>
<tr>
<td>Bus</td>
<td>5.6</td>
<td>3.6</td>
<td>3.6</td>
<td>5.2</td>
</tr>
<tr>
<td>Walk only</td>
<td>17.9</td>
<td>12.8</td>
<td>14.5</td>
<td>17.0</td>
</tr>
<tr>
<td>Other modes</td>
<td>2.3</td>
<td>2.1</td>
<td>1.9</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Unless otherwise stated, estimates are for an average weekday (HTS 2005)

JTW data from the latest census in 2006 are still being processed, due to be released within the next few months.

JTW data supplied by TDC may produce slightly different counts to those obtained directly from ABS for the same geographic level due to:

- ABS confidentializing process (randomization of small cells due to privacy concerns)
• Further validation and adjustment of the data by TDC (e.g., allocating incorrect or incomplete destinations to correct addresses)
• Imputation (i.e., random allocation) of unknown destination counts across zones within an SLA to eliminate locality “dump” codes

When using JTW data, it is also important to understand that the recorded information may be specific to the census day. For example, the travel mode is the mode(s) used to travel to work on Census day, which may be different from the mode(s) usually taken by the person. The origin of the trip refers to the place where the person spent Census night, which may be different from the person’s usual residence. Also, the destination represents the employer’s address, which may be different from the usual workplace.

2.2.3 Commercial Transport Study (CTS)

The Commercial Transport Study (CTS) provides estimates of commercial vehicle trips between origin–destination zone pairs within the Sydney Greater Metropolitan Area (GMA). The 2002 CTS marks the release of a second set of base year estimates from the CTS, originally produced for 1996. The 2002 CTS output comprises trip data for three separate commercial vehicle types – rigid trucks, articulated trucks, and light commercial vehicles, by day, type, and time-of-day. This data is in the form of origin by destination trip tables (matrices) at the TDC travel zone level. The 1996 “average day” estimates have also been updated, but without a day-type/time-of-day breakdown. The data available comprises:

• 1996 (revised data) – for “average day” only
• 2002 “average day”, “average weekday”, “average Saturday” and “average Sunday”
• 2002 “time-of-day” data for weekdays and weekend days:
  • AM peak – between 7AM and 9AM (2 h)
  • Midday – between 9AM and 3PM (6 h)
  • PM Peak – between 3PM and 6PM (3 h)
  • Night – between 6PM and 7AM (13 h)

The 2001 travel zone system used is based on the TDC’s Strategic Travel Model (STM) zone numbering system. This uses a modified numbering for the same zone coverage to improve model processing efficiency. Five additional CTS “zones” represent external trip ends at the boundary of the GMA (TDC 2008f).

2.2.3.1 Overview of the Estimation Procedure

The CTS estimation consists of separate processes for the estimation of heavy and light vehicles. The heavy vehicle estimation is driven by three key inputs, which are
consolidated in the MVESTM matrix estimation module of the TRIPS transport modelling software to produce optimal trip tables:

- Trips ends are generated using the FDF FreightInfo tonnages within and between five FDF regions in the GMA. These tonnages are converted to trip ends using loading, backloading, and trip-chaining factors derived from the TDC 2002 Industry Survey. The resulting trips are then distributed to the travel zones using the Journey-to-Work employment data.
- A seed (prior) matrix based on a trip table derived from the 1991 CVS was used for the 1996 estimation to guide the distribution of trips across travel zones during the matrix estimation. For 2002, the prior matrix used was the estimated 1996 matrix.
- Heavy vehicle screenline counts were also a major input to the matrix estimation process. In 1996, there were no available actual heavy vehicle counts; therefore, most of the counts were simply estimates of likely proportions of rigid and articulated trucks to available AADT counts. During the reestimation of 1996 CTS data, TDC revised some count estimates based on the 2002 Classified Vehicle Count Study. However, it was not possible to assess the degree of improvement in the revised estimates as differences in the level and route choices between 1996 and 2002 could not be determined.

For the estimation of Light Commercial Vehicle (LCV) trips, an alternative approach was required due to the lack of information about light commercial vehicles providing services, rather than carrying commodities. These service vehicles are excluded from the FreightInfo tonnage data, and traffic count data does not distinguish light commercial from passenger vehicles. Without accurate LCV traffic counts, there is little purpose in using the MVESTM matrix estimation module.

LCVs are separated into commodity-carrying or Light Goods Vehicles (LGV) and Service Vehicles (SV), that is, those providing services but not carry goods. Trip ends for LGVs are produced in the same way as those for heavy vehicles, using the FreightInfo tonnage data and trip conversion process. For the SVs, data on trip generation rates for SVs to households and businesses was obtained from TDCs Service Vehicle Attraction Rate Study (SVAR). These rates were applied to household and employment distributions to obtain trip end estimates.

### 2.2.3.2 CTS Validation

As the CTS process is designed to use all available data, the availability of independent data for the validation of output is limited. CTS estimates were compared with data from the 1991 CVS and the Australian Bureau of Statistics (ABS) Survey of Motor Vehicle Usage (SMVU), to assess the consistency with other data sources. Although 1991 CVS data is also an input to the process, it acts as a seed matrix and is given less significant weighting in the process than the other inputs.
2.2.3.3 Scope of the CTS

In the CTS, “trip” is restricted to movements directly related to the movement of commodities, including trip-chaining and backloading components. Trips that are not represented in the CTS include those made by vehicles not counted in the FreightInfo data, such as waste disposal vehicles. Trips for a commercial purpose not related to commodity-carrying (e.g., driving to a meeting by truck or travel to/from home to start or finish work) are also not counted in the CTS.

2.2.3.4 Recent Developments – The New Freight Movement Model

There are quality issues with some of the external inputs used in the CTS model. TDC is currently developing a new Freight Movement Model (FMM) to minimize reliance on these inputs.

The framework underlying the Sydney FMM is made up of a number of integrated models: Production Models that estimate tonnes of road freight on the basis of employment, gross state product, and ABS freight movement data; Distribution Models that estimate distribution patterns on the basis of accessibility between different areas within and beyond the GMA, and taking into account known movements between different industry segments; Loading Models that estimate the number of rigid and articulated vehicle trip movements given the number of tonnes; and Assignment Models that estimate the volume of freight vehicles on sections of the GMA road network.

The Sydney FMM developed for this project is made up of 61 SLA-based internal freight areas, seven special generator freight areas (such as ports), and 17 external freight areas (country NSW and interstate regions). The model has 13 production-based industry classes and nine redistribution-based industry classes. Parameters used for the Sydney FMM have been a combination of the Melbourne FMM parameters, and new parameters derived from the surveys conducted in Stage 2 of this project.

2.2.4 Travel Zone Population and Employment Forecasts

In New South Wales, the Department of Planning (DoP) is responsible for producing the official New South Wales population projections for State government planning purposes. Projections are first produced at a State and regional level and then developed for each statistical local area (SLA) within each region.

The Transport Data Centre produces population forecasts at the Travel Zone level across the entire Greater Metropolitan Area (GMA) by age and sex at 5-yearly intervals from the 2001 base year till 2031, as inputs to other modelling processes, in particular TDC’s Strategic Travel Model. These forecasts are compatible with, but should be differentiated from, the official population projections, which
are produced by the NSW Department of Planning and subscribed to by the NSW Government (TDC 2008d).

The July 2007 Population Forecasts is the latest set of TDC travel zone population forecasts. Key Inputs to the July 2007 Population Forecasts:

- ABS 2001 Estimated Resident Population by CD
- November 2006 TDC TZ Population Forecasts
- Official NSW 2005 SLA Population Projections (NSW Dept Planning)
- 2004 Metropolitan Development Program (NSW Dept Planning)

The population projections are distributed across the travel zones within each SLA using a dwelling stock model. This model produces dwelling numbers within each travel zone and then converts these to population counts based on assumptions about:

- Household size
- Dwelling commencements in established areas and lot releases in greenfield areas for both multiunit and detached dwellings
- Vacancy and replacement rates
- Occupancy rates

The model adjusts dwelling demand in each Statistical Local Area iteratively until the totals match the official SLA population projections by which they are constrained.

Employment forecasts at travel zone level are estimated using the TDC’s Small Area Employment Forecasting Model (SAEFM). These forecasts are provisional small area estimates, which are normally used as inputs in other modelling processes such as in the Sydney Strategic Travel Model (TDC 2008e).

Forecasts to 2031 at 5-yearly intervals from the 2001 base year are available for each travel zone (TZ) in the entire Greater Metropolitan Area (GMA).

The November 2006 Employment Forecasts is the latest set of travel zone employment forecasts available. These estimates are based on TDC’s November 2006 Small Area Population Forecasts for the GMA, which are in turn based on the New South Wales Department of Planning (DoP) 2005 Population Projections at Statistical Local Area (SLA) level. In addition, updated labor force participation and unemployment rate projections have been used in the November 2006 employment forecasts.

The DoP population projections for the GMA are used as the basis for calculating workforce projections for the region. This is done by applying the projected labor force participation and unemployment rates to the projected population. The resulting projections of the workforce within the GMA (at 5-yearly intervals, corresponding to Census years) are used as control totals for subsequent forecasts of employment at lower geographical levels.

Once projections of total employment for the GMA are estimated, this total region employment is then disaggregated using three forecast review modules in SAEFM: the Industry Review module, the Regional Review module, and the Job Node Review module.
The Industry Review module provides the capability to forecast the share each industry category has of the total number of jobs at the GMA level at each 5-yearly period. There are 27 Industry categories used in SAEFM.

The Regional Review module provides the capability to forecast the share each subregion has of the total number of jobs in each industry (as previously determined in the Industry Review module) at each 5-yearly period.

The Job Node Review module provides the capability to forecast the share each job node has of the total number of jobs in each Industry × Subregion (as previously determined in the Industry Review and Regional Review modules) at each 5-yearly period. There are 94 Job Nodes used in SAEFM.

For some areas, the review modules cannot properly take into account known or expected changes in employment. As a result, SAEFM also includes a “New Developments” module, which allows for adjustments to forecasts for individual zones.

2.2.5 Strategic Travel Model (STM)

The Sydney Strategic Travel Model (STM) is designed for the evaluation of transport policy and planning options within the Sydney Greater Metropolitan Area. The STM can be used to estimate the effects on future travel demand of

- Major infrastructure changes
- Different population and/or employment growth and distribution scenarios
- Various Travel Demand Management scenarios, such as alternative public transport, parking, and pricing policies

The STM is a four-step gravity type multimodal travel demand model implemented in the Emme transport modelling software platform. The study area covers the Sydney GMA, divided into 1,129 origin–destination Travel Zones. It aims to model the travel demand on an average weekday in the base year (currently 2001) and every 5 year for the next 30 years: 2006, 2011, 2016, 2021, 2026, and 2031.

The model is developed, calibrated, and continuously updated using the travel behavioural characteristics obtained from the HTS and JTW databases and the population and employment projections at the travel zone level. The behavioural models of frequency of travel, mode, and destination choice are using population segmentation into 4 income groups, 4 employment type, and 8 car availability groups, giving 128 segments altogether. The models also consider the prediction of license holding and car ownership, conditional on license holding (TDC 2008b).

The STM produces zone level origin by destination base year trip matrices as well as forecasts by travel mode (car driver, car passenger, train, bus, taxi, bicycle, walk), purpose (work, nonwork), and time of day (am peak, mid-peak, pm peak, evening). It can also provide future estimates of travel times, average speeds, public transport patronage, station usage, number of vehicles using each
Data Collection, Use and Provision at the Transport Data Centre

road link, and vehicle-km of travel (VKT) for various alternative scenarios. Besides the standard model outputs by travel zones, customised outputs are provided on request for a fee.

It is important to note that the STM is designed for broad strategic level demand forecasting. The following limitations must be considered in interpreting the outputs of the model:

- External trips are not included, e.g., trips from Lithgow to Sydney or Taree to Newcastle.
- Intersection-related factors such as traffic signal phasing, adequacy of turning lanes, and the interaction of other road links using the same intersection are not specifically considered.
- Each potential public transport user within a travel zone has the same access time. This limits the evaluation of bus routes to strategic corridors rather than specific routes.
- The nonwork trips are calculated using an expansion process based on commute travel.
- The model is not calibrated for toll delays.
- It does not incorporate any procedures to model the effects of future changes in peak-spreading, a likely behavioural response in areas where congestion increases.
- The model does not have a capacity constraint for public transport.
- It does not allow for ramp up when assessing the impacts of the introduction of new services.

The STM is in continuous development to improve output accuracy and reliability. Current model improvements in progress include the implementation of a more detailed travel zone system (more than doubling the number of travel zones in the study area), the reestimation of the model based on the latest 2006 JTW data, and the implementation of more accurate procedures for the prediction of nonwork travel purposes. While the STM is an excellent strategic modelling tool, it is unable to inform some local area, yet strategically important, transport issues requiring more detailed operational model specifications, such as how to get the maximum potential of bus operations in the Sydney CBD with minimal effect on traffic, or the effects of potential light rail services on bus and private vehicle traffic. In order to improve the quality of STM outputs provided for our clients and stakeholders, both at strategic and operational levels, TDC has started developing a microsimulation model of the Sydney CBD, using the AIMSUN simulator platform, directly linked to the STM Emme model. Stage 1 of the microsimulation model has been used to provide advice on alternative management options of bus services between the Sydney Harbour Bridge and CBD. In parallel with the continued extension of the AIMSUN microsimulation model, TDC has also started experimenting with mesoscopic transport models, using the DYNAMEQ (INRO) and AIMSUN meso software platforms.
2.3 TDC Use of Traffic Count Data

While TDC is primarily a provider of transport data, in developing its datasets, TDC is also a user of traffic count data. Traffic volume counts are used for the estimation of commercial vehicle trips, and for the calibration and validation of the STM and microsimulation models.

The New South Wales Roads and Traffic Authority (RTA) undertakes strategic traffic volume surveys on a 3 year repeating cycle, which covers the whole of the State of NSW by Region. Annual Average Daily Traffic (AADT) data is available free of charge for certain years and regions from the RTA website (http://www.rta.nsw.gov.au). TDC obtains more detailed, hourly and daily traffic counts from the RTA as part of a service agreement between the two organisations. TDC staff then analyse these data and develop appropriate statistics for calibration and validation purposes. While sometimes there are concerns about the reliability of the RTA counts, TDC can only address these issues by personal consultation with RTA staff.

2.4 Issues Related to Data Access and Use

The Transport Data Centre provides a wide range of transport-related data that is available to the public in a range of formats to suit users’ needs. There are standard “off the shelf” products ranging from printed summary publications and predefined electronic data tables. Most of these are published on the TDC website: http://www.transport.nsw.gov.au/tdc, some available free of charge, others can be purchased for a fee. In March 2008, TDC implemented a new pricing and data dissemination policy. Clients who wish to purchase data from TDC now have the option to pay a fixed registration fee to get complete access to TDC’s various datasets – data that are normally for sale and not already available for free from the TDC website. This new policy provides substantial savings to users and allows easier access to TDC data. The annual access fees, depending on the type of organisation, are

- Government agencies: $300
- Private companies: $1,000
- If the company wishes to commercialise the data: $3,000
- Not-for-profit companies, libraries, academic, and research institutions: free

Alternatively, on request, TDC Client Services can provide a customised data set tailored to specific needs of the client.

Users of TDC data include all levels of government, research, and academic institutions, libraries, private consultants in all sectors and not just transport, such as planning, insurance, banking, and finance sectors. There is great interest in using the TDC data, especially at state and federal government level. The following list includes some recent important applications of TDC data by NSW government agencies:

- Transport Infrastructure Planning: Roads (M7, Lane Cove Tunnel, M4 East), Rail (Epping–Chatswood Rail, North West Metro, South West Rail Link), Bus
2 Data Collection, Use and Provision at the Transport Data Centre

(Liverpool–Parramatta Transitway, North West Transitway), Interchanges (Bondi Junction, Parramatta, Chatswood)

- Transport Policy: Bus reform, TravelSmart, Setting of public transport fares, Integrated ticketing and fare reform, Rail market expansion study, Travel demand management, Monitoring of State Plan Targets

There is somewhat less demand from local councils and private companies, but this is mostly due to lack of knowledge by people working in these organisations of what is available – once they know more about the available data, interest grows rapidly. The TDC is continuously working to further publicize its products and services.

TDC has a strict policy to comply with privacy policy. Raw data (e.g., unit records from the HTS, which would allow identification of an individual or organisation) or exact geocoded locations are never made available to the public or other organisations. During the 12 years of operation of the TDC, there have never been any problems related to privacy.

Inappropriate usage of the data by nonexperts is prevented by TDC’s policy to the extent possible: all data are processed and analysed internally by TDC staff to make sure that the published standard tables and summaries are correct, clearly explained and easy-to-understand. TDC staffs put significant efforts in making sure that the data provided are correct and free of any deficiency. While there are some limitations of the data, e.g., due to sample size, survey, and/or modelling methodology, these limitations are always clearly stated and explained in the documentation for the users. On-demand customised requests are also processed by TDC staff to avoid misuse of the data. However, further usage of the data published on the TDC internet site is beyond the control of TDC staff.

One specific concern in Australia is the acknowledged deficiency in data availability on commodity and freight movements. This, combined with the lack of official classified traffic counts, makes the estimation of commercial vehicle trips more difficult and less accurate. The new Freight Movement Model attempts to overcome these deficiencies by using a model concept already proven successful in other states in Australia, and by organizing specific surveys in NSW to collect information required for the calibration of the model.

2.5 Summary

This paper presented an overview of the data collection and provision activities of the NSW Transport Data Centre. The key elements of the data available from the TDC are the Household Travel Survey (HTS) providing information on the travel behavior of residents in the Sydney Greater Metropolitan Area, the Journey-to-Work origin–destination tables by Travel Zones, the Commercial Transport Study (CTS) providing estimates of heavy vehicle trips, the population and employment projections
by Travel Zones, and outputs from the Sydney Strategic Travel Model (STM) providing estimates of future travel demand based on various assumptions of future transport supply alternatives. These datasets are highly sought after by government organizations, academic and research institutions, and private consultants. During the 12 years of its existence, the TDC datasets have been, and are being used for a large number of important transport infrastructure projects, transport and land use planning studies, and transport policy decisions. Notwithstanding the acknowledged quality of its products so far, TDC is continuously working to improve the accuracy and reliability of its datasets and to promote its products to the public.

Acknowledgment This paper is based on some internal TDC documents as well as information available from the TDC website (http://www.transport.nsw.gov.au/tdc). Comments and suggestions received from several TDC colleagues are gratefully acknowledged.

References

Chapter 3
Data Collection for Measuring Performance of Integrated Transportation Systems*

Wei-Bin Zhang, Alex Skabardonis, Meng Li, Jingquan Li, Kun Zhou, and Liping Zhang

3.1 Introduction

More than ever, traffic congestion is plaguing our heavily populated metropolitan areas. Transportation professionals have recognized that we cannot build our way out of this ever-increasing congestion. The challenge over the next decade is to get more out of the existing transportation system by improving its productivity. To address this challenge, we must evolve into “system managers”: agencies and individuals who manage the system through operational strategies, complemented by targeted expansion investments. The concept of system management has been embraced by many agencies at both state and federal levels. For example, the California Department of Transportation (Caltrans) and most of its stakeholders adopted the concept of the System Management pyramid, as depicted in Fig. 3.1. The foundation of system management is “System Monitoring and Evaluation”. This foundation provides support for a variety of informed investment decisions.

Monitoring and evaluation of transportation systems require high quality data, a set of Measures of Effectiveness (MOEs), and modeling or simulation tools, which lead to a comprehensive understanding of the performance of a system or set of systems. These data, measures, and tools allow for the identification of problems (e.g., bottlenecks, high incident locations), their causes, and for the development of new approaches, technologies, and investment strategies to eliminate or ameliorate the congestion caused by these problems. There has, however, been a lack of high quality data and analysis tools for supporting system monitoring, evaluation and research, particularly for arterial highways and transit. Furthermore, studies thus far have primarily focused on individual transportation networks, and efforts are needed to collect data and develop tools for addressing transportation systems as a whole.

*The paper is dedicated to the memory of former PATH director Robert E. Parsons.

W.-B. Zhang (✉)
California PATH, University of California at Berkeley,
1357 S. 46th Street, Richmond, CA 94804, USA
e-mail: wbzhang@path.berkeley.edu

J. Barceló and M. Kuwahara (eds.), Traffic Data Collection and its Standardization,
International Series in Operations Research & Management Science 144,
This chapter discusses the data needs for monitoring and evaluating integrated multimodal transportation systems and the efforts undertaken by the California PATH program to collect high quality data for supporting comprehensive evaluations. Particularly, it describes the Parsons Traffic and Transit Laboratory and, through a few examples, how it supports research at PATH.

### 3.2 Data Needs for Multimodal Transportation Systems

#### 3.2.1 Data for Measuring Freeway Performance

Freeway performance measurement requires the availability of real-time data from surveillance sensors and tools for data storage, processing, and analysis. Surveillance data include traffic volume (number of vehicles traveling over the detectors) and occupancy (the percentage of time that a vehicle “occupies” the detector) in 20- or 30-s resolution.

Researchers at California PATH developed the freeway Performance Measurement System (PeMS) (Choe et al. 2002; Varaiya 2006). PeMS receives 30 s detector data from 23,000 sensors in California freeways. It processes the data in real time to perform detector diagnostics, speed calculations, aggregations at various time intervals, and calculations of performance measures. It includes an extensive set of reports.
Data Collection for Measuring Performance of Integrated Transportation Systems

PeMS collects and stores data on freeway incidents. PeMS can be accessed online at http://pems.eecs.berkeley.edu/. PeMS calculates the amount and causes of freeway congestion based on the analysis of surveillance and incident data for individual freeways, districts, or the entire State (Fig. 3.2). This provides a quantitative estimate of the freeway congestion that can be attributed to different causes, sets congestion reduction targets, and monitors the effectiveness of congestion relief strategies. PeMS also produces contour plots of speeds and occupancy to assist in the identification of freeway bottlenecks and their spatial and temporal impacts. It also includes an algorithm for automatic bottleneck identification.

PeMS calculates travel time statistics on user defined freeway segments (routes). Figure 3.3 shows the travel time statistics for the southbound I-880 freeway by time of day. The figure illustrates that 15% of drivers experience delay of about 10 min during the midday and 20 min during peak hours.

3.2.2 Data for Measuring Performance of Arterial Highways

The operation of arterial highways is more complex due to the presence of signalized intersections and, as a consequence, the performance of arterial highways is more difficult to measure. Studies are being conducted to develop performance measures for arterials, but no deployment has yet been reported. Arterial performance can be measured at intersection, corridor, and network levels. At the intersections, performance measures may include:

- Intersection delay per cycle, average vehicle waiting time at the intersection
- V/C Ratio, measured by the ratio of volume to capacity
- Queue size, number of vehicles queuing before the intersection
- Signal cycle failure (queued vehicles cannot depart due to insufficient capacity)
- Phase utilization (percentage of green time used during cycle)
Progression quality (percentage of vehicles arriving in green interval)
Number of stops
Arterial travel time

The data needs for supporting signalized intersection and arterial performance measures are traffic detection inputs on the intersection approaches (volume, occupancy, presence, and speed of individual vehicles) and signal settings (cycle length, phase sequence, green times, and offsets) plus event data for actuated controllers (gap out, max out, and force off).

An analytical model was developed at California PATH for arterial travel time estimation (Skabardonis and Geroliminis 2005). The analytical model calculates travel times on each signal cycle. Input to the model consists of loop detector flow and occupancy data in each cycle and the signal settings. The travel time is modeled as the sum of the free flow time and the delay at the traffic signal. The delay at the traffic signal is calculated as the sum of (a) the delay of a single vehicle approaching a signalized intersection without any interaction with other vehicles, (b) the delay because of the queues formed at the intersection, estimated according to the kinematic wave theory to explicitly consider the temporal and spatial formation of queues, and (c) the oversaturation delay, the additional delay caused when the arrival rate is greater than the service rate at the signal. The application of the model on two arterial sites and comparisons of the estimated travel times with simulated and field data showed that the model accurately predicts travel times at the selected sites. This model was extended to account for the effects of detector placement and
29

Data Collection for Measuring Performance of Integrated Transportation Systems

queue spillovers (Geroliminis and Skabardonis 2006). The results of the model application show that it accurately estimates travel times for a wide range of operating conditions.

In another study (Li et al. 2008, 2009), an online travel time estimation method was developed to utilize the detailed signal infrastructure data collected at PATH Parsons Traffic and Transit Laboratory. This method also accounts for the typical detection layout situation in closed-loop signal control systems, i.e., under the scenarios that only arrival presence detectors for the arterial through traffic (National Electrical Manufacturers Association, or NEMA, movements 2 and 6) are available. Arterial performance measures were studied in the context of arterial travel time, trip reliabilities, and number of stops per arterial trip. The arterial model creates an imaginary trajectory from an origin to a destination with a constant headway (e.g., 5 s). The link travel times for each trajectory on different links were estimated by finding the closest virtual arrival time. For travel time estimation, a two-level hierarchical approach was developed to estimate the dynamic route travel time along multiple arterial links. The travel time estimation results are shown in Fig. 3.4. The results were validated by the field data. For each 5-min interval, the estimated travel time was compared with the baseline average travel time for traversing the length of the six-intersection arterial corridor. The root mean square error (RMSE) and root mean square percentage error (RMSP) for the estimation were 9.5 s and 5.9%, respectively.

Research is also underway at PATH on the use of probe vehicles for arterial travel time estimation. These investigations are conducted under the condition that

Fig. 3.4 Measured versus estimated arterial travel times
the probe penetration is high enough to be representative and include using less frequent transit vehicles as probes to estimate travel time, when data are filtered to adjust for specific bus operating conditions.

### 3.2.3 Data for Measuring Performance of Transit Operations

Transit performance measures include quality of service aspects such as the overall measured or perceived performance of transit service from the passengers’ point of view. Also, quantitative transit service measures, or MOEs are commonly used. These measures are indicative of transit access and use.

From a planning perspective, transit routing and frequency, coordination of transit service, and ridership have been used for measuring transit accessibility, coverage, mobility, and the quality of transit service. They have also been used to assist transportation agencies in allocating investments for improving transit services to reduce auto trips, to improve air quality, and to make informed decisions regarding land use.

Transit operating performance is typically examined at the route level. Travel time, trip length, wait time, and dwell time are used to determine the quality and effectiveness of transit service. Comparison is made between scheduled and actual service, such as hours of service, number of trips, miles traveled, number of operators, and speed of the vehicle. Additional passenger information can be measured including counts of passengers carried, boarding, and alighting. From this data, other measures such as average passenger load during each trip and number of passengers per mile can be calculated. Additional performance measurements include the number of service hours, number of trips, load factor, miles traveled, and number of operators.

The essential data needs for supporting performance measures for evaluation transit operations include AVL data from each individual bus with a sampling rate desirable every second. Door opening and closing data from transit vehicles and automated passenger count data are useful for assessing the bus stop level operations and ridership. In situations where there is an interaction between transit vehicles and traffic control devices, traffic and signal status data as defined in the previous two sections becomes very helpful.

### 3.2.4 Data for Measuring Performance of Integrated Corridor Management

An increased emphasis on the movement of people and goods, and the recognition of the importance of performance measures of the transportation as a whole, which address the needs and interests of the audiences for mobility information, will result in a very different set of procedures for evaluating transportation infrastructure
and land use policies. Under the recent USDOT sponsored Integrated Corridor Management (ICM) initiative, studies are being conducted to investigate the interactions and coordination among different transportation networks, i.e., freeway, arterial, and transit networks. Two consortia of agencies in the San Francisco Bay Area and San Diego, CA, have been investigating how existing ITS technologies can be used to facilitate integrated operation. The studies’ goals are to encourage mode shift, to reduce congestion and, ultimately, to achieve higher efficiency from the existing transportation infrastructure.

The benefits of the Integrated Corridor Management System (ICMS) are challenging to measure directly because they are, in effect, a “system of systems”. The benefits gained from use of the independent systems are not in fact the benefits of the ICMS, but rather the ICMS benefits are the benefits gained from the integration with other systems and balanced use of these systems. These benefits are evaluated by a set of performance measures, which are based on the performance measures of individual systems described above and capture both the individual benefits from use of the individual systems in isolation and the benefits of the integrated systems throughout the corridor. Depending on the characteristics of each corridor, the performance measures for each ICMS can be different, but may be established along the following aspects:

- Total corridor throughput, measured by both people and vehicle throughput
- Average and distribution of congestion delay throughout the corridor, measured from delays of all networks and modes
- Frequency, duration, and handling time of incidents
- Emergency response time
- Balanced use of the corridor network resources, including changes in mode split of bus transit, rail transit, and shared ride automobile
- Fuel consumption savings and pollutant emissions reductions associated with changes in mode split and reductions in traffic flow disturbances.

The data needs will include all data described above for individual networks and these data need to be synchronized in order to support the integrated evaluation.

### 3.3 California PATH Parsons Traffic and Transit Laboratory

California PATH is an applied research program with expertise ranging from concept development and proof-of-concept field testing to application development. Since PATH was established in 1987, research based on scientific data and a thorough understanding of the subject matter to be studied has been the tradition and strength of the program.

Starting in 1999, in support of Caltrans’ efforts to quantitatively evaluate freeway performance, PATH developed the PeMS described in Sect. 3.2.1. In parallel with the development of freeway performance measures, PATH has been developing performance measures for arterial and transit systems. PATH also conducted
research on enabling technologies for improvements of arterial traffic signal control and transit operations. There are growing needs for high quality data and performance measures on arterial traffic and transit systems to facilitate understanding their operational characteristics and interactions.

Starting in 2004, PATH, with substantial support from Caltrans and San Mateo Transit District, began to collect detailed traffic and transit data on a few intersections along El Camino Real corridor in the San Francisco Bay Area as part of a research project on transit signal priority (TSP). This effort has significantly helped PATH researchers to understand traffic behaviors on arterial corridors, the operational characteristics of arterial traffic signal control systems, and the interaction between transit operations and traffic signals. Building on the success of this effort, PATH and Caltrans determined to expand the data collection capabilities to support research on development of ITS technologies, and field testing and implementations of research products. This effort led to the expansion of the data collection effort to include several arterial corridors and several dozen transit vehicles, as well as the development of data processing tools. In 2006, PATH consolidated the arterial and transit data capabilities and established the Parsons T² Lab. The Lab was named after Robert E. Parsons, the founding director of the California PATH, who devoted his career to developing innovative transportation systems in the USA. Mr. Parsons established a solid foundation for PATH’s success and was one of the few people in the USA who was an early and strong supporter of the U.S. national ITS program.

The Parsons T² Lab has become a unique transportation laboratory that features high quality traffic and transit data from arterials and corridors, a growing number of data management and analysis tools, and a full range of experimental environments. It supports a wide range of research projects in variety of project areas including TSP systems, improvement of efficiency and safety of urban grade crossings, red-light-running (RLR) collision avoidance, performance measures for arterial highways and transit operations, Vehicle Infrastructure Integration (VII), ICM Initiative, environmental friendly driving, and other areas.

### 3.3.1 Data Collection

Parsons T² Lab has been collecting detailed, high-resolution, real-time traffic and transit data from a number of arterial corridors and transit systems. Currently, as shown in Fig. 3.5, five data collection sites in the San Francisco Bay Area and San Diego region have been established. In the San Francisco Bay Area, data from two segments of the El Camino Real Corridor including over 50 intersections at San Mateo and Santa Clara, CA and 15 intersections along the San Pablo Avenue Corridor at Alameda, CA are continuously being transmitted to the Lab. In San Diego, CA, data are collected from the downtown area. The freeway data collected and stored in PeMS are also available to the Parsons T² Lab. Table 3.1 lists the data available at Parsons T² Lab.
Collaboration with the Caltrans engineers produced interface software with 170/2070 type traffic responsive field master (TRFM) and traffic signal controllers. Such development facilitates collection of second-by-second traffic signal status and control status data, loop detector count, and occupancy data. Additional traffic data collection methods include traditional pneumatic road tubes and counters to collect traffic volume and vehicle speed data and Doppler radar to collect continuous vehicle speed data. Autoscope video cameras have been heavily instrumented at some testing intersections along El Camino Real Corridor. They send traffic volume and speed data processed from images to the Parsons T$^2$ Lab.

Second-by-second AVL data from more than 100 transit vehicles are transmitted to the Parsons T$^2$ Lab. PATH implemented an innovative and cost-effective data collection system using GPS-enabled cell phones connected with Motorola iDen Network. More than 100 such devices have been installed on buses operated by Alameda-Contra Costa Transit Agency (AC Transit), San Mateo County Transit District (SamTrans), and Santa Clara Valley Transportation Authority (VTA) on San Diego trolleys, and on probe vehicles to be operated in the San Francisco Bay Area. Parsons T$^2$ Lab also developed a PC104-based data collection system, which in addition to AVL location data, collects door open/close status and wheelchair lift operating status through digital inputs.

### 3.3.2 Data Management

Data management is critical for supporting evaluation and analysis. Archiving and managing the large set of data efficiently is very challenging since there are many different types of data including the traffic data, transit data, freeway data, underlying...
### Table 3.1 Data collected at Parsons T² Lab

<table>
<thead>
<tr>
<th>Network</th>
<th>Data source</th>
<th>Data types</th>
<th>Data resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Freeway data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3 – North Central</td>
<td>Loop detectors</td>
<td>Traffic volume</td>
<td>30 s values</td>
</tr>
<tr>
<td>D4 – Bay Area</td>
<td>Caltrans TMC</td>
<td>Occupancy</td>
<td>5 min aggregation</td>
</tr>
<tr>
<td>D5 – Central Coast</td>
<td>CHP/CAD</td>
<td>Speed</td>
<td></td>
</tr>
<tr>
<td>D6 – South Central</td>
<td>PeMS</td>
<td>Incidents</td>
<td></td>
</tr>
<tr>
<td>D7 – LA/Ventura</td>
<td>Toll tag readers</td>
<td>Travel time</td>
<td>Fastrak receivers along 880 installed</td>
</tr>
<tr>
<td>D8 – Riverside/San Bernardino</td>
<td>Bay area D4</td>
<td>Traffic volume</td>
<td>5 min aggregation</td>
</tr>
<tr>
<td>D10 – Central Valley</td>
<td>511 system</td>
<td></td>
<td>less than 10 miles</td>
</tr>
<tr>
<td>D11 – San Diego/Imperial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D12 – Orange County</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Arterial data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>El Camino Real (SR 82)</td>
<td>Caltrans D4 (Bay area)</td>
<td>Vehicle counts, presence, velocity</td>
<td>Presence detection at each intersection and advanced detection for most intersections at 2 s interval</td>
</tr>
<tr>
<td>San Pablo Ave (SR 123)</td>
<td>ACCMA TMC</td>
<td>Surveillance video</td>
<td>Between intersections of major arterials within the smart corridor</td>
</tr>
<tr>
<td>E 14th St (SR 195)</td>
<td>San Diego downtown</td>
<td>Signal status/loop detector counts</td>
<td>Detailed signal status data and loop detector data at 1 min interval</td>
</tr>
<tr>
<td>San Diego downtown</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Bus data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AC transit</td>
<td></td>
<td>Bus location coordinates</td>
<td>AVL data at every second</td>
</tr>
<tr>
<td>SamTrans</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VTA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Trolley data</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>San Diego trolley</td>
<td>GPS-based cell phones and AVL data logger</td>
<td>Vehicle AVL data, door opening and close event data</td>
<td>AVL data at every second</td>
</tr>
<tr>
<td>North County Transit</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sprinter Rail</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Probe vehicles</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Individual drivers</td>
<td>GPS-based cell phones</td>
<td>Vehicle AVL data30+ probe vehicles</td>
<td>AVL data at every second</td>
</tr>
</tbody>
</table>

Road network data, and detector data. The data from these different sources is strongly related to each other, as shown in Fig. 3.6. This database architecture reflects this correlation. For example, to implement the TSP, both the current bus location and signal status of the approaching intersection are needed. Moreover, the amount of traffic and transit data is substantial. For instance, in the case of the GPS location of each bus, data are sent to the transit center once every second. The database may contain billions of records when the data for hundreds of buses for a year is stored. Hence, our goal is to design flexible, scalable, and efficient ITS databases. The relational database (e.g., MySQL) was selected to store the data.
To improve flexibility and reduce complexity, PATH has decomposed the data management into several layers. The lowest layer is the underlying road network including intersections and road segments. The information from each intersection, or a node in the network term, is stored in a database table. The table columns include a specified ID, intersection name (usually two crossing road names), latitude, and longitude. Each road segment or a link in the network term contains the information of a specified link ID, starting and ending node IDs. Road length and number of lanes are also stored.

The second layer is designed for storing the arterial traffic and transit data. The traffic controller information is associated with a node and is stored in a database table. The real-time and historical information of the signal status are stored in another table including the current date and time, phase information, force-off point, and other attributes. The signal status table is associated with the node table by using the node ID. For the transit data, bus stop information is stored in a table of the transit database, with the specified stop ID, agency name, bus stop name, longitude and latitude, and associated link ID. The bus route is then defined as a sequence of bus stop IDs in the route and stored in another table. The real-time GPS information from buses is stored in a table with the current time, position, and speed.

To support real-time data queries, we set two query indexes that include an index for each bus and for each specific date. Additionally, we have designed the application programming interfaces (API) for each layer so that the users can just call an API rather than access the data directly.

### 3.3.3 Experimental Environment

The Parsons T² Lab also supports field research and operational tests (FOT). A hardware-in-the-loop (HiL) simulation environment has been developed (Fig. 3.7). In this HiL simulation environment, the signalized intersections are controlled by...
real traffic signal controllers through the controller interface devices (CID). Traffic is generated by a microscopic traffic simulation tool, e.g., PARAMICS or VISSIM. By using HiL, we can easily debug developed methodologies, verify their applicability, and evaluate their impacts. A testing intersection at UC Berkeley’s Richmond Field Station supports initial field testing before FOT. Except for flowing traffic, all of the communication and control hardware and software, roadside and on-vehicle detection sensors, and the movement of testing vehicles are consistent with the field situation. In the Lab, monitoring programs, database tools, and data analysis software have been developed to help researchers and engineers to debug and improve the testing system. During the FOT, the Parsons T^2 Lab is able to monitor testing progress, store and organize testing data, and support online data analysis.

3.4 Role of Parsons T^2 Lab in Supporting PATH Research

The Parsons T^2 Lab has become an effective tool for conducting research in traffic and transit operations at PATH. The following sections describe example projects that have been supported by data collected at the Lab. Without such data support, it would be very difficult to obtain the findings reported from these projects.

3.4.1 Evaluation of TSP and Development of New TSP Approaches

TSP systems are increasingly deployed in the USA. In reported TSP evaluation studies, the benefits due to the introduction of TSP vary significantly among implementations (Smith and Hemily 2005). This large variation in reported impacts is because the selected MOEs, the types and quality of the collected data, and the methods used for analyzing the data vary significantly. As a result, the findings of
these studies cannot be compared directly. For example, most of the evaluation studies focused on savings in transit trip time to evaluate the benefits of TSP. Because TSP systems are in many cases deployed in conjunction with other changes, including, but not limited to, changes of schedules, routes, and operation policies, the time savings do not reflect the benefit of just TSP implementation. Reported savings may also include the results of all the improvements rather than the TSP benefits in isolation. Furthermore, data collection methods vary considerably, ranging from manually recording transit trip times to using AVL data to compare arrival times at selected route locations. This also adds to the uncertainties of how the data can be compared. Despite the fact that the possible interruption on traffic by frequent TSP calls is a common concern to many traffic engineers, assessment of the negative impacts of TSP on traffic has been very limited, and the results in many cases are qualitative and subjective.

To assess the benefit of a TSP system and its traffic impact on nonpriority streets and to determine the specific conditions under which TSP is most cost-effective, PATH has developed a set of detailed MOEs and a comprehensive evaluation method to support an objective and comprehensive evaluation of the performance of a TSP system (Zhou et al. 2008). An evaluation system has been established that can collect, with second-by-second resolution, the bus location data as well traffic detection and signal status data (Fig. 3.8).

A number of TSP systems, including those deployed by several transit agencies in the San Francisco Bay Area including AC Transit, Valley Transportation Authority (VTA), and Samtrans have been evaluated. Tables 3.2 and 3.3 show the

---

**Fig. 3.8** Data collection system architecture for TSP evaluation
results obtained from the evaluation of AC Transit’s Rapid 72R line operated along a 13.5-mile segment of the San Pablo Avenue (State Route 123) corridor. Under this TSP deployment, a total of 37 out of 82 signalized intersections were TSP enabled to reduce the transit travel time along the corridor. The TSP system utilizes 3M’s Opticom TSP system to detect the presence of transit vehicles and to request TSP operations from the signal controller. Enhancements for TSP operations developed by Caltrans have been incorporated by updating the C-8 signal controller software to provide early green and green extension treatments. The evaluation results not

<table>
<thead>
<tr>
<th>TSP strategy</th>
<th>Early green Arterial (transit route)</th>
<th>Cross street</th>
<th>Green extension Arterial (transit route)</th>
<th>Cross street</th>
<th>Change (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without TSP (s/veh)</td>
<td>17.5</td>
<td>36.3</td>
<td>17.2</td>
<td>38.1</td>
<td></td>
</tr>
<tr>
<td>With TSP (s/veh)</td>
<td>16.3</td>
<td>38.1</td>
<td>16.4</td>
<td>38.2</td>
<td></td>
</tr>
<tr>
<td>Change (%)</td>
<td>−7.0%</td>
<td>5.0%</td>
<td>−4.8%</td>
<td>0.1%</td>
<td></td>
</tr>
</tbody>
</table>

### Table 3.2 San Pablo TSP: Comparison of traffic delays

<table>
<thead>
<tr>
<th>TSP strategy</th>
<th>Early green</th>
<th>Green extension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arterial (transit route)</td>
<td>AM peak 45.5</td>
<td>44.1</td>
</tr>
<tr>
<td></td>
<td>Midday 50.2</td>
<td>48.9</td>
</tr>
<tr>
<td></td>
<td>PM peak 53.5</td>
<td>54.0</td>
</tr>
<tr>
<td>Cross street</td>
<td>AM peak 38.4</td>
<td>37.5</td>
</tr>
<tr>
<td></td>
<td>Midday 40.6</td>
<td>40.0</td>
</tr>
<tr>
<td></td>
<td>PM peak 41.7</td>
<td>42.7</td>
</tr>
</tbody>
</table>

### Table 3.3 San Pablo TSP: Arterial performance measures

<table>
<thead>
<tr>
<th>Travel direction</th>
<th>MOE</th>
<th>Time period</th>
<th>AM peak</th>
<th>Midday</th>
<th>PM peak</th>
<th>AM peak</th>
<th>Midday</th>
<th>PM peak</th>
<th>AM peak</th>
<th>Midday</th>
<th>PM peak</th>
<th>AM peak</th>
<th>Midday</th>
<th>PM peak</th>
<th>AM peak</th>
<th>Midday</th>
<th>PM peak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northbound</td>
<td></td>
<td>Travel time (min)</td>
<td>45.5</td>
<td>50.2</td>
<td>53.5</td>
<td>38.4</td>
<td>40.6</td>
<td>41.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Running time (min)</td>
<td>7.1</td>
<td>9.6</td>
<td>11.8</td>
<td>6.5</td>
<td>9.0</td>
<td>11.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total intersection stopped time (min)</td>
<td>17.4</td>
<td>20.6</td>
<td>22.7</td>
<td>16.7</td>
<td>20.5</td>
<td>23.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of stops at signals</td>
<td>20.6</td>
<td>20.5</td>
<td>23.6</td>
<td>20.5</td>
<td>23.6</td>
<td>23.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Southbound</td>
<td></td>
<td>Travel time (min)</td>
<td>48.6</td>
<td>53.3</td>
<td>53.6</td>
<td>47.3</td>
<td>49.6</td>
<td>52.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Running time (min)</td>
<td>39.8</td>
<td>42.3</td>
<td>42.6</td>
<td>39.4</td>
<td>40.3</td>
<td>42.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Total intersection stopped time (min)</td>
<td>8.7</td>
<td>11.0</td>
<td>11.1</td>
<td>7.8</td>
<td>9.3</td>
<td>10.3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of stops at signals</td>
<td>18.7</td>
<td>21.0</td>
<td>20.9</td>
<td>18.3</td>
<td>19.6</td>
<td>21.0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Changes statistically significant
only showed the time savings specific to the TSP deployment but also the impact of traffic delays due to the TSP activation. The evaluation also revealed that the TSP detection rate at certain intersections was very low which resulted in maintenance improvements for the system.

3.4.2 Developing Optimized Control for Urban Railway Crossings

Urban rail is an effective and popular solution for traffic congestion mitigation along major urban corridors. However, frequent signal preemptions and priority treatment at rail/highway grade crossings and their effects can significantly interrupt coordinated traffic flows and threaten the safety of pedestrians and other vehicles. PATH researchers have been working with the San Diego Association of Governments (SANDAG), San Diego Trolley and the City of San Diego to investigate the interaction and conflicts between urban/suburban rail and auto traffic, and develop integrated and practical solutions for minimizing delays to motor vehicle traffic while improving schedule adherence for rail operations.

The research team developed and installed GPS-based train movement data acquisition systems. Road sensors are used to collect traffic volume data at ten grade crossings. Traffic signal data were collected through the Traffic Management Center (TMC) located in downtown San Diego. The data have been analyzed to identify any inefficiencies in traffic operations. Figure 3.9 shows the number of times a train has to stop between stations.

Based on the findings from the data analysis, optimization models were developed to minimize intersection delays for trolleys by providing signal priority and to minimize impacts to the rest of the traffic triggered by the priority action. As opposed to existing signal priority studies, which typically focus on isolated intersections, the proposed models develop optimal timing plans for a system of intersections and provide an optimized green band for an incoming train. The optimized green band would start at the right time to cover the predicted train arrival time at intersections and would be wide enough to accommodate prediction errors. The proposed model adjusts the green band by changing the signal offset and green phase lengths.

A comprehensive simulation model, consisting of accurate simulation networks, data collection and analysis tools, has been developed using PARAMICS, a microscopic traffic simulator and its API. To evaluate existing conditions and compare different improved signal control logics and signal timings, MOEs, including average traffic delays at those light rail grade crossings have been developed and analyzed. Analysis and simulation results shown that the new approach can significantly improve trolley operation while reducing by as much as 40% delays for other traffic. SANDAG, San Diego Trolley and the City of San Diego have provided funding for field testing and are very interested in the potential for future deployment.
In a separate grade crossing project, PATH is working with Sprinter Rail in the San Diego Region to develop countermeasures that would minimize the impact of SPRINTER operations on local traffic. An optimized signal timing and control algorithm is under development. It utilizes real-time AVL data, time-to-arrival at grade crossing prediction, traffic volume data as well as the preemption and safety requirements for optimizing the street signal timing before the preemption is initiated. This algorithm is explicitly designed to provide better mitigation of traffic impacts.

### 3.4.3 Development Red-Light-Running Collision Avoidance System

RLR at signalized intersections is increasingly becoming a major safety issue. In 2004, 2.5 million or 40% of all police-reported crashes in the USA occurred at or near intersections. Of these intersection-related crashes, 8,619 or 22.5% were fatal accidents and 848,000 or 46% resulted in injury. The RLR aspect of this problem is known and serious, as approximately 20% of all intersection crashes occur due
Data Collection for Measuring Performance of Integrated Transportation Systems

Under the sponsorship of USDOT and Caltrans, PATH has conducted studies on RLR over the past few years. To understand the RLR behavior, a number of data collection efforts were made. These included data collection using conventional traffic loops, temporary pneumatic tube detectors and counters, and a specially instrumented Autoscope video detection system. The data were collected at intersections along the El Camino Real (State Route 82), in the cities of San Mateo and Palo Alto. Applying the tracking algorithm developed under RLR project, the vehicle trajectories are reconstructed. Figure 3.11 shows a total of 4,885 vehicle trajectories, among which 152 are involved in RLR. Data analysis of these RLR behaviors reveals that 44 are from the first-to-stopped vehicle.

Figure 3.10 shows the 152 extracted trajectories of three interesting maneuvers, to include the first-to-stop, go-through yellow and RLR, from a total of 4,885 trajectories. Among the 152 interesting samples, 44 were from the first-to-stopped vehicles (blue solid lines), 84 were from vehicles that went through yellow (yellow solid lines) and the other 24 were from RLR vehicles (red solid lines). Approximately 92% of RLR occurrences entered the intersection within 2 s after the start of red. Figure 3.11 shows the relationship between the percentage of

---

**Fig. 3.10** Reconstructed vehicle trajectories
vehicles stopped and the travel time to the stopline. It clearly shows the probability of stopping before the stopline increases significantly when the travel time to the stopline is greater than 5 s.

Through further data analysis, we also found that increased arrival flows during the green phase have a strong correlation with increased RLR. Because the arrival flows were measured at the advance loops, the green arrival flow accounts for the majority of the total arrival traffic during the cycle. Furthermore, as a result of the detailed cycle-based analysis, the yellow arrival flow has shown a much larger influence on the probability of RLR and should receive greater attention. The results from the subsection analysis strengthen the notion that certain sections of the signal cycle are the source of RLR. Higher flow during the yellow subsection indicates a higher chance that those vehicles are staying with a platoon (or cluster). Consequently, truncating a platoon in the middle could increase the probability of RLR. As a result of these findings, it was concluded that yellow arrival flow is a critical influencing factor for RLR at intersections. Based on this finding, PATH developed an enhanced signal timing approach that optimizes signal offsets based on the patterns of arrival flows while still maintaining signal progression. This approach significantly reduces yellow arrivals and consequently reduces RLR probability and subsequent collisions.

PATH is furthering its data collection efforts by extensively instrumenting an intersection along El Camino Real with several Autoscope® video detection systems. This allows us to configure various emulated detection and speed loops along each approaching leg. A multihypothesis tracking (MHT) algorithm was developed to associate the data from the emulated discrete loop detectors to build vehicle trajectories. These emulated detectors, combined with the tracking algorithm, make it possible to collect detailed traffic and vehicle movement data and to optimize the

Fig. 3.11 Stopping probability versus travel time to the stopline
location of the detectors based on empirical observations. Calibration was conducted using a test vehicle instrumented with high accuracy GPS at the test intersection and has indicated that the accuracy of the detection is adequate for the RLR detection purpose. Using the data collected from the test intersection, analysis was conducted to model RLR movements and to predict the RLR for arterial intersection collision avoidance. The prediction of RLR is a crucial part for RLR collision avoidance systems, which must be able to sense the dynamic characteristics of a vehicle approaching the intersection from a reasonable range. RLR needs to be predicted to enable the engineering measures such as issuing a warning or dynamically adapting the signal to protect vehicles affected by the red light runner.

PATH researchers developed an RLR prediction algorithm, which is a combination of predicting the vehicle stop-or-go motion and estimation of the arrival time using the speed detections at different advance locations (Zhang et al. 2009). In this algorithm, we developed a multidimensional model where the distance to intersection and measured speeds of the approaching vehicles and time into yellow are important contributory factors (example shown in Fig. 3.12). The prediction algorithm is based on the Neyman–Pearson criteria, which gives the algorithm the capability of controlling the balance of the two types of prediction errors, namely the missed report error and the false alarms, to address for the different consequences of different types of errors in operation. A system operating characteristics (SOC) function can then be defined as the achievable correct prediction rate versus the allowable false alarm rate. Based on an empirical data set containing over 6,000 vehicles collected during March 2008, we set the parameters for the prediction

![Empirical model of first-to-stop (left peak) and going through yellow/red vehicles](image)
model and applied the algorithm to data collected during May and June 2008. Simulation showed that for the newly collected data, the algorithm provided correct prediction rate of 80% when the allowable false alarm rate is 3% (Fig. 3.13).

### 3.5 Concluding Remarks

To meet the evaluation needs of integrated multimodal transportation systems and to support advanced research needs, California PATH has been collecting a comprehensive set of data from freeways, arterial highways, and transit systems. The Parsons T² Lab was established to collect, store, and process high quality traffic and transit data from freeways, arterials, and corridors. A growing number of data management and analysis tools and a full range of experimental environments are being developed to support in-depth understanding of driver behavior, and traffic and transit operations. These data and analysis tools are available to the research community at large and greatly facilitate the development of advanced ITS solutions and enable field implementation of new methodologies and new technologies.

**Acknowledgments** The authors thank Kai Leung of Caltrans’ Division of Traffic, Paul Chiu, James Lau, and Lindy Cabugao with Caltrans District 4, Sonja Sun, Don Dean, and Greg Larson of Caltrans’ Division of Research and Innovation, Frank Burton of Samtrans, Fanping Bu, Yue Li,
and Scott Johnston of California PATH, Guoyuan Wu, Offer Grembek, Amy Wang, Peter Lau, Joe Liu, and Yao Ma of University of California at Berkeley for their contributions and support toward the Parsons T² Lab.

References

Choe T, Skabardonis A, Varaiya P (2002) Freeway performance measurement system (PeMS): an operational analysis tool, transportation research record #1811, Washington, DC
Li M, Zou ZJ, Bu FP, Zhang WB (2008) Application of vehicle-infrastructure integration (VII) data on real-time arterial performance measurements, under review by Transportation Research, Part C
Li M, Song MK, Gu GY (2009) An online performance measurement method based on arterial infrastructure data. Paper presented at the 88th annual meeting of the transportation research board, Washington, DC
Smith HR, Hemily B (2005) Transit signal priority, a planning and evaluation handbook. ITS America, Washington, DC
Varaiya P (2006) Freeway performance measurement system (PeMS), PeMS 6. Final report for CCIT TO 15
Zhou K et al (2008) Field evaluation of San Pablo Corridor Transit Signal Priority (TSP) system, PATH research report. Institute of Transportation Studies, University of California, Berkeley
Chapter 4
International Traffic Database: Gathering Traffic Data Fast and Intuitive

Marc Miska, Hiroshi Warita, Alexandre Torday, and Masao Kuwahara

4.1 Introduction

Gathering real life data, for whatever type of use, is a time consuming job. A lot of data is measured and stored in several places and different formats around the world. While a lot of it is not used, other institutions gather similar data on different locations or, worse, on the same ones. In this way, a lot of money and time is spent unnecessarily (Miska and Lint 2006). Thus, the aim of the International Traffic Database (ITDb) project is to provide traffic data to various groups (researchers, practitioners, public entities) in a format according to their particular needs, ranging from raw measurement data to statistical analysis.

The idea of providing traffic data or data storage standards is not new and is done in several parts of the world. Among others, the following projects and institutions provide such information, just to name a few:

- The National Roads Authority, Ireland (Miska et al. 2007)
- Department of Transportation (DOT), Federal Highway Administration (FHWA), the Federal Motor Carrier Safety Administration (FMCSA), the National Highway Traffic Safety Administration (NHTSA), and NHTSA’s National Center for Statistical Analysis (NCSA), USA
- Traffic Management Data Dictionary (TMDD), Institute of Transportation Engineers, USA
- Next Generation Simulation (NGSIM), USA
- Region Laboratory Delft, Netherlands
- ASTRA, Switzerland
- Clearing house for transport data and transport models, DLR, Germany
However, the traffic data available to the public is provided in changing formats, different aggregations, and with a varying density of Meta information (when existing), which allows just a limited reuse of the data or a time consuming process of formatting or understanding the data. To overcome this drawback, the United Kingdom started a project called “ITS Metadata Registry” to allow users of such data to match variables from different sources and to understand different formats more easily. Instead of just trying to make the data understandable for other users, ITDb incorporates this translation or data name matching and applies it to actual data to provide a comprehensive data pool to the user.

These different formats and the many places to check for useful data are slowing down the process of data gathering and therefore the studies using them (research, model calibration, etc.). To tackle this problem, the ITDb promotes a flexible traffic data provision format on the basis of user needs and standard habits, instead of another in-house solution for sharing data. Further, the collection of worldwide data sources and making them accessible via one single platform accelerates projects and decision making in any data sensitive field.

Reflecting all this, the ITDb is founded on three pillars: Data, platform, and quality management (see Fig. 4.1). Handling them independently in the design and development allows focusing on technical aspects on the one hand and the fulfillment of user needs on the other. The ITDb is not based on existing data, tailored for its provision, but based on the final operation. The sustainable development has to take into account different user groups, transfer efficiency and a high quality standard. This is the major difference to other existing implementations of data platforms, where the design is purely based on a single set of data types and has to be changed or adapted with every revision of the content.
4.2 Contributions to Traffic Engineering

The ITDb project is contributing to the field of traffic engineering by providing a standardized Metadata with an underlying traffic data structure and providing the whole range from raw to processed data at various levels for the usage in research, model calibration and validation, performance analysis, and so on. Using traffic data has usually as final goal analyzing or assessing network performances, a wide variety of data type has to be available in order to get the full picture of the traffic behavior. For this reason, not only conventional traffic data but also accidents, traffic events, traveler information messages, parking, and environmental data are stored. This data can be collected and downloaded in a whole package, which is transformed to a requested format (based on the availability of data) or common standard used in traffic engineering. This means that every user can format the requested data according to personal preferences. Additionally, it provides rules for sustainable data storage as guideline for quality management in the provision of traffic data for future projects.

ITDb stores traffic data, environment data, and incident data from various sources in a common format. Stored Meta information, which can be linked together regionally, can be browsed and used to make a collection of data of interest for the user of the database. This allows a high flexibility, since the data sets are separated from their description. Instead of downloading project-specific files or browsing online databases of traffic data in various formats, ITDb allows a regional search for all available data and data type-specific searches worldwide. This data will be presented to the user in various selectable formats to meet his personal needs. The ITDb database together with the storage facility for nontext-based data, such as maps, videos, and photograph series, creates a powerful tool for gathering a comprehensive data collection of study areas.

Concerning the data storage, it is a necessity to provide standardized Metadata to enable several parties to use the available data and to ensure that this data can be shared among them. Metadata have two main purposes, including support of improved data quality and ensuring the longevity of data by documenting who, when, where, why, and how the data was collected. The major contribution of this work is the development of a standardized Metadata set for traffic measurement data and flexible formats for providing the measurement data, which allows the translation to commonly used existing standards such as (Miska et al. 2007):

- The Traffic Management Data Dictionary (TMDD),
- The P1512 Incident Management Data Dictionary (P1512-IEEE),
- Traffic Model Markup Language (TMML),
- Geographic Markup Language (GML),
- Universal Traffic Data Format (UTDF),
- Digital Geospatial Metadata (FGDC-STD-001-1998),
- European traffic information exchange standard (DATEX2), and others,

to ensure compatibility and easy usage.
Since traffic data is valuable information and storage as well as provision includes costs, the ITDb structure is designed to include data selling and brokerage. This means that data provider can determine if the download of their data is free or if they would like to charge a usage fee. The costs for this data are shown on the website and are marked differently. With brokerage, ITDb wants to make data exchange between parties more attractive. While companies might be more interested in the selling, research institutes might want to use the chance to exchange data for data. It allows users to offer their data in exchange for similar data sets from different location or from other sources. Additionally, providers can offer data with a price to be negotiated bilateral or data requests can be posted on the platform.

Finally, quality management has to be guaranteed as using traffic measurement data for research, calibration, or validation is very sensitive to the data source reliability. Algorithms and tools can only work properly with correct inputs and quality management is therefore vital to avoid “garbage-in/garbage-out” scenarios.

To gain the user’s trust in the ITDb, all data sets available on this platform are strictly checked according to the quality management rules, which require a minimum set of Metadata information that allows the user of the data to reconstruct the exact locations of the measurements, the conditions under which the measurements took place, and the equipment used for these measurements. Next to this technical information, the International Traffic Database requires administrative Metadata including a contact institution or person for further inquiries about the data.

While a strict handling may raise the need for postprocessing existing data, it is supposed to be a guideline for future traffic measurement experiments and installations, to store the data in a format that allows further usage for research projects, or simulation calibration tasks.

4.3 Overall System Design

The design of the database consists of a Meta database per country that stores the Meta data and can be browsed for network descriptions, projects, measurements, environment, and incident data in certain locations and time spans. The structure of this Meta search database can be found in Fig. 4.2.

These Meta data is linked to the actual data sets in the data storage of ITDb. This design allows an efficient search for the user and fast access to the wanted information, because the download is not limited to certain projects, but can be accumulated from various uploaded elements. Further, it allows to link data from different data sources together in the case that the user is looking for regional data provided by various institutions.

The data flow in the ITDb system has two parts. A public front end where registered users can browse through the Meta database to search for useful data sets, and a second, hidden part, in which a scheduler will trigger the collection of data from the ITDb data storage and provides it to the user as a downloadable file via FTP or HTTP (Fig. 4.3).
Data protection and security is a major problem for their owners who do not want to open their database to any usage. Thus, depending on the aggregation level, the data can be disclosed to any user, or to a subset of registered users. In order to give total flexibility on the control of the data provision process, various option of shared responsibility between data owner and the ITDb are available.

4.4 Metadata Search Engine

To be able to search efficiently through various data sets worldwide, ITDb features a metadata search engine. Metadata for each data set is essential to give the user a complete knowledge about how to use the data and to avoid the garbage-in garbage-out
phenomenon. ITDb requires the following meta information for each data set as a mandatory requirement:

- Georeferenced location of the measurement in WGS84
- Time stamp of the measurement
- Description of the measurement equipment
- Description of the measurement installation
- Information about if and how the raw measurement is processed
- What values are measured and in which form
- Aggregation level of processed data
- Contact information of the data owner

Further meta information will be stored, but is not required. The metadata search engine enables users to request specific data they are interested in. One could for example look for raw, disaggregated loop detector data, also known as pulse data. Then the search engine will locate all measurement spots known to the system and visualize them for the user in the commonly known environment of Google Maps, which allows an intuitive browsing through the found matches.

### 4.5 Data Storage

The actual data is stored directly in ITDb, or outside of ITDb, and linked to the meta information. The differences between these two ways of storage are explained in more detail below, since especially links to third party databases can become a security concern.

#### 4.5.1 Data Stored in ITDb

All continuous point measurement data are stored in a SQL database, to easily extract the data for a specific timeline or to search for data under specific circumstances like accidents, rainfall, and so on. One time measurements, such as video observations, or one time measurements of emissions, are stored in their original file format and a description file is attached to allow a unified handling of the data, disregarding the format in the later process. The description file contains all information on how to handle the original file and which download options are available (i.e., streaming for video formats, bulk download for data files, etc.).

For security reasons, the data is not stored on the web server and can only be reached by the software modules handling the data gathering. In this way, the possibilities for unauthorized manipulation of data are eliminated. The software for gathering the data from the ITDb storage will only allow a request generated locally, which secures the data repository from outside attacks. However, ITDb is
logging every user action to detect unusual behavior, and if unusual activity is found, the user accounts are shut down for further protection of the integrity of the whole system.

### 4.5.2 Data Stored in Third Party Locations

To avoid redundancy of data storage, ITDb is using linked data sources from third parties whenever possible. In these cases, ITDb just stores the meta information locally and the data gathering module will receive the data directly from the data provider. If the provider agrees, often requested data will be mirrored at ITDb to a defined maximum amount to reach better performance.

The merit of no redundancy in storage, however, raises more security concerns – this time from the data owner themselves. Needless to say that any interference with the original databases of the data owner can lead to difficulties in operations, which should be avoided under all circumstances. To meet the needs of the data providers, ITDb offers several ways of data acquisition on the fly. The possibilities range from a direct access to their database, via database mirrors controlled by ITDb, to software modules that run in house with the data provider, so that the data flow is fully under their control. With this setup, every data provider can decide individually if and how much control over their data they want to allow to be taken over by ITDb.

### 4.6 Universal Data Translator

To make data widely accessible, ITDb features a universal translator. This translator creates a dictionary for every new format described for ITDb and therefore is able to translate the stored data in ITDb back to the known formats. Certainly, formats are not consistent, and some information in one format is not sufficient for the other, or a format has too much information to be fully exported. In the latter case, it is the user’s choice to change the format, if he is interested in the additional information, but given that he uses his best know format. The information should be sufficient for his needs. More attention is needed for the case that not all information of a format is available in ITDb. Considering that formats are used to serve specific applications or methods, it is unacceptable to transmit empty data or unexpected data. To avoid this, the user will be informed about the missing data and he is required to input a default value for the missing data. This seems appropriate, since every robust method or software should be able to deal with default values (i.e., −1 for missing data).

In the long run, we hope to enable with this method the foundation for a common standard of traffic data storage. By accumulating many formats, we will be able to identify the common basis and the “extras,” and by combing them, ITDb
will be satisfying for the users and, with enough feedback, has the potential of rising to a new standard.

4.7 User Front End

The prototype application is a dynamic website that allows browsing all data content via keyword-based searching and visually by browsing through the network and selecting the data sources of interest. The latter one is realized using the map interface provided by GoogleTM (see Fig. 4.4) and allows an intuitive “shopping” for traffic data based on maps or satellite images.

The found and selected data by the user is stored in a DataCart, which acts as a shopping basket. The user can add or delete items from it and when finishing check out if the data is prepared well enough to be downloaded. After the system has collected all the data and created a downloadable package with the desired content, the user will receive an e-mail with a download link. This link will be available for 48 h, while the data will be available until fully downloaded.

To upload data to the ITDb, the user has to fill out the necessary Meta information concerning the data and choose between uploading via FTP or giving a download location for a remote upload. The data will be available on the website after the Meta information has been checked for consistency. This is necessary to keep a high standard of data quality on the website. If the information is incomplete, ITDb will request further information. The prototype application can be reached via http://www.trafficdata.info.

Fig. 4.4  Web front end of the International Traffic Database
4.8 Services and Tools

Obviously, the primary service of ITDb is to provide an overview of available traffic data around the world and to offer access to this data. Taking into account data that is costly to acquire due to the measurement methods, ITDb will additionally offer data brokerage. Data that is not freely accessible can still be browsed with ITDb, but the access to the data set itself will be restricted for paying customers. This service will require a password, provided from the data owner to the user after the payment in the first stage, while preparations are ongoing to allow direct payment via ITDb.

Next to the fundamental service of data provision, ITDb will in the near future offer several services for users and providers to enhance the attractiveness of ITDb for both. These services will include:

- Data project management (access for closed groups)
- Data hosting (data storage solution)
- Metadata generation (outsourcing of the data preparation)
- Added value generation (ITDb will gather and compile additional information about the data)
- Data cleansing (preparing raw data for direct usage)
- Data mining (analyzing incoming data and provide results)

4.9 Future Developments

The database currently includes data samples from different parts of the world to give a broad coverage. Using various data source types and worldwide locations, a standardized format for storing traffic data has been defined, which allows providing them based on the needs of several user groups. While researchers prefer raw, disaggregated data sets for specific analysis, practitioners are interested in processed data (e.g., speed–density relations, lane occupancy, queue discharge behavior, etc.) for example to calibrate and validate simulation models. Public entities on the other hand might be interested in the quality of service in networks. The developed standardized format is based on existing approaches of standardization and contains next to the traffic data values a broad selection of Meta information to assure the usability of the provided data.

Registered users are able to browse through the data in different process levels. Uploading new data is restricted by the quality management guidelines provided to ensure the users of the database high quality and comprehensive data sets.

Future work includes the collection of more data sets and to use other access methods for downloading data, like the REST (representational state transfer) protocol. This would allow a direct use of traffic information in simulation models or other applications without the manual transfer of the data and would increase the convenience of using the ITDb system for research and practice.
References

Miska MP, Lint Hans van (2006) A web based data source platform for modeling and simulation. Proceedings of the EWGT2006 joint conferences, advances in traffic and transportation systems analysis, Technical University of Bari, Faculty of Engineering, Bari, Italy

Chapter 5
Data Mining for Traffic Flow Analysis: Visualization Approach

Takahiko Kusakabe, Takamasa Iryo, and Yasuo Asakura

5.1 Introduction

Data mining has attracted considerable attention as a method that can be used to discover certain characteristics from large amounts of data. In traffic flow analysis, a large amount of traffic flow data is continuously collected and stored over several years.

Among the methodologies for observing traffic flow, one methodology that can be used to observe wide-area networks for a long period is a traffic detector system. Although there are many methods for observing traffic flow, they are sometimes not suitable for wide-range observation. For example, through video images, we can detect the detailed trajectory of each individual vehicle; however, installing many cameras on the roads and constructing a comprehensive image processing system may be expensive. On the other hand, traffic detector systems have been installed in many road network systems in many countries. Some of these systems store a considerable amount of historical data on entire road networks.

Such extensive data may provide information on many phenomena that cannot be easily predicted. Data-oriented approaches are an effective way to obtain the unknown characteristics of traffic flow from the data. Data mining techniques help us to obtain information on these characteristics from the large amount of data.

Visual data mining is one of the techniques that detect the characteristics of data. This technique converts data into pictures by using certain methodologies (for example, Keim 2000). This technique provides analysts with a process to overview a large quantity of data. The characteristics of the data are obtained from these pictures by visual perception. Moreover, this technique can depict information more efficiently than using numeral letters without decreasing the amount of information. One may assume that this technique is not reliable because it is based on human perception; however, it should be noted that many other data mining methodologies
rely on some given algorithm which may not necessarily be suitable for finding unknown characteristics of data. On the other hand, because the visual data mining technique depends on visual perception, the knowledge and experience of the observer can be utilized for discovering and refining the unknown characteristics of the data.

In this study, we develop a visualization system for long-term traffic flow data of freeways or expressways and apply it to data collected on the Hanshin Expressway and Metropolitan Expressway in Japan. To build a system suitable for discovering the characteristics of traffic flow from data collected over a long period, the following two requirements should be fulfilled:

- Accurate indication of information with an appropriate color expression
- Fast and seamless display

Traffic flow data collected by a detector on a section of the road usually consists of traffic volume, speed, and occupancy. Because they are continuous numbers, they should be represented by a continuous change in colors during the visualization process. Although there are many conventional schemes that represent such data by a change in color (for example, red-yellow-blue corresponds to jam-slow-free flow in a speed contour map), a color scheme that changes continuous numbers to appropriate colors should be developed by considering the color model.

Fast and seamless display is another requirement when analysts represent data in various conditions from a large quantity of data to refine their knowledge. If a large amount of process time is needed to visualize data, it becomes difficult for analysts to examine entire range of a large amount of data. This study proposes a system that can quickly visualize data and efficiently discover characteristics.

### 5.2 Characteristics of Traffic Detector Data

The visualization techniques in this study are used to examine traffic flow data collected by traffic detectors over a period of several years. In general, data on the traffic volume, average speed, and occupancy are collected in a time period of 1–5 min by a detector or point (a point consists of a few detectors installed at the same position but in different lanes) along the expressway (as shown in Fig. 5.1). These data can be considered time series data for each point or detector.

Generally, a large amount of data is collected when the observation period is long. The size of data collected in 1 year by a traffic detector, which contains three indices, namely volume, speed, and occupancy, is 1 MB at each point (it is assumed that the data is saved as 32 bit integer data for each index). For example, the size of data acquired in 1 year at 2,000 observation points is over 2.5 GB and that acquired in 4 years is 10 GB. This implies that visual data mining is required to process the data quickly.
5.3 Process for Visualizing Traffic Detector Data

This section explains the methodology of the visualization process proposed in this study. Each index of the detector’s data can be recognized as a single-dimensional time series. However, the changes in such data depend on both the time of a day and days. The data collected will be depicted in three-dimensional space, which contains a within-day time axis, a date axis, and an observed value axis. A conceptual representation of this system is shown in Fig. 5.2. In this figure, the time-date plane represents the day-to-day changes in the traffic flow, and the observed value-time plane represents within-day changes. Such an image system that visualizes the data in a two-dimensional plane is referred to as a traffic contour map (TCM). In a TCM, each observed datum is changed into a certain color by using a pre-determined relationship between the value of the data and color. Section 5.3.1 defines this relationship by considering perceptions of color. There are two ways of drawing a TCM; these are discussed in Sects. 5.3.2 and 5.3.3. Section 5.3.4 discusses the implementation of this system.
5.3.1 Color Mapping System for TCM

TCM represents the change in traffic flow by colors. To describe a quantitative change in the observed values precisely, a perceptive difference in colors should correspond to the difference in the observed values. This study employs the CIE 1976 ($L^*$, $a^*$, $b^*$) color space (also referred to as CIELAB) and its color difference formula (International Standards Office 2008). CIELAB is a color model in which a change in the amount of a color value produces a corresponding change in visual perception. As shown in Fig. 5.3, the color space of CIELAB consists of three axes – $L^*$, $a^*$, and $b^*$. Every color is uniquely placed in this space. The $L^*$ component corresponds to the human perception of lightness, the $a^*$ component represents the position between red and green, and the $b^*$ component represents the position between yellow and blue.

The color difference between color $i$ and $j$ is calculated by the color difference formula:

$$\Delta E_{i,j} = \sqrt{(L_i^* - L_j^*)^2 + (a_i^* - a_j^*)^2 + (b_i^* - b_j^*)^2}$$

(3.1)

where $L^*_i$, $a^*_i$, and $b^*_i$ are the coordinate value of color $i$ in CIELAB.

This formula gives the Euclidian distance in CIELAB between two colors indicating the color distance in human perception. The color difference and the difference in the observed values should be proportioned in order to represent a larger difference of the value in a larger color difference in perception. This implies that the colors are derived from an indeterminate linear curve in the color space. Thousands of linear lines can be drawn in the color space; however, each line should be as long as possible. Therefore, each line should be allocated considering the range of colors that is available on the device. Liquid crystal displays (LCD) and CRT displays are the most popular output devices for computers, and sRGB is the most common color specification for these devices (International Electrotechnical Commission

![Fig. 5.3 CIE 1976 ($L^*$, $a^*$, $b^*$) color space](image)
If an LCD or CRT display is chosen as the device, the line should be chosen by considering the available color area of sRGB. Then the line passes through the points that represent deep purple, pale purple, light grey, pale green, and vivid green in the color space.

There are many ways of assigning colors to the observed values. For example, some methodologies use colors that have a relation of hue rotation, such as blue, yellow, and red. However, these colors are not in a straight line in CIELAB; that is, the difference between the colors is not proportional to the difference between the observed values. This causes an error in determining the characteristics of traffic flow because a small difference in the observed values can have a larger perceptive difference. The methodology of this study can minimize this disproportion in order to have a better understanding of the characteristics of traffic flow.

5.3.2 Time-and-Day TCM

This section proposes a time-and-day TCM that has time and day axes. In a time-and-day TCM, each observed value is mapped by a pixel corresponding to its observed time of day and date, as shown in Fig. 5.4. The within-day changes in the value are represented along the time axial direction, and its day-to-day changes are represented along the day axial direction.

The time-and-day TCM can also be divided into several sub-maps according to the range of the values in order to express a detailed change in the data, which are displayed on the same screen at the same time. Such a method is referred to as multiple TCM method. This method allows us to display all ranges of observed values on one screen while keeping the color difference as large as possible. Figure 5.4 shows a multiple TCM. The multiple TCM displays the range between

![Diagram of Time-and-Day TCM](image-url)
Fig. 5.4 (continued)
5th and 95th percentile traffic volume with the help of four sub-maps that show different ranges of the observed values, that is, 5–27.5th percentile, 27.5–50th percentile, 50–72.5th percentile, and 72.5–95th percentile. Values that are larger than the range of each sub-map are represented by a color indicating the maximum value of the range, and those that are less than this range are represented by a color indicating the minimum value.

### 5.3.3 Time-and-Value TCM

A time-and-value TCM shows the distribution of the observation values at every point of time within a day. This is suitable for checking the within-day changes in the observed values in detail. A time-and-value TCM is shown in Fig. 5.5. The vertical axis represents the observed values and the horizontal axis represents the time in a day. The colors represent the frequency of an observed value corresponding to the observed time of the value. The numerical value of this frequency is normalized by the number of observations collected at each point of time in a day, and the colors are determined by the frequency using CIELAB.

### 5.3.4 Implementation of the Visualization System

This section discusses how to implement a visualization system that fulfils the requirements described in Sect. 5.1. There are many ways to implement the visualization methodology on computers. For example, existing application software, such as spreadsheet and graph-drawing software, can be used to visualize observed data. However, there is also an option of developing new software. It is important to fulfil the requirements that are necessary to implement the methodology of this study. These requirements are as follows:

1. Appropriate assignment of CIELAB colors for the observed data
2. Pre-processing of data and mapping them on the TCM
3. Quick access to data and seamless display

Usually, it is not easy to find software that satisfies all these requirements. Software satisfying requirement 2 may be easily available; however, a technique for the allocation of colors in requirement 1, which is the novel technique proposed by this research and software for implementing CIELAB directly may not be readily available. Furthermore, it is difficult to fulfill requirement 3. As discussed in Sect. 5.2, a large amount of data is handled in this research. Therefore, it is difficult for general-purpose software to process such large-scale data because of a low transaction speed and low memory.
A simple way to fulfill the abovementioned requirements is to develop new software exclusively for this purpose. By using this software, it will be easy to satisfy requirements 1 and 2. Moreover, because the data format and processing method have already been fixed, it is relatively easy to satisfy requirement 3 by adopting the data structure and the algorithm.

The adopted data structure is shown in Fig. 5.6. The traffic detector data file is separately made by each detector or point, and the observed values for each detector or point are arranged in a time series. The file access speed can be improved by adopting this data structure because the observational data of each observation place can be continuously read in time sequence. The starting position of the meta-information of each observation point is calculated as:

\[ M_i = (i - 1) \times (m + n \times o \times 1440) / s \]  

(3.2)

where
- \( i \): Identification number of the observation point or the detector (\( i = 1,2,\ldots,k \))
- \( s \): Time interval between two successive observations (min)
- \( n \): Number of observed indices (when volume, speed, and occupancy are observed, this value is 3)
- \( o \): Number of observed days
- \( m \): Size of meta-information

The position of the value observed at time \( t \) on day \( d \) is derived as follows:

\[ V(i,d,t,c) = M_i + m + ((c - 1) \times o + d) \times 1440 / s + t / s \]  

(3.3)

where
- \( d \): Days counted from the beginning of observation (\( d = 0,1,\ldots,o - 1 \))
- \( t \): Within-day time (min) (\( t = 0,1,\ldots,1440 \))
- \( c \): Identification number of the observed index (\( i = 1,2,\ldots,n \))

The process of drawing a TCM consists of data extraction, color mapping the observed values with the methodology discussed in Sect. 5.3.1, and drawing the TCM according to the procedure discussed in Sects. 5.3.2 and 5.3.3. The software is written in C language and works on X window systems (including MS Windows and Macintosh). We used a computer having Pentium 4, a 3.40 GHz CPU and a 1.25 GB RAM with Windows XP.
5.4 Empirical Analysis with Visualization

This section gives examples of empirical analysis of the actual traffic detector data. Section 5.4.1 discusses an analysis of the detector data taken on the Hanshin Expressway, and Sect. 5.4.2 discusses an analysis of the detector data taken on the Metropolitan Expressway.

5.4.1 Analysis of Data Taken on the Hanshin Expressway

Data of traffic flow on the Hanshin Expressway was collected by 827 detectors at 369 points; this covers almost the entire network. A detector is installed at different locations in each lane and it collects data on the traffic volume, average speed, and occupancy every 5 min. This data was collected from 2 March 2003 to 31 May 2006 (except on weekends, holidays, 30 December–3 January, and 12–16 August). This analysis visualizes the observed values of the traffic volume from all observation points. In the following sections, several characteristics obtained by the visualization will be explained.

5.4.1.1 Detection of Detector’s Error

Figure 5.7 gives a visual representation of the observed values of the traffic volume at 2.5 KP in the inbound direction of the Osaka-Ko Line. Several errors of the detectors are seen at this point. An intermittent lack of data is seen after a large data

![Traffic Detector Data File](image)

**Fig. 5.7** Detection of detector’s error
deficit in June 2004. Moreover, it is observed that the observation values after the data deficit become smaller than before.

Such errors cannot be easily found in collective data. Figure 5.8 shows the daily traffic volume. Errors occurring intermittently between November and December 2004 cannot be seen in the graph, whereas errors around June 2004 can be easily observed.

In such a situation, the period and the point having many observation errors can be detected by using the technique of visualization proposed in this study. Once these error points are discovered, analysts can omit such data before another analysis.

5.4.1.2 Detection of Effect of Toll Gate Operation

Figure 5.9 gives a visual representation of the observed values of the traffic volume at 9.0 KP in the inbound direction of the Matsubara Line that connects the suburban areas with the city center of Osaka. In this figure, a distinguishable blue vertical line appears from March 2003 to June 2005 at approximately 9.00 a.m. This implies that there is relatively less traffic at this time. Figure 5.10 represents the weekly mean traffic volume observed from 8.55 to 8.59 a.m. and that observed from 9.05 to 9.09 a.m. Such a decrease in traffic volume, as shown in Fig. 5.9, can also be observed in Fig. 5.10. This phenomenon might have been caused by the operation of a throughway toll plaza at a point slightly upstream of the detector. However, the number of vehicles equipped with the electronic toll collection system (ETC) rapidly increased between December 2004 and December 2005, and the operational work of the toll plaza might decrease. Evidently, from Fig. 5.10, it can be observed that the difference in the traffic volume between 8.55–8.59 a.m. and 9.05–9.09 a.m. decreased with an increase in the penetration rate of ETC.

5.4.1.3 Detection of Traffic Capacity Reduction after Sunset

Figure 5.11 visually represents the traffic volume at 0.9 KP in the leftmost lane of the loop line. This figure shows a characteristic sine-curve-like shape for traffic volume from 4.00 to 7.00 p.m. This curve corresponds to the sunset time in Osaka city. A detailed objective analysis carried out by Kusakabe et al. (2006) implies that this traffic volume reduction is caused by a reduction in the road capacity at the bottleneck during sunset.

5.4.2 Analysis of Data of Traffic on the Metropolitan Expressway

Data collected at an observation point on a line of the Metropolitan Expressway was analyzed. Data on volume, speed, and occupancy were collected every 1 min.
Fig. 5.8 Change in traffic volume

Date
2007-05-07
2007-04-02
2007-03-01
2007-02-01
2007-01-04
2006-12-01
2006-11-02
2006-10-02
2006-09-01
2006-08-01
2006-07-03
2006-06-01
2006-05-08
2006-04-03
2006-03-01
2006-02-01
2006-01-04
2005-12-01
2005-11-01
2005-10-03
2005-09-01
2005-08-01
2005-07-01
2005-06-01
2005-05-06
2005-04-01
2005-03-01
2005-02-01
2005-01-04
2004-12-01
2004-11-01
2004-10-01
2004-09-01
2004-08-02
2004-07-01
2004-06-01
2004-05-06
2004-04-01
2004-03-01
2004-02-02
2004-01-05
2003-12-01
2003-11-04
2003-10-01
2003-09-01
2003-08-01
2003-07-01
2003-06-02
2003-05-06
2003-04-01
2003-03-03

Time (hour)
8.0 12.0 16.0 20.0
(veh/5.0 min)
101.0 < 87.3 73.7 60.0 46.3 32.7 19.0 ≤
The period of analysis was from 1 June 2004 to 31 May 2007 (except on weekends, holidays, 30 December–3 January, and 12–16 August).

Figure 5.12 shows the traffic volume observed at the upstream end in the inbound direction of the Shibuya line, which connects the suburban areas and an intercity motorway with the city center of Tokyo. This point is near a throughway toll plaza between the intercity motorway and the Shibuya line.

From Fig. 5.12, distinct vertical lines can be observed around 12.00 a.m. and 10.00 p.m. from March 2003 through June 2005. This implies that there is a rapid increase in the traffic flow. To analyze this phenomenon in detail, the time-and-value TCM from 8.00 p.m. to 12.00 a.m. is shown in Fig. 5.13. The increase in the traffic volume at around 10.00 p.m. can also be observed in Fig. 5.13. The increase in the traffic volume begins at 10.05 p.m. and lasts for approximately 5 min. The increase in the traffic volume is approximately 10–20 veh/min. This phenomenon might have been caused by a discount service which targets ETC-equipped vehicles that pass through the toll plaza after 10.00 p.m. The increase in the traffic volume at around 12.00 a.m. may also be attributed to another discount service that begins at 12.00 a.m.

5.5 Conclusions

This study proposes a visualization technique suitable for discovering the trends in the traffic flow from data taken over a long period. Certain empirical analyses are carried out by this technique. This visualization technique allows researchers to
Fig. 5.10 Traffic volume around 10.00 p.m. and penetration rate of ETC (Utilization rate of ETC is released by Hanshin Expressway Company 2008)
overview the data collected on traffic flow over a long term. Moreover, as discussed in Sect. 5.4.1, this technique can find errors due to operation problems of detector systems. As mentioned in Sects. 5.4.1 and 5.2, this technique can be used to evaluate various controls and policies on expressways that may cause a change in the traffic flow pattern. This implies that the visualization system can be a useful tool for road administrators or highway officials. Moreover, the system can be used as the communication tool between researchers and them.

The visualization system that is developed in this study focuses on determining the trends in traffic flow. This system can be made into a more convenient tool by combining it with statistic tools and image processing techniques. The visualization technique can be used for finding the unknown characteristics to be evaluated precisely by using more quantitative statistical analyses.

Acknowledgments The traffic detector data was provided by the Hanshin Expressway Company and the Metropolitan Expressway Company. This study was partly supported by the JSPS Grant-in-Aid #17656163.
Fig. 5.12 Time-and-day TCM
References


Chapter 6
The Influence of Spatial Factors on the Commuting Trip Distribution in the Netherlands

Tom Thomas and Bas Tutert

6.1 Introduction

Traffic flows are the result of movements of people and goods. They are modeled with the help of behavioral patterns that are supposed to remain relatively constant over time. In traditional transport modeling, some of these patterns are described by trip distribution functions, which represent the propensity to make trips with certain costs. The distribution functions (DF) are used to estimate a priori origin-destination (OD) matrices.

Distribution functions (DF) have been part of spatial interaction models for a long time. They have been applied successfully in (double constrained) gravity models (e.g., de Ortuzar and Willumsen 2001). For the DF, different shapes have been used, e.g., a power-law, a log-normal distribution, or an exponential distribution (e.g., de Ortuzar and Willumsen 2001, de Vries et al. 2004). The gravity model with an exponential DF is of particular interest, because it also follows from entropy maximization (Wilson 1970), in which trips are distributed according to the most likely state or configuration (with maximum entropy). The entropy maximization method was developed independently from gravity models and actually originated from statistical mechanics.

In the last two decades, disaggregated random utility models (e.g., Ben-Akiva and Lerman 1985, McFadden 2001) have been gaining popularity in the academic world. In these models, trip distribution is estimated by maximizing the utility of users. Trips are allocated to alternative destinations for which the (perceived) utility is maximal. More recently, activity-based models (e.g., Arentze and Timmermans 2004) have also been made operational (e.g., Bhat and Guo 2005, Bradley and Vovsha 2005). The advantage of these models is that they are able to model sequences of decisions in which the order of attributes may be important.
Disaggregated choice models, such as random utility and activity-based models, can model several aspects of travel behavior in general and model policy interventions simultaneously.

However, disaggregated models are complex. It is impossible to model how each individual values utility. The utility is influenced by several unknown factors, and the utility function may vary from individual to individual. As a result, random components are usually added, and assumptions about the model structure have to be made in advance. The evaluation of the results is also not straightforward, because model parameters may be rather sensitive to changes in observed choices.

Due to these problems, many policy makers in the Netherlands still prefer to use the less sophisticated gravity method when they model traffic flows. Aggregated models provide another advantage. They can be used to obtain the distribution of observed choices as a function of certain attributes. This function can be fitted directly to the observations. In other words, the observations can be parameterized by an empirical model. Although this type of modeling is not often used in traffic engineering, we will adopt it in this study. We will show that it provides results that can be interpreted in a straightforward way.

Parameterization is especially useful when considering spatial characteristics, which have long been ignored in distribution estimates. Recently, spatial behavior has found its way into discrete choice models (e.g., Walker and Li 2007, Magidson et al. 2003). More generally, studies on network structures (e.g., Patuelli et al. 2007, Albert and Barabasi 2000) and urban and regional structures (e.g., Ma and Banister 2006, Cörvers and Hensen 2003) have shown the importance of the influence of spatial structures on commuting trips. In a study of the RPB (Ritsema van Eck et al. 2006) for example the gravity model for the Netherlands has been improved by the extension of spatial components, i.e., including measures for urban and regional structures.

In this chapter, we study the influence of spatial characteristics on the trip distribution of commuters. We focus on commuting trips because they still contribute most to the total demand during the peak hours. In Sect. 6.2, we describe the survey data that we use. In Sect. 6.3 we explain the method with which we estimate the DF (within the concept of the gravity model). In Sect. 6.4 we present the results and Sect. 6.5 provides conclusions.

### 6.2 Data

This study is based on data from the Dutch national travel survey, MON (Rijkswaterstaat 2004–2006). The main objective of this survey has been to obtain annual national statistics on travel behavior. Recently, the survey has been adapted to become more suitable for developing traffic models. In the survey, each member of a household fills in which trips were made during 1 day. For each trip, motive, origin and destination, departure and arrival times, mode, submodes, and distances are reported. The zones are on a postal 4 level. There are about 4,000 postal 4 zones in the Netherlands. These zones typically have a diameter of a few hectometers in urbanized areas and several kilometers in rural areas. In general, the resolution is thus quite
The Influence of Spatial Factors on the Commuting Trip Distribution in the Netherlands

high although it will be difficult to estimate the distribution function for subkilometer scales. Weight variables, which are included in the database, can be used to correct for underrepresentation of certain social groups and days. However, no corrections are possible for the underreporting of short noncommuting trips. Like in other mobility surveys (e.g., Stopher et al. 2007), underreporting of short trips is a serious problem. It may lead to flawed estimates of the DF for noncommuting trips. Fortunately, selection effects are limited for commuting trips, which is the focus of this study.

Despite its limitations, the MON is an impressive database. For this study, we used MON data from 2004, 2005, and 2006, which include about 85,000 commuting trips. In this study, we want to estimate the probability that a worker who lives in zone $i$ has a job in zone $j$. We therefore only selected the first commuting trip of the day for each respondent. We thus excluded trips made from work, and trips that were made to work for a second or third time (a small, but significant fraction of workers leave their working place during lunch). About 90% of the selected commuting trips are made from home. The remaining trips are made for example after shopping or after a child has been taken to school. If we include these trips, our sample is contaminated by nonhome-based trips. However, it would also complete our sample, because these trips are made among others by working parents. Because most of these nonhome-based trips are made from locations that are near home, we decided to include these trips in the sample. Since nonhome-based trips only form a small part of the sample, it does not influence the final results in a significant way. We also excluded so-called false reports, which are described in Appendix A1. This leaves us with about 40,000 commuting trips.

For each trip, not only the origin and destination of the zones, but also the reported travel distance, $d_r$, and travel time are known. In addition, we used the centroids of the zones to estimate the Euclidean distance, $d_e$, and we used a network to determine the network distances, $d_n$, and free flow travel times between zones. In Appendix A2, we show that network and reported distances are on average quite similar for trips longer than 2 km. We therefore argue that both estimates can be used for trips larger than 2 km. For trips smaller than 2 km, however, $d_r$ becomes significantly larger than $d_n$. We argue that this difference is caused by an overestimation in $d_r$ because respondents tend to “round off” their reported distances. Reported distances are thus unreliable for short trips. In this study, we will focus on network distances.

There is however a problem: we do not have network distances for internal trips (within zones). Because these trips are small (typically smaller than 2 km), we cannot use reported distances. In Appendix A3, we show that we can estimate internal distances by using the average relation between the Euclidean and network distance.

6.3 Method

OD matrices are essential in many applications of traffic management and planning. The (generalized) gravity model, e.g., de Ortuzar and Willumsen (2001), is an expected model for estimating the OD matrix $T_{ij}$. In the gravity model, the number
of trips between $i$ and $j$ is estimated by multiplying the production $O_i$ (total number of trips generated in origin $i$) and attraction $D_j$ (total number of trips attracted by destination $j$) with the distribution value $f_{ij}$. The gravity model can be summarized as follows.

$$T_{ij} = f_{ij}O_i D_j / \sum_i O_i = f_{ij} H_{ij} \tag{6.1}$$

With the following conditions,

$$O_i = \sum_j T_{ij} \tag{6.2}$$

$$D_j = \sum_i T_{ij} \tag{6.3}$$

$$\sum_i O_i = \sum_j D_j \tag{6.4}$$

The distribution function $f$ can be seen as a measure for the attractive force (like in the gravitational law of physics) between origin and destination per unit of production (in the origin) and unit of attraction (in the destination). If the attraction between origins and destinations is constant then $f=1$, satisfying conditions (6.2) and (6.3). The OD matrix is then equal to the production attraction matrix $H_{ij}$ (see (6.1)). In reality, $f$ is not constant, but decreases with travel distance, travel time, and/or travel costs. The attraction is relatively high ($f>1$) for small distances, while $f \approx 0$ for very large distances. The shape of the distribution function can be estimated from large amounts of survey data, but rarely will fulfill conditions (6.2) and (6.3) exactly. These conditions, which make the gravity model double-constrained, may be met by applying the method of Furness (Furness 1970; usually expressed by two coefficients $A_i$ and $B_j$ in (6.1)).

The distribution can be modeled in different ways. The traditional approach is to adopt a function form and to calibrate the parameters subsequently. Several methods have been applied to calibrate the distribution, e.g., maximum-likelihood (e.g., Fotheringham and O’Kelly 1989) and neural networks (e.g., Black 1995, Gopal and Fischer 1996). We adopted a different approach in which the form of the distribution function follows directly from the survey data. In this approach, trips with similar attribute values are aggregated into the same group or bin. Distribution values are then calculated for each bin, and they are plotted versus average attribute values. From these distribution plots, the form of the best fitting distribution function is determined. We will show that the evaluation and interpretation of the results becomes straightforward in this graphical approach.

Before we can aggregate, we first need to estimate $T_{ij}$ and $H_{ij}$ from the survey data. For each OD pair, the MON survey provides us the number of trips $T_{ij}^{\text{MON}}$. Because the MON only surveys a small part of the population, the number of trips is actually very small for individual OD pairs, and many cells are equal to 0.
However, the row totals $O_{i}^{\text{MON}}$ and column totals $D_{j}^{\text{MON}}$ are often significantly larger than ten trips, and in more than 99% of the cases greater than 0. We used the row and column totals to estimate the production attraction matrix, $H_{ij}^{\text{MON}}$

$$H_{ij}^{\text{MON}} = O_{i}^{\text{MON}} D_{j}^{\text{MON}} / \sum_{i} O_{i}^{\text{MON}}.$$  

(6.5)

For the observed OD matrix and production attraction matrix, we then aggregated trips with similar Euclidean distances. We aggregated $T_{ij}^{\text{MON}}$ and $H_{ij}^{\text{MON}}$ into the following bins: $0 < d_{i} \leq 1$, $1 < d_{i} \leq 2$, $2 < d_{i} \leq 3$, $3 < d_{i} \leq 4$, $4 < d_{i} \leq 6$, $6 < d_{i} \leq 8$, $8 < d_{i} \leq 10$, $10 < d_{i} \leq 15$, $15 < d_{i} \leq 25$, $25 < d_{i} \leq 40$, $40 < d_{i} \leq 80$ km, and $d_{i} > 80$ km. For those bins, the observed distribution values $f_{i}^{\text{MON}}$ are calculated by:

$$f_{i}^{\text{MON}}(d_{\text{bin}}) = T_{i}^{\text{MON}}(d_{\text{bin}}) / H_{i}^{\text{MON}}(d_{\text{bin}}).$$  

(6.6)

With $T_{i}^{\text{MON}}$ being the observed trip length distribution, $H_{i}^{\text{MON}}$ the observed production attraction distribution and $d_{\text{bin}}$ the average (weighted by $T_{i}^{\text{MON}}$) distance in each bin. We used Euclidean distance because Euclidean distances can be estimated in any circumstance and their estimates are very robust. Euclidean distances are also strongly correlated with travel times and distances. Per Euclidean distance bin, the average network distance $d_{n}$ can be estimated, and here we will describe the distribution as function of network distance. It is worth stressing that the distribution function does not depend on the number of origins or destinations (production attraction) because the trip length distribution $T$ is divided by the production attraction distribution $H$. The latter distribution corrects for the fact that the number of destinations increases with distance.

The distribution also depends on other attributes, i.e., it depends on spatial and social characteristics. The MON data show that high-income commuters for example make on average longer commuting trips than low-income groups. However, if the origin zones (where the workers live) become larger, the population mix becomes more heterogeneous and social differences will average out. On a postal 4 level (postal 4 zones are the smallest zones in the MON), we have checked that different social groups only have a moderate effect on distribution. This effect is negligibly small at a city level, because within a complete city the population mix is even more heterogeneous.

There is another reason why we do not focus on social characteristics in this study. This has to do with the way surveys are conducted. Most surveys and census databases give information about users rather than about activities. With the MON survey it is quite easy to distinguish different social groups, but it is not possible (without additional information) to distinguish between different types of jobs. This can lead to severe limitations in OD estimates, at least at the local level. For example, take two areas that lie in the proximity of each other. The first area is a living area for low-income people. The second area is a working area with mainly high quality jobs. According to the DF, one would expect many commuters traveling between the two areas. However, because of the mismatch between workers and
jobs, only few commuters will travel between these areas. Related to this problem is how workers and jobs need to be classified. Medical doctors for example belong to the high-income group and therefore are expected to make relatively long commuting trips according to most common models. In reality, medical doctors often need to live nearby the hospital and therefore may make shorter commuting trips.

Many of these examples exist. It therefore seems unpractical to distinguish between individuals who have different social characteristics. We suggest that a limited classification between for example high-income and low-income areas may be more practical. The main focus of this study, however, will be on spatial factors because on larger scales they have got a much larger effect on the distribution. Examples are not only the level of urbanization, but also the spatial orientation of cities with respect to each other. In the next paragraph, we will include these spatial characteristics and we will compare results for different urbanization levels and configurations of cities.

6.4 Results

We investigated which distribution function best describes the data. In Fig. 6.1, we plotted the natural logarithm of the observed distribution values $f_{\text{MON}}$ versus the average network distances to the power $0.4$, $d_n^{0.4}$. According to the figure, the observations are fitted well by a linear function (dashed lines). We tried other powers than $0.4$, but according to the least-square method, these powers provide less superior fits. We therefore conclude that the DF can be described in the following mathematical form.

$$\ln f(d) = a + b \times d_n^{0.4}$$

(6.7)

Fig. 6.1 Distribution versus network distance
Figure 6.1 shows the average DF for the whole of the Netherlands. The slope of the DF, \( b \), equals \(-1.5\). We tested whether these results are robust. First, we distinguished different years. We found similar results for all years. We also distinguished between different income groups and urbanization levels. In a subsequent paper, we will discuss the results in detail. The most important finding is that the distribution function for all income groups, urbanization levels, and years has the same form as described in (6.7).

However, one DF is not sufficient for describing the distribution in the Netherlands because the slope \( b \) of the DF depends on spatial characteristics such as the degree of urbanization. We therefore estimated \( b \) for locations with different urbanization levels. The degree of urbanization is strongly correlated with the size of a city or village. The city-size we quantified with \( N_{\text{work} + \text{job}} \) which is the sum of the number of workers who live within a city and the number of jobs that are located within that city. A city or village is defined as an urban area with only one central core. Suburbs are therefore considered to be separate cities or villages. We aggregated commuting trips within about 100 different \( N_{\text{work} + \text{job}} \) bins. Each \( N_{\text{work} + \text{job}} \) bin contained approximately an equal number of trips. Cities for which \( N_{\text{work} + \text{job}} > 100,000 \) form their own bin. We further used the same distance bins as described previously. We estimated the distribution twice. For each \( N_{\text{work} + \text{job}} \) bin, we determined the distribution for trips that were made from the residences. This is the distribution of the “worker-side.” We also estimated the distribution for trips that were made to the cities. This is the distribution of the “job-side.” For each \( N_{\text{work} + \text{job}} \) bin, we fitted the model given by (6.7) to the distributions of both the worker and job-side.

In Fig. 6.2, the results are shown. The left panel of Fig. 6.2 shows the relation between the slope \( b \) of the DF (see (6.7)) for the worker-side and the number of workers \( N_{\text{work}} \) who live in the residence city. In the right panel of Fig. 6.2 the same relation is shown for the job-side with \( N_{\text{job}} \) the number of jobs. The estimated slopes for each \( N_{\text{work} + \text{job}} \) bin are shown by the solid symbols. For each \( N_{\text{work} + \text{job}} \) bin, the average \( N_{\text{work}} \) and \( N_{\text{job}} \) were estimated, and these estimates are given on the \( x \)-ordinates of the left- and right-hand panels of Fig. 6.2, respectively. Note that there obviously is a very strong correlation between \( N_{\text{work}} \) and \( N_{\text{job}} \). From both panels, we can conclude

![Image](image-url)
that the slope \( b \) becomes shallower toward larger cities. This means that the attractive force spreads out to a wider range for large cities. The correlation between \( b \) and city-size is in fact quite strong and can be parameterized in the following way. For the worker-side: \( b = -2.1 \) for \( N_{\text{work}} < 1,000 \) and \( b = -2.1 + 0.39 \times \log(N_{\text{work}}/1,000) \) for \( N_{\text{work}} \geq 1,000 \). For the job-side: \( b = -2.2 \) for \( N_{\text{job}} < 1,000 \) and \( b = -2.2 + 0.46 \times \log(N_{\text{job}}/1,000) \) for \( N_{\text{job}} \geq 1,000 \). With \( \log \) being the \( \log_{10} \). These mathematical models are shown by the solid lines in Fig. 6.2.

The relation between the slope of the distribution and city-size is slightly different for the worker-side than for the job-side. In villages (with small \( N \)), the distribution is steeper for the job-side. Most jobs in villages are taken by locals, while relatively more villagers have jobs outside their village. In large cities, the opposite is the case. Here, the distribution is shallower for the job-side. This means that relatively many jobs are occupied by workers from outside the city. These effects are linked to the spatial distribution of workers and jobs in the Netherlands. Compared to workers, relatively few jobs are located into villages, while there is a large concentration of (better) jobs in the large cities, which attract many workers from outside.

Apart from this small difference, both panels show a similar steep relation between the slope of the distribution and city-size. City-size therefore is a very distinctive parameter. In fact, the remaining scatter of the residuals (“observed slopes of DF – parameterized slopes of DF”) in Fig. 6.2 is relatively small. For each \( N_{\text{work}+\text{job}} \) bin, the random error in \( b \) was estimated from the least-square fit statistics. The average random error in \( b \) is about 0.10, while the root-mean-square (rms) of the residuals is about 0.12. Most of the scatter can therefore be contributed to random errors. Some of the scatter is however caused by the fact that distribution depends on other spatial parameters than city-size alone.

We conclude that the distribution for all locations in the Netherlands can be described by a simple model with only a few free parameters. To test the predictive power of the model and at the same time look for deviations from the general trend, we compared the expected and observed trip frequencies between individual cities. We did not distinguish between years because of the limited number of cases per OD relation. Also, no OD relations were kept outside the model for validation purposes, since their exclusion would not have a significant influence on the model parameters. The number of free parameters in the model is after all very small compared to the number of compared OD relations. We therefore argue that the results from this comparison can be used for general conclusions.

The observed trip frequencies follow directly from the MON. The expected trip frequencies were estimated with the parameterized DF of the worker-side, with the constraint that the row totals are equal to the observed row totals. The observed row total is the total (weighted) number of commuting trips in the MON that were made from each resident city or village.

In Fig. 6.3, the results are shown. In the left panel, the observed and modeled trip frequencies (\( T_{\text{obs}} \) and \( T_{\text{mod}} \) respectively) are plotted for each relation, i.e., each pair of cities or villages. The trip frequency is often smaller than 10. However, there are quite a few relations for which \( T_{\text{mod}} \geq 10 \). These relations are shown by the solid symbols. The upper panel shows internal relations, i.e., commuting trips made within
The Influence of Spatial Factors on the Commuting Trip Distribution in the Netherlands

The center panel shows relations for which at least one city has got \( N_{\text{work+job}} > 200,000 \) (these are the cities of Amsterdam, Rotterdam, The Hague, Utrecht, and Eindhoven). The bottom panel shows the remaining relations. The solid line in the middle of each plot is the line \( y = x \). The two other lines are the 3\( \sigma \) upper and lower limits.

The \( \sigma \) is an estimate of the random error in the observed frequency, which we approximated in the following way. For a given resident city, the trip frequency toward a certain destination is given by a binomial distribution, with \( p \) the probability that a trip is made to that destination and \( (1-p) \) the probability that the trip is made to another destination. When \( p \) is small, the binomial distribution can be approximated by a Poisson distribution with expected frequency \( T \) and standard deviation \( \sqrt{T} \). The latter is the approximation of the random error in \( T_{\text{obs}} \) with \( T_{\text{mod}} \) being the estimate for the expected \( T \). Note that this approximation is valid for trips between cities. It is however an overestimation of the error for internal trips, for which \( p \) is not very small and lies typically around 0.5 (which means that about half of the trips made from a city are internal trips). In the right panel, the distribution of residuals \( (T_{\text{obs}}-T_{\text{mod}}) / \sqrt{T_{\text{mod}}} \) for which \( T_{\text{mod}} \geq 10 \) divided by their approximated random uncertainties is plotted. If the residuals would only contain random errors, these distributions should follow the Normal distributions (dashed lines).

From Fig. 6.3, we can conclude that most relations are described well by the model. This is clearly the case for relations between smaller cities and villages (bottom panel). However, there are some significant systematic differences between model and observations. The distribution of the residuals is shifted to the left.
(\(T_{\text{mod}}\) is on average larger than \(T_{\text{obs}}\)) for external relations, especially for those between small and large cities (center panel). The distribution will shift to the right when model values are estimated for the job-side, but in that case the scatter in Fig. 6.3 will also increase. These effects cannot be solved by applying the method of Furness or by just averaging both estimates (for the worker and job-side). These effects therefore need to be studied in more detail, so that they can be dealt with in a correct way. Note that they are related to the spatial distribution of workers and jobs.

The most interesting systematic deviations in Fig. 6.3 are the so-called outliers, which lie beyond the 3\(\sigma\) levels. In Table 6.1, we show the frequencies for internal relations of the five largest cities. The model values correspond well with the observations for Amsterdam, Eindhoven, and Utrecht. However, the model underestimates the number of trips for Rotterdam and The Hague significantly (>3\(\sigma\)). It appears that nearly all positive outliers (\(T_{\text{obs}} > T_{\text{mod}}\)) in the upper panel of Fig. 6.3 are on internal relations of cities that are located in the greater area of The Hague and Rotterdam. Related to these positive outliers are the negative outliers (\(T_{\text{obs}} < T_{\text{mod}}\)) in the center panel. These are mainly external relations between The Hague or Rotterdam on the one hand and other cities in the region on the other hand. These results are quite remarkable. They suggest that most cities in the region of Rotterdam – The Hague (including Rotterdam and The Hague) are oriented inwards and that strong relations between cities within the region are much less common than for example in other regions (such as the area around Amsterdam).

There are also a number of positive outliers (\(T_{\text{obs}} > T_{\text{mod}}\)) in the center panel of Fig. 6.3. These are the relations between large cities and their satellite cities. The most well-known examples are the relations Almere – Amsterdam and Zoetermeer – The Hague. Almere and Zoetermeer are clearly oriented to Amsterdam and The Hague, respectively, and the number of observed trips is significantly larger than the number of expected trips for these relations. We conclude that for most outliers, except for the internal relations of Rotterdam and The Hague, there typically is a factor 1.5–2 between observations and model. These are thus rather large differences, which should be studied in more detail.

### 6.5 Conclusions

In this chapter, we parameterized the spatial distribution in the Netherlands using trips from the national travel survey (MON). With our approach we created a straightforward model with only a few parameters. The general distribution function can

---

**Table 6.1** Expected and observed frequencies within the five largest cities

<table>
<thead>
<tr>
<th>Origin</th>
<th>Destination</th>
<th>(T_{\text{mod}})</th>
<th>(T_{\text{obs}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>Amsterdam</td>
<td>1,050</td>
<td>1,085</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>Rotterdam</td>
<td>642</td>
<td>752</td>
</tr>
<tr>
<td>The Hague</td>
<td>The Hague</td>
<td>606</td>
<td>721</td>
</tr>
<tr>
<td>Utrecht</td>
<td>Utrecht</td>
<td>348</td>
<td>325</td>
</tr>
<tr>
<td>Eindhoven</td>
<td>Eindhoven</td>
<td>330</td>
<td>325</td>
</tr>
</tbody>
</table>
be described by a negative exponential-to-the-power law, as a function of distance and city-size. This empirical relation follows from direct fits to the measurements. Other nonlinear functions may also describe the data accurately (e.g., Levinson and Kumar 1996). In our study, however, we tried to find the most simple function form that still fits the data well.

The model enables us to make accurate estimates of trip frequencies between and within cities. By comparing model results with survey data, we found several relations with deviant trip frequencies. Examples are the relations Almere – Amsterdam and Zoetermeer – The Hague. These relations show significantly more trips (from the survey) than expected (by the model). This effect, caused by the strong link between main cities and their spillover satellite cities, is well known, but not a priori included in traditional traffic models. The number of relations with deviating trip frequencies may not be substantial, but they are important for the prediction of (future) traffic flows. In a follow-up paper, we will study these relations in more detail. In the ideal case, the incorporation of variables concerning geographic (political, cultural, and economic) history will take away the deviations.

Recently, disaggregated models gained a lot of popularity for predicting travel flows. These models focus on individual travel behavior. However, for prediction of flows between cities individual differences are cancelled out. We therefore argue that aggregated models are a more direct and effective way for investigating spatial relations on city level. In our opinion, aggregated models can still play an important role in future transport modeling.

Acknowledgment This research has been partly funded by Transumo.

### 6.6 Appendix 1. Removal of False Reports

Most false reports are detected and corrected for within the MON acquisition process. Sometimes, however, errors remain unnoticed; for example, when a respondent accidently fills in the wrong postal zone for the origin or destination. We checked for these false reports by comparing the reported travel distances \( d_r \) (in kilometers) with the Euclidean distance \( d_f \) between the centroids of the origin and destination zone. In Fig. 6.4 we show the results. Note that less than 4% of the distances were imputed in the MON acquisition process, so the distances reflect the answers of the respondents.

Figure 6.4 shows the correlation between \( d_r \) and \( d_f \). The average relation is shown by the solid line. We suggest that trips for which \( d_r < (d_f - 5) \) km or \( d_r > 2 \times (d_f + 5) \) km (shown by the dashed lines) are unrealistic and can be regarded as false. The fraction of false reports is only 6% of the distances were imputed in the MON acquisition process, so the distances reflect the answers of the respondents.

Figure 6.4 shows the correlation between \( d_r \) and \( d_f \). The average relation is shown by the solid line. We suggest that trips for which \( d_r < (d_f - 5) \) km or \( d_r > 2 \times (d_f + 5) \) km (shown by the dashed lines) are unrealistic and can be regarded as false. The fraction of false reports is only 6% of the distances were imputed in the MON acquisition process, so the distances reflect the answers of the respondents.

The identified cases were subsequently left out of the sample. Note that the figure includes internal trips, for which the origin and destination are in the same zone. The estimate for the free internal distance is described in Appendix 3.
Appendix 2. Network Versus Reported Distances

Here, we address the question whether to use reported or network distances. Network distances can be estimated from macroscopic models that assign trips to the network. The network distance $d_n$ (between two postal zones) is estimated as the distance along the fastest route in free flow conditions. Since only commuting trips are considered, it is expected that many trips are affected by congestion. We claim, however, that the distances in free flow and congested conditions will often not be very different. To validate this claim, we compared $d_n$, provided by a transport model, with $d_r$. In Fig. 6.5 (upper panel), we have plotted the average values of the aggregates within free distance bins. We used the bins, as described in Sect. 6.3. For $d_r$ and $d_n$, the averages are plotted for car trips (filled symbols) and all trips (open symbols). We excluded internal trips.

According to Fig. 6.5, both distances are in general more or less equal for large distances (the symbols lie around the dashed lines). This is also illustrated by the lower panel of the figure, in which the logarithmic difference between both distances is shown. The lower panel shows that the network distance is on average slightly larger than the reported distance for distances between 2 and 10 km, although this is mainly so for non-car trips. This can be explained by the fact that $d_n$ is calculated for car trips, while cyclists and pedestrians can take short cuts. Note that the network distances are also slightly larger for car trips, because cars can use streets that are not included in the transport model.

Although there are some small differences for $d_n > 2$ km, network and reported distances are in general quite comparable. For very small distances, however, $d_n$ becomes significantly smaller than $d_r$. We argue that this difference is caused by an overestimation in $d_r$ because respondents tend to “round off” their reported distances. More generally, “accidental” errors in reported distances (due to flawed estimates from respondents) may contribute to inaccuracies.
To obtain internal network distances, we first estimated the relation between network and free distance. This relation is shown in Fig. 6.6, in which we have plotted the averages of aggregates (filled symbols) within free distance bins. Note that we used different bins here, since in this case we were able to extend our bins to larger distances (because we were not restricted to the survey sample size). According to the figure, the relation between $d_n$ and $d_f$ can be described by $d_n = 2.04 \times d_f^{0.90}$ (solid line). Note that this relation implies that the “detour” ratio $d_n/d_f$ becomes significantly larger.
than 1 for very short distances. Such a trend is also reported in (Chalasani et al. 2004), albeit weaker for a high-resolution network. The network of the transport model, which was applied for the whole of the Netherlands, is not very detailed. As a result, travel distances may still be slightly overestimated for the very short distances.

With the relation between $d_n$ and $d_f$, it becomes possible to estimate average internal network distances. For this, we need an estimate of the internal Euclidean distance. If the pools of production and attraction are distributed homogeneous (but not necessarily uniformly) throughout a postal zone, the average Euclidean distance for internal trips is equal to $0.5\sqrt{(A/\pi)}$ with $A$ being the area of the zone. We adopted this estimate as the Euclidean distance for internal trips.

References

Ben-Akiva M, Lerman S (1985) Discrete choice analysis. MIT, MA


Chapter 7
Dynamic Origin–Destination Matrix Estimation Using Probe Vehicle Data as A Priori Information

Rúna Ásmundsdóttir, Yusen Chen, and Henk J. van Zuylen

7.1 Introduction

For most Origin–Destination (OD) matrix estimation methods, a priori information in the form of a matrix (so-called a priori matrix) is necessary as an initial guess. In the estimation process, this matrix is updated with traffic counts until a final estimated matrix has been found. The more this a priori matrix matches the real matrix, the better the final outcome of the estimation will be.

Until this date, a priori information has been acquired from noncurrent data sources (i.e., data that is neither from the estimation time period nor real-time) such as household surveys and road interviews. Due to possible changes in the network or the traffic demand, there is no guarantee that such a priori information is still accurate for the estimation time period. In order to increase the reliability of the a priori information, current data (i.e., data from the estimation time period) should rather be used.

This paper describes a methodology to derive a priori information about origins and destinations from Probe Vehicle Data (PVD). The PVD come from the same time period as the traffic counts used for the estimation, and are therefore current data. The goal is to improve the a priori information for OD matrix estimation, and hence improve the final estimated matrices. Additionally, the route choices within the PVD are analyzed. That is done in order to utilize them for the mapping of OD flows in a complete OD matrix estimation and to see how the Trip Length Distribution (TLD) within the PVD looks like. This is based on the work of Ásmundsdóttir (Ásmundsdóttir 2008) where more detailed explanations and examples can be found.

Unfortunately, the real OD matrices for the study area are not known in this case. Thus the validity of the estimations can so far not be shown. However, importantly, the clear feasibility of the method will be demonstrated and the estimated OD matrices are compared to a reference matrix.

Y. Chen (✉)
CYStone, The Netherlands
e-mail: YS.Chen@CYStone.NL
PVD come from so-called probe-vehicles, i.e., vehicles that are equipped with the necessary devices to transmit geographic position data to a traffic center at regular time intervals. The data comprise information on the status of the vehicle, for instance, its location and speed (Coëmet et al. 1999). The difference between these data and data from local traffic sensors\(^1\) is that the PVD sensors actually measure traffic quantities over road sections while the local quantities can only be generalized over space, at the price of assuming that vehicle operations are both homogeneous and stationary during the observation period and over the considered road section. Van Zuylen et al. (2006) show how different vehicle information can be obtained from PVD.

The PVD used in this paper come from the City of Chengdu in southwest China. The probe-vehicles are GPS-equipped taxis driving within the city centre. The time interval between measurements is 1 min and there were on average 1,960 taxis detected within the study area on the days that were examined.

The PVD are very rich source of information but still there are two serious questions that are worth raising regarding their reliability. First of all, the probe cars are just a fraction of the whole traffic and possibly not a random and representative sample for all the traffic. Thus, the PVD are only sample data where information such as traffic volume is missing. This has to be taken into account when constructing the a priori matrix and examining the route choice and the TLD. Hence the first question is:

*Do PVD comprise enough information to build a good, representative a-priori OD matrix and give sufficient information about the route choice and the TLD?*

The second question deals with the fact that the data come from taxis and that taxi drivers might behave differently in the traffic than other drivers:

*Are data coming from taxis representative for the whole traffic, and if not, can the bias be estimated and adjusted?*

In this paper, answers to these two questions will be given.

When using GPS data, there might also be limitations associated with the equipment: the signal might include errors and in some locations the reception is interrupted. An alternative for the use of GPS is cell phone data (Krygsman and Schmitz 2005; Krygsman et al. 2008; Caceres et al. 2007), with the advantage of a larger sample of drivers that can be followed and the disadvantage that the location accuracy is much less than for GPS.

### 7.2 Literature Review

Several publications have been issued regarding the usage of PVD. These publications discuss, e.g., how PVD can be used to estimate travel times, to detect queues and to analyze route choices. However to the authors’ knowledge, none of them

\(^1\)Data from local traffic sensors are, for instance, vehicle counts, flows, time or harmonic mean speeds, “local” densities, proportions of vehicle types, vehicle lengths, etc.
covers the subject of OD matrix estimation. Following is a discussion about a few of these papers.

### 7.2.1 PVD: Part of “Roads of the Future” Research Program

In The Netherlands, the Ministry of Transport, Public Works and Water Management (Ministrie van Verkeer en Waterstaat) has carried out an experiment with PVD. The purpose of this experiment was to investigate the usefulness of PVD and to get an understanding of the possibilities and problems with PVD. The experiment was part of a large innovation program called “Roads to the Future.” Approximately 60 vehicles in the city of Rotterdam were equipped with GPS and GSM devices and the data were used to estimate travel times. The results of the experiment were satisfactory. After the data had been filtered, about 75% of all the measurements could be used to estimate the travel times. The accuracy of the estimated travel times lies within 1% of the actual travel times for relatively larger road sections. In the report, it is mentioned that the PVD can be used for other useful things as well, such as deriving OD and route-choice information (Coëmet et al. 1999; Taale et al. 2000).

### 7.2.2 Deriving Road Networks from PVD

For all applications of PVD, it is essential to know on which roads the vehicles are travelling. For that purpose a digital network is used in most applications. The production and maintenance of these networks requires a lot of work and resources. Furthermore, the current digital networks have an inherent static nature while the real road networks are dynamic by nature – new roads are built and old ones reconstructed. Temporary changes such as road works and accidents also influence the network. In order to overcome this problem, Hamerslag and Taale (2001) suggest an algorithm that derives road networks from PVD. The idea behind this is: “where there are vehicles, there must be a road.”

### 7.2.3 PVD for Traffic Monitoring

Torp and Lahrmann (2005) proposed a complete prototype system that uses PVD for both automatic and manual detection of queues in traffic. The system consists of small hardware units placed in mobile traffic report units (taxis were used) and backstage databases that collect all the data from the report units. The automatic detection was based on analyzing GPS data from the taxis. The manual detection was based on taxi drivers reporting traffic queues by using the equipment in the taxis. One month field test, where ten taxis were used, showed that the system is operational and that the communication costs are very low. The field test also
provoked new questions, such as how many taxis are needed to do real-time queue detection, how to combine automatic and manual queue detection, and how to integrate the PVD with existing queue detection systems.

### 7.2.4 Local MAD Method for Probe Vehicle Data Processing

Ban et al. (2007) presented a local Median Absolute Deviation (MAD)\(^2\) method that processes travel times from raw probe vehicles data. Travel times generated by probe vehicles may contain a significant amount of outliers that must be filtered. For this filtering, the local MAD method is applied locally to each time window (band) with a fixed duration. A sensitivity analysis showed that for data with more than 2,000 data samples per day, a bandwidth of 15–30 min should be used.

### 7.2.5 Real Time Route Analysis Based on PVD Technology

Zajicek and Reinthaler (2007) proposed a system that uses PVD to calculate detailed routes and travel times for hazardous goods transport in the Austrian road network. Furthermore, the PVD is used to calculate historical time series and actual travel times.

Other papers on the topic are, e.g., *The application of Probe Vehicle system in Beijing* by Wen and Chen (2007) and *Validating travel times calculated on the basis of taxi Probe Vehicle data with test drivers* by Brockfeld et al. (2007). These two papers present methods in calculating travel time with probe vehicles.

### 7.3 Methodology

The study area used for this study is the city centre of Chengdu in southwest China and the probe-vehicles are GPS equipped taxis. The time interval between measurements is 1 min and there were on average 1,960 taxis detected within the study area on the days that were examined.

The PVD comprise information about the vehicle’s locations at each time. These locations are given in GPS coordinates. In order to be able to utilize the PVD, the coordinates need to be translated into zone- and link numbers within the network. That was done using a method developed by Chen (Chen et al. 2007).

Since the PVD come from taxis, the origins of trips are defined when passengers enter the taxis and the destinations when they exit the taxis. The PVD comprise no

\(^{2}\)A statistical measure for capturing the variation of a given set of data points.
information regarding the vehicle’s occupancy. Thus, the first step in the data processing is to construct rules that determine where origins and destinations of single trips lie within the dataset. The sensitivity of these rules is tested in a later section. When the origins and destinations have been identified, the OD matrices can be built. Furthermore, the possibility of exploiting the route choices within the PVD for mapping of OD flows to a network is examined and the TLD within the PVD is studied. The a priori matrix estimations and the route choice analysis are both done for time slices of 10 min.

7.3.1 Rules for Determining Origins and Destinations within the PVD

The rules that define the origins and destinations within the PVD are based on assumptions regarding the speed of the vehicles and the times of the measurements. Hereafter, these rules and the corresponding assumptions are described in detail.

7.3.1.1 Rule 1: Real Stop Versus Intermediate Stop

Stopping at traffic lights should under normal circumstances not exceed 2 min that was also the case when the data collection was carried out at the network, and when a vehicle is stopped in a traffic jam, it can be assumed that its speed is not completely 0 km/h for more than 2 min. When a trip in a taxi ends, the driver has to print out a receipt and the passenger has to pay for the ride. This process is assumed to take at least 2 min. These assumptions lead to the first rule of the PVD processing:

A stop is considered to be a real stop if the measured speed is 0 km/h for 2 minutes or more. Stops that last less than 2 minutes are considered to be intermediate stops. Thus, the first measurement after a real stop is an origin and the last measurement before a real stop is a destination.

7.3.1.2 Rule 2: Break Versus Lost Measurements

When a vehicle is driving, the GPS equipment usually sends out measurements every 1-min. When the driver takes a break and turns off the vehicle, and hence the GPS equipment, it can be assumed that the last measurement before the break is a destination and the first measurement after the break is an origin. It is assumed that a disruption of data emissions due to high-rise buildings, tunnels, and flyovers should not, under normal conditions, exceed 2 min. This assumption leads to the second rule of the PVD processing:

---

3This variable was tested in a sensitivity analysis.
When the time between two measurements exceeds 2 minutes it can be assumed that the driver has taken a break. Thus, the last measurement before the break is a destination and the first measurement after the break is an origin.

7.3.1.3 Rule 3: Vehicles Entering/Leaving Study Area

The first and last measurements with speed larger than 0 km/h that are detected from a vehicle before/after leaving/entering the study area are considered as an origin or a destination for the study area. The driver could, however, also be leaving the study area for a short time within a trip that begins and ends inside the study area. It is assumed that when a driver leaves the study area for 2 min or longer, he has actually left the study area. This assumption leads to the third rule of the PVD processing:

A vehicle is defined to have left the study area if it dwells outside it for 2 minutes or longer. The first measurement with speed larger than 0 km/h that is detected from a vehicle after it enters the study area is an origin. The last measurement with speed larger than 0 km/h that is detected from a vehicle before it leaves the study area is a destination.

7.3.1.4 Rule 4: First and Last Measurements from a Vehicle

The first and last measurements with speed larger than 0 km/h that are detected from a vehicle are an origin or a destination. This leads to the fourth rule of the PVD processing:

The first measurement with speed larger than 0 km/h that is detected from a vehicle is an origin. The last measurement with speed larger than 0 km/h that is detected from a vehicle is a destination.

Surely, more than one rule can apply for one measurement, for instance, there can be a real stop both before and after a measurement, a break right after the first measurement from a vehicle or just before the last measurement from a vehicle. In those cases, these measurements are according to the rules considered to be both an origin and a destination. That is, however, not logical; thus these measurements are neither assigned an origin nor a destination.

If information regarding the taxis’ occupancy would be added to the PVD, rules based on assumptions would not be needed to detect the origins and destinations, since the additional information would show clearly when a trip starts and ends. However, if other vehicles than taxis were equipped as probe-vehicles, these kind of rules would be necessary.

---

4 This variable was tested in a sensitivity analysis.
5 This variable was tested in a sensitivity analysis.
7.3.2 A Priori Matrix Estimation with PVD

When the origins and destinations within the PVD have been determined, OD matrices can be derived directly from the dataset (these matrices are hereafter referred to as OPVD matrices). As previously mentioned, one of the disadvantages of the PVD is that they are only from a sample of the whole traffic. Hence, these OPVD matrices include only a sample of the total traffic volume. A first question is whether taxi trips are sufficiently representative for all trips. Since in Chinese cities taxis are cheap and are often used by many people, it can be assumed that the OD pattern of taxis is similar to the OD pattern of all traffic. When the estimation time periods are small, it is thus highly likely that the OPVD matrices have no measured trips for OD pairs with low values. Due to this, the OPVD matrices might not include enough information to serve as good a priori matrices. In order to amend that, their missing values need to be replaced and they should be scaled up to match the real traffic volumes.

A rough method to do this would be to replace all the missing values with a low value, like 1 for instance (these matrices are hereafter referred to as PPVD). But since the values in the OPVD matrices are already very low, that would make the structure of the PPVD matrices rather uniform and thus probably different from the actual OD matrices, plus this does not solve the problem of the too low traffic volumes.

Another solution is to use a complete OD matrix from the area, both to fill in the missing measurements and to scale the matrix so that it matches the total traffic volumes. That matrix can, for instance, come from a historical database, a traffic survey, or another estimation method. In this way, additional information is added to the matrices that are traditionally used as a priori matrices. The resulting matrix will hereafter be referred to as CPVD. In Fig. 7.1, a procedure for this replacement of missing measurements and scaling is suggested.

7.3.3 Route Choice Analysis with PVD

After the PVD have been mapped to the network, using the method of Chen (Chen et al. 2007), they contain information about the link on which the vehicles are driving as well as the driving direction and location on link. Thus, the paths of the vehicles can be traced through the network. The measured paths, however, are sometimes not complete. Due to, e.g., missing measurements, measurement errors, or just the fact that there is normally a whole minute between two measurements, links are often missing to connect two consecutive measurements. In order to construct a complete route, the computed shortest paths were inserted between the links when needed.

When all the used routes have been constructed, a route choice analysis can be made. This route choice analysis can then be used for the mapping of OD flows to the network. There are three potential situations that exist for all the OD pairs in the
network and for those situations three corresponding measures are used to analyze the routes. These situations and measures are the following:

- **There is no detected route between an OD**: A route is found to compensate for the missing information. That route can, e.g., be found using either static or dynamic traffic assignment. Here the shortest calculated paths are used.
- **There is only one detected route between an OD**: That particular route is used.

**Fig. 7.1** A suggested process for deriving CPVD
– *There are more than one route detected between an OD:* For all the links in those routes the parameter $p_{ij}^a$ needs to be found.

The parameter $p_{ij}^a$ indicates the fraction of traffic between zones $i$ and $j$ that travel via link $a$.

### 7.3.4 Trip Length Distribution Analysis with PVD

The TLD can be obtained in two different ways, which are described below.

#### 7.3.4.1 TLD Obtained Directly from PVD

When the paths within the PVD are completed, the trip lengths can be calculated. The length of all the links in the network is known. Hence, the lengths of the links used for each trip can be added up.

#### 7.3.4.2 TLD Calculated from Estimated OD Matrices

Final estimated OD matrix is used to calculate the trip lengths. In the OD matrix, the number of trips between each OD pair is given, and with that information and the computed shortest paths, the TLD can be calculated.

In this paper, the TLD is only used in order to estimate whether the distribution of trip lengths is logical. It is, however, possible that the TLD can be used as an additional constraint for the OD matrix estimation (Van Zuylen 1981). That approach will later be reported in another paper.

### 7.4 Results

In this section, the sensitivity of the defined rules is tested, the driving behavior of taxis is examined, and the results of the estimations done with the PVD are described. From that discussion, answers to the two questions regarding PVD that were raised in the beginning of this paper are given.

#### 7.4.1 Sensitivity Analysis

Since the rules that determine the origins and destinations within the PVD dataset are based on assumptions, one might consider the reliability of the outcomes.
Due to this, a sensitivity analysis is performed for the parameters real stop and break. The parameters are tested separately with data recorded during two days, May 9 and 10, 2007. The trip distribution, i.e., the distribution of the number of trips per hour of the day, is examined as well as the total number of trips. In order to test either parameter, several calculations are done with a different value of each parameter while the other one is fixed to 2 min.

### 7.4.1.1 Sensitivity of the Parameter Real Stop

The parameter break is fixed to 2 min while the parameter real stop is changed for each calculation. Seventeen calculations are performed with the value of real stop ranging from 1 min to 1.440 min (24 h). Figure 7.2 shows a comparison between the trip distributions for different values of the parameter real stop for the data collected on May 9.

Figure 7.3 shows how the number of detected trips on May 9 changes with different values of the parameter break (both in the case when intrazonal trips are included (all trips) and for interzonal trips only).

---

6 In order to simplify the calculations and decrease the size of the dataset, all measurements from outside the study area were deleted. By doing that, rules 2 and 3 were combined. Hereafter the parameter for these combined rules will be referred to as break.

7 Intrazonal trips are trips that start and end in the same zone.

8 Interzonal trips are trips that start and end in different zones.
From these figures, it can be concluded that variation of the parameter real stop will change the structure of the trip distribution only slightly, while the change in number of trips is quite substantial. The number of trips decreases when the parameter is increased until it stabilizes. The number becomes more or less stable when the parameter is set as 500 min and larger, and it becomes completely stable when the parameter approaches the size of 24 h. The results are very similar for May 10.

7.4.1.2 Sensitivity of the Parameter Break

The parameter real stop is fixed to 2 min while the parameter break is changed for each calculation. Seventeen calculations were performed with the value of break ranging from 1 min to 1.440 min (24 h). Figure 7.4 shows a comparison between the trip distributions for different values of the parameter break for the data collected on May 9.

Figure 7.5 shows how the number of trips on May 9 changes with different values of the parameter break, both including and excluding intrazonal trips.

From these figures, it can be concluded that variation of the parameter break will change the structure of the trip distribution only slightly (though a bit more than when changing the parameter real stop) while the change in number of trips is quite substantial. Interestingly, the number of trips increases in the beginning when the parameter break is rather low. This is due to the definition that a measurement is neither an origin nor a destination if there is a break or a real stop both before and after it.
Hence, when the break parameter is small, several origins and destinations are cancelled out. The number of trips becomes more or less stable when the parameter is set as 500 min and larger, and it becomes completely stable when the parameter approaches the length of 24 h. The results were very similar for May 10.
7.4.1.3 Conclusions from the Sensitivity Analysis

The sensitivity analysis of the parameters real stop and break shows that the values of both parameters do not have a great effect on the trip distribution. They, however, affect the total number of trips considerably.

When an a priori matrix is constructed, it can be scaled up to match the total amount of traffic. Hence, the trip distribution is of more importance than the total number of trips. The route choice analysis is based on the location of the vehicles and the values of the parameters do not have any effect on that. It is therefore concluded that the values of the parameters should be kept according to the previously mentioned assumptions, both as 2 min.

In order to eliminate all doubts about when the trips are really beginning and ending, additional information from the PVD is required, i.e., information about the taxi’s occupancy. That would considerably increase the reliability of the information derived from the PVD. This is, however, only possible when the probe-vehicles are taxis.

7.4.2 The Driving Behavior of Taxis

One of the two important questions that were raised in the beginning of this thesis is whether data from taxis are representative for the whole traffic. A large part of the traffic in Chinese cities consists of taxis, yet it is highly likely that their driving behavior differs from the driving behavior of the common driver.

The trip distribution derived from the PVD of four different days (May 9, May 10, June 6 and June 7, 2007) can be seen in Fig. 7.6. The general trip distribution for this area is not known. However, it can be seen that the PVD trip distribution is rather consistent between days. The distribution of May 9 differs from the other distributions in the afternoon period, but apart from that, the difference between those days is not large. Thus, it can be concluded that these distributions are “typical” taxi trip distributions and therefore outcomes from the PVD analyses should be consistent between days.

If a survey were done for the study area and a “typical” trip distribution constructed, it would perhaps be possible to use that trip distribution to scale the PVD so it would match the total traffic. Another possible way to correct this difference would be to adjust the OD matrices using the TLD of the PVD and the real TLD.

7.4.3 Estimations with the PVD

In this section, the two different estimations made with the PVD are discussed, i.e., the a priori matrix estimation and the route choice analysis.
Earlier in this paper, three different ways of building an a priori matrix from the PVD were discussed:

- OPVD: The original PVD OD matrix
- PPVD: The original PVD OD matrix with the value 1 inserted where values are missing.
- CPVD: The original PVD OD matrix with scale and missing measurements adjusted by using another matrix (apOD) from the study area.

The impact of those a priori matrices and additionally a unit matrix, UOD (with value 1 for all cells), on an estimated OD matrix can be compared with a reference matrix. The reference matrix, apOD, is a complete OD matrix, based on a survey of all major junctions in the study area where all turning fractions are available for given time slices. It is estimated with a dynamic OD matrix estimation method called REMDOE (Chen 1993).

In Table 7.1, all main outcomes from the estimations when different a priori matrices were used can be seen. Interesting remarks can be drawn from the table:

- UOD: it systematically underestimates the total matrix. Counts matching is also the worst.
- OPVD: the original PVD matrix, though partial, does capture the total reference matrix apOD. Though its effective OD pairs are just 40% of the reference one, it performs well.

Fig. 7.6 The trip distribution of the taxis from four different days
Table 7.1 Estimation indicators for various a priori matrices

<table>
<thead>
<tr>
<th>A priori matrix</th>
<th>Initial counts mapping</th>
<th>OD matrix update</th>
<th>Output Counts matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Deviation</td>
<td>DifMax</td>
<td>DifMin</td>
</tr>
<tr>
<td>UOD</td>
<td>94.8</td>
<td>0.0</td>
<td>−99.2</td>
</tr>
<tr>
<td>OPVD</td>
<td>87.9</td>
<td>671.4</td>
<td>−100.0</td>
</tr>
<tr>
<td>PPVD</td>
<td>90.0</td>
<td>678.0</td>
<td>−94.2</td>
</tr>
<tr>
<td>CPVD</td>
<td>73.1</td>
<td>611.8</td>
<td>−81.1</td>
</tr>
<tr>
<td>apOD</td>
<td>60.4</td>
<td>546.6</td>
<td>−87.4</td>
</tr>
</tbody>
</table>
- PPVD: extra information to the OPVD improves the estimation as depicted. It is comparable with the reference matrix.
- CPVD: this is the combination of PVD with apOD. Though its performance is better than PPVD, it needs the information from apOD.

To get further insight into the quality of estimated matrices, a dynamic simulation is performed with Dynasmart, to check the traffic states. Table 7.2 shows the main indicators, and provides the following observations:

- UOD and OPVD alone do not provide sufficient information to obtain a complete OD output. Its combination, PPVD, however, does lead to a sufficient quality.
- PPVD, when combined further with apOD, provides even better results than apOD, when looking at simulation and output.

Interestingly, this shows that PVD does provide very good information to estimating a dynamic OD matrix. That can lead to the following findings:

- A partial PVD matrix, when combined with a unit matrix, can approach to the performance of a high-quality survey a priori matrix. This is significant, as PVD info is readily available while it is very costly to organize a high-quality dynamic survey.
- When a partial PVD matrix is combined with a survey a priori matrix, it outperforms the quality of the latter.

### 7.4.3.2 Route Choice Analysis

The numbers of trips detected on the 4 days that were considered in this study are the following:

- May 9: 13,561 trips
- May 10: 15,540 trips
- June 6: 4,427 trips
- June 7: 12,769 trips

For all the OD pairs detected in the PVD within each time slice (10 min), the route choice was analyzed. After missing paths between OD pairs have been compensated (here with the calculated shortest paths) and the value of $p_{ij}$ has been calculated, the routes can be used directly as an input for an OD estimator.

### 7.4.4 Trip Length Distribution

Earlier in this paper, two methods were suggested to derive TLD from PVD, firstly directly from the data and secondly from a complete estimated OD matrix. In both cases, short trips were overestimated compared to a typical TLD. That can be
<table>
<thead>
<tr>
<th>Matrix</th>
<th>Total vehicles</th>
<th>Tagged vehicles (out)</th>
<th>Total travel times</th>
<th>Average travel times (min)</th>
<th>Total entry queue times (h)</th>
<th>Total stop time (h)</th>
<th>Total trip distance (miles)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UOD</td>
<td>35,332</td>
<td>31,895</td>
<td>2,444.55</td>
<td>4.15</td>
<td>97.52</td>
<td>927.96</td>
<td>39,305.19</td>
</tr>
<tr>
<td>OPVD</td>
<td>60,991</td>
<td>44,863</td>
<td>6,622.34</td>
<td>6.51</td>
<td>1,573.60</td>
<td>3,462.93</td>
<td>53,416.51</td>
</tr>
<tr>
<td>PPVD</td>
<td>60,981</td>
<td>39,739</td>
<td>7,050.89</td>
<td>6.94</td>
<td>2,430.67</td>
<td>4,087.80</td>
<td>48,526.17</td>
</tr>
<tr>
<td>CPVD</td>
<td>59,160</td>
<td>37,352</td>
<td>7,418.40</td>
<td>7.52</td>
<td>2,065.96</td>
<td>4,326.79</td>
<td>48,896.23</td>
</tr>
<tr>
<td>apOD</td>
<td>57,892</td>
<td>37,668</td>
<td>7,379.70</td>
<td>7.65</td>
<td>2,290.43</td>
<td>4,286.49</td>
<td>47,967.47</td>
</tr>
<tr>
<td>UOD</td>
<td>61.0%</td>
<td>84.7%</td>
<td>33.1%</td>
<td>54.3%</td>
<td>4.3%</td>
<td>21.6%</td>
<td>81.9%</td>
</tr>
<tr>
<td>OPVD</td>
<td>105.4%</td>
<td>119.1%</td>
<td>89.7%</td>
<td>85.2%</td>
<td>68.7%</td>
<td>80.8%</td>
<td>111.4%</td>
</tr>
<tr>
<td>PPVD</td>
<td>105.3%</td>
<td>105.5%</td>
<td>95.5%</td>
<td>90.7%</td>
<td>106.1%</td>
<td>95.4%</td>
<td>101.2%</td>
</tr>
<tr>
<td>CPVD</td>
<td>102.2%</td>
<td>99.2%</td>
<td>100.5%</td>
<td>98.4%</td>
<td>90.2%</td>
<td>100.9%</td>
<td>101.9%</td>
</tr>
<tr>
<td>apOD</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
explained with the fact that the study area is merely 6 km². Thus, all trips within the area are rather short (the longest calculated shortest path is 3.5 km). In order to get a better TLD, the study area needs to be enlarged. With a better TLD it would be possible to scale the a priori matrices or the estimated matrices so that they fit a typical TLD, and hence the whole traffic. For a proper dynamic mapping of the OD flows into the network, relatively many OD pairs need to have detected routes and preferably more than one.

Table 7.3 shows, for the 4 days, the proportion of OD pairs that have detected routes during each time period.

The highest percentages in the table above are just around 25%, which means that minimum 75% of OD flows are mapped into the network in a traditional way.

From this discussion, it can be concluded that these PVD alone do not comprise enough information to sufficiently determine the route choice. They can, however, be used in combination with traditional mapping methods, by which the information is maximized.

### 7.4.5 Answers to Stated Questions

In the beginning of this paper, two questions were stated. In the previous sections the following answers to those questions have been found:

*Data from taxis may not be representative for the whole traffic, but the data is consistent and the bias can be adjusted for.*

PVD alone do neither comprise enough information to build a good a priori matrix nor analyze the route choice. Other information sources need to be combined with the PVD, in that way the used information is enhanced. PVD, when combined with a survey a priori matrix, can provide high quality input for the estimation of an OD matrix.
7.5 Conclusion and Further Work

The results of this paper are very promising. It is shown that using PVD for a priori matrix estimation and mapping OD flows to a network is well feasible. When the derived information from the PVD is fused with traditional methods of a priori matrix estimation and mapping of OD flows to a network, the quality of the used information and consequently the quality of the estimated OD matrices is increased. Simple methods to fuse the different data are also suggested and tested. This gives good results, though the validity of the estimated matrix and route choices has not been proven yet since the real OD matrices and routes are not known.

It is likely that the driving behavior of taxis might differ from the total traffic (although in Chinese cities taxis are so frequently used by so many people that taxi trip patterns might be very similar to the pattern of the total traffic). In this paper, this bias is not adjusted for but two methods are suggested, firstly to adjust the trip distribution of the PVD to the real trip distribution of the area and secondly to use the TLD to do the same. It is recommended that these topics will be examined in further studies of the matter. Thus, the real trip distributions and TLD need to be found and, in order to avoid overestimating short trips and getting a good TLD from the PVD, the study area needs to be relatively larger than the one used here.

References


Brockfeld E, Passfeld B, Wagner P (2007) Validating travel times calculated on the basis of taxi floating car data with test drivers. Proceedings of 14th world congress on intelligent transport systems, Beijing, China


Chen YS, Van Zuylen HJ et al (2007) Chengdu feasibility study for traffic information and route guidance, project report


Hamerslag R, Taale H (2001) Deriving road networks from floating car data. 9th world conference on transport research, Seoul, Korea


Taale H, De Hoog A, Smulders SA, Tool O (2000) The results of a Dutch experiment with floating car data. 9th IFAC symposium on control in transportation systems, Braunschweig, Germany
Torp K, Lahrmann H (2005) Floating cat data for traffic monitoring. 5th European congress and exhibition of intelligent transport systems and services, Hannover, Germany
Van Zuylen HJ (1981) Some improvements in the estimation of an OD matrix from traffic counts. Proceedings of 8th international symposium on transportation and traffic theory, University of Toronto Press, Toronto, Canada
Van Zuylen HJ, Van Lint JWC, Hoogendoorn SP, Bliemer M, Muller THJ, Chen YS, Taale H, Zuurbier F (2006) ITS and dynamic traffic management, lecture notes, code CT5804, Delft University of Technology
Wen H, Chen F (2007) The application of floating car system in Beijing. Proceedings of 14th world congress on intelligent transport systems, Beijing, China
Zajicek J, Reinthaler M (2007) Real time route analysis based on floating car technology. Proceedings of 14th world congress on intelligent transport systems, Beijing, China
Chapter 8
Using Probe Vehicle Data for Traffic State Estimation in Signalized Urban Networks

Henk J. van Zuylen, Fangfang Zheng, and Yusen Chen

8.1 Introduction

Probe Vehicle Data (PVD) is becoming more and more common for the collection of information about the traffic state. In most cases, the information that can be obtained from a probe vehicle refers to the position, the speed and the direction of movement at certain time intervals. Especially in urban networks, the raw GPS data needs a cleaning process to map the measured position to the road network. The cleaned information about positions in the network at fixed moments where GPS signals are collected can be used to derive travel time along certain routes.

Even though such information is very useful for monitoring purposes, it has less value if one wants to predict travel times. Simple extrapolation of travel times does not sufficiently reflect the complicated processes in signalized urban networks. Furthermore, travel times are intrinsically uncertain in urban networks. The delay at a signalized intersection depends on the arrival moment: a car arriving at the beginning of the red phase will experience a longer delay than a vehicle that arrives at the end of the green phase after the queue has disappeared. If the travel time of two consecutive probe vehicles would be measured where the first one passes the intersection without delay and the second one arrives just at the beginning of the red phase, the extrapolation of the travel time would give completely useless and invalid results. This chapter clarifies this problem and gives a method to deal with the intrinsic uncertainty of probe vehicle data for state estimation.

From the position and speed of vehicles, an estimate can be made of the delay experienced by a vehicle at an intersection. Figure 8.1 shows the simplified trajectory and the way in which the delay at the signalized intersection can be defined.

H.J. van Zuylen (✉)
Delft University of Technology, Delft, The Netherlands
e-mail: H.J.vanZuylen@tudelft.nl

The delay depends on the difference between the arrival time \( t \) and the start of the green phase, and the queue at the moment of arrival.

From empirical data on positions, the actual delay can be estimated, as shown in Fig. 8.1. The positions of the PVD that are used for the estimation of the delay should be far enough from the stop line to estimate the delay properly. The delay between two PVD points is calculated as:

\[
T_{\text{delay}} = (T_2 - T_1) - T_{\text{free}}
\]  

(8.1)

In (8.1), the free flow travel time can be estimated as:

\[
T_{\text{free}} = \frac{(L_1 \times (1 - P_1) + L_2 \times P_2)}{V_{\text{ff}}} \]  

(8.2)

where \( T_1, T_2 \) are consecutive PVD time stamps when information (e.g., positions or speeds) are recorded; \( L_1, L_2 \) are length of link 1 and link 2; \( P_1, P_2 \) are fraction positions on link 1 and link 2, respectively; \( V_{\text{ff}} \) is the free flow speed (speed limit was chosen as the free flow speed in our study 40 km/h).

This estimated delay can be used to estimate the state of the intersection, especially the degree of saturation.

A more sophisticated method has been developed by Hellinga et al. (2008) (Fig. 8.2).

The distribution of travel times between position 1, position 2 and position 3 over link 1, 2 and 3 is done by maximizing a likelihood function that Hellinga et al. derived from some a priori assumed probabilities for stopping, distribution of
congestion over consecutive links, etc. Also Hellinga’s method can be used to derive the delay at the signalized intersection similar to (8.1):

$$\text{Delay} = TT_{\text{estimated}} - TT_{\text{ff}} = T_1 (\text{decomposed}) + T_2 (\text{decomposed}) - T_{\text{free}}$$  

(8.3)

where $T_{\text{free}}$ is the free flow travel time.

In Sect. 8.2, the characteristics of the delay at a signalized intersection are elaborated. This section shows that a specific traffic state corresponds with a range of possible delays, so that a single delay cannot determine the traffic state with certainty. The delay experienced can be seen as evidence for the status of the intersection but not as an indicator to uniquely determine the status. Basically, only probabilities of a traffic state can be obtained. A fuzzy state approach is proposed: each traffic state can have several vehicle delays and a measured delay is the input for a membership of the present state to a certain fuzzy state class. Section 8.3 analyzes the validity of the PVD data by a comparison with simulation. Section 8.4 gives a comparison of the delay estimated from the probe vehicles and simulation. The concept of fuzzy traffic states is elaborated in Sect. 8.5. In Sect. 8.6, some initial applications are described of the analysis of empirical data.

The empirical data used in this chapter have been obtained from the Chinese city Chengdu, the capital of Sichuan province (Chen et al. 2007).

### 8.2 The Delay Probability Distribution at Signalized Intersections

This section gives a formula for the probability distribution of individual delays at a signalized intersection. This distribution can be used to derive a membership function of a certain measured delay to a traffic state, as is described in Sect. 8.5. For the calculation of the delay distribution, we assume that the red phase starts at $t = 0$.
The queue length is determined by the initial queue at the start of the red phase and by the number of arrivals between \( t = 0 \) and the arrival time of a vehicle at time \( t \). Van Zuylen and Viti (2006) developed a model for the initial queue.

Given a certain queue length \( n \) at arrival, the time until the vehicle can pass the stop line is the time until the green phase plus the time for the vehicles standing in front has passed the stop line. If the queue in front is longer than the number that can pass in the green phase, the vehicle has to wait for the next green phase:

\[
W(t \mid n) = (t_r - t) + \left[\frac{(n+1)}{s_{gt}}\right]C + \left(n - \left[\frac{(n+1)}{s_{gt}}\right]\right)/s \quad \text{if } t < t_r \tag{8.4}
\]

and

\[
W(t \mid n) = \left[\frac{(n+1)}{s_{gt}}\right]C + \left(n - \left[\frac{(n+1)}{s_{gt}}\right]\right)/s \quad \text{if } t \geq t_r \tag{8.5}
\]

\( C \) is the (fixed) cycle time, \( t_g \) the green time \((t_g + t_r = C)\). The square brackets \([ \cdot ]\) mean the integer value of the expression inside the brackets. The expression \( \left[\frac{(n+1)}{s_{gt}}\right] \) is the number of full cycles the arriving vehicle has to wait until the waiting vehicles in front and the vehicle itself can depart. It is assumed that at the moment no vehicles are waiting any more in front and the signal is green, the delay has ended.

The expectation value \( E(W \mid t) \) of the delay \( W \) for a vehicle arriving at time \( t \) is given by

\[
E(W \mid t) = \sum_{n=0}^{\infty} P_n^n(t)W(t \mid n) \tag{8.6}
\]

where \( P_n^n(t) \) is the probability that there is a queue of \( n \) vehicles waiting at time \( t \).

We want to know the probability for a certain delay \( W \) for a vehicle arriving at \( t \), \( P^n(W \mid t) \). The delay functions (8.4) and (8.5) have to be converted to a function of \( t \) only. As a simplified first step, we can look at the following case of an intersection that is not oversaturated and with a neglect of random effects. Let us assume that the initial queue is zero and the queue builds up proportional to the time in the red phase and decreases proportional to the time in the green phase. In that case, the delay is given the function

\[
W(t \mid n = qt) = (t_r - t) + qt/s = t_r - t(1 - q/s) \quad \text{if } t < t_r \tag{8.7}
\]

and

\[
W(t \mid n = qt - (t - t_r)(s - q)) = (qt_r - (t - t_r)(s - q))/s = t_r - t(1 - q/s) \quad \text{if } t_r < t \leq t_r + (1 - q/s) \tag{8.7a}
\]

\[
W(t \mid n = 0) = 0 \quad \text{for } t_r / (1 - q/s) < t < C \tag{8.7b}
\]
The delay as a function of the arrival time is given graphically by Fig. 8.3.

In this simplified case, the arrival probability is uniform. Figure 8.4 shows that the probability that a vehicle has a delay between $d$ and $d + \Delta$ is given by the chance that vehicles arrive in the interval $t = W(d)$ and $t + \Delta t = W^{-1}(d + \Delta)$, where we use the inverse function of the delay function $W(t)$. It is clear that

$$\Delta t = -\Delta [dW(t)/dt]^{-1}.$$ 

Given that the arrivals are uniform over the cycle, the number of vehicles that have a delay between $d$ and $d + \Delta$ is equal to the number of vehicles arriving between $t - \Delta dt / dw$ and $t$, i.e. $-q\{dt(W)/dW\}$. The normalized probability density for vehicles $P$ is given by $P(W) = -\{dt(W)/dW\}/C$. Both expressions are valid for $0 < W < t$. The inverse mapping of the delay $W$ to the arrival time is not a single valued function and as can be seen from Fig. 8.4. The derivative does not exist at $W=0$. This is a complication that can rather simply be solved by introducing the Dirac delta function $\delta(x)$ with the following properties:

$$\delta(x) = 0 \quad \text{if } x \neq 0,$$

$$\delta(0) = \infty \quad \text{if } x = 0$$

and

$$\int \delta(x - x_0)f(x)dx = f(x_0)$$

Fig. 8.3 Queue and delay as function of the time for uniform arrivals and departures.
which makes the probability density function

$$P(W) = \alpha \delta(W) + \beta \quad \text{for} \quad 0 \leq W \leq t_r$$

(8.8)

where $\alpha = 1 - t_r / \{C(1-q/s)\}$ and $\beta = \{C(1-q/s)\}^{-1}$. 

Expression (8.8) has two terms, the delta function as representation of the derivative at $t=0$ and the second term as the derivative $dt(W)/dW$ for $t>0$.

The average delay is then

$$E[W] = \int_0^{t_r} xP(x)dx = 0.5\beta t_r^2 = \frac{t_r^2}{2C(1-q/s)}$$

(8.9)

Equation (8.9) is the well known expression for uniform delay at an under saturated intersection.

As a next step let us assume that an initial queue exists at the start of the red phase and that the green phase is still long enough to handle all traffic.

$$W(t \mid n = n_0 + qt) = (t_r - t) + (n_0 + qt) / s$$

$$= t_r + n_0 / s - t(1-q/s) \quad \text{if} \quad t < t_r$$

(8.10)
and

\[ W(t | n = n_0 + qt - s(t - t_r)) = \{ (n_0 + qt - s(t - t_r)) / s \}
= \{ t_r + n_0 / s \} - t \left( 1 - q / s \right) \quad \text{if} \quad t \geq t_r \]  

(8.11)

i.e. the expressions (8.10) and (8.11) are the same as expression ((8.7), (8.7a) and (8.7b)) except that the red time is replaced by \( t_r + n_0 / s \).

The problem becomes slightly more complicated when the initial queue becomes so large that the green phase becomes oversaturated. Figure 8.5 shows the behavior of the queue if there is an initial queue while the intersection still is undersaturated. Whether an arriving vehicle can depart in the first green phase or has to wait for the next cycle to depart, depends on the cumulative arrivals in the cycle plus the initial queue. As soon as this quantity exceeds the number of vehicles that can depart in the green time, the vehicle has to wait for a following cycle. In a similar way, one can deduce when an arriving vehicle has to wait for two or more cycles. Now the delay becomes according to (8.4) and (8.5),
The transition occurs at 
\[ t = \left( s_{tg} - n_0 - 1 \right)/q \] 
and the delay at that moment is

\[ W_i = \frac{n_0}{s} - \frac{(s_{tg} - n_0)(1/q - 1/s)}{t_r + n_0 / s - t_{tg} s / q + t_{tg} / q - n_0 / s} = C + \frac{n_0 - t_{tg}}{q} \] (8.14)

If the intersection is oversaturated, the formulas (8.12) and (8.13) still hold and the following probability distribution for the delay is also applicable.

The probability distribution of delays for the case of an undersaturated intersection is two block shape functions that may overlap as shown in Fig. 8.6:

The height is given by 
\[ D(d) = \frac{d}{dW} / C = \{C(1-q/s)\}^{-1} \].

The beginning and end points of the distribution are:

\[ W_i \] \begin{align*} 
W_1 &= t_r + n_0 / s - (s_{tg} - n_0)(1/q - 1/s) \\
&= t_r + n_0 / s - t_{tg} s / q + t_{tg} / q - n_0 / s \\
&= C + \frac{n_0 - t_{tg}}{q} \] (8.14)

Fig. 8.6 The probability distribution for an intersection with initial queue where the queue at the end of the green phase is larger than zero
\[ W_1 = C + \frac{(n_0 - t_s)}{q} \]
\[ W_2 = \frac{t_r + n_0}{s} \]
\[ W_3 = 2t_r + \frac{n_0}{s} - C(1 - \frac{q}{s}) \]
\[ W_4 = C + \frac{(n_0 - t_s)}{q + t_r} \]

The block function will be denoted by \( B(x, x_1, x_2) \) with the properties
\[
B(x, x_1, x_2) = 0 \quad \text{if} \quad x < x_1 \\
B(x, x_1, x_2) = 1 \quad \text{if} \quad x_1 \leq x \leq x_2 \\
B(x, x_1, x_2) = 0 \quad \text{if} \quad x > x_2.
\]

Using this function the probability distribution is expressed by
\[
P^W(W | n_0) = \left\{C\left(1 - \frac{q}{s}\right)\right\}^{-1}\left\{B(W, C + \frac{(n_0 - t_s)}{q}, t_r + \frac{n_0}{s}) + B(W, 2t_r + \frac{n_0}{s} - C\left(1 - \frac{q}{s}\right), C + \frac{(n_0 - t_s)}{q + t_r})\right\} \quad (8.15)
\]

If the initial queue is a stochastic quantity, the probability distribution of the delay has to be composed as a weighted sum of probability functions (8.15):
\[
P(W) = \sum_{n_0=1}^{\infty} P(W | n_0)p(n_0) \quad (8.16)
\]

where \( p(n_0) \) is the probability that the initial queue consists of \( n_0 \) vehicles. As Van Zuylen and Viti (2006) show, the probability distribution of the initial queue can be found from a Markov model. No analytical expression for this distribution could be derived yet.

If we now find that a vehicle experiences a certain delay at the intersections, we can determine the possible status of the initial queue, the overflow queue of the previous control cycle. This quantity can be seen as a characteristic of the status of the link. In Sect. 8.3, we will explain how this status can be estimated. Before we analyze that problem, we will first give an example of the probability distribution of the delay in the case that the initial queue is stochastic and has a certain probability distribution.

**8.2.1 An Example with a Stochastic Initial Queue**

As an example, we calculate the probability distribution of the queue length and delay for an approach to a controlled intersection as will be described also in the next section. (The difference is that in this case we assume, for simplicity reasons, a constant traffic situation over 90 min, while in the next section the flows are more dynamic and differ every 10 min.) The approach has three lanes, \( s = 5,400 \text{ veh/h} \). The inflow is 1,662 veh/h, the cycle time is 143 s, and the effective green time is 46 s.
In each cycle 66 vehicles arrive and 69 depart. The degree of saturation is 0.96. Due to the stochastic character of the arrivals, some cycles will be oversaturated and the queue at the start of the red phase will not always be empty. Figure 8.7 gives the probability distribution of the queue length at the start of red.

The graphical presentation of the delay distribution is given below in Fig. 8.8:

The structure of the delay distribution function is a sum of block functions, where each block corresponds to the status of the overflow queue. Furthermore, there is the delta function at delay 0. If no overflow queue or only a small one exists, no double stops occur and the probability distribution ends at the duration of the red phase (i.e. 97 s in this case). If overflow queues are longer and double stops occur, the probability distribution starts at an integer times the cycle time and ends at a time that is a red phase longer. This structure will be found empirically as described in the next section.

8.3 Delay Monitoring from Probe Vehicles

In this section, the delays calculated from PVD data are analyzed with the objective to use this information to recognize the traffic state. The PVD are collected in Chengdu, the capital of the Chinese province Sichuan. In a research project, traffic
Using Probe Vehicle Data for Traffic State Estimation in Signalized Urban Networks

Data have been collected in a central area of the city: volume counts, video with number plate recognition and Probe vehicle Data. A daily average of 2,100 taxis were driving with GPS and radio communication in the study area of 3 by 4 km as depicted in Fig. 8.9. Every minute the position and speed are recorded and communicated to a central computer.

In order to test whether the delay can be measured, taxis were identified when they drove on the links feeding into an intersection and going out of the intersection. The links between nodes 57_4_61 as shown in Fig. 8.9 are used as the test area. The delay is computed for each PVD crossing these nodes, according to the method shown in (8.1) and this delay is plotted in Fig. 8.10.

The red time of the intersection 4 is 100 s and the cycle time 144 s. The intersection was oversaturated during the peak hour, so that double stops occurred for more than 90% of the peak traffic. This explains the very wide and irregular delay distribution. The delay distribution derived with Hellinga’s method looks more similar to the theoretical distribution than the delay derived from (8.1). The distribution as shown in Fig. 8.11 looks like a combination of a sharp peak at delay zero, a more or less uniform distribution up to 40 s and some block shaped components. Apparently, we can distinguish several states in Fig. 8.11, the first part until about 50 s (the duration of the red phase) are states without oversaturation, between 50 and 90 we have light oversaturation and above that we have the situation that drivers have to wait more than two cycles.

Fig. 8.8 Probability distribution of delays with stochastic overflow queue
The measurements have been compared with simulated data as described in Sect. 8.4.

This trial shows that the theory described in Sect. 8.2 seems applicable to the situation of the Chengdu network with probe vehicles as monitoring tool but that some further improvements are needed.
Fig. 8.10  Delays estimated from PVD data at different times of the day

Fig. 8.11  The delay probability distribution (probability per 5 s time interval) on link 57_04 on 9 May, 2007 (according to (8.1) and Hellinga’s method)
8.4 Comparison with Simulation

In order to verify the quality of the PVD, the intersection and connected links were simulated and from this simulation, data are extracted from individual vehicles similar to the PVD data. Figure 8.12 shows the delay observations from the simulation and Fig. 8.13 the comparison of the simulated and observed delay distribution.

Comparison of Fig. 8.10 with Fig. 8.12 shows that there are significant differences between simulated and estimated delays. The simulated delays show a more regular pattern with a uniform distribution up to 80 s and two block functions: 90–120 s and 170–240 s. This could correspond again to the state that there is no oversaturation, slight oversaturation (one additional red phase waiting time) and severe oversaturation (two cycles additional delay).

The discrepancy between simulated and observed data is probably due to the inaccuracy in the Hellinga’s method to derive delays on one link from three consecutive probe positions. However, the qualitative picture is very similar: a diffuse pattern of the delay distribution, which makes it impossible to characterize a traffic state with a more or less unique, deterministic delay.

8.5 The Estimation of the State of an Intersection

The state of an approach to an intersection might be characterized by

1. Light traffic: no double stops
2. Intermediate traffic condition: substantial overflow queues which often cause double stops
3. Heavy congestion: more than two stops are likely.
The states 1, 2 and 3 can be considered as fuzzy states when the states are measured by the delay. For each state, several delays are possible, as explained in Sect. 8.2, depending on the moment of arrival of a vehicle at the intersection. Therefore, a certain measured delay can correspond to different states of the intersection. This can be represented by considering each state as a fuzzy one, where a certain range of delays can occur in each state and every measured delay can give evidence of more than one state.

When range $A$ of the experienced delays for a certain traffic state is represented by a fuzzy set, the membership function of $A$, $h_A(x) \times X$, represents the degree that a value ($x$) in range $A$ is compatible with the status $A$. The probability density function as introduced in Sect. 8.2 and the membership function of a fuzzy set are fundamentally different. The former refers to the expectation of occurrence of a certain random delay; the latter is the representation of a set of traffic states with soft boundaries. It can be considered as just the inverse of the probability distribution of delays. It gives the possibility that a certain traffic state is present, given a certain experienced delay. We refer to various textbooks (e.g. Zadeh 1978; Klir et al. 1997; Klir and Folger 1988; Yager and Filev 1994; Zimmermann 1996) for detailed explanations of the differences and interpretations of the membership function.

![Fig. 8.13 Comparison of the probability distribution of delays (probability per 5 s time interval) as derived from PVD data and simulation for a peak hour. The number of PVD measurements were insufficient to apply Hellinga’s method](image)
Figure 8.14 gives the membership functions for the intersection as described in Sect. 8.2.1, where the membership functions have been normalized to a maximum of 1. For instance, if a vehicle passes the intersection with a delay less than 7 s, the actual traffic state has only a membership with the light condition. If the delay is 35 s, the actual traffic state is equally member of the light as the medium condition. If a car experiences a delay of more than 1 min, the only significant membership is with the state with heavy traffic.

The height of the membership function is determined by the number of possible ways that a certain delay can be realized by different initial overflow queues. One could discuss this concept and replace it by other but possible more simple membership functions.

8.6 Some Initial Experience with the Fuzzy State Estimation

In Fig. 8.15, the distribution is shown for the two peak periods and the off-peak. It is clear that there is no sharp distinction between the traffic state characteristics in the peak and the off-peak, although during the peak hours the higher delays are more frequent than in the off-peak hours.

The classification based on light, medium and heavy traffic conditions does clearly not fully coincide with the peak/off-peak period. This shows that heavy traffic conditions exist during different times of the day, also in off-peak periods and that light traffic conditions occur during the whole day.
The intrinsic large variations in delay of vehicles arriving in a certain short time interval at the intersection and the low number of probe vehicle data make it still difficult to get a straightforward state estimation. A single observed delay does not give much evidence about the traffic state. However, the membership of the different states, calculated per time interval is a promising indicator of the traffic state.

In Fig. 8.16, the division of the travel times over distinct classes of the delay is clearly visible. Assuming that the clustering values at around 50 and 100 s are the middle of fuzzy classes and that delays above 150 s constitute the third fuzzy set, the assignment of vehicles to a delay class is easy. The state of the intersection in a certain time interval can be determined from the membership to delay classes of all vehicles arriving in that interval. Still one has to keep in mind that the traffic state is a fuzzy one with a range of possible delays, so that even if a prediction can be made of the future state, this will not imply that we can also predict travel times.

The cause of the variation of the estimated delays in Fig. 8.16 is probably due to the fact that overflow of queues happen irregularly and thus that traffic states vary within a short time period. A verification of this phenomenon is still necessary and will be done in future research.

Finally, since taxis as PVD can be seen as a sample of the total traffic, it is not unlikely that the number of PVD per time period is also a measure of the traffic volume. However, this has not been verified yet in the present data set.
8.7 Conclusion and Discussion

Probe vehicles give very rich data that can be converted into specific information about the status of the traffic. If Probe vehicle data is converted to travel time or travel speed only or when the delay data of different probe vehicles is averaged, relevant information might be ignored. With some manipulation of the individual Probe vehicle data, an estimate can be made of the delay experienced at a single intersection on the route. This delay can be seen as an indicator of the traffic state. By defining traffic states as fuzzy sets, the Probe vehicle data can be used to estimate the most applicable traffic state.

The more traditional traffic state estimation requires the measurement of traffic volumes. In cities without the possibility to install detectors at the road, the method proposed in this chapter could be a good alternative to get a state estimation and a basis for travel time prediction. It is also of particular interests to estimate junction delay with PVD, knowing that junction delay is important but difficult to estimate and is ignored in most methods till now.

The step from state estimation to state prediction still has to be made for fuzzy traffic states. The possibilities are to use a traffic model, but more simple would be to set up a transition scheme for networks of intersections of the form:

IF [(intersection. Upstream = light) AND (intersection. Downstream ≠ heavy) AND (state = medium)], THEN next. State = light

Such transition rules can be extended with a probability, so that for the next state a mixed set of state memberships can be determined and thus can be used to calculate the range of travel times for the next time period.

Fig. 8.16 Measured delays on different times of the day for the intersection in Chengdu
Much work still has to be done to fully develop the use of PVD for travel time prediction. This chapter is just a first attempt to develop a method for this purpose. The authors will proceed with that by further collecting more detailed information from other sources such as video cameras with video recognition and the analysis of video registration to derive trajectories, and fusing these data with the PVD in order to calibrate and validate the methodology.

**Acknowledgments** This research is made possible by subsidies from the Chinese National Science Council and the Transport Research Centre Delft.

**References**

Chen YS, Van Zuylen HJ et al (2007) Chengdu feasibility study for traffic information and route guidance, project report


Chapter 9
Floating Car Data Based Analysis of Urban Travel Times for the Provision of Traffic Quality

Jan Fabian Ehmke, Stephan Meisel, and Dirk Christian Mattfeld

9.1 Introduction

The management of urban traffic systems demands information for the real-time control of traffic flows as well as for strategic traffic management. In this context, state-of-the-art traffic information systems are mainly used to control varying traffic flows and to provide collective and individual information about the current traffic situation. However, the provision of information for strategic traffic management as well as for traffic demand dependent planning activities (e.g., in city logistics) is still a potential field of research due to the former lack of reliable city-wide traffic information. Recently, historical traffic data arising from telematics-based data sources provided information for time-dependent route planning, for the improvement of traffic flow models as well as for spatial and time-dependent forecasts. In this chapter, we focus on the analysis of historical traffic data, which serves as a background for sophisticated real-time applications.

Information for strategic traffic management and traffic demand dependent route planning is provided by macroscopic measures such as traffic quality. Common approaches for deriving and evaluating traffic quality are described in the Highway Capacity Manual (HCM; Transportation Research Board 2000). Traffic quality is provided by the evaluation of mean travel times in terms of a level of service concept. National traffic authorities usually rely on the HCM’s evaluation concept by definition of levels of service depending on the specific infrastructure. For example, in Germany, the HBS (FGSV 2005) provides standard approaches for deriving infrastructure specific traffic quality.

The determination of traffic quality requires traffic data as well as infrastructure data. Up to now, urban traffic data has not been available to a sufficient extent due to prohibitive census costs. Furthermore, infrastructure data comprises physiognomic attributes such as vertical alignment (length and longitudinal gradient), profile
(the number and width of lanes) as well as the function and the location of the road section. This data is not available city-wide but is necessary for setting up microscopic traffic flow models. Today, a vast amount of telematics-based traffic data allows for the determination of traffic quality.

In this contribution, we refer to the city-wide collection of traffic data based on Floating Car Data (FCD) technology, also known as Probe Vehicle Data (e.g., Lorkowski et al. 2004). We present an information system that analyzes FCD-based travel times and visualizes average traffic quality. Mean travel times are evaluated and provided for strategic traffic management. Furthermore, the use of the evaluated travel times for traffic quality dependent route planning in city logistics is demonstrated. The demonstrations are based on FCD collected in the area of Stuttgart, Germany, from 2003 to 2005.

In Sect. 9.2, we contrast the traditional approach for traffic quality determination to a telematics approach based on FCD. The main focus is on data collection. In Sect. 9.3, we analyze FCD by Data Mining methods and derive mean travel times. Section 9.4 illustrates our data evaluation approach by means of two real-world examples from the city of Stuttgart in terms of strategic traffic management and traffic quality dependent route planning. Finally, a conclusion is given in Sect. 9.5.

### 9.2 Data Collection for Traffic Quality Determination

In this section, standard approaches for the collection of traffic data and for the determination of traffic quality are described. We first illustrate the traditional approach for determining traffic quality in urban road networks. Then this approach is extended with respect to emerging technologies contributing to the field only recently. In particular, we identify the role and the functionality of telematics-based systems such as FCD for long-term traffic data collection.

#### 9.2.1 Traditional Approach

Basically, traditional approaches for the determination of traffic quality consist of three steps (cf. Fig. 9.1):

<table>
<thead>
<tr>
<th>step</th>
<th>DATA COLLECTION</th>
<th>DATA ANALYSIS</th>
<th>DATA EVALUATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(manual) traffic census</td>
<td>conventional data analysis</td>
<td>measure evaluation</td>
</tr>
<tr>
<td>method</td>
<td>collection method</td>
<td>data aggregation</td>
<td>evaluation scheme</td>
</tr>
<tr>
<td>result (goal)</td>
<td>traffic data</td>
<td>traffic measure</td>
<td>traffic quality</td>
</tr>
</tbody>
</table>

Fig. 9.1 Traditional approach: Process steps for the determination of traffic quality
1. **Data collection**: The first step in determining traffic quality refers to the collection of (empirical) traffic flow data in terms of (manual) traffic census. Several collection methods for the provision of microscopic traffic data exist.

2. **Data analysis**: Microscopic traffic data have to be aggregated to macroscopic traffic measures in the data analysis step, which comprises conventional data aggregation and statistical analysis. Average travel times are calculated or derived from traffic flow models.

3. **Data evaluation**: Strategic traffic management demands an aggregated, system-wide view of macroscopic traffic measures. To this end, traffic measures have to be evaluated. Evaluation can be done with evaluation schemes such as the HCM leading to evaluated macroscopic measures in terms of traffic quality.

The data collection step is usually carried out by stationary sensors (e.g., infrared sensors, induction loops, video surveillance) or by manual short-time census. Due to cost issues, these methods usually do not cover the whole road network but are located at only a few, significant points of the network (Gühnemann et al. 2004).

Traffic flows in urban road networks are highly fluctuant with respect to different network links and times of the day. To derive traffic quality in terms of city-wide travel times, area-wide data collection is necessary. To this end, conventional data collection methods are of limited use because they require a tremendous amount of effort. Actually, for vast parts of the city road network, no data samples are collected. In sum, city-wide traffic data collection is a costly and challenging task so far.

### 9.2.2 Data Collection by Telematics

The weaknesses of the traditional approach may be alleviated by the use of emerging technologies. In particular, data collection can be improved by using telematics systems such as FCD (cf. Fig. 9.2). FCD collection enriches or substitutes manual traffic census by the automated traffic census of travel times. For an overview of FCD projects worldwide in the context of Intelligent Transportation Systems, see Bishop (2005). A discussion on FCD applications and its recent extensions can be found in Messelody et al. (2009).

![Fig. 9.2 Telematics approach: Telematics complements traffic data collection](image-url)
FCD provide traffic data in terms of travel times for a single vehicle. A fleet of vehicles must be equipped with communication devices enabling area-wide collection and transmission of huge volumes of traffic data. Generally, aggregated FCD have become a promising approach for monitoring varying travel times (Brockfeld et al. 2007a, b; Gössel 2005). For an overview on FCD technology, see Lorkowski et al. (2005) and Breitenberger et al. (2004).

FCD are supposed to enrich or substitute traditional sensor or census-based traffic data. Therefore, FCD-based travel time data and flow data from, e.g., stationary detectors have to be merged (“fusion”). This is possible by means of the already mentioned level of service concept, which is used for the standardization and aggregation of different measurement methods’ outputs. Schmidt et al. (2008) exemplify the data fusion process in order to provide a consistent picture of the current traffic state.

This research is based on the Taxi-FCD system run by the German Aerospace Center (DLR). Taxi-FCD implements the idea of using taxis as mobile data sources (“probe vehicles”) for the collection of FCD. For a general overview on Taxi-FCD see Lorkowski et al. (2004).

The architecture of the data collection system is shown in Fig. 9.3. GPS positionings are sent to the control room, where vehicles’ individual trajectories are derived from the positionings. These trajectories are matched to a commercial digital map, and travel times are assigned (“data processing”). The raw data is filtered to eliminate bad GPS signals, implausible travel times, and nonrepresentative data, e.g., for special bus and taxi lanes where regular traffic is prohibited. In this context, DLR has developed several map-matching and data handling algorithms.

![System architecture of the Taxi-FCD system](image)

**Fig. 9.3** System architecture of the Taxi-FCD system (in: Gühnemann et al. 2004)
Usually, the data is then utilized to depict the current traffic situation for off-board navigation, for disposing fleets, and for the construction of digital roadmaps (Lorkowski et al. 2005). For the determination of the current traffic situation and short-term traffic prediction, FCD have to be smoothed because raw FCD are usually very noisy. If the penetration rate is sufficient, this can be done by simple aggregation. Otherwise, more sophisticated smoothing approaches are available (e.g., Sohr and Wagner 2008).

For more details regarding the Taxi-FCD collection method and its evaluation, we refer to Brockfeld et al. (2007a) and Brockfeld et al. (2007b). Brockfeld et al. (2007a) discuss requirements regarding the sufficient penetration rate of Taxi-FCD systems. Therefore, they give a detailed report on an urban measurement campaign and conclude that “the few taxi data may be able to represent the characteristics of the whole traffic stream”. Furthermore, Brockfeld et al. (2007b) conduct test drives to demonstrate the reliability of travel times provided by the Taxi-FCD approach. The conclusion is that differences between the travel times from the Taxi-FCD system and the test drives conducted are “in average in the range of system immanent variation”.

In sum, a Taxi-FCD record provides the travel time of a single vehicle being part of the current traffic flow. Given a fleet of taxis operating in a certain metropolitan area, it is possible to collect a huge amount of single records for almost all of the links of the traffic network considered. The structure of a resulting FCD record serving as input for the following data analysis is shown in Table 9.1.

| Time of positioning: 2003-08-01 07:01:22 | Road segment ID: 54362718 | Calculated speed (km/h): 50.73 |

In contrast to today’s FCD applications, we focus on strategic traffic management and planning purposes, i.e., traffic quality determination. Therefore, we consider huge amounts of FCD originating from long-term data collection to provide typical traffic quality city-wide. To this end, we use Data Mining as an advanced data analysis method. In the next section, we provide a Data Mining approach for the analysis of historical Taxi-FCD.

9.3 Data Analysis for Traffic Quality Determination

In traditional traffic quality determination, data analysis is usually carried out by conventional data aggregation and statistical analysis. For example, data samples are utilized for the instantiation of traffic flow models (e.g., speed-flow diagrams, hydrograph curves, cf. Daganzo 1997). However, traffic flows on urban main roads are subject to a large variety of influences leading to modeling obstacles. A detailed reconstruction of travel times from traffic flow samples is not always possible.
Further, it is often virtually impossible to generalize travel times determined for one certain network link to other links without detailed infrastructure information in terms of physiognomic information (e.g., vertical alignment or road profile) as well as the function and the location of the road section (Gössel 2005).

In contrast to the traditional approach, we rely on Data Mining for traffic data analysis (cf. Fig. 9.4). As a precondition, Data Mining requires a huge volume of network-wide traffic data. Data Mining methods allow for the sophisticated analysis and aggregation of this data. The analysis results in area-wide traffic information in terms of application-dependent mean travel times. For urban areas, GPS-based Taxi-FCD provide the necessary data volume at low costs.

Data Mining requires the implementation of a process known as Knowledge Discovery (Fayyad et al. 1996; Han and Kamber 2000). An overview of the Knowledge Discovery Process and our implementation for FCD analysis is provided in Fig. 9.5. The process comprises three phases.

- **Preprocessing**: The collected Taxi-FCD records are integrated into a single database containing historical travel time data (data integration). Erroneous data records are removed (data cleaning).
- **Data Mining**: The preprocessed FCD are then aggregated and analyzed by means of Data Mining algorithms resulting in mean travel times. Standard
algorithms are available in software packages such as WEKA (Witten and Frank 2005).

- **Verification**: The travel times generated are subject to evaluation and presentation. Furthermore, mean travel times are evaluated for the use in travel time-dependent applications such as strategic traffic management and traffic quality dependent route planning.

The three steps of the Knowledge Discovery Process are explained in detail in the next sections.

### 9.3.1 Preprocessing

The goal of the preprocessing step is the determination of data for Data Mining by handling incomplete or apparently erroneous data records. Within this phase, FCD records are revised (*data cleaning*) and therefore investigated in terms of consistency, completeness, and erroneous data. A vast amount of methods for the identification of outliers and inconsistencies as well as for smoothing values exist (cf. Han and Kamber 2000). Furthermore, traffic data (FCD) and infrastructure data (based on, e.g., a digital roadmap) must be integrated into a single travel time database (*data integration*).

All attributes are checked in terms of range, completeness, and denomination. Concerning the given Taxi-FCD records (cf. Table 9.1), the following main statements can be made:

- **Time**: Missing entries occur and cannot be replaced by estimated values. This leads to the deletion of the affected data sets because no timely interpretation of the speed value is possible. In the following Stuttgart example, only a few data sets had to be removed. The majority of data records were in the expected time range.
- **Link**: The given link ids must match with a common digital roadmap of the spatial area. Digital roadmaps represent the road network in the travel time database and provide spatial and physiognomic data for each network link. In the following Stuttgart example, some FCD records did not match due to export errors and had to be removed because a spatial interpretation of the measured speeds was not possible without compatible link data.
- **Speed**: This attribute has to be analyzed in detail. Statistical measures are calculated to identify erroneous data records, e.g., the arithmetic mean of values, minimum and maximum values as well as the values’ distributions. An example of a speed distribution for a downtown link is given in Fig. 9.6, distinguishing the relative frequencies of all Monday measurements versus measurements in the morning rush hour versus measurements at night. While all measurements depict a normal distribution around the legal speed, morning rush hour and free flow at night feature noticeable different distributions. Implausible speed observations have been discarded. In the following Stuttgart example, about 3% of all
speed values were identified as outliers (e.g., speed measurements larger than legal speed × 1.5).

Next to traffic data in terms of FCD, infrastructure data resulting from a common digital roadmap has to be preprocessed. For FCD analysis, the integration of the digital roadmap data schema with occurring link ids in FCD records is crucial for the spatial interpretation.

The result of the preprocessing step is a consistent historical travel time database supporting the following Data Mining step.

### 9.3.2 Data Mining

The Data Mining step allows for the sophisticated analysis of huge amounts of empirical data. In this context, FCD have to be aggregated to provide macroscopic traffic information in terms of mean travel times. Depending on the intention of the analysis, several Data Mining tasks have been defined in the literature (cf. Hand et al. 2001). Mean travel times can be provided with methods from the field of descriptive modeling. Next to simple aggregation (cf. Sect. 9.3.2), we use the...
cluster analysis method for limiting the volume of input data for strategic traffic management visualization and traffic quality dependent planning methods (cf. Sect. 9.3.2).

9.3.2.1 FCD Aggregation

Generally, FCD speed averages can be calculated for each link \( l \) according to

\[
v_{lw} = \frac{n}{\sum_{i=1}^{n} 1/v_{lw_i}},
\]

with \( n \) being the number of speed measurements for \( l \) within time window \( w \) and \( v_{lw_i} \) being the single vehicle speed. The result is a mean FCD speed \( v_{lw} \) for each link \( l \) and time window \( w \). The total number \( W \) of time windows has to be determined. On the one hand, the total number of \( W \) depends on the amount of data available for each time window \( w \). On the other hand, requirements of the specific application regarding time window granularity have to be considered.

In the literature several choices with respect to the setting of \( W \) have been defined. For example, Eglese et al. (2006) use 15 time windows of different length per day (\( W = 15 \times 7 \)) for time-dependent route planning, whereas Fleischmann et al. (2004) refer to 217 time windows per day (\( W = 217 \times 7 \)). In this contribution, we use 24 time windows of equal length per day (\( W = 24 \times 7 \)) due to the following reasons:

- The penetration rate compared with the time window length is sufficient to allow for an almost city-wide travel time analysis.
- The results can be compared to common analysis methods from the area of traffic research (e.g., Pinkofsky 2006). Furthermore, our approach can be demonstrated very easily.
- Typical traffic states, such as e.g. rush hours, can be obtained, which allows for the utilization of Taxi-FCD records for strategic traffic management as well as for traffic quality dependent route planning.
- The following cluster analysis approach requires a homogeneous time window granularity on all links considered.

FCD collection is characterized by varying spatial and temporal distributions. Thus, the length of the time window \( w \) should reflect the temporal and spatial variation within the associated FCD records. In Fig. 9.7, the total amount of Taxi-FCD records per hour and per weekday is shown, which allows for conclusions on the penetration rate of taxis in terms of the temporal distribution of Taxi-FCD. The penetration rate at night is usually best on weekends, whereas the penetration rate during the day is satisfactory on working days. Obviously, the temporal distribution for demand for taxi services in urban areas is reflected.

Furthermore, the spatial distribution of FCD records has to be considered. Links, which feature the highest penetration rate, are distinguished from the ones with relatively low, but still sufficient amount of data records. Generally, travel time data for trunk roads are available in a comprehensive way, whereas travel time data for minor roads are rather sparse.
Preprocessed FCD records are aggregated. Only links, which contain at least five FCD records in each time window, are considered. Thus, we make sure that the calculated average is at least meaningful for a general overview on city-wide traffic quality. An example for the relative frequencies of links regarding the total number of FCD records (2003–2005) in the Monday morning rush hour is given in Fig. 9.8. Most of the speed averages are based on at least 25 or much more FCD records, ensuring the significance of calculated averages. To overcome difficulties regarding the significance of travel time averages due to e.g., low penetration or high variation, one can also utilize the median as a representative for mean travel time.

The resulting speed averages are referred to as *FCD hourly average* (FH) (cf. Table 9.2 for an example). They are supposed to cover expected fluctuations in travel times during 24 h of the day and 7 days of the week. FCD originating from
Floating Car Data Based Analysis of Urban Travel Times for the Provision of Traffic

Public holidays are treated as Sunday data. There is no consideration of further attributes as, e.g., school holidays or special events. Therefore, additional data and preliminary analysis would be needed. Here, we focus on the calculation of general FCD averages that can be used to determine typical travel times by consideration of link lengths under the assumption that the main (but surely not all) traffic flow characteristics are covered by 24×7 time windows.

### 9.3.2.2 FCD-Based Cluster Analysis

Traffic quality dependent applications based on FH values must cope with significantly more values (e.g., up to 24×7×100,000 values for the city of Stuttgart) than in the common digital roadmap case, where every link is characterized by only one (estimated) travel time. Thus, limited memory capacity and the desire for fast and efficient management activities require the reduction of data input without accepting a significant decrease of reliability. The following Data Mining approach responds to these requirements by providing weighted FCD averages (FW).

For traffic quality dependent applications, weighted FCD averages reduce the volume of input data depending on a specified number \( k \) of groups of similar links. We cluster links according to their daily speed variation into preferably homogeneous groups. To this end, the 24 FH values of a weekday per link are normalized to values between “0” and “1” (min–max normalization). The resulting link vectors (cf. Table 9.3) are clustered with a \( k \)-means algorithm (MacQueen 1967).

The \( k \)-means algorithm is a partition-based clustering algorithm, requiring the number \( k \) of desired clusters and a distance function as input. The algorithm then iteratively minimizes the error sum of the data objects’ distances to the cluster centers. We parameterize the \( k \)-means algorithm by a Euclidean distance function. The number of clusters is determined by means of experiment. The trade-off is as follows: on the one hand, \( k \) must be large enough to give a good approximation of the actual link travel time variations. On the other hand, \( k \) should be kept as small as possible to minimize input data for use in, e.g., routing algorithms. According to Sohr and Wagner (2008) a good starting point is \( k=6 \).

<table>
<thead>
<tr>
<th>Day</th>
<th>0–1</th>
<th>1–2</th>
<th>...</th>
<th>3–4</th>
<th>...</th>
<th>8–9</th>
<th>...</th>
<th>16–17</th>
<th>...</th>
<th>22–23</th>
<th>...</th>
<th>23–24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>47</td>
<td>52</td>
<td>...</td>
<td>55</td>
<td>...</td>
<td>50</td>
<td>...</td>
<td>42</td>
<td>...</td>
<td>47</td>
<td>...</td>
<td>50</td>
</tr>
<tr>
<td>Fri</td>
<td>47</td>
<td>52</td>
<td>...</td>
<td>52</td>
<td>...</td>
<td>33</td>
<td>...</td>
<td>31</td>
<td>...</td>
<td>44</td>
<td>...</td>
<td>46</td>
</tr>
</tbody>
</table>

### Table 9.2 FCD hourly averages example for a link in km/h

<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>0–1</th>
<th>1–2</th>
<th>...</th>
<th>3–4</th>
<th>...</th>
<th>8–9</th>
<th>...</th>
<th>16–17</th>
<th>...</th>
<th>22–23</th>
<th>...</th>
<th>23–24</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun</td>
<td>Sun</td>
<td>0.4</td>
<td>0.8</td>
<td>...</td>
<td>1.0</td>
<td>...</td>
<td>0.7</td>
<td>...</td>
<td>0.1</td>
<td>...</td>
<td>0.4</td>
<td>...</td>
<td>0.6</td>
</tr>
<tr>
<td>Fri</td>
<td>Fri</td>
<td>0.8</td>
<td>1.0</td>
<td>...</td>
<td>1.0</td>
<td>...</td>
<td>0.2</td>
<td>...</td>
<td>0.1</td>
<td>...</td>
<td>0.6</td>
<td>...</td>
<td>0.7</td>
</tr>
</tbody>
</table>

### Table 9.3 Normalized FCD hourly averages example for a link
An example of clustering results is given in Table 9.4. Each cluster represents a group of links. Each link is associated with its groups’ representative vector of 24 discount factors. Altogether, 24 discount factors per weekday are provided for each link group. They can be used for the calculation of expected link speeds by time-dependent weighting of a robust average link speed.

### Table 9.4 From the cluster analysis resulting discount factors (example with six groups)

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Time 0–1</th>
<th>1–2</th>
<th>…</th>
<th>3–4</th>
<th>…</th>
<th>8–9</th>
<th>…</th>
<th>16–17</th>
<th>…</th>
<th>22–23</th>
<th>23–24</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.46</td>
<td>0.46</td>
<td>…</td>
<td>0.49</td>
<td>…</td>
<td>0.44</td>
<td>…</td>
<td>0.42</td>
<td>…</td>
<td>0.47</td>
<td>0.46</td>
</tr>
<tr>
<td>2</td>
<td>0.77</td>
<td>0.79</td>
<td>…</td>
<td>0.80</td>
<td>…</td>
<td>0.31</td>
<td>…</td>
<td>0.32</td>
<td>…</td>
<td>0.69</td>
<td>0.72</td>
</tr>
<tr>
<td>3</td>
<td>0.69</td>
<td>0.73</td>
<td>…</td>
<td>0.67</td>
<td>…</td>
<td>0.28</td>
<td>…</td>
<td>0.28</td>
<td>…</td>
<td>0.60</td>
<td>0.64</td>
</tr>
<tr>
<td>4</td>
<td>0.80</td>
<td>0.79</td>
<td>…</td>
<td>0.77</td>
<td>…</td>
<td>0.45</td>
<td>…</td>
<td>0.60</td>
<td>…</td>
<td>0.76</td>
<td>0.78</td>
</tr>
<tr>
<td>5</td>
<td>0.46</td>
<td>0.65</td>
<td>…</td>
<td>0.74</td>
<td>…</td>
<td>0.19</td>
<td>…</td>
<td>0.16</td>
<td>…</td>
<td>0.34</td>
<td>0.38</td>
</tr>
<tr>
<td>6</td>
<td>0.73</td>
<td>0.76</td>
<td>…</td>
<td>0.81</td>
<td>…</td>
<td>0.20</td>
<td>…</td>
<td>0.16</td>
<td>…</td>
<td>0.59</td>
<td>0.66</td>
</tr>
</tbody>
</table>

### 9.3.3 Verification

The verification step deals with the presentation, verification, and exemplary application of aggregated data resulting from the Data Mining step. More on verification approaches in general can be found in Han and Kamber (2000).

In this context, resulting travel time data sets are investigated in terms of the expected temporal and spatial variations. For example, a more or less “free traffic flow” is expected at night with lower travel times in contrast to supposed longer travel times in “rush hours”. The approach is supported by expert interviews in terms of discussions with local representatives controlling the traffic information system. Mean travel times provided by the information system are supposed to provide typical traffic quality as good as possible.

Next to validation, the verification step comprises application-dependent presentation and the exemplary use of Data Mining processed data.

On the one hand, we use the geographical information system Google Earth (Google Earth KML 2.0; Google Earth Overlays) for explorative data analysis and the support of strategic traffic management. Here, high-resolution satellite pictures allow for the investigation of hypotheses on variances of travel times considering the links’ geographical positions. Furthermore, the evaluation in terms of a level of service concept helps revising assumptions regarding the expected traffic quality for the whole traffic network as well as for specific time windows.

On the other hand, the benefits of the provided travel times for planning activities are exemplified. Both examples for the verification step in terms of application-oriented traffic data evaluation are presented in the following section.
9.4 Example Application: Travel Times for the City of Stuttgart

In this section, we illustrate the data evaluation step for traffic quality determination with FCD collected for the city of Stuttgart. From a Knowledge Discovery perspective, we verify traffic information resulting from the Data Mining step (cf. Fig. 9.9).

The evaluation of macroscopic traffic measures is traditionally carried out by evaluation schemes based on HCM, which is most suitable for highways. In urban areas, evaluation schemes are available for specific parts of the infrastructure, e.g., for traffic quality evaluation on trunk roads. The general, city-wide provision of traffic quality is difficult because of the large variety of influences on urban traffic flows (cf. Sect. 9.2.1) that can hardly be described by a single evaluation scheme so far.

Two real data examples arising from the analysis of historical FCD are investigated:

- First, mean travel times are evaluated to provide traffic quality for strategic traffic management by means of an FCD-based common measure evaluation. The traffic quality of the network links is presented as a city-wide overview. Our approach provides and evaluates hypotheses regarding temporal and spatial variation of traffic quality (cf. Sect. 9.4.1).
- Second, mean travel times are utilized for the planning of itineraries in a city logistics context. Our approach supports the planning of faster and more reliable itineraries (cf. Sect. 9.4.2).

We base our experiments on Taxi-FCD collected in the urban area of Stuttgart, Germany. About 9 million itineraries are realized in the wider area of Stuttgart every day. Up to 2010, the traffic flows in Stuttgart will increase about 18% to a level of 119 million passenger-kilometers per working day (Verband Region Stuttgart 2002). Due to this challenge, Taxi-FCD are continuously collected since 2003 for Stuttgart. The data is used as one of several inputs for a traffic monitoring system. However, the system is used for monitoring and control purposes only. The DLR provided Taxi-FCD collected throughout the years 2003–2005, making a total of about 500 million data records, which have been collected and analyzed as described in the previous sections.

<table>
<thead>
<tr>
<th>DATA COLLECTION by telematics</th>
<th>DATA ANALYSIS by Knowledge Discovery</th>
<th>DATA EVALUATION by verification</th>
</tr>
</thead>
<tbody>
<tr>
<td>(automated) traffic census</td>
<td>data analysis</td>
<td>measure evaluation</td>
</tr>
<tr>
<td>floating car data</td>
<td>data mining</td>
<td>evaluation scheme</td>
</tr>
<tr>
<td>traffic data</td>
<td>traffic information</td>
<td>traffic quality</td>
</tr>
</tbody>
</table>

Fig. 9.9 Data evaluation for traffic quality determination
9.4.1 Provision of Travel Times for Strategic Traffic Management

In this section, mean travel times are evaluated for the support of strategic traffic management. The evaluation of mean travel times leads to an area-wide overview of typical traffic quality and allows as well for detailed insights in traffic quality evolvement for specific parts of the city.

We refer to a six-step evaluation scheme by Busch et al. (2004) and instantiate it with FH values (cf. Sect. 9.3.2). Traffic quality is then derived from the ratio of the “current” speed and the “ideal” speed of a network link for every hour of a weekday. The evaluation of the ratio leads to link specific traffic quality. We parameterize the ratio for every link $l$, considering the link specific daily maximum of its 24 FH values ($\max(v_{lw})$) and the current time window’s FH value $v_{lw}$ (cf. Fig. 9.10 left-hand side).

For a typical Monday, the result of the evaluation is depicted in Fig. 9.10 (right-hand side). At night, more than 80% of the network links are evaluated with levels A and B, illustrating a “free flow” network. In contrast to night hours, traffic quality generally decreases in rush hours (8–9 a.m. and 4–5 p.m.). Here, only 30% of the network links provide a traffic quality of A or B, whereas about 28% of the network links suffer from relatively high average travel times implying “bad” or “very bad” traffic quality. For a detailed spatial examination of traffic quality and typical traffic states, our information systems allow for a visualization of certain time windows by means of the geographical information system Google Earth.

For detailed insights regarding the temporal evolution, Fig. 9.11 exemplifies weekday dependent speed curves depicting the variation of average speeds along the network links of the road “Cannstatter Straße”. Here, the outbound lane features an average speed of about 40 km/h decreasing sharply on all workdays except Fridays. On weekends, no late rush hour can be observed. The inbound lane is characterized by sharply decreasing speeds in morning and afternoon rush hours on

![Levels of traffic quality in the course of an average Monday for the city of Stuttgart](image)

Fig. 9.10 Levels of traffic quality in the course of an average Monday for the city of Stuttgart
working days. The development on Fridays is slightly different; on weekends, no sharp falls can be noticed.

Altogether, well-known variations of traffic quality on weekdays can be observed from historical FCD. The evaluated travel times are feasible and interpretable and offer an aggregated view on overall traffic quality as well as local traffic quality development. Typical traffic quality evolution can be found. Thus, strategic traffic management is supported, especially the supervision and advancement of traffic flow control strategies.

9.4.2 Provision of Travel Times for Planning in City Logistics

As a microeconomic example, we utilize mean travel times for the planning of traffic quality dependent itineraries in a city logistics context. Therefore, we introduce a simple city logistics scenario located in the area of Stuttgart.
City logistics is logistics in urban areas. The focus is on concepts for fast and reliable transportation of goods in terms of cost-efficient and environmentally acceptable pickup and delivery routes. Nowadays, service providers have to consider dynamics within logistics processes, e.g., shorter delivery time, higher schedule reliability, and delivery flexibility (Windt and Hülsmann 2007). The importance of traffic quality dependent travel times for increased routing performance is widely acknowledged, but only a few approaches to the provision of traffic quality dependent planning for city logistics exist (e.g., Fleischmann et al. 2004; Ichoua et al. 2003; Van Woensel et al. 2008).

Realistic travel times for the links of the traffic network are one of the most crucial factors for the quality of planned itineraries, because reliable planning strengthens the competitiveness of a city logistics service provider (Eglese et al. 2006). Thus, we plan itineraries with different travel time data sets and compare the resulting itineraries with respect to the realization of the fastest itineraries and the most reliable travel time prediction. The itineraries’ realization is done by simulation of the planned itineraries. Here, simulation means the recalculation of planned itineraries based on “true” travel times for specific days and time slots. The required “true” travel times have its seeds in the travel time database, resulting from calendar date specific FCD of about 40 Mondays (workdays).

### 9.4.2.1 Evaluating Travel Times by Simulation

The test setting for the exemplary evaluation and application of the several travel time data sets is as follows. We investigate the traffic quality dependent planning of one example itinerary by means of four representative times of the day and three travel time data sets. The example itinerary comprises a trip from Stuttgart airport (located on the outskirts of the city) to Stuttgart main station (downtown). This route is heavily frequented. Hence, a high fluctuation of the travel times is to be expected, and planning activities should consider the varying traffic quality to ensure reliable itineraries.

Route planning is carried out for the following times of the day, which have been identified within the scope of strategic traffic management (cf. Sect. 9.4.1):

- “free flow network” (Monday 3–4 a.m.),
- “early rush hour” (Monday 8–9 a.m.),
- “average traffic” (Monday 11–12 a.m.), and
- “late rush hour” (Monday 4–5 p.m.).

Thus, we consider varying traffic quality and varying travel times for route planning. Furthermore, the following travel time data sets are taken into account:

- **Roadmap travel time** (DR): Estimated travel times associated with the links of a NAVTEQ digital roadmap (NavTeq-2005-Q1), which represents the traffic network of Stuttgart by means of 100,000 links and 128,000 nodes, serve as a benchmark for the following FCD-based travel times ($W = 1$).
Floating Car Data Based Analysis of Urban Travel Times for the Provision of Traffic

- **FCD hourly average (FH):** Travel times resulting from the aggregation of FCD in 24 time windows per day ($W=24 \times 7$).
- **FCD weighted averages (FW):** Travel times resulting from a weighted average speed ($W=24 \times 7 = 168$). With regard to the number $k$ of groups in cluster analysis, $k=6$ is chosen. We do not consider experiments with a larger number of clusters as the implied increase in data volume contrasts to the basic idea of the approach (cf. Sect. 9.3.2.2).

We provide time-shortest itineraries for each combination of representative time slot and available traffic data set. The shortest routes are calculated using Dijkstra’s algorithm (Dijkstra 1959). They are characterized by their traffic network links and their anticipated duration $d_a$. To evaluate the several travel time data sets, we compare the anticipated durations of an itinerary to its mean simulated durations based on “true” travel times. These travel times are gained from the historical travel time database. The evaluation of the certain travel time data sets works as follows:

- A simulated duration $d_{si}$ of an itinerary corresponds to a realization of the itinerary’s duration from the historical travel time database.
- The mean simulated duration $\overline{d}_i$ of an itinerary is given by $\overline{d}_i = \frac{1}{n} \sum_{j=1}^{n} d_{si}$, with the number of simulations $n$ and the simulated durations $d_{si}$.
- The comparison of anticipated versus simulated durations is done in terms of the mean difference $\overline{c} = \frac{1}{n} \sum_{j=1}^{n} |d_{si} - d_{pi}|$ (absolute value and percentage). The mean differences allow for conclusions about the reliability of the route planning performed.
- Further, the standard deviation $s = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (a_i - \overline{a})^2}$ with $a_i = |d_{si} - d_{pi}|$ is considered. The standard deviations of the average differences are taken into account for drawing conclusions about the robustness of anticipated durations.

### 9.4.2.2 Computational Results

The computational results of the travel time evaluation are presented in Fig. 9.12. Note that in the digital roadmap case (DR), there is only one anticipated duration $d_a$ resulting from only one general travel time value per link, whereas FH and FW

<table>
<thead>
<tr>
<th>data set</th>
<th>digital roadmap (DR)</th>
<th>FCD speed average (FH)</th>
<th>FCD weighted average (FW)</th>
</tr>
</thead>
<tbody>
<tr>
<td>airport – main station</td>
<td>anticip</td>
<td>sim</td>
<td>diff</td>
</tr>
<tr>
<td>3-4 am</td>
<td>11.4</td>
<td>14.8</td>
<td>3.4</td>
</tr>
<tr>
<td>8-9 am</td>
<td>22.2</td>
<td>10.8</td>
<td>95%</td>
</tr>
<tr>
<td>11-12 am</td>
<td>17.3</td>
<td>5.9</td>
<td>51%</td>
</tr>
<tr>
<td>4-5 pm</td>
<td>23.5</td>
<td>12.1</td>
<td>106%</td>
</tr>
</tbody>
</table>

**Fig. 9.12** Computational results
data allow for traffic quality dependent travel time anticipations (cf. columns “anticip”). For example, the use of DR data for route planning leads to a traffic quality dependent travel time anticipation of 11.4 min and an average simulated route duration of 22.2 min (cf. “simulated”, based on FCD “true” travel times) in the early rush hour (8–9 a.m.). The simulated route durations differ 10.8 min (or 95%, cf. “diff”) from the anticipated route duration and are characterized by a standard deviation (cf. “std-dev”) of 2.75 min.

Concerning the realization of the actual fastest itinerary, FH and FW data sets are superior to a large extent (e.g., late rush hour: 23.5 (DR) versus 18.1/18.4 min (FH/FW)) in comparison with DR data. FCD averages (FH/FW) provoke considerably smaller differences (cf. columns “diff”; DR: up to 12.1 min, FH/FW: up to 1.5/3.4 min only). The anticipated itineraries are more reliable. Note that in the DR case, simulated durations differ up to 106% from anticipated values, possibly resulting from the fact that DR-based speed values are derived by legal speeds only. In contrast, the maximum difference for FH and FW itineraries is 8% and 16%, respectively.

A comparison of the data sets in terms of robustness of anticipated durations shows that decision making based on FH data provokes relatively robust routes (cf. columns “std-dev”); the use of FW data results in slightly worse routing decisions due to cluster analysis based aggregation.

In sum, this small experiment demonstrates that route planning by means of traffic quality dependent travel times results in shorter and/or more reliable itineraries. Simultaneously, planning on FW data sets tremendously reduces the volume of data required. The calculated travel time is applicable in the city logistics context due to supporting traffic quality dependent as well as reliable route planning.

Altogether, planning itineraries in city logistics can be supported by the provision of traffic quality dependent travel times. A more sophisticated simulation experiment could gain deeper insights into the several travel time data sets, but this is not the intention of this paper.

9.5 Conclusion

In this contribution, two examples for the city-wide provision of traffic quality are investigated. The traditional approach of traffic quality determination is extended by telematics-based traffic data collection, data analysis based on Data Mining methods, and application-oriented traffic information evaluation. In particular, mean travel times are provided and evaluated for the support of strategic traffic management and for traffic quality dependent route planning in city logistics. The mean travel times result from an information system, which allows for the analysis of large amounts of historical FCD and application specific aggregation.

Traffic quality for strategic traffic management is provided by the aggregation and evaluation of FCD with a common evaluation scheme. This leads to an overview of typical traffic states. Furthermore, daily courses of speed give detailed insights for certain areas of the network.
Traffic quality dependent route planning is demonstrated by evaluating travel time data sets for city logistics. Three approaches to travel time determination are compared and evaluated for a simple routing example. The Data Mining approach enables traffic quality dependent route planning in a memory efficient way without a significant reduction of planning reliability and robustness.

This contribution focused on the static analysis of a huge volume of traffic data. In real-world applications, travel times evolve due to the change of, e.g., traffic strategies, traffic restrictions, or network configuration (infrastructure). To enhance the analysis regarding real-time application, the data analysis step must be complemented by adapting to the changing actors’ behavior in the traffic network:

- Data collection and processing must be carried out continuously. The approaches presented must be extended to the inclusion of and fusion with other data sources than FCD (e.g., stationary detectors) as well as to the inclusion of environmental data influencing the traffic situation, because the coverage of FCD technology apart from trunk roads seems to be too low for real-time applications yet.
- The presented historical travel time models must be validated and – if necessary – be adapted to the evolved traffic system behavior by implementing continuous and automated data analysis. Therefore, approaches from real-time Data Mining have to be taken into account, describing potentials and requirements for the integration of real-time data and updating information models (Schneider 2007; Cohen et al. 2008).
- The forwarding of improved traffic management and control information has to be assured by the provision of application-dependent interfaces (e.g., for road users or business applications). Modern web-based geographical information systems such as “Google Earth” provide open source frameworks and can be integrated in the presented data analysis process easily.

Finally, future work is concerned about more elaborate ways of performing the cluster analysis required for data reduction and explorative data analysis. Regarding the city logistics example, a more sophisticated comparison of travel time data sets by, e.g., a higher number of representative itineraries and data sets should be examined.

Acknowledgments A Floating Car Data base as well as managerial insights into the Floating Car Data collection approach was provided by the German Aerospace Center, Institute of Transport Research, Berlin. This support is gratefully acknowledged.

References


Chapter 10
A Cost-Effective Method for the Detection of Queue Lengths at Traffic Lights

Thorsten Neumann

10.1 Introduction

Limited road capacities and an increasing traffic volume are, or will become a serious problem for urban mobility in many regions of the world such as Europe, China, Japan, or the USA. To ensure an acceptable level of traffic quality, local authorities need reliable traffic state information which can be used for the optimization of traffic management, e.g., for improvements in the control of traffic lights.

So, with regard to the inner-city problems, the detection of delay times or queue lengths at traffic lights becomes very important. Unfortunately, traffic monitoring by classical means such as loop detectors has several drawbacks, not least the high costs for infrastructure. Besides approaches like video observation or other local detection methods, the so-called floating car (also probe vehicle) technology has become an interesting candidate for future traffic monitoring.

Several cars from the overall traffic flow communicate their positions (GPS, Global Positioning System) together with the corresponding time stamps to a traffic management center within periodic time intervals (Schäfer et al. 2002) or at predefined locations of the road network (Kerner et al. 2002). Additional information such as speeds or weather conditions observed by sensors of the vehicles (XFCD, Extended Floating Car Data) can be transmitted as well (Linauer 2005).

Many different types of floating car systems have been implemented during the past 10 or 15 years. Most of them use speed data as original information (Skwarek and Lampl 2007) or are based on travel times (Schäfer et al. 2002; Kerner et al. 2005), which cannot be measured by other detection methods in such an easy way. For the given two consecutive data points (positions) of the same floating car, the trajectory of the vehicle through the road network between these two points can be reconstructed by suitable routing algorithms. Accordingly, the time interval between the corresponding time stamps directly yields the travel time for the driven trajectory.

T. Neumann (*)
German Aerospace Center, Institute of Transportation Systems, Rutherfordstr. 2,
12489 Berlin, Germany

J. Barceló and M. Kuwahara (eds.), Traffic Data Collection and its Standardization,
International Series in Operations Research & Management Science 144,
So, the easiest and perhaps the most common way to implement a traffic information system on the basis of FCD (Floating Car Data) is that the floating cars communicate their positions (and speeds etc.) just every few seconds or minutes.\(^1\) Floating car systems of this type are working at some German cities like Berlin, Hamburg, Stuttgart, or Nuremberg, for example.

Of course, travel times as described above can directly be used for navigation decisions and route guidance. They make it easy to find the fastest way between two given points of the road network in respect of the actual traffic situation. However, the requirements of traffic management applications such as traffic light control are slightly different. In this context, not a complete travel time map but detailed information about the traffic states at single junctions such as delay times or queue lengths at traffic lights is needed.

Several approaches have already been developed, which, however, are mostly based on loop detector data (Mück 2002; Bernhard and Riedel 1999) and which in many cases have some limitations regarding the maximum measurable queue length. Furthermore, because of financial reasons, it seems to be unrealistic to get whole cities with their complex road networks covered by that kind of detection infrastructure. Hence, other methods are needed, which rely on more cost-efficient data sources and which allow for the integrated use of all available traffic information.

In the following section, a new data fusion approach based on FCD is described which at least provides estimations of queue lengths for urban road sections with traffic light controlled outflow. Thus, the requirements on the used floating car systems are minimal in a way so that the new method can handle most of the existing and currently collected floating car data. Accordingly, it embodies a very cost-effective way to get area-wide information needed for typical traffic management applications and can be implemented without any additional infrastructure.

### 10.2 Description of the New Method

With regard to inner-city traffic, it is an important fact that clustering of vehicles is typically observable in front of traffic lights, which are mainly responsible for the occurring delay times. Assuming that the sensor vehicles of a given floating car fleet are distributed (nearly) homogeneously amongst all cars, it is clear that there also is a kind of clustering of floating cars at the traffic lights. If now the floating cars communicate their positions independently of traffic situation and predefined

\(^1\)The frequency mostly depends on the communication costs. For example, position data from taxis or busses are collected by many cab or bus companies anyway. Therefore, no additional communication costs arise when implementing a floating car system based on these data. However, the time interval between two consecutive FCD messages of the same vehicle will typically be much larger than in the context of commercial FCD fleets with sometimes one or even more data points every 5 s from all floating cars.
locations, e.g., in periodic time intervals as is usual for most of the existing floating car systems, then this clustering effect should be observable in real data independent of possible nonsystematic GPS errors.

Indeed, Fig. 10.1 shows such a typical distribution of registered floating car positions for a road link in Nuremberg (Germany) with a traffic light at its downstream end (approximately at 465 m). The single columns specify the total number of registered floating cars at each point of the observed road section during June 2007.

Now, clustering of cars also means that at traffic lights a higher local density than in the noncongested free flow regions of the road network should be measurable. If the observed floating car positions are assumed to be stochastically independent, that implies that the probability of registering a special floating car position is directly correlated with the actual local density at the corresponding location of the road. In other words, after normalization, the profiles of local traffic densities for a given road section can be interpreted as density functions of possible probability distributions for the observed floating car positions.

Accordingly, the traffic density profiles play an important role for the structure of floating car data sets. Based on a simple traffic flow model (Nagel and Schreckenberg 1992) and using some previous results (Neumann and Wagner 2008), an analytical approach for the mathematical computation of such density profiles under stationary traffic conditions was developed (Neumann 2007). At that, using a concrete Markov chain queuing model similar to the one by Brilon and Wu (1990), which, however, especially takes the spatiotemporal structure of the traffic light queue into account and also helps to derive the queue length accurately.

Thereby, besides some parameters such as road length, cycle times of the traffic lights and maximum velocity, the density profiles $K(q)$, as well as the queue length $L(q)$ uniquely depend on the (virtual) traffic demand $q$ (vehicles per second).

Figures 10.2 and 10.3 depict a selection of analytical density profiles for several values of $q$ given some sensible road parameters, respectively, a comparison between analytically computed and simulated queue lengths for the considered traffic
flow model. Obviously, there is the typical transition from under- to oversaturation (Neumann and Wagner 2008) at the critical traffic demand ($q_{\text{crit}} \approx 0.29$).

Now, the interesting question is how to find the correct actual (virtual) traffic demand $q^*$ from a set of observed floating car positions $x_1, \ldots, x_n$ since the knowledge of $q^*$ will finally allow for the estimation of the actual queue length $L$ by setting $L := L(q^*)$.

Assuming that all registered positions result from the same $q^*$, i.e., assuming stationary traffic conditions, which can easily be realized by means of a classical or

---

2 Subject to the concrete derivation of the density profiles $K(q)$, stationarity has to be interpreted in the sense of Markov chain theory in this case.
generalized maximum-likelihood estimation. Therefore, let $K_q$ be the normalized version of $K(q)$ for all $q$, i.e., let $K_q(x)$ be the probability for registering a single floating car position at location $x$ given traffic demand $q$. Then, $q^*$ is defined in such a way that it maximizes the so-called (generalized) likelihood function (given $x_1, \ldots, x_n$)

$$f(q \mid x_1, \ldots, x_n) := w_q \times \prod_{i=1}^{n} K_q(x_i)$$

where $w_q \geq 0$ are suitable a priori weights depending on $q$. In other words, $q^*$ makes the observation set $x_1, \ldots, x_n$ to have the maximum probability among all feasible values of $q$, possibly weighted by a given a priori distribution. Thereby, this freely selectable a priori distribution allows for the incorporation of traffic data from arbitrary other sources such as video or loop detectors by increasing the weights of those traffic demands $q$, which seem to more likely due to the available additional information. Because of the great flexibility of this approach, even human intuition can be taken into account when estimating the traffic demand $q^*$ if desired.

Accordingly, by defining suitable a priori distributions, the described method can be (and to a certain degree already is) extended to a profound data fusion methodology, which combines various types of data for a highly reliable and complete traffic monitoring system.

### 10.3 Results and Discussion

For a first practical implementation, the junction B4/Nordring in Nuremberg (Germany) was chosen, which is the crossing point between the city ring road and the route from the city center to the airport. For several reasons, the analysis focused on the northern branch. The data set was given by the position data from June 2007 of approximately 500 taxis operating in Nuremberg.

Figure 10.1 has shown the distribution of all data points among the observed road section. However, the shape of this distribution varies when subdividing the data set according to weekdays and/or times of day. Figure 10.4 depicts this effect for the two most significant weekdays Friday and Sunday. Furthermore, the corresponding analytical density profiles computed and estimated by the above described methods are plotted, too. This especially shows how well the analytical results fit in with the real traffic data. Aside from stochastic effects, the shapes are nearly the same.

Obviously, the new method is able to detect temporal changes of the queue length, which is reflected by the different estimations of $q^*$ (see Fig. 10.4). The time series of queue lengths for a whole week is displayed in Fig. 10.5. As can be seen, the shortest jams are observed at the weekend while the widest congestion can be found on Friday. Since this agrees with the typically expected behavior at traffic

---

3 The classical maximum-likelihood estimation is realized when $w_q = 1$ for all $q$.
4 For example, one could expect a low traffic demand at night and a large $q$ during typical rush hours.
lights of crucial junctions among the road network, the results can be assumed to be at least qualitatively correct.

However, with regard to typical traffic management applications like traffic light control, a higher resolution for such time series is needed. For this, the data set was rearranged and subdivided hourwise according to times of day. By aggregating all data from Monday to Friday, a plausible time series of a typical working day (see Fig. 10.6) could be derived by means of the above described FCD method although the number of data points per time slice was very small.\(^5\) Figure 10.7 gives an

\(^5\)The used a-priori weights are \(w_q := 1\) for all \(q \leq q_{\text{crit}}\) and \(w_q := 0\) for all \(q > q_{\text{crit}}\) where the critical traffic demand \(q_{\text{crit}}\) in this case is \(q_{\text{crit}} \approx 0.262\) vehicles per second.

Fig. 10.4 Distribution of registered floating car positions and assigned analytical density profiles at a typical (a) Friday and (b) Sunday (Nuremberg, Germany. Junction B4/Nordring, northern branch)
impression of very low data densities, which nevertheless can be handled by the new approach in an effective way resulting in an estimated average queue length of 20.7 m in this special case.

As can be seen, the time series in Fig. 10.6 shows two significant rush hours and small jams at night. Since this again reflects the expected traffic dynamics at important crossing locations of urban road networks, the results in general can be assumed to be qualitatively correct.

Nevertheless, there are some peculiar fluctuations during the afternoon rush hour, which point to some problems of the described detection method. Firstly, traffic situations during time intervals with traffic flow near to saturation are highly unstable. For statistical reasons, this means that much more data points than during other times of day are needed to get a reliable picture of the average distribution of

---

**Fig. 10.5** Time series of queue length (weekdays) (Nuremberg, Germany. Junction B4/Nordring, northern branch)

**Fig. 10.6** Time series of queue length (times of day) (Nuremberg, Germany. Junction B4/Nordring, northern branch)
floating car positions. So, either additional data points from a historical database could be used\(^6\) or the size of the operating fleet of floating cars had to be increased to reduce fluctuations in the time series of queue lengths.

However, there is another problem, which is due to the basic assumption of the new method that traffic has to be stationary during the observation interval in a certain sense. Unfortunately, sometimes this condition is not fulfilled, i.e., especially during the afternoon rush hour. In which way a violation of this stationarity influences the estimation of queue lengths can be seen from the results at another junction in Nuremberg (Germany).

For that reason, the location Bayernstr./Münchener Str. in the south-eastern part of the city ring road was chosen because of the opportunity to compare the estimated queue lengths with pictures from a video camera, which was installed at that place. Again, the northern branch was selected for a detailed study. The data set comprised registered floating car positions from October 8, 2007 to November 18, 2007.

Then, a typical working day during this observation time interval was chosen more or less randomly, and the maximum number of congested cars was estimated manually from the corresponding video images hourwise. By this, it was possible to get a kind of ground truth for the validation of the maximum queue lengths computed by the new FCD method afterwards.

Unfortunately, the two types of data, i.e., number of cars and queue length in meters were not comparable directly. However, by scaling to the maximum value in both cases, i.e., by interpreting the maximum value of each time series as 100%, this problem could be overcome easily. Figure 10.8 depicts the resulting curves.

\(^6\)This, of course, implies a loss of actuality. Furthermore, one has to be very careful when selecting the additional data points since the described detection method expects stationary traffic conditions.
As can be seen, there is a very good agreement between real and estimated queue lengths for most of the time. Especially, the morning rush hour and the time of the afternoon jam dissolution are located extremely well. Nevertheless, something seems to go wrong during the afternoon hours.

Obviously, the FCD method does not detect the slowly growing congestion from 12:00 to 18:00 h correctly, but estimates to large jams in the early afternoon. Fortunately, there is a simple explanation for this effect. For, the continuous growth of the jam length directly implies that the basic assumption of stationarity is violated. Especially, there is no (even not approximately) equilibrium traffic during the afternoon hours. Accordingly, none of the used model based density profiles $K(q)$ adequately reflects the real traffic situation in this case. As a consequence, the new method cannot interpret the slow increase of clustering of the registered floating car positions at the traffic lights properly, but misreads it to be evident for a long traffic light queue.

Actual research activities are dealing with this problem by the approach to expand the set of relevant density profiles by those (heuristic or model based) that are more appropriate for instationary traffic conditions such as in the example above. Furthermore, extensive simulations (Neumann 2009) and measurement campaigns will allow for a detailed analysis and evaluation of the new method beyond the scope of the initial results presented here.

**10.4 Conclusions**

Aside from the described problems, which most probably will be overcome by future improvements, the new FCD method provides a very promising tool for an intelligent traffic monitoring. By using already existing and/or currently collected floating car data, it allows for an area-wide observation of urban road networks.
according to the queue lengths at traffic lights. Thereby, no additional infrastructure
is needed so that the new approach is very cost-effective compared to other detection
methodologies.

Furthermore, it can be used as a real data fusion algorithm in a very flexible way
by integrating arbitrary types of additional traffic information such as speeds, traffic
flow, or even weather conditions in the context of the described generalized maxi-
mum-likelihood estimation. Thus, the new method has got the potential to form the
backbone of a highly reliable urban traffic monitoring system.

However, mostly very sparse data densities in the context of current floating car
systems presently necessitate the aggregation over several hours of data to achieve
sensible time series as depicted in the above paragraphs. Therefore, actual results
are mainly based on historical data and only allow for several offline applications.
Nevertheless, the expansion of available floating car data, for example, by means
of innovative intervehicle or car-to-infrastructure communication and/or the opti-
mized use of the described opportunity of data fusion can be expected to make
real-time traffic monitoring possible, too.

Acknowledgments The author wants to thank Dr. Peter Wagner for fruitful discussions about the
topic of traffic data collection and traffic modeling. By interesting ideas, he and Dr. Rüdiger
Ebendt initialized the work on the method described above.

References

Heureka ’99. FGSV Verlag, Köln
Brilon W, Wu N (1990) Delays at fixed-time traffic signals under time-dependent traffic condi-
ein Verkehrsnetz mit verkehrsgeregelten Netzknoten durch Meldefahrzeuge. Patent specification
DE “100 18 562 C 1”
networks. ITS 2005 Proceedings, 44–49
Linauer M (2005) Generierung streckenbezogener Verkehrsdaten als Basis für den Einsatz in
Verkehrstelematiksystemen. University of Natural Resources and Applied Life Sciences, Vienna
haltelinienernaher Detektoren. In: Tagungsbericht Heureka ’02. FGSV Verlag, Köln
2:2221–2229
Neumann T (2007) Profile lokaler Verkehrsrichtungen auf Straßenabschnitten mit lichtsignalgesteu-
Neumann T (2009) Efficient queue length detection at traffic signals using probe vehicle data and
data fusion. ITS World Congress 2009, Stockholm, Sweden. Submitted
means of real-time floating-car data. ITS World Congress 2002, Chicago, USA
DE “10 2005 060 856 A 1”
Chapter 11
Extended Floating Car Data in Co-operative Traffic Management

Thomas Scheider and Martin Böhm

Abbreviations

ABS Anti-lock braking system
C2C Car to car
CALM Communications, air-interface, long and medium range (for TS)
CAN CAN-bus (controller area network)
CCTV Closed-circuit television
COOPERS Co-operative systems for intelligent road safety
CVIS Co-operative vehicle-infrastructure systems
DAB Digital audio broadcast
DSS Decision support system
ESP Electronic stability program
FCD Floating car data
FP6 6th Framework programme
GPRS General packet radio service
GPS Global positioning system
I2V Infrastructure-to-vehicle
ICT Intelligent communication technologies
ISA Intelligent speed adaptation
ITS Intelligent transport systems
OBD On-board diagnostic system
TCC Traffic control centre
TISP Traffic information service provider
TMC Traffic message channel
UMTS Universal mobile telecommunications system
V2I Vehicle-to-infrastructure
V2V Vehicle to vehicle

T. Scheider
AustriaTech, Donau-City-Straße 1, 1220 Vienna, Austria
11.1 Introduction

In the last 10 years, the tasks of motorway operators increased from tendering the construction, operation, and maintenance to motorway operation with a strong link to traffic management and provision of traffic information to the drivers of the single vehicles. This information was in the beginning roadside information available through Variable Message Signs (VMS) and broadcast information, e.g., by traffic status messages in the radio. In the last years, digital information has been transmitted to the single driver via the Traffic Message Channel (TMC), which informs the driver directly within the vehicle about the road- and traffic status.

11.2 Co-operative Traffic Management

Current research projects of the European 6th framework programme (FP6) like COOPERS (Co-operative Systems for Intelligent Road Safety) or CVIS (Co-operative Vehicle-Infrastructure Systems) focus on the next level of traffic management and driver information: Co-operative traffic management (Böhm et al. 2007). With such kinds of systems, it will be possible to compute and transmit safety relevant information about the current traffic and road-surface status directly to the drivers by using infrastructure-to-vehicle communication techniques. This way, the driver only gets the information needed to perform the driving task for the segment he is currently driving on. Figure 11.1 depicts the concept of Co-operative traffic management.

1. Traffic data is monitored roadside and transmitted to the Traffic Control Centre (TCC).
2. EVENT FILTER detects traffic disturbances and analyzes the nature of the incident.
3. Result is fed into the Decision Support System (DSS), which provides the road operator with a set of recommended mitigation strategies and their predicted impacts.
4. Road Operator decides on the most suitable strategy. In this example, it is an updated variable speed limit profile to reduce the speed of the vehicles approaching the incident scene including information on the cause of speed limit reduction to raise driver awareness.
5. The driver gets informed via an on-board unit of an incident ahead and consecutively decreasing variable speed limits. This way, the driver adopts the vehicle
speed to the optimal level without being stressed or forcing other drivers to hazardous manoeuvres due to abrupt reactions.

Having the wireless communication link already deployed, it is another huge benefit for co-operative traffic management to use the backward channel for vehicle-to-infrastructure communication to transmit extended floating car data (xFCD). This way, road safety relevant information that is generated by sensors in the car can be transmitted to the Traffic Control Centre. This in-vehicle generated data can be used to get an accurate overview of the whole traffic and road surface situation on the road network, by improving Step 1 – traffic data monitoring. Vice versa this data will be used to generate very specific information for the vehicles on a road segment improving Step 5 – Driver information, thus making traffic flowing more safely and effectively.

Below, the services of COOPERS are listed (McDonald 2007). Each individual service can benefit either directly or indirectly from xFCD due to improved quality and quantity of the data base. Furthermore, xFCD can help to close the gaps of static sensor infrastructure and achieve complete coverage of roadside monitoring.

1. Accident/Incident/wrong-way driver warning
2. Weather condition warning
3. Roadwork information
4. Lane utilization information (lane banning, lane keeping, auxiliary lane)
5. In-vehicle variable speed limit information
6. Traffic congestion warning
7. Intelligent speed adaptation (ISA) with infrastructure link
8. Road charging to influence demand
9. International service handover
10. Route navigation – estimated journey time
11. Route navigation – recommended next link
12. Route navigation – map update
11.3 Extended Floating Car Data

Conventional active floating car data (FCD) systems are already established and deployed all over Europe. Its main component is an on-board unit consisting of a GPS receiver and a wireless communication module, for example GPRS. This combination ensures good position accuracy and reasonable communication costs at a good deployment coverage. On the other hand, a lot of vehicles need to be equipped with extra equipment to provide a solid FCD basis. Therefore, partners such as taxi or public transport fleets are needed that report their location data in real-time. The location reports are then fed into models to derive traffic information like traffic flow or travel times (http://www.arsenal.ac.at/products/products_mob_fcd_was_de.html, accessed 06 Dec 2008). These systems have a major drawback as their equipped fleets have a restricted operational area to specific road corridors (transport fleets), cities (taxi fleet), or do not reflect the real traffic status (dedicated lanes for public transport).

Floating Cellular Data do not need any special equipment in the car beside an activated cell phone. Rough positions can be estimated from cell handovers as the vehicle progresses on the road. Beside the cell phone data of the mobile network provider, complex map matching processes and models are needed to extract reliable traffic information http://www.itisholdings.com/cfvd.asp, accessed 06 Mar 2008.

Both systems can be operated either stand-alone or fused with infrastructure sensors to enhance the traffic picture. However, they still can provide only speeds, travel times, and deviations while information for improved incident detection is missing.

Extended floating car data (xFCD) take the concept further to address exactly this issue. The location and timestamp information are enriched with vehicle status information derived from the in-vehicle bus system or additionally attached sensors. For example, reported ABS and ESP activities could indicate slippery road surface conditions or – depending on the temperature – black ice. The key aspect of xFCD is that they have the potential to indicate hazardous conditions BEFORE they turn into real incidents. Here, the road operator can already run risk mitigation strategies. Conventional floating vehicle data systems can only report travel time deviations or vehicle speeds, which indicate only the impact after the incident has happened. A key goal of COOPERS is to increase safety by preventing congestion and accidents. Therefore, the road operator needs to know about potential hazardous conditions before they turn into incidents, or if still an incident occurs, at least immediately after. Moreover, there are still certain motorway stretches with sparse sensor coverage because the deployment does not pay off. xFCD enable incident detection also on the very last black spots provided wireless communication infrastructure is available within the range of tolerated latency.

A huge amount of different sensors and corresponding vehicle status information is in the vehicle. From this range, many sensors could provide safety relevant information. Unfortunately, although most information is available via CAN interface,
its representation is neither standardized nor made public. OBD-II is the most potential standardization attempt up to now, but as it focuses on engine control information, the provision of all other status information is up to the car manufacturer (http://www.obd-2.de, accessed 06 Mar 2008). Therefore, at the common denominator, only vehicle speed, wheel revolutions per minute, and temperature are available. To fully utilize the potential of xFCD standardization of the representation formats on a wider basis would be needed.

Table 11.1 indicates the potential of vehicle sensor data contributing to a better data basis for COOPERS services. C2C remarks indicate xFCD, which is also of direct use for other vehicles.

### 11.4 Advantages for Road Operators

Even high-level road networks that already have good static sensor coverage can benefit from xFCD integration. While static sensors perform very well on section-based measurements of vehicles such as traffic flow, travel times, or road occupancy levels, they lack in spatial flexibility and therefore detection of local events such as accidents, bad road surface status, or tail end of congestions. In many cases, higher density of static sensors or CCTV can bypass the problem of spatial flexibility, but this results in significant cost increase. Furthermore, like conventional FCD, many kind of static roadside sensor types can only monitor the aftermath of an incident, i.e., traffic congestion in case of an accident. The aim of the road operator must be to prevent not only the consequences resulting in congestion but ideally also the incident as well. Using both, static roadside sensors and floating cars providing extended sensor data, the road operator gets the complete picture due to the complementary nature of those data sources.

From a cost-benefit point of view, xFCD systems have even more potential on roads where no or not much roadside sensor infrastructure is deployed. This could be urban areas, rural roads, or even whole countries, which just start to deploy ITS.
With currently emerging and fast deploying ITS using in-vehicle systems such as tolling, eCall, navigation systems, real-time traffic, and travel information, the deployment of xFCD in-vehicle functionality could be easily achieved as a by-product.

Besides deployment, lots of operational issues have to be clarified in advance. Two main economic issues are the communication costs from the vehicle to the central and the data maintenance. For both issues, it is crucial to restrict floating car data exchange and storage to only those packets that are of real added value. Besides economic aspects, also legal ones have to be clarified in advance: who is the owner of the data? Who is responsible in case of incorrect data? Is the maintainer also allowed to sell data?

11.5 xFCD Transmission Strategies

Due to the costs involved with the transmission, storage, and maintenance of xFCD, intelligent algorithms are needed that select and compose from the vast amount of available floating car data only those which are of added value for improved service provision.

Extended floating car data systems have to provide complete, timely, and cost-effective information, which are three conflicting requirements. Therefore, it is necessary to define exactly what kind of and how much xFCD is needed for the provision of each service.

11.5.1 Attribute Categorization

In COOPERS, vehicle status information is categorized in continuous data and event-triggered data. Typical continuous data are the vehicle’s position, heading, and speed. Usually, these attributes provide continuous data that are of added value. Event-triggered data provide information about a discrete number of states. It is also possible that an attribute (e.g., break force) consists of two states, where zero indicates inactivity and any other value represents the attribute in case of activity. Therefore, only the changes in states that are triggered by events are of added value. Table 11.2 shows some potential vehicle attributes and their classification.

<table>
<thead>
<tr>
<th>Continuous vehicle data</th>
<th>Event-triggered data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vehicle position</td>
<td>Break status</td>
</tr>
<tr>
<td>Vehicle heading</td>
<td>ABS/ESP status</td>
</tr>
<tr>
<td>Vehicle speed</td>
<td>Wiper status</td>
</tr>
<tr>
<td>Vehicle temperature</td>
<td>Hazard warning flasher</td>
</tr>
</tbody>
</table>

Table 11.2 Continuous and event-triggered data examples
When the composition of an xFCD packet is triggered by an event, it naturally contains at least position, timestamp, and the attribute, which triggered the composition. In many cases, the assessment of the reason for the event requires more than only the trigger attribute. For example, ESP activity could be a result of icy road, aquaplaning, bad roadwork, or speeding. Central data fusion with weather or road geometry information can tackle this issue, but if additional vehicle status information is provided, the result will be more reliable. Therefore, a matrix can be configured, which contains the necessary additional information to be transmitted together with each respective trigger attribute.

### 11.5.2 Intelligent Communication Media Selection

The implementation of the vehicle-to-infrastructure link in COOPERS is realized with two different communication types: third party cell-based (GPRS, UMTS, WiMAX) and gantry-based (CALM-IR, CALM-M5). Gantry-based communication is only possible in the range of fixed gantries, which are owned by the road operator and therefore impose no additional costs of data transmission. Cell-based communication systems provide almost full coverage, but data transmission is costly as the communication infrastructure is owned by a third party.

Intelligent transmission algorithms need to make the best use of the available communication types according to the requirements. Below, a transmission strategy on the example of COOPERS is shown:

1. Transmit safety-related information immediately via GPRS
2. Buffer continuous and event-triggered data
3. If gantry communication is available, transmit buffer via CALM-IR/M5
4. If buffer information is out-of-date, discard it

Regarding rule number 4, different configurations can be implemented. For example, instead of discarding the information it could still be transmitted via GPRS if no free-of-charge communication spot was available. This fulfils the requirement of a complete, accurate data basis, but implicates trade-offs to timeliness and costs.

The strategy above requires the vehicle to determine if an event-triggered package composition is safety-relevant for immediate transmission or not. Furthermore, it requires a routine to determine if buffered information is out-of-date and therefore not necessary anymore.

A simple way of out-of-date handling is to realize the in-vehicle packet storage as a ring-buffer. When the ring-buffer is full and a new packet is received, it discards the oldest information. It is an easy implementation, which requires no extra logic and allows implementing a small and fixed buffer size. However, in case of many event-triggered message compositions caused either naturally or by malfunction, even valid information will be discarded.
11.5.3 Feedback Channel Referencing

Despite intelligent selection of communication media, other useful strategies should be implemented additionally. “To avoid unnecessary repetition of in-vehicle message feedback-channel-referencing has to be implemented. Every message sent over air (for example: digital audio broadcast-DAB, traffic message channel-TMC) into the vehicles contains what the centre already knows and does not need to be informed about. An intelligent message management in the vehicles with feedback channel referencing will minimise the number of sent messages” (Euroregional Monitoring Expert Group 2008) (Fig. 11.2).

This strategy does not only minimize the total number of messages but prevents flooding of the communication infrastructure in case of large incidents such as wide-area congestions and hazardous weather conditions.

11.6 Data Quality Aspects

Beside engine-control information, each vehicle type and brand uses proprietary attributes and different data formats. Furthermore, data differs in temporal resolution and accuracy.

For a final implementation of an extended floating car data system relying on a variety of passenger car types and brands, it is crucial to standardize the CAN-matrix for safety-relevant vehicle status information. Unfortunately, standards and common quality definitions for vehicle bus data are unlikely to be solved in the near future. Therefore, a strong deployment of vehicles acting as probes as well as intelligent fusion algorithms are needed to create at least the overall traffic picture in a
defined and assessable quality. This is necessary as traffic data exchange becomes more and more important in the future due to the trend to co-operative data sharing and common data platforms for service provision.

Even in extended floating car data systems, the positioning information including the timestamp is the most important one. Besides providing the same information as conventional floating car data systems, a reliable and accurate position or location information enables the use of more restrictive filters, which results in improved and more reliable event detection and verification. One aim of COOPERS is to realize lane-accurate positioning in the vehicle to provide lane-dependent information to the driver as well as lane-accurate tagging of extended floating car data packets.

Improved automatic incident detection via xFCD is a huge advantage, but to exploit the full safety potential, it is necessary that the mitigation strategies are implemented in a proactive way in order to prevent hazardous conditions to turn into real incidents. This requires automatic decisions and execution of control strategies on the basis of xFCD. The key problem for this application is to assess and ensure the quality of the incident detection on basis of extended floating car data. As a starting point in COOPERS, single extended floating car data packets need to be validated via a second independent source. Even if it speeds up the automated incident detection, it still requires information from static roadside sensors for validation; therefore, the field of application is limited once again to roads with good sensor infrastructure deployment.

Another pragmatic strategy is to verify events reported by single xFCD packets with other, equal xFCD packets to ensure reliable detection without necessary support of fixed sensor infrastructure. In COOPERS, an event detected by xFCD is considered reliable when reported equally by three vehicles. “Equally” is defined by acceptance borders in the location, time, and value domain. Narrower acceptance borders increase the reliability of a detected event, but they cause a trade-off with detection rate due to timeouts until equal messages are received. If the timeout is too long or not existing, the detection system will become less dynamic due to the increased latency. While filtering mechanisms in the time domain can be established quite easily due to the common timestamps of the positioning signal they alone do not improve event detection significantly. More important are filter mechanisms in the spatial domain, but here acceptance borders are limited to the positioning accuracy of the vehicles. A basic requirement to even enable comparisons in the value domain is to have common data elements and formats. Therefore, a standardized or common intermediate format needs to be defined and implemented.

Major aspects when assessing data quality are the timely and spatial availability. Already deployed floating car data systems offer usually restricted data availability to areas (taxi fleet, public transport), corridors (transport fleet), or time (weekend lorry ban, public transport during night). To assure continuous data availability, it is necessary to additionally involve passenger cars as floating vehicles in order to cover the last gaps. Co-operative safety systems capable of positioning and bi-directional infrastructure-to-vehicle communication like COOPERS provide a solution to deploy harmonized xFCD in-vehicle functionality to all kind of vehicles.
11.7 Outlook

COOPERS and SAFESPOT, an FP6 Integrated Project focusing on vehicle-to-vehicle communication, make efforts to harmonize their sets of messages from the vehicle. This enables COOPERS to adapt its vehicle-to-infrastructure communication to the SAFESPOT protocol when it is available. A common protocol for V2V and V2I ensures deployment and acceptance on a wider basis for both systems and is a first step towards an overall co-operative safety system.

11.8 Conclusion

This paper gives an overview about co-operative traffic management in COOPERS where information from the motorway infrastructure operator is directly provided into the vehicle and adapted to the current situation. In return, the vehicle feeds the infrastructure with extended floating car data to enhance its traffic status data base.

Even for road operators with already deployed monitoring infrastructure, extended floating car data are of added value as they have the potential to identify hazards before the turn into incidents, to close the gaps in deployment of monitoring infrastructure, and to use as an additional, complementary data source. Eventhough xFCD systems relying on passenger cars would overcome spatial and temporal restrictions of FCD from commercial fleets, further standardization effort is needed to enable its deployment at a large scale.

Low transmission costs are crucial for acceptance of the system. This paper presents transmission strategies on the example of COOPERS: attribute categorization, intelligent media selection, and feedback channel referencing. Finally, data quality considerations for automated processing of events are outlined.

References

12.1 Introduction

Driving behavior observed at traffic networks varies considerably depending on the type of road section. At signalized junctions, drivers are taught to moderate their speed, and to comply with the priority rules set by the traffic light. Therefore, vehicles stop and queue up during the red phase, and they leave the junction during the green and amber phases. During these operations, vehicle driving patterns vary significantly. The way they decelerate, stop at the back of the queue and accelerate changes from driver to driver. Aggressive drivers may show to operate stronger accelerations and to respond more quickly to the right-of-way signal. During the amber phase, some aggressive drivers accelerate to clear the intersection faster, while risk-averse drivers may decide to decelerate earlier, or even brake hard to avoid passing the stop-sign after the start of the red phase. Moreover, the observed trajectories will depend on whether a queue is actually present and on its length, as well as on the road characteristics, e.g., how clear is the view upstream of the junction, whether one or more lanes are dedicated to a traffic stream, etc. Therefore, individual vehicle trajectories are found to be highly variable at signalized intersections, as individual speeds and speed variations have strong dynamic and stochastic patterns. In applications that require accurate estimates of vehicle driving modes, like when estimating concentration levels of emissions, it is fundamental to provide realistic estimates of these trajectories, and, more importantly, to derive, from these trajectories, realistic speeds and speed variations.

Collecting microscopic data at these road sections is argued, in this paper, to be fundamental to test and enhance existing traffic models. Microscopic emission models are often combined with microscopic traffic simulation models to obtain...
estimates of emission levels, for instance when assessing ITS measures or new planning strategies. However, simulation models tend to oversimplify vehicles trajectories. Even if they can model vehicle movements at fractions of a second and account for many behavioral aspects (vehicle acceleration and deceleration characteristics, car-following behavior, degree of aggressiveness, etc.), speed and speed variations are often found quite unrealistic, due to the time discretization and the limited set of behavioral and network parameters.

Microscopic simulation models are often calibrated and validated using aggregated data, such as mean speeds, densities, or travel times. In this way, differences between reality and model results in terms of speeds and speed variations cannot be captured completely. The main reason for the inconsistency between models and reality is the lack of data able to describe with sufficient accuracy the traffic process in relation with the many parameters involved in this process. In this paper, we describe a procedure to obtain a very detailed picture of the traffic process near a traffic control signal. In particular, we focus on acquiring a dataset suited for studying the longitudinal behavior of drivers at these road sections, i.e., their speed and speed variation behavior, when driving unconstrained, or in combination with their car-following behavior, or as a function of the queue and the traffic signal time and phase. We do so by collecting individual vehicle trajectories using image processing techniques, which allow one to obtain the position of vehicles at very high definition and to measure, under some degree of reliability, their speed and speed variations at fractions of a second. This dataset can be exploited in many ways. It can, and will, be used to study the variability of drivers’ behavior at these signals, to gain insight into which factors determine their actual trajectories, to calibrate and validate traffic flow models, as input for emission models, and so forth.

This paper is structured as follows. The next section describes the main parameters of drivers’ behavior at signalized intersections and how it is modeled in currently available simulation software programs. Later, we give an overview of the collected dataset, and the description of how it has been processed. Then, we analyze the microscopic data and investigate the drivers’ interactions with the traffic control operations, looking especially at their speed and speed variations upstream and when passing the traffic control. Finally, we conclude this paper and present future steps in this research direction.

### 12.2 Modeling Microscopic Driving Behavior at Signals

Traffic behavior can be modeled at macroscopic, mesoscopic, and microscopic levels. Macroscopic models provide direct relationships between macroscopic variables (e.g., average speeds, flows, and densities). In mesoscopic models, traffic flow and performance variables are instead represented through probability distributions. Examples of widely used mesoscopic programs are Dynasmart (Mahmassani 1997), DynaMIT (Ben-Akiva et al. 1998) and the Cellular Automata model (Nagel and Schreckenberg 1992). Both macroscopic and mesoscopic models have a limited
Microscopic Data for Analyzing Driving Behavior at Traffic Signals

application range and they simplify the behavior of vehicles by using specific distributions to model individual trajectories. Moreover, these models are bound to satisfy macroscopic rules between aggregated measures like speed, flow, and density.

Microscopic models have been developed with the main aim of simulating the movement of vehicles on the roads at the individual level, i.e., each vehicle movement is determined through the simulated network infrastructure at fractions of a second, and interactions with other vehicles are simulated on the basis of a few driving rules (e.g., car-following, overtaking, etc.), modeled through mathematical relationships. Examples of widely used microscopic software programs are Paramics (Cameron and Duncan 1996), Aimsun (Barceló et al. 2004), VISSIM (PTV 2003), MITSIM (Yang et al. 2000), and INTEGRATION (Van Aerde 2005). Each vehicle is characterized in these programs by a speed profile that depends on vehicle and driver behavior characteristics, on the interaction with the other vehicles, and on the traffic rules in the system. However, not all aspects in microscopic simulation models are completely realistic. For instance, in the context of signalized intersections, acceleration and deceleration behavior play an important role, not only in the interaction between one vehicle and the other, but also in their interaction with the signal control. However, simple models of motion are often used to generate the trajectories of each vehicle. In these models, vehicle movements are updated at discrete time steps. Drivers’ desired speeds, accelerations, and decelerations are randomly simulated and assigned to each vehicle loaded onto the network. Inter-driving behavior variability is thus modeled in microscopic programs. Intra-driving behavior, i.e., the capacity of a driver to maintain a constant speed is, however, deterministic in all microscopic programs. Moreover, there is always a limited number of different classes describing the heterogeneity in driving behaviors and vehicle types, often user-defined and based on rules of thumb.

Microscopic models should be tested using real world data to ensure that model results can capture the important aspects of the analyzed traffic process. Therefore, the calibration and validation are, in a microscopic model, of paramount importance, since the effect of modeling errors made at the microscopic level can grow considerably when looking at the results from a macroscopic point of view, i.e., at the network level, or, inversely, the consistency of aggregated parameters does not imply that instantaneous measures are modeled realistically. All model parameters contribute to the determination of each vehicle position as much as each individual vehicle characteristic, e.g., desired speed, acceleration, deceleration, etc. Individual vehicle positions depend, therefore, on a large set of parameters and modeling rules. However, not all parameters are directly observable in reality, and they are often strongly correlated.

Despite the importance of accurate calibration and validation of microsimulation results, default behavioral parameters are often used. These parameters may fit the context in which they have been calibrated, but they may change considerably in other road sections or traffic conditions. The Aimsun default parameters, for example, have been calibrated using observed flows and speeds of the Barcelona Ring Roads and main accesses to the city, while concerning VISSIM, the default values have been set after calibration on German highways (Fellendorf and Vortisch 2001).
These parameters have been adapted for better fitting the traffic performances in e.g., Californian freeways (Gomes et al. 2004), Texan urban and suburban areas (Kim 2006). The adaptation of VISSIM parameters for matching the performance of urban signalized intersections was investigated in a study in Puerto Rico (Gonzales Velez 2006). Microsimulation results were matched in this study with travel times collected at urban corridors.

The large majority of studies compared microscopic model results with average flows, speeds and travel times, both in the urban and extra-urban contexts, but very few stress the importance of comparing instantaneous variables such as instantaneous speeds and speed variations. Nesamani et al. (2005) used GPS signals in California to match the variability of driving patterns observed on urban corridors with the ones simulated with the VISSIM microscopic program. The study for the first time discusses the importance of calibrating microsimulation programs with microscopic data to estimate emission rates in areas where flow interruptions are determinant, such as urban signalized intersections. However, since GPS positions were recorded for only a subset of vehicles in the platoons, comparison could only be made using individual travel times.

The use of macroscopic data for calibrating microscopic simulation models has been justified in the past with the difficulty of collecting microscopic trajectory data. Thanks to the application of advanced video imaging techniques, behavioral aspects that play fundamental role especially at microscopic levels could be analyzed, and the models representing these aspects could be benchmarked in past studies (Ossen and Hoogendoorn 2007). The principle of aerial data was used to calibrate and compare car-following models at uninterrupted flow sections by observing traffic from a helicopter or from high-rise buildings (Kovvali et al. 2007). The method chosen in our study to collect data is similar to this last approach. A new research stream is currently being developed since 2003 in California, to overcome this data shortage and to provide suitable data for calibration and validation procedures to traffic researchers and practitioners (NGSIM 2003). Many studies nowadays use the NGSIM dataset, which contains individual vehicle trajectories collected in several freeway and arterial roads in the US (NGSIM 2008). By using the trajectory data provided by the NGSIM project, new approaches to calibration of microscopic models could be followed (Skabardonis 2005). Recently, the NGSIM dataset has been used in an air quality study by Treiber et al. (2008), in which a new instantaneous fuel consumption and emission model has been proposed and evaluated with simulated and real trajectories, but no real comparison between the simulation results and the results estimated with the NGSIM trajectories was done in this study. The use of microscopic trajectory data provides the opportunity to better understand the origin of noise in traffic flow, since it allows for a more detailed differentiation between noise which is simply in the parameters, and noise which is generated by the dynamics itself (Brockfield et al. 2005). However, no study has so far focused on collecting and analyzing trajectories specifically at urban signalized intersections with sufficient accuracy to analyze speed and speed variations, and to test the validity of microscopic models to simulate and evaluate measures to reduce emissions (PM, NOx, and CO₂) on urban roads. These reasons motivate the study presented in this paper.
12.3 Data Collection and Processing

In our study, we aim to obtain trajectories through which the many parameters used in microscopic models near signalized intersections can be analyzed. More specifically, we focus on the longitudinal driving behavior, i.e., the way drivers interact with the traffic control, in combination with the car-following behavior, which is expected to vary considerably with respect to motorway sections due to lower speeds and to the queuing process at the signal. To obtain this dataset, we need to: (1) choose a suitable study area to analyze the longitudinal driving behavior near signalized intersections, (2) collect (microscopic) data for calibration and validation of microscopic simulation models for generating realistic emissions and to do that we need to (3) process and smooth this data to obtain individual trajectories with a sufficiently small percentage of errors.

Therefore, the data collection process is subdivided into the following sub-tasks:

1. Specification of the basic data requirements
2. Choice of the study area
3. Data collection
4. Data conversion
5. Cleaning process
6. Data smoothing

We describe in detail how we undertook these phases in the following of this paper.

12.3.1 Data Requirements and Choice of the Study Area

To study the longitudinal behavior near and passing a signalized intersection, we choose a location and a dataset with the following requirements:

1. The road section should not have any significant change in the road geometry (lane width reduction, speed limit change, etc.) to reduce any other dynamic source in the system than the traffic light system and the queuing process.
2. It should mainly contain longitudinal movements. Overtaking and lane changing maneuvers are behavioral aspects that go beyond the scope of this research. To obtain this feature in the dataset, an appropriate study area has to be selected, possibly where overtaking is not allowed.
3. The dataset should contain a sufficient number of (consecutive) individual trajectories. This is due to the complexity of the functional relationships used to simulate the vehicle movement and for the variability of vehicles’ real driving behavior.
4. It should contain trajectories that can be traced back to a sufficient number of meters, upstream the stop-sign, to study the passing and deceleration behavior, the car-following behavior and the behavior of drivers in the presence of different queue lengths.
5. It should contain trajectories that can be traced forward up to a sufficient number of meters, after the stop signal, in order to evaluate the acceleration behavior from the stop sign.
6. The recorded time of individual vehicle positions should be as short as possible, and at least the minimum updating time of commercial microscopic software programs.
7. It should contain a small number of measurement errors and missing data to be able to derive accurate estimates of individual speeds and accelerations.
8. The dataset should contain a complete picture of the whole study area at each time step.

Especially because of this last point, we will use camera detection techniques, according to the methodology applied in Ossen and Hoogendoorn (2007).

A place that fits the above requirements has been identified in Rotterdam, the Netherlands. The Euromast Space tower is a high rise building, where at 100 m from the ground is located a restaurant with a view of the Maas River (Fig. 12.1). Aside of the tower, a traffic light controls the flows coming from two two-lane roads that merge into one two-lane section that traverses the river through a tunnel. The two carriageways run in parallel and have small difference in terms of geometry. Overall the study area lane changing is prohibited. This means that we can study a four-lane road with only longitudinal driving operations.

Fig. 12.1  Google Earth™ view of the study area and the Euromast tower
Vehicles can be seen, from the tower, up to 180 m upstream of the stop sign, while the view downstream is less clear due to the presence of trees. The traffic light is vehicle-actuated, thus green is assigned alternatively to the two carriageways dynamically. Maximum and minimum green times are set to respectively 40 and 6 s. Overall the study area speed limit is set to 50 km/h and a speed control camera is installed aside of the road. The view from the tower is quite limited downstream of the traffic signal, but it is sufficient to analyze the acceleration behavior of vehicles. High resolution pictures have been taken from the Euromast tower at 15 Hz frequency with a digital camera.

The choice of this study area has inevitably a number of limitations for the image processing method. The first is that pictures had to be taken from behind a safety glass. Shadows sometimes affect the light balance of some picture areas, and this is a source of errors that needs to be cleaned in the filtering process. Moreover, some tree hinders the view for a part of a lane downstream the signal, while the view on the left lane is quite clear and allows the collection of full trajectories for the complete length of the road analyzed. Signal phases have been recorded manually using a portable counting device and synchronized with the pictures.

Traffic can be taken from these pictures up to 120 m upstream and to 60 m downstream the stop-sign. A second problem in using these pictures to automatically detect individual vehicle trajectories is that brightness changes very often and this is a major source of error especially to obtain a good background image, as it will be explained in the next section. We explain later, in the filtering process, how we partly solved these issues.

12.3.2 Data Acquisition and Conversion Processes

To automatically detect vehicles as moving objects, a background image is generated through superposition of a number of frames. The background image is obtained by finding the grey-level that is most likely to be observed at each pixel, i.e., the color of the road pavement on the carriageways. In this way, moving objects are removed. The process was done using software programmed in MATLAB®. Only a limited number of frames can be loaded contemporarily to obtain the background image, and so it is important to choose (1) the time window for which the background image is calculated, and (2) the number of frames used for this calculation. Differently from the highway system, vehicles may in fact stand still for many frames near the stop-line. In saturated green phases, this can occur for more than 40 s. To avoid errors in the background image we loaded a number of frames randomly in 2 min intervals. We could not use longer time periods because ambient light changes and the color of the background image could not match sufficiently the color of the loaded frames, resulting in a very noisy tracking.

To correct the image distortion due to the camera lens, we apply image transformation, so that the longitudinal movements of the vehicles occur parallel to the horizontal axis Fig. 12.2. Images are later compared frame by frame to find the
consistent color differences, and the shapes and areas of the moving objects are tracked through clustering. This will allow us to estimate the vehicle type. The center of the area is recorded and will be used to derive vehicle positions.

The method described has, however, some critical limitation. On the one hand, it enables one to get a complete picture of the area. On the other hand, many vehicles tend to be grouped into one cluster when they drive too close or they can be recognized as multiple clusters typically when their color is close to the pavement color; cluster aggregation occurs often in two cases: (1) when vehicles are standing in queue, they wait in very short distances to be detected clearly by the tracking method (longitudinal cluster error), and (2) when driving in parallel on two adjacent lanes (lateral cluster error). This occurs only in the downstream view since we could subdivide the upstream parts into different pictures for each lane.

Figure 12.3 gives an example of recorded vehicle positions using image processing and clustering, and some examples of noise are found in the dataset. The picture refers only to one lane. For instance, there is clearly an isolated group of
dots, which should not belong to any vehicle. These are due to, e.g., errors when calculating the background image, or because of shadows, etc. These types of errors should be easily detected and removed from the system by considering (1) the area of the cluster and (2) a minimum number of data points describing the trajectory of one vehicle. Furthermore, the position of, e.g., the first vehicle in queue is not smooth while standing still. This error is attributable to the clustering algorithm, which determines a different cluster area for each time frame. Even a small difference in terms of number of clustered pixels can reflect into an error, and certainly this is more visible when vehicles stand still. A third type of error shown in the picture is the longitudinal clustering. If two vehicles are grouped into one cluster, the position of the centroid of the clustered area lies somewhere in between the two actual centroids. This results in a number of errors, which needs to be identified and corrected. The cleaning process is described in the next subsection.

### 12.3.3 Data Cleaning and Recording Individual Trajectories

Some of the noise in the dataset can be easily identified by looking at the area size of each cluster. If clusters are too small, or too large, they are not likely to be vehicles and they should be removed from the dataset. Maximum and minimum cluster sizes can be defined by looking at the histogram of recorded areas. We can also use information on the signal phase sequences to clean the dataset and to automatically recognize the carriageway to which each cluster belongs. This is particularly needed for the upstream direction, where the separated carriageways merge into one. The errors are further detected and cleaned by converting moving objects into vehicle trajectories, as explained in the following of this section.

So far, moving objects have been detected at each time frame. However, there is yet no information about individual vehicle trajectories, since cluster detection is done by analyzing pictures separately. The tracking of individual trajectories is here done in two steps: (1) a one-time ahead prediction and (2) prediction based on the last recorded speeds. In the first part of the process, we look at each moving object at one time frame and find the next point in time and space. A trajectory is therefore identified if a point is found, assuming a minimum (zero) and a maximum speed at which a vehicle may have moved forward. We define maximum speed \( v = 80 \text{ km/h} \).

To allow small errors at low speeds due to the object clustering, small negative speeds are also considered valid by us. These errors will be later corrected, and bound to be non-negative, using a proper smoothing technique.

Detecting trajectories by looking at only one time step interval may result in finding many partial trajectories of the same vehicle. For example, there can only be one time step where a vehicle has not been detected correctly (due to, e.g., aggregation with another vehicle), and later this error does not occur. This single error results in two separated trajectories of actually the same vehicle. We thus estimate the speed during the last second before losing the trajectory. By using quadratic
interpolation we can thus find, by comparing partial trajectories in pairs, partial trajectories belonging to the same vehicle.

Figure 12.4 gives a visual example of the result of this procedure, the thinner line showing the connection between partial trajectories. In this application, we have used the last recorded speeds to generate a range of possible positions for the vehicle in later time steps. However, this method can be equally used to connect trajectories by looking back in time, i.e., by estimating the speed at the start of each trajectory and compare it with the last recorded position of all trajectories in previous time frames.

### 12.3.4 Data Smoothing

The data processed so far still contains position errors due to, primarily, the clustering algorithm. If these errors look small on a space-time diagram, they become sensibly larger when looking at first and second derivatives, i.e., at speed and acceleration profiles. We could have used different video detection algorithms, for instance, the double-camera detection used in the NGSIM dataset, to partly overcome this error. In fact, automatic tracking of vehicles based on background subtraction has already been found in past studies to generate quite some noise in the data, with a range of error of 0.5 m on 0.2 s time intervals (Hidas and Wagner 2004). However, the algorithms used in NGSIM are also not error-free and, even if they avoid the aggregation of nearby vehicles, allowing more correct matches of detected vehicles (Kim et al. 2005), they cannot guarantee a perfect measurement of vehicle positions. Kesting and Treiber (2008) and Treiber et al. (2008) observed, when analyzing the NGSIM dataset, that it contains around 2/3 of the accelerations beyond the interval $[-3, +3]$ m/s$^2$, which is clearly an unrealistic statistic even in heavy congestion and
stop-and-go traffic. To solve this problem, they also relied on a form of moving average as we do in our approach.

Smoothing the data is therefore a fundamental step to obtain realistic speed and acceleration profiles. However, traditional smoothing algorithms (like linear or quadratic fitting) may oversimplify the real behavior of the vehicle. Some actual variations of speeds and accelerations may be smoothed out, changing the actual vehicle behavior. Advanced smoothing methods, like (Extended) Kalman Filtering, may be more opportune. However, their underlying assumption is that errors are normally distributed and independent. As pointed out by Punzo et al. (2009), such methods can be effective only when the noise is white, while the smoothing will suffer any bias or systematic error. This is the case of our dataset. Errors can be correlated and biased during a certain interval of time (e.g., a shadow, temporary longitudinal clustering, etc.). In these cases, errors cannot be considered to belong to the same distribution, nor can be considered independent. To overcome this problem, Punzo et al. (2009) propose to deal in the correction procedure with pairs of trajectories, as most of the systematic errors are caused by grouping/splitting of vehicles. This correction procedure would not, however, be effective in case of shadows and partial occlusion of vehicles.

For this reason, we apply Locally Weighted Regression, which has been demonstrated to be an effective method to smooth trajectories in a reliable way (Toledo et al. 2007). Locally Weighted Regression corrects any data point by using information from its neighboring points also. The principle is based on Robust Fitting, i.e., a type of Least Square Error estimation where for each point, used to estimate a best-fitting curve, a weight is assigned. Given \( T \) points in the neighborhood of one vehicle position, the correct trajectory value can be determined by considering that this position is dependent on all other \( T \) recorded positions. In this way, we smooth position measurements and estimate smoother trajectories, speed and acceleration profiles, based on fitting a local curve at the points of interest. The methodology consists of two steps that are repeated for each point: (1) estimation of a smooth time-continuous trajectory function from the discrete position observations using weighted local regression, and (2) Estimation of instantaneous speeds and accelerations by differentiating the fitted trajectory function.

Let \( x(t), t=1,\ldots,T, \) denote the time series of measurements of the position of a given vehicle. At a point \( t_0 \), a local trajectory function is estimated using only observations in the neighborhood of \( t_0 \). The trajectory function in the neighborhood of \( t_0 \) is assumed to be a function of time:

\[
y(t) = f_{t_0}(t, \beta_{t_0}) + \epsilon_{t_0,t},
\]

where \( f_{t_0}(t, \beta_{t_0}) \) is the fitted position at time \( t \) estimated by the local regression function entered at time \( t_0 \). \( \beta_{t_0} \) is the vector of parameters of the fitted curve to be estimated, and \( \epsilon_{t_0,t} \) are normally distributed error terms. Local regression then uses weighted least squares estimation of the parameters of the local function \( f_{t_0}(t, \beta_{t_0}) \) with the \( T \) observations in the window around \( t_0 \). The observations weights are usually based on some measure of the time difference between the observation and \( t_0 \).
Hence, the problem of applying local regression to position data in order to develop a local trajectory function centered at \( t_0 \) is formulated as a minimization problem:

\[
\min_{\theta_0} \left[ X_{t_0} - f(t, \theta_0) \right]'
\]

where \( X_{t_0} \) is the column vector of \( T \) position observations used to estimate a trajectory function centered on \( t_0 \). \( f(t, \theta_0) \) is the corresponding vector of fitted values. \( W_{t_0} \) is a \([T \times T]\) diagonal matrix, with elements corresponding to the weights of the observations used for the local estimation.

We are concerned about the validity of this method for generating realistic speeds and accelerations. Unfortunately, we cannot use the collected dataset to clearly assess the correctness of the smoothed dataset, since we do not have any information on the real correctness of each point and therefore the actual vehicle speeds and accelerations. To operate a quantitative analysis of these errors in such type of problems, Punzo et al. (2009) provide a set of indicators to evaluate the correctness of the data based on e.g., percentage of unrealistic jerks (derivatives of the acceleration) and platoon consistency, i.e., percentage of unrealistic inter-vehicle distances.

Alternatively, to obtain an equally fair estimate of the smoothing quality, we randomly perturbed the trajectories of GPS data collected in another project (Hoogendoorn et al., 2006), for which speeds were also logged at a frequency of 0.1 s directly from the probe vehicles. Speeds and accelerations are therefore considered to be very reliable in this dataset. The differences between the real and estimated values of position, speed, and acceleration are estimated using the traditional root mean squared error (RMSE):

\[
\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}{n}}
\]

Since we do not know the real degree of reliability of the image processing dataset, we perturbed the GPS dataset using different resulting mean squared errors, in order to introduce errors in the acceleration at least in the same order of magnitude as in the camera data.

Figure 12.5 gives a visual impression of the error randomly generated in the floating car dataset (bottom), resulting in a RMSE of 0.2 m/0.1 s against the acceleration values computed with the image processing method. As one can see, the corrupted floating data shows much more illogical accelerations than in our dataset, i.e., a larger number of outliers is expected to be introduced. As a result, the smoothing algorithm seems to perform really well with the generated noise, and it recovers the original dataset very accurately.

In Toledo et al. (2007) a tri-cubic function is recommended as weight function, thus the information level of each point in the neighborhood is simply based on its (temporal) distance. By applying clustering technique to process the data, we have information also on the area of each cluster, and this estimated area is determinant.
of part of the error contained in the positions recorded at each time frame (i.e., the position of the centroid of these areas). We can therefore use this information as a measure of reliability of each point in the neighborhood. This measure is calculated with the following function:

\[ R_A = \frac{|A(t) - \bar{A}(t_0)|}{2\sigma_A} \]

where \( A(t) \) is the area of the point recorded at time \( t \) belonging to the neighborhood of \( t_0 \), \( \bar{A}(t_0) \) is the average value of the area, and \( \sigma_A \) is the standard deviation of the area in the neighborhood. The weight is thus computed by \( W_0 \times R_A \), where \( W_0 \) is the value calculated with simply the tri-cubic function as in Toledo et al. (2007). The value of \( R_A \) is set to zero if \( |A(t) - \bar{A}(t_0)| > 2\sigma_A \). Figure 12.6 gives an example of a computed weight function.

Since it is not possible to quantify the accuracy of the regression algorithm with our dataset, and since we have no information about the actual position of the vehicles, we have to rely on its ability to generate smooth and realistic speed and acceleration profiles. Table 12.1 gives an overview of the percentage of corrected data points, which still generate illogical speeds (i.e., negative or higher than 80 km/h, accounting that there is a speed limit of 50 km/h and a clearly visible speed camera), or accelerations. It should be pointed out that the large majority of illogical speeds are only slightly negative (in the order of -1 m/s) and occur while the vehicle is standing still. Therefore, they can be easily corrected.
Figure 12.7 shows how the Extended Locally Weighted Regression (ELWR) method proposed in this paper succeeds in smoothing the data and give more realistic speeds even for short time frequencies. In this example, we estimated the best-fitting cure keeping the data at 15 Hz frequency.

As already Toledo et al. (2007) stressed, this method is also robust with missing data. The minimum requirement is simply a sufficient number of data points recorded in the neighborhood of a missing point. Obviously, smaller the number of

<table>
<thead>
<tr>
<th>Table 12.1 Illogical speeds and accelerations after the smoothing process</th>
<th>Total: 21,498 points</th>
<th>Number</th>
<th>Ratio</th>
<th>Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Illogical speeds</td>
<td>717</td>
<td>3.34%</td>
<td>$v &gt; 80$ km/h or $v &lt; 0$</td>
<td></td>
</tr>
<tr>
<td>Illogical accelerations</td>
<td>5</td>
<td>0.0233%</td>
<td>$a &gt; 5$ m/s$^2$ or $a &lt; -5$ m/s$^2$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 12.7 shows how the Extended Locally Weighted Regression (ELWR) method proposed in this paper succeeds in smoothing the data and give more realistic speeds even for short time frequencies. In this example, we estimated the best-fitting cure keeping the data at 15 Hz frequency.

As already Toledo et al. (2007) stressed, this method is also robust with missing data. The minimum requirement is simply a sufficient number of data points recorded in the neighborhood of a missing point. Obviously, smaller the number of
available points, or larger the chosen neighborhood, the worse the correction capability expected. Therefore, these two parameters need to be carefully chosen and may be different from case to case. Figure 12.7 also shows a few points where the ELWR method does not improve the data, although this number is very small. Most of these errors disappear when the data is interpolated to larger time intervals. However, in other cases this might not result in good prediction. In these cases, one should use different intervals of neighboring points or a larger minimum number of points in the neighborhood. If this does not result in improving the prediction, then other smoothing methods should be tried, like traditional polynomial fitting.

### 12.4 Microscopic Analysis of Driving Behavior

The described data has enormous potentials for gaining insight into the process of traffic at signal controls. In this paper, we just list a number of possible directions of research, which will be undertaken in the future, and to give some examples of how empirical speed and acceleration distributions can be derived and analyzed, and we will draw some preliminary conclusion.

During the recorded time period, the signal was considerably busy, but hardly oversaturated, thus no exceptionally long queues were observed. After the data processing and cleaning, we could extract more than 1,000 vehicle trajectories with a sufficient accuracy. From each vehicle position, we could estimate a number of parameters such as instantaneous speed, acceleration, and distance from its preceding

![Smoothed speeds calculated after the correction of each point using extended locally weighted regression](image)
vehicle. These parameters describe measurable characteristics of each vehicle and its movement in the system, and can be used, e.g., to calibrate the main behavioral parameters in microscopic models, such as the desired speed and acceleration functions, or for calibrating and validating their car-following behavior in the presence of a signalized intersection system. We argued in fact that longitudinal driving behavior in these systems is significantly different from uninterrupted flow sections, as drivers can anticipate the behavior of their preceding vehicles and operate the stop-and-go operation in a different way than in freeway congestion. When observing a red signal, they may in fact start decelerating earlier, in order to operate a smoother decrease of speed and reduce the waiting time in idle.

Figure 12.8 shows an example of trajectories in a car-following situation that can be observed from the dataset (top) and the observed speeds (bottom). We can observe from this picture that the first vehicle approaching the signal starts decelerating very early and very smoothly, since it has detected that there is no right of way upstream. Moreover, the following driver starts decelerating after around 2 s since the leading driver decreased his speed, and maintains a large gap with the leader (around 25 m) for most of the deceleration phase. This behavior suggests that car-following behavior in this system might be significantly different during the green and the red phase.

Figure 12.9 shows an empirical distribution of the car-following behavior at this system. Drivers’ distance from a preceding vehicle is described in function of the speed difference in the same time instant. The bold continuous curve shows the process of closing in to the leading driver, and the oscillations caused by the imperfect distance-speed keeping of real drivers.

The continuous line describes also qualitatively the behavior of vehicles approaching the back of the queue. As one can observe the deceleration phase starts

![Figure 12.8](image-url)  
*Fig. 12.8* Example of car-following situation while approaching the signal in red
already at a considerable distance (around 70 m), and decelerations are normally below 3 m/s², thus decelerations are very smooth in this driving operation.

The difference in driving behavior between green and red phases is also confirmed when looking at speed and acceleration distributions. Figure 12.10 draws the distribution of speeds for different road section intervals, both upstream and passing the stop-line. As one can see, speed distributions are very different, especially right upstream the stop-sign. The effects of the signal and the queue formed by the signal are shown when approaching the stop-line, and the chance to observe vehicles at very low speeds increases. At 40 and 20 m from the stop-line there is a wider range of speeds observed. This is due to mainly the presence of a queue, and of vehicles decelerating during the red phase.

Figure 12.11 shows the distribution of vehicle accelerations. It shows that drivers hardly show accelerations or decelerations stronger than 1 m/s², thus they tend to drive smoothly through the signal control. However, while in red, drivers show a tendency of decelerating early (the larger skewness factor is observed around 40 m upstream of the stop-sign). On the other hand, as one can expect, strong accelerations are observed mainly near the stop-sign.

The speed-acceleration diagram gives a good indication of the dynamic and stochastic behavior observed in the traffic light system. Figure 12.12 displays the observed couples of speeds and accelerations. Again the size of the dots is proportional to the frequency of observations in the neighborhood of each dot, so that we can better visualize the dynamics of deceleration and acceleration phases. From this figure, we can make a few observations. First, that observed data with nearly zero speed variation are highly observed near zero speed and in between 20 and 50 km/h. These are respectively those vehicles waiting in queue and those passing the signal
at high speeds. In between 20 and 30 km/h there is a denser area in the deceleration phase. This can indicate that drivers approach the signal by moderating the speed. It is rather unlikely to observe low speed variations in between 5 and 20 km/h. This is easily explained as drivers in the analyzed system use these speeds only in transition between cruise driving and stopping in the queue. The largest speed variations are observed in the range of 15–20 km/h for both accelerations and decelerations.
In conclusion, by having a complete picture of the traffic process at signals in terms of individual vehicle trajectories and traffic light sequences, we can undertake many research questions, which will be presented in future papers, e.g.:

- **Driving behavior analysis:** By studying the behavior of individual vehicles at microscopic levels, we can acquire deeper knowledge on the way traffic is processed at interrupted flow sections such as signalized intersections. The data allows one to observe and quantify the various aspects that determine the longitudinal behavior of vehicles, i.e. their speed and speed variations approaching and near the signal, under different traffic conditions and signal phase sequences (e.g., behavior at start and during the green phase, discharge times, driving behavior during the amber phase, etc.);

- **Calibration of traffic flow models:** We can obtain from this dataset a complete picture of the individual driving behavior. Knowing the signal phase sequence and times, we can gain insight into this behavior in the two conditions, for instance, to check whether approaching vehicles decelerate earlier when red or a queue has been observed, or if they modify their speed near the intersection. Knowing the position of vehicles in the vicinity of each vehicle, we can analyze the behavior in different driving regimes, i.e. in free driving or car-following regimes. These are only examples of behavioral aspects that, for instance, can be used to calibrate the various parameters defined in microscopic software programs.

- **Traffic impact analysis:** The possibility to analyze vis-a-vis microscopic (instantaneous speeds, accelerations, car-following distances, and relative speeds), mesoscopic (number of vehicles in queue, delay, speed and acceleration distributions), and macroscopic parameters (mean flow, traffic composition) allows the analysis of a large number of traffic assessment studies. For instance, it can be used to estimate emission levels as already mentioned, or to derive and analyze safety measures like time-to-collision.
12.5 Conclusions and Future Research

This paper has described the collection and processing of microscopic data obtained using image processing and clustering techniques. Furthermore, we presented the first results of a study of driving behavior at signalized intersections using this dataset. Speed and acceleration distributions have been presented near and upstream the stop-sign, and in the dynamic process of stop and go at the signal. Analysis of these distributions has shown that speed variability increases considerably closer to the stop-sign when one observes these distributions. Moreover, indications on how car-following behavior is influenced by the presence of the traffic signal could be visualized.

The described data has enormous potentials for gaining insight into the traffic process at signal controls. This unique dataset can be used further in many ways, e.g., to study the variability of drivers’ behavior at these signals, to gain insight into factors that determine their actual trajectories and to benchmark and calibrate the car-following parameters in microscopic models. Other ways of exploiting this information will be, e.g., to use it as input for emission models to study the levels of concentrations and to compare different emission models. These applications will be done and presented in future works.

Acknowledgments The authors would like to thank Minwei Li, Janneke van Baalen, and Jie Li for the contributions made within this project. This project is funded by TNO, Delft University of Technology, and the TRAIL Research School (project name T3).

References


Index

A
 Acquisition, 5, 39, 53, 83, 177–179
 A-priori matrix, 5, 6
 Automatic, 27, 91, 92, 169, 177, 179, 180

B
 Behavior, 6, 23, 32, 41, 44, 49, 55, 73, 74, 83, 97, 101, 107, 115, 147, 155, 171–190
 Bottleneck, 8, 25, 27, 66

C
 Cleansing, 3, 55
 Collection, 1–44, 48–50, 55, 66, 93, 109, 130–133, 137, 146, 147, 175–185, 190
 Congestion, 3, 8, 21, 25, 27, 31, 39, 84, 111, 122, 155, 159, 163–165, 168, 180, 186
 Controller, 28, 33, 35, 36, 38, 161
 Cost-effective, 33, 37, 151–160

D
 Dataset, 11–22, 24, 93, 95, 97, 98, 172, 174–176, 178–183, 186, 189, 190
 DATEX2, 7, 49

Detectors, 2, 3, 5, 8, 26, 28, 29, 33, 34, 41–43, 52, 57–66, 70, 126, 132, 147, 151, 152, 155

Diagnostics, 26

E
 Electronic toll collection, 3, 66
 Emissions, 4, 31, 52, 93, 171, 172, 174, 175, 189, 190
 Estimation, 5, 6, 16–17, 21–23, 28, 29, 43, 75, 81, 84, 89–107, 109–127, 152, 154, 155, 158, 160, 181, 182
 Extended Floating Car Data, 151, 161–170

F
 Fusion, 132, 147, 152, 155, 160, 165, 167, 168

G
 Geographic Markup Language (GML), 7, 49
 GSM, 91

I
 Incident, 7, 8, 25, 27, 31, 34, 49, 50, 162–165, 168–170
 Induction loop, 2, 131
M
Macroscopic, 84, 129, 131, 136, 141, 172–174, 189
Measurement, 2–4, 8, 26, 30, 47, 49, 50, 52, 55, 83, 90–95, 98, 99, 102, 123, 126, 132, 133, 135–137, 159, 165, 176, 180, 181
Metadata, 7, 48–52, 55
Microscopic, 36, 39, 130, 131, 171–190
Mining, 55, 57–72, 130, 133–141, 146, 147
Modeling, 1, 6, 25, 73, 74, 83, 133, 136, 172–174
Monitoring, 6, 14, 23, 25, 26, 36, 91–92, 109, 118–121, 132, 141, 151, 155, 159, 160, 163, 168, 170

O
Online, 27, 29, 36, 49, 83
Operation, 1, 4, 5, 7, 9, 10, 21, 23, 25, 27, 30–32, 35–40, 43, 44, 48, 53, 66, 68, 70, 73, 90, 91, 162, 164, 166, 171, 172, 176, 186, 187
Optimization, 1, 39, 151
Origin-Destination (OD) matrix, 5, 6, 75–77, 89–91, 95, 97, 102–104, 106
Outliers, 82, 92, 135, 136, 182

P
Performance, 2, 4, 8, 25–44, 49, 53, 104, 144, 172, 174
Placement, 28
Platform, 7, 20, 21, 48, 50, 169
Privacy, 3, 4, 8–9, 11, 15, 23
Probe vehicle, 2, 5, 6, 8, 29, 33, 89–107, 109–127, 130, 132, 151, 182

Q
Quality management, 48–50, 55

R
Real-time, 26, 32, 35, 40, 89, 92, 129, 147, 160, 164, 166

S
Search, 49–52
Sensors, 2, 26, 36, 39, 90, 131, 151, 163–165, 169
Simulation, 1, 3, 4, 25, 35–36, 39, 44, 47, 50, 55, 104–105, 111, 122–123, 144–146, 159, 171–175
Smoothing, 133, 135, 175, 179–185
Software, 7, 17, 20, 21, 33, 36, 38, 52, 53, 63, 64, 135, 172, 173, 176, 177, 189
Storage, 4–9, 26, 47, 49, 50, 52–53, 55, 166, 167
Surveillance, 26, 27, 131
Survey, 3, 5, 12–14, 17, 18, 22, 23, 74–77, 82, 83, 85, 89, 95, 101, 102, 104, 106

T
Telematics, 129–133, 146
Tracking, 9, 41, 42, 178–180
Traffic Control Centre (TCC), 161–163
Trajectories, 41, 42, 127, 132, 171–177, 179–182, 185, 186, 189, 190
Transport, 1–5, 7–44, 47, 73, 83, 84, 86, 91, 92, 129, 131, 144, 161, 164, 169

U
Utilization, 27, 66, 69, 137, 163

V
Variable Message Signs (VMS), 7, 162
Video, 2, 8, 33, 41, 42, 49, 52, 57, 119, 127, 131, 151, 155, 158, 159, 174, 180
Visualization, 5, 57–72, 137, 142

W
Weather, 3, 151, 160, 163, 167, 168