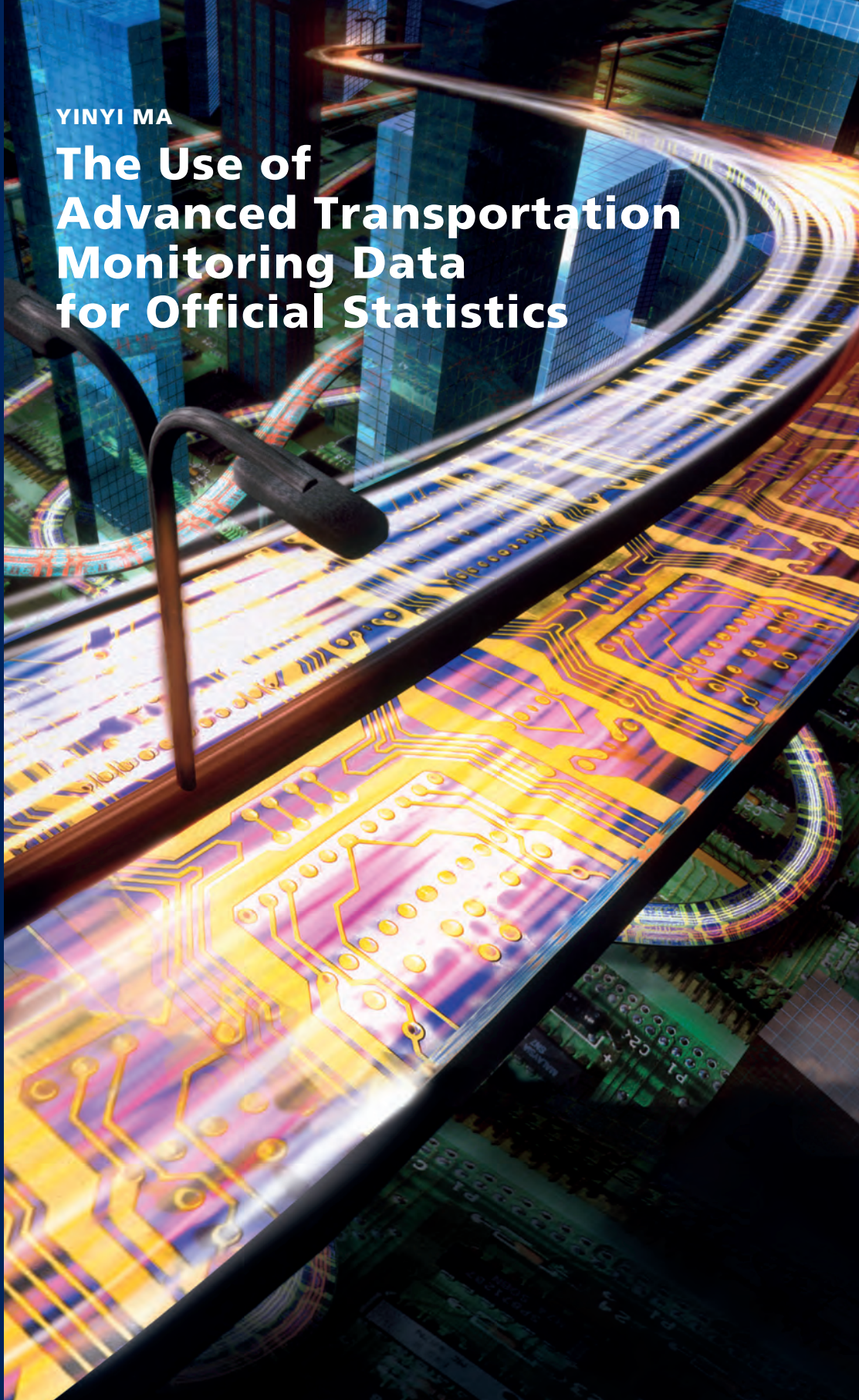
YINYI MA

The Use of Advanced Transportation Monitoring Data for Official Statistics



The Use of Advanced Transportation Monitoring Data for Official Statistics

The Use of Advanced Transportation Monitoring Data for Official Statistics

Het gebruik van elektronisch verzamelde transportgegevens voor officiele statistieken

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. H.A.P. Pols

and in accordance with the decision of the Doctorate Board

The public defense shall be held on

Friday 3 June 2016 at 9:30 hrs

by

Yinyi Ma
born in Nanjing, China.

Erasmus University Rotterdam

Doctoral Committee

Promotor:	Prof.dr. L.G. Kroon
Other members:	Dr. R. Kuik
	Prof.dr. H. Mahmassani
	Prof.dr. H.J. van Zuylen
Copromotor:	Dr. J. van Dalen

Erasmus Research Institute of Management - ERIM
The joint research institute of the Rotterdam School of Management (RSM)
and the Erasmus School of Economics (ESE) at the Erasmus University Rotterdam
Internet: http://www.erim.eur.nl

ERIM Electronic Series Portal: http://repub.eur.nl/pub

ERIM PhD Series in Research in Management, 391
Reference number ERIM: EPS-2016-391-LIS
ISBN 978-90-5892-449-0

Design: B&T Ontwerp en advies www.b-en-t.nl
This publication (cover and interior) is printed by haveka.nl on recycled paper, Revive®.
The ink used is produced from renewable resources and alcohol free fountain solution.
Certifications for the paper and the printing production process: Recycle, EU Flower, FSC, ISO14001.
More info: http://www.haveka.nl/greening

To My Parents

Acknowledgments

My research was fully funded by Statistics Netherlands. It is grateful to have their support for these years, allowing me to visit international conferences and to visit Northwestern University, USA, as a scholar.

I would like to express my deepest gratitude to my promotor, Prof.dr. Leo Kroon, for his excellent guidance, patience, encouragement and support.

I would like to thank my co-promotor, Dr. Jan van Dalen, for his time, efforts and invaluable advice on my dissertation. In the past seven years, I have learned so much from him, and for this, I will always be grateful.

I also would like to thank the other committee member in Statistics Netherlands, Dr. Chris de Blois. My dissertation benefited a lot from his comments about methodology and practice. Besides, he helped me to set up many interviews in Statistics Netherlands, connecting me to the people from Statistics Offices in other countries.

I would like to express my gratefulness to my co-author, Dr. Roelof Kuik, who gave many valuable comments to our papers and my dissertation.

Thanks to my advisor in Delft University of Technology, Prof.dr. Henk van Zuylen, for his consistent encouragement and suggestions within nine years.

Also thanks to Prof. Hani Mahmassani at Northwestern University, who provided a great opportunity connecting me to his research group during my four-month stay in Evanston as a visiting scholar.

My special thanks go to the late Prof.dr. Jo van Nunen, who provided me this great opportunity to explore my research ability.

I am thankful to many friends who have made Erasmus University, Statistics Netherlands, Delft University and Northwestern University fun places to work.

Finally, my parents deserve my deepest gratitude for their encouragement, support and love.

Yinyi Ma

Rotterdam, 2016

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Official Statistics

Traffic and transportation statistics, as produced and published by Statistics Netherlands, aim to describe relevant properties of real-world traffic states and transportation processes. Statistics Netherlands, like other statistics-producing institutions worldwide, collects traffic and transportation data from various sources, e.g., Rijkswaterstaat, Water, Traffic and Environment (WVL), Customs, and private companies, using a variety of formats such as surveys and secondary databases. After collection, Statistics Netherlands processes the data and publishes statistics on an annual, quarterly or monthly basis. The target audience includes various stakeholders, such as governments, employer and employee organizations, industry associations, interest groups, and science and education organizations.

The data collection process is widely perceived as an administrative burden by data providers, notably firms such as carriers and forwarders. User requirements with respect to the detail and frequency of publication further strain the data processing and statistical reporting. The traditional, questionnaire-based approach is rather time-consuming to collect and process, and involves publishing periods of at least three months up to and including one year. The use of secondary data sources partly copes with these challenges, but has the downside of making the data collection and processing dependent on the operations and objectives of data suppliers. If such data suppliers for some reason implement changes in their activities or abandon data collection, then the production of official statistics could be seriously hampered. For example, the formation of the European Union in 1993 led to the abolition of trade declaration forms, and consequently to the canceling of an important source of information about international trade.

Moreover, the reporting of official statistics is largely on a sector-by-sector basis. This format is strongly rooted in the overarching framework of the national accounts, but has the disadvantage that related economic activities are often independently observed. In particular, the interrelated activities by various supply chain partners are not fully observed, which may leave economically relevant information unused and ignores a potential source of observational efficiencies. Finally, current traffic and transportation statistics are of a rather static kind, embodying highly aggregated information published at large time intervals. This feature is in contrast to the highly dynamic nature of traffic and transportation processes, which may demand flexible means of observation and frequent rates of publication.

1.2 Various Types of Transportation Data

In parallel to the survey-based data collection, recent years have seen the massive generation of dynamic electronic traffic and transport information. This is a result of the rapid diffusion of devices, such as loop detectors, Weigh-in-Motion (WiM), Automated Number Plate Recognition (ANPR) cameras, Bluetooth, Global System Mobile Communication (GSM), and Global

Positioning Systems (GPS). These various data capturing technologies differ with respect to the quality of the collected data and usage cost. Some, like Bluetooth and GPS, allow collection of information about travel times while others do not. Some can be flexibly used at any location, while others, like WiM-systems, are bound to a particular location. These data are grouped into three categories based on the characteristics of the different data sources: point data such as loop detectors and WiM; point-to-point data or path data such as cameras, Bluetooth and GSM; and route data such as GPS. Table 1.1 gives an overview of the various data capturing technologies and their characteristics. These advanced monitoring systems provide Statistics Netherlands with opportunities for collecting more detailed and more dynamic data.

Table 1.1: Summary of Traffic Data

	Point Data		**Point to Point Data / Path Data**			**Route Data**
	Loop Detector	WiM	Camera	Bluetooth	GSM	GPS
Counts	yes	yes	yes	yes	yes	yes
Travel Time	no	no	yes	yes	yes	yes
OD matrix	no	no	yes	yes	yes	yes
Vehicle Type	partially	yes	yes	no	no	no
Location	highway	highway	highway	everywhere	everywhere	arbitrary point
Accuracy	1m	1m	2m	50m	100m	15m
Authorities	NDW	ILT	VID/ ARS	Rotterdam Port/ VID	Telecom Companies	TomTom Garmin

1.2.1 Loop Detectors

Loop detectors are the most common technology used to collect traffic data. Loop detectors are installed mainly in the pavement of the highway. In the Netherlands, a loop detector is installed every 500 meters on a highway. Loop detectors are rare on urban or provincial roads, making the traffic situation there quite hard to observe. Loop detectors offer information about traffic flows, time mean speed (the average speed of a traffic stream passing a marked point along a roadway measured over a specific period of time) and vehicle length. Based on the vehicle length from loop detectors, one can distinguish three to five vehicle types, such as trucks and trailers. But vans and cars are counted together, since the lengths of vans and cars are roughly the same, which complicates the separation of vans from normal traffic. Loop detectors cannot observe the origin-destination (OD) information directly. Their data are used to estimate OD matrices. But the associated under-specification problem (Van Zuylen and Willumsen, 1980) leads to an infinite number of feasible solutions, which renders its exclusive usage for OD matrix estimation unsatisfactory. In the Netherlands, NDW (national data warehouse) collects and stores the loop detector data.

1.2.2 Weigh-in-Motion (WiM)

Weigh-in-Motion devices capture and record the axle weights and gross vehicle weights of vehicles. Once a vehicle passes a WiM device, information such as speed, time, axle weight and axle distance is measured. But the WiM device does not store data for vehicles with a length of less than 7 meters. Based on this information, trucks are distinguished from passenger cars and the types of trucks can be estimated. The data from a WiM device does not reveal OD information. The data accuracy of WiM systems is within the range of one meter, which is relatively high, compared to the other data capturing systems. Usually, the WiM device is combined with a camera, which is called a WiM system in the Netherlands. The WiM system captures the pictures of all trucks, including the number plates. In the Netherlands, ILT (Inspectie Leefomgeving en Transport) collects the data of the WiM system. Only a few Dutch WiM systems have been installed. Eighteen WiM systems have been installed on nine Dutch highways (one for each direction).

1.2.3 Automated Number Plate Recognition (ANPR)

Automated number plate recognition (ANPR) data, commonly referred to as camera data, can identify individual vehicles based on their number plates. Combining the recordings of different cameras allows one to determine the routes of vehicles through the network, depending, of course, on the coverage of the cameras. Video camera data can capture detailed information of vehicles passing in a certain lane. Vehicle travel times can be estimated if two or more cameras have been installed along a route. Obviously, ANPR offers quite detailed information including traffic counts, OD information if sufficient cameras have been installed, vehicle types, and so on. In the Netherlands, VID (the traffic information department), ARS T&TT (Traffic & Transport Technology) and HIG Traffic Systems collect these data.

1.2.4 Bluetooth

Bluetooth scanners are becoming popular as a means of collecting traffic information since they are widely used on board, and capturing the signals is relatively inexpensive. The flexibility of Bluetooth scanners allows them to be installed throughout the road network. They capture the Bluetooth signals of any Bluetooth devices on board of vehicles, once carrier vehicles are within a range of approximately 50 meters. This spatial accuracy of Bluetooth scanners is lower than that of loop detectors, Weigh-in-Motion, cameras and GPS, as illustrated in Table 1.1. This relatively low accuracy may result into data outliers. In order to remove the outliers, statistical detection methods are used, such as the moving standard deviation algorithm (Quayle and Koonce, 2010), the smoothed histogram algorithm (Haghani et al., 2010), and the box plot filtering method (Schneider et al., 2009). Quayle et al. (2010) show the usefulness of Bluetooth scanners for travel time measurements for a 2.5-mile corridor in Portland, USA. Bluetooth

information can be used to estimate an OD matrix. But it is difficult to estimate the truck flow, because not all vehicles have Bluetooth devices installed, and those that have them cannot be identified by type of vehicle. In the Netherlands, VID and the Port of Rotterdam collect and use Bluetooth data.

1.2.5 Global System Mobile Communication (GSM)

GSM data is provided by mobile phones within a certain radius using an antenna tower as the center-point. The data includes time, antenna tower, and phone identification. Normally GSM data is obtained only if phones are in use. At that moment, the signal is strong enough to be received. However the phone signal may also be received even if the phone is not calling, though it may be too weak to be accurate. Travel time and route information can be assessed from GSM data while OD data is partially observable. Vehicle types are hard to separate from GSM data and data accuracy is not that high, only about 100 meters. In the Netherlands, TomTom collects GSM data from Vodafone and stores the data in its servers. Mezuro uses the Vodafone network data for statistical research on mobility.

1.2.6 Global Positioning System (GPS)

The Global Positioning System is a space-based satellite navigation system that provides spatial data of single vehicles, such as location (latitude, longitude), speed, time and direction. Such detailed individual real-time data can describe traffic trajectories along temporal-spatial dimensions, with a high observation frequency of 1575MHz (Raju, 2003). Through this, travel time, origin-destination, and route information of each observed vehicle are available. However, at present the penetration of GPS installed vehicles is quite low. Therefore the subset of vehicles with GPS installed may not provide a representative sample of the actual traffic states on the roads. In addition, the accuracy of GPS data is within 15 meters. Many factors affect the accuracy: surrounding conditions, number of satellites in view, distance from reference receivers, and so on. Nevertheless, OD information derived from GPS data is much more accurate than that from other data sources. Although registered GPS data can identify vehicles, it cannot recognize the type of vehicle unless the registration of the data is linked to a vehicle specification. Thus, without this specification information, separating trucks from passenger cars based on registered GPS data is difficult.

There are two main reasons which restrict the availability of GPS data: commercial value and privacy. Commercial value implies a cost to users. In the Netherlands, TomTom and Garmin collect GPS data, including cars and trucks, and stores the data in their servers. This information from vehicles with GPS is not offered for free. In other countries including China, Germany and Austria, vehicles with GPS are a common method for collecting information about traffic conditions, mainly from taxi companies. Their GPS data is not freely accessible. Additionally,

privacy issues may arise either through customer confidentiality agreements or legal restrictions. In most of the countries, companies that collect GPS data as part of the navigation services they provide, are not willing to share this information.

1.3 Representation of Multiple Data Sources

The availability of multiple data sources in the road network attracts the attention of many traffic and transportation companies. For instance, TomTom wants to use this rich data to generate accurate traffic information, and to direct travelers to drive on roads with the lowest travel time. Google buys various traffic data from different companies and maps the real-time traffic information on Google maps, informing travelers about the current traffic situation in the road network.

Besides road traffic information, the Port of Rotterdam is interested in the number of vehicles traveling from origins to destinations to gain insight into the freight truck demand (NMmagazine, 2016). For that, they installed multiple data sensors in the areas of the A15 motorway which are important to them, to capture flow data and to further estimate the traffic demand.

The A15 motorway in the Netherlands connects the so-called Maasvlakte - the most Western port area of Rotterdam on the sea shore - with the German border. The A15 is heavily used by trucks to transport cargo from the port region to destinations elsewhere in the Netherlands, abroad and vice versa. The A15 motorway provides us with a concrete example of how different data capturing technologies are used in combination with the aim of gaining insight into transport behavior in the road network. The loop detector and WiM data have been obtained from the Dutch Ministry of Infrastructure and the Environment. The Bluetooth data have been collected during a pilot project called Roportis, which was initiated by the Port of Rotterdam.

Figure 1.1 is a map of a part of the A15 area, showing where different data capturing technologies are located: loop detectors, WiM systems and Bluetooth scanners. Camera locations have not yet been included in 2012. Two WiM systems are placed on the secondary roads near the A15. Bluetooth scanners are located near the motorway and parallel roads. There are 24 scanners in total. The locations of the Bluetooth scanners have been selected by the Port of Rotterdam as the main points for freight transport in the port area. With the patterns plotted from this available data, we can get a valuable insight into the travel behavior of vehicles.

1.3.1 Information from Weigh-in-Motion and Loop Detectors

Due to location restrictions, the Weigh-in-Motion systems in the Dutch highway network mainly provide transport cargo weight data for each passed vehicles. The system cannot present network information, although cameras are involved in the WiM systems. In contrast, loop detectors cannot identify trucks, but they do give a general impression of truck flows and traffic flows. The flows from loop detectors are representative for the network. Assuming the truck types in

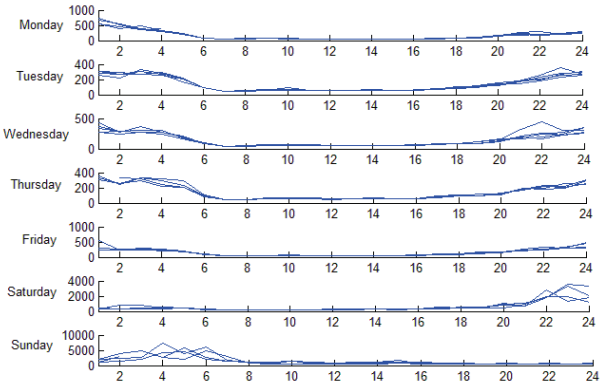
Figure 1.1: The locations of Bluetooth devices, WiM systems and loop detectors around the A15 motorway. (Source: Reportis.com)

the network are similar to the truck types in WiM locations, we can have the average cargo weight per truck per hour. By way of illustration, we present graphs of the average cargo weight per truck and the occupancy rates of trucks, by hours of the day and days of the week in Figure 1.2(a). The results are obtained from WiM data collected in March 2010. For weekdays, we find from Figure 1.2(a) that the ratio of weight and truck flow is much higher during the period from 20:00 to 05:00 am the next day than for the rest of the time. The patterns are a consequence of the fact that during daytime, a range of commercial vehicles, including vans, light trucks and lorries, travel on the roads, while during the evening hours and night time, there are mainly heavy lorries with full loads, such as international transport from and to the Port of Rotterdam. Additionally, Figure 1.2(b), plotted based on the WiM data in March and April, 2009, shows that the occupancy rate of trucks on the A15 highway from the port area to the hinterland is high during the early times of the day (00:00-07:00) and then gradually lowers as the network presence of passenger cars increases.

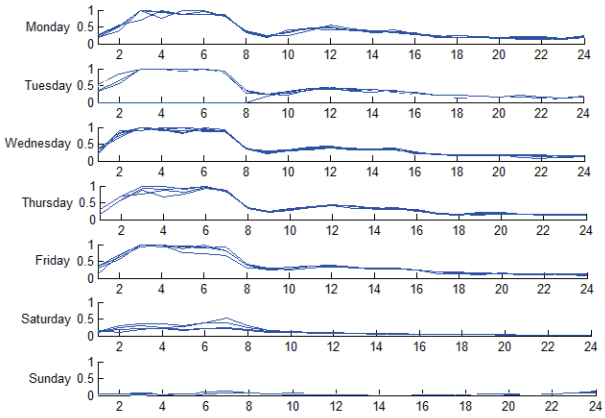
1.3.2 Truck Trajectory Analysis based on Bluetooth Data

Bluetooth data may be used to improve the accuracy of the OD estimation of freight trucks, since they identify individual vehicles and provide information about vehicle routes. Still, it may be challenging to extract truck routes from these data. We illustrate this with the Bluetooth recorded trajectories of two vehicles in Figure 1.3. Both charts have the scanner ID on the vertical axis and the recording time on the horizontal axis.

Vehicle 1 is first detected in the early morning of July 5^{th} 2011 by scanner 5 on a secondary road. It is observed to travel via the A15 (scanner 10) to another secondary road. After arriving at a transit location, it waits for about two hours, possibly for loading and unloading, and then travels back to the location nearby scanner 9 for another task. At about 12:50, the vehicle



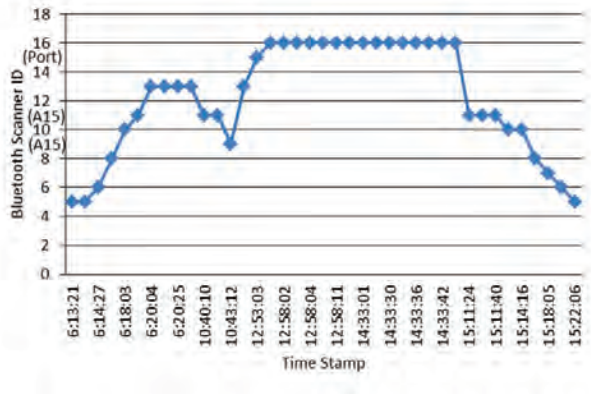
(a) Average total weight per truck



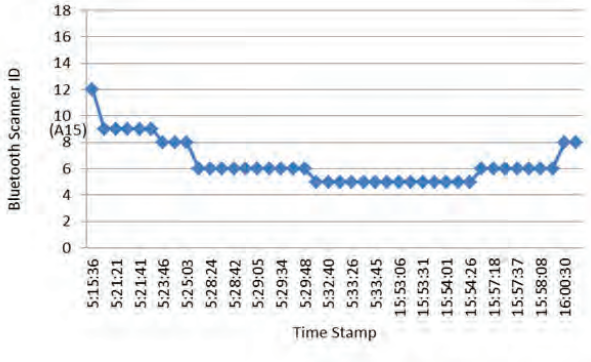
(b) Ratio between truck flow and traffic flow

Figure 1.2: Relation among weight, truck flow and traffic flow per hour.

drives to the Port of Rotterdam (scanner 16) and in the afternoon returns via the A15 to the departure location. Based on the journey and the locations visited, one may infer that the vehicle actually is a truck. In contrast, the trajectory of vehicle 2 is relatively simple. In the early morning it drives in the direction opposite to the port area via a road parallel to the A15 and in the afternoon, around 16:00, it leaves the A15 area. In this case, there is no way to infer anything about the type of vehicle. However, the Bluetooth data is sufficiently precise to



(a) Trajectory of Vehicle 1



(b) Trajectory of Vehicle 2

Figure 1.3: Two examples of truck trajectories from Bluetooth data on July 5th, 2011

determine individual routes and to distinguish between the usage of motorways or secondary roads.

1.4 Freight Demand Management

All these data can support decision making, taking the research of freight truck demand management as an example. The interpretation of freight truck demand in this dissertation mainly

refers to the number of trucks travelling from an origin to a destination. The concept of an origin and destination matrix (OD) is introduced in transportation planning to represent the trip production and attraction from one zone to another during a certain time interval. Usually, an OD matrix is taken as an input to the traffic assignment to obtain flow observations in the road network.

1.4.1 Origin Destination Matrix

The trip production and attraction of freight vehicles, as part of the general vehicles, can be represented in a freight truck origin destination matrix as illustrated in Table 1.2. The origins are denoted as O; the destinations are denoted as D. The T^{ij} is the number of vehicles from origin i to destination j in a certain time period.

Table 1.2: A Freight Truck Origin Destination Matrix

Zones	1	2	$\cdots$	j	$\cdots$	β	Production
1	T^{11}	T^{12}	$\ldots$	T^{1j}	$\ldots$	$T^{1\beta}$	O^1
2	T^{21}	T^{22}	$\ldots$	T^{2j}	$\ldots$	$T^{2\beta}$	O^2
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
i	T^{i1}	T^{i2}	$\ldots$	T^{ij}	$\ldots$	$T^{i\beta}$	O^i
$\vdots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\vdots$
α	$T^{\alpha 1}$	$T^{\alpha 2}$	$\ldots$	$T^{\alpha j}$	$\ldots$	$T^{\alpha\beta}$	O^α
Attraction	D^1	D^2	$\cdots$	D^j	$\cdots$	D^β	T

note: $D^j = \sum_i T^{ij}$, $O^i = \sum_j T^{ij}$, and $T = \sum_{ij} T^{ij}$

The estimation and forecasting of freight truck OD matrices has become increasingly important over time (Afandizaden and Yadi, 2006; Li, 2009; Shan and Li, 2008). Freight vehicles contribute significantly to network congestion and air pollution (Stanley et al., 2009), and their activities adversely affect the pavement substantially more than other road users. Insight into their journeys greatly enhances our understanding of which links in the road network, both primary and secondary, are influenced by developments at origins and destinations and to what extent. Traffic operators can use freight truck OD information to guide freight vehicles arriving at easily congested destinations, such as port areas and other logistic hubs, or to alleviate traffic congestion in the network through demand management. For instance, they may provide logistics companies with information about optimal time slots for loading and unloading trucks, thus enabling them to align their departure times and routing decisions (Ma et al., 2010). Furthermore, freight truck OD information can be useful for future infrastructure design, such as decisions whether and where to build extra terminals, or to evaluate the logistic consequences of the economic development of certain regions.

1.4.2 Origin Destination Tuples

The vehicles traveling along an OD pair to fulfill demand are usually assumed to be homogeneous and the trips of vehicles are asumed to be independent. However, in reality vehicle trips are inter-related. One vehicle may contribute to time dependent OD matrices several times a day, according to its schedule or travel plans. Commuters travel from home to work in the morning and back home in the afternoon. Trucks with multiple tasks drive from a distribution center to a store and later to a port area, for instance. The drivers have to find a rest area after driving for two hours.

In practice, the traditional definition of OD matrix lacks a behavioral basis and trip-based model structure (Kitamura, 1996). This traditional setup ignores the fact that people plan ahead and choose attributes of each trip (including mode, destination, and departure time) while considering the entire trip chain, not each individual trip separately (Kitamura, 1996). To estimate OD matrices in the field of traffic engineering, link flows observed from loop detectors are taken as a main data source. The loop detector on each link simply counts the number of vehicles passing by. Thus, one reason for the ignorance of the trip chain in the OD matrices could be the anonymous loop detector data, which cannot identify vehicles.

Meanwhile, another research stream of travel activity-based research digs into individual travel behavior, such as activity schedules and travel choice. Jones et al. (1990) provide a comprehensive definition of activity analysis: it is a framework in which traveling is analyzed as daily or as multi-day patterns of behavior, related to and derived from differences in life styles and activity participation among the population. They take into account the fact that travelers have travel plans as a trip chain, such as HWH (from home to work and back) and HWH+(work tour with at least one additional stop for another activity) (Bowman and Ben-Akiva, 2001). Survey data (Stavins, 1999; Kroes and Sheldon, 1988) is the main information source supporting this research. Although surveys may demonstrate some trip chains of travelers, the sparseness of survey data is an issue restricting the presentation of travel behavior. It is consequently hard to estimate OD matrices from patterns of behavior and to use the survey data as input for dynamic traffic assignment. Hence, the scope of the patterns of behavioral research is normally limited to the demand side, ignoring the road network.

Due to conceptual differences and data capturing limitations, there is no link between OD matrices in transportation planning and trip chains in behavioral activity-based research, although many similarities and potential benefits have been shown. We introduce the concept of *Origin Destination Tuple* (ODT) to represent traffic demand. A tuple as used in set theory is a sequence of elements. An Origin Destination Tuple is a sequence of OD pairs within a certain time period, which represents a trip chain in the road network. The traditional OD pair is obviously the simplest case of an ODT.

An ODT is an ordered set of OD pairs: a number of vehicles with the same entry and exit points on the road network. Clustering travelers with the same travel pattern from a geographic

point of view during a certain time period actually takes some individual travel behavior into account. ODT brings the travel demand from the aggregated level to the disaggregated behavior level. It does not only address the issue of the anonymous vehicles from an origin to a destination, but also focuses specifically on the trip chains of vehicles with the same travel pattern.

For example, in the network of part of the A15 motorway in the Netherlands, illustrated in Figure 1.4, the traditional OD data for a whole day could be 6000 vehicles for in3 → out4 and 5000 for in6 → out7. But in practice there are 500 vehicles among these demand data traveling with the trips of first in3 → out4 and then in6 → out7 as an ODT. Consequently, the demand data should be with three OD tuples instead of two OD pairs: 5500 vehicles for in3 → out4, 4500 for in1 → out4, and 500 for in3 → out4 ∼ in6 → out7.

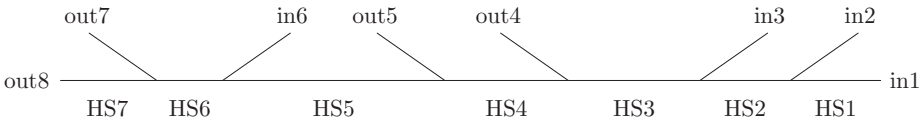


Figure 1.4: Part of A15 motorway from Hoogvliet to Havens

In the short term, forecasting ODTs can help to better understand the interaction between discrete trips and real travel behavior in the network. For the long term, transport policies such as road pricing or tolling systems to improve the travel situations may also benefit from the use of ODTs.

1.4.3 Research Questions

The two main data sources for statistics offices to generate annual official statistics are surveys and secondary databases. Also for freight statistics, they send out questionnaires and analyze the sample data. In the field of transportation management, there are several types of monitoring data available. Thus, the main general research question in this thesis is: how to use these captured monitoring data of freight transport to generate official statistics?

Narrowing down to freight truck demand management, research sub-questions arise from two aspects. One is related to the freight truck OD matrix. What are the proper methods, such as information methods and Bayesian inference, to estimate the freight truck OD matrix based on available information? If we want to forecast the freight truck OD matrix, which method can be aligned with the estimation methods? How to evaluate the estimated and forecasted demand?

The other aspect is about the multiple data sources. How to integrate multiple data sources either from surveys or from monitoring systems? Do multiple data sources help to improve the estimation accuracy of the demand?

1.5 Contributions of this Dissertation

This dissertation has theoretical, methodological and practical contributions to the field of freight transportation statistics.

1.5.1 Theoretical Contributions

The main issue this dissertation addresses is the connection between official statistics and transportation management. Official statistics are mainly based on bottom-up methods. Organizations which produce official statistics usually obtain the data, plot the data to understand the patterns, and publish the patterns. Transportation management is usually carried out through a top-down approach. Researchers observe phenomena, model the phenomena, and use data to validate the proposed model. This dissertation combines these two approaches and takes advantage of the observations for transportation management to update traditional statistics.

In addition, the concept of Origin Destination Tuple as a sequential dependence of the OD matrices is introduced taking advantage of path flows. This concept bridges transportation modeling which considers only OD pairs, and activity-based model research (Cascetta, 1984; Bell, 1991; Bierlaire and Toint, 1995) that focuses on travel behavior with the trip chain concept. ODTs actually bring extra uncertainty to the under-specification problem for the estimation and forecasting. Taking advantage of monitoring systems that are able to identify trip chains of vehicles, the path flows from identification devices such as cameras, significantly decrease the uncertainty due to the OD tuples. To our knowledge, this concept is introduced and tested in the freight transportation field for the first time.

1.5.2 Methodological Contributions

This dissertation has several methodological contributions. Firstly, it shows how data fusion brings benefits to freight transport, particularly when different data sets are involved, such as survey data and monitoring data. Traditional surveys consist of a rich content. Taking surveys for logistics companies for instance, there is information on departure time, arrival time, location, travel mode, cargo weight and so on. But it requires manpower and time to fill in data. In 2010 surveys are sent out twice per year by Statistics Netherlands. The response rate is around 60% in total. In 2015, the response rate has increased to 80% due to on-line surveys. Comparing with surveys, advanced monitoring systems can capture data automatically, thereby providing data in seconds for the whole day. With a certain capture error, the road network is full of flow information. However, the variety of advanced monitoring systems is limited. Loop detectors can provide link flow data during a certain time, mean flow data and length of vehicles. Cameras with identification function can capture more information, but limited to number plates and path flow data. Combining survey data and monitoring data extends the

data collection methods. Surveys collect data with various variables and advanced monitoring systems collect data continuously.

Secondly, the method of Kullback-Leibler divergence proposed in Chapter 3 generalizes the information minimization method to estimate OD demand. Comparing with the information minimization approach (Van Zuylen and Willumsen, 1980), the Kullback-Leibler divergence method better represents the underlying concept of information minimization for estimating the OD matrix and relaxes the assumption implied by Stirling's approximation. The Stirling's approximation is feasible when flow data is large. For the situation where the flow data is small, the Kullback-Leibler divergence method has lower average deviations between the estimated demand and the ground truth demand than the information minimization method. Thus, the Kullback-Leibler divergence model is regarded as a generalized approach of the information minimization method.

Thirdly, the dissertation demonstrates the feasibility of applying hierarchical Bayesian networks to estimate and forecast freight truck demand. In Chapter 4, the normal distribution and the log-normal distribution are applied in the framework of hierarchical Bayesian networks. An analytical approach is applied in the case of normal distributions, which leads to very fast computations. But the symmetric shape of the normal distribution may under-present the probability of a large number of trips. Log-normal distributions avoid these disadvantages. But a simulation approach has to be applied to the log-normal model, such as the Gibbs sampling nested by the Metropolis-Hastings sampling. This requires extensive computing resources. In addition, Chapter 5 contributes the approach to forecast the next-day demand. The multi-process model associated with auto-regressive dynamics is first applied to forecast transportation demand, to our knowledge. This model has the advantage of taking the demand in several previous days into account, giving each pre-defined scenario of combination of the previous days' demand a prior probability, and coming up with posterior probabilities for each scenario. Usually, one of the scenarios gains a unit probability, and the rest has zero probability. In the case of same scenarios, they will share the unit probability with the same proportions of the prior probabilities. With this multi-process model, people's experience is taken into account in the pre-defined scenario. The observed data together with errors update the prior probability and end up to the most likely scenario.

Lastly, this dissertation explicitly takes measurement errors into account when conducting the modeling. The observation error in most relevant research (Maher, 1983; Hazelton, 2000; Castillo et al., 2008a; Zhou and Mahmassani, 2006) is additive. The underlying assumption of additive errors is that whatever the problems are in the detectors, the errors stay the same. In our model of log-normal distributions, multiplicative errors are proposed, which actually relaxes the underlying assumption. These multiplicative errors represent the percentage failure of each device.

1.5.3 Practical Contributions

The results of this dissertation can be used by statistics offices to update transportation databases with more accurate and recent data. The OD demand of on-ramps and off-ramps can be aggregated to a certain area with the same postal code or even a city. The aggregation could be along space and time, such as monthly demand or yearly demand. The relationship between official statistics and transportation planning is illustrated in Figure 1.5. The prior freight demand information from the statistics offices, associated with observations in the road network, is taken as an input to the OD estimation approach. The output of the estimated OD matrix is aggregated to get freight demand statistics. The use of the observation data could reduce the administrative burden by statistics offices and produce accurate information.

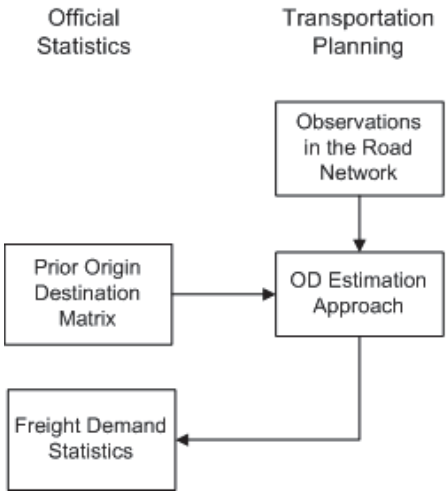


Figure 1.5: Relation between Official Statistics and Transportation Planning

From the product point of view, statistics offices process the historical data to get the useful information published, so they may not be interested in forecasting. But from the methodology point of view, forecasting methods can be applied to deal with missing data and to smoothen historical data, which is aligned with the goal of statistics offices. In addition, statistics offices nowadays would like to have information reported or published in a more efficient way. That is why they are interested to make a step forward to more detailed real time information. In the near future, if statistics offices want to publish real time data, they must do the short term forecasting first. With the forecasting model and the coming data, statistics offices are able to have an efficient data reporting system. With the real time information, statistics offices can do more than right now.

1.6 Outline

The outline of this dissertation is presented schematically in Figure 1.6 and discussed in more detail in this section. This dissertation consists of six chapters.

Chapter 2 gives an overview of the methods used to estimate freight truck demand. There are two main approaches: one is point estimation, such as maximum likelihood, least squares and information methods; and the other is distribution estimation, such as the Bayesian inference methods. This chapter compares these two approaches. The distinction between the anonymous link flows observed from loop detectors and the path flows from cameras or Bluetooth scanners which can identify vehicles are described.

Chapters 3 and 4 address the issue of freight truck demand estimation. Chapter 3 extends the work of Van Zuylen and Willumsen (1980), in which an information minimization method was utilized. In order to get insight into their research and relax their assumptions, we introduce the Kullback-Leibler divergence approach and show that these two methods are strongly connected from a mathematical point of view. The Kullback-Leibler divergence approach generalizes the information minimization method with fewer assumptions.

In order to relax the assumptions in Chapter 3 even further, such as independence of the link flows and no measurement errors, a stochastic approach is taken. Chapter 4 applies hierarchical Bayesian networks to estimate the stochastic freight truck demand, taking historical data as prior information and multiple sensor data as evidence. To model the freight truck demand, two situations are discussed. One is that all quantities follow normal distributions, where an analytical approach is sufficient to obtain the posterior demand. The symmetric shape of the normal distribution may under-present the probability of a large number of trips. Therefore, the log-normal distribution is applied as another alternative associated with multiplicative errors. The multiplicative errors represent uncertain scales. The case study on the A15 motorway illustrates the proposed hierarchical Bayesian networks.

The goal of Chapter 5 is to forecast day-to-day freight truck demand. The framework setting is with hierarchical Bayesian networks, and further involves a multi-process model to do the forecasting. The model is able to identify the right demand weights among several proposed scenarios. A concept of origin destination tuple additionally is introduced to represent the trip chains of vehicles. Chapter 6 concludes the dissertation.

1.7 Cooperation Contributions in Each Chapter

Statistics Netherlands funded this research. Dr. Chris de Blois is the supervisor from Statistics Netherlands, helping me to conduct interviews with the right people in Statistics Netherlands, providing the necessary data, and coordinating the relevant support.

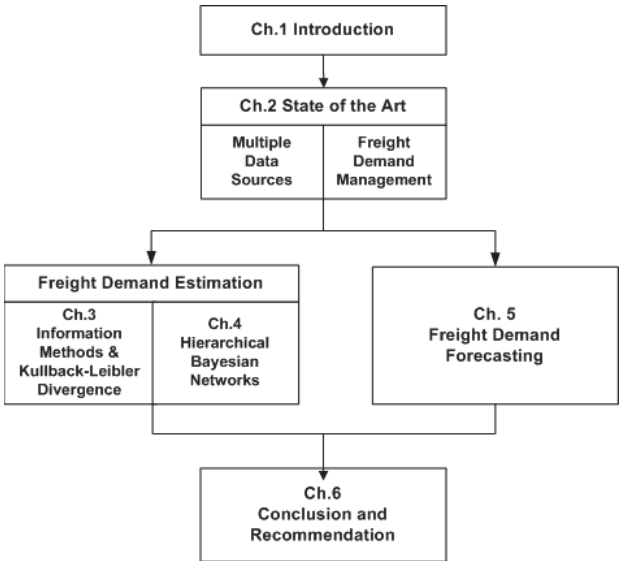


Figure 1.6: Outline of the Dissertation

In Chapter 2, I have summarized the statistical methods of OD estimation, associated with different types of detector data. On top of the static route proportion for link flows, Dr. Roelof Kuik and I proposed the route proportion associated with the path flows, from cameras for instance. This approach is based on the Hadamard product of the route proportions for link flows, from loops for instance. Prof. Leo Kroon, Dr. Jan van Dalen and Prof. Henk van Zuylen gave comments on this chapter and polished my texts.

In Chapter 3, I have reviewed in detail the information minimization method for OD estimation (Van Zuylen and Willumsen, 1980). To relax the assumptions in the paper, Dr. Roelof Kuik and I have proposed another information measure, called Kullback-Leibler divergence method. Mathematically, I have shown the connection between these two methods. Practically, the results from the case study demonstrates the feasibility of the Kullback-Leibler divergence method. Prof. Henk van Zuylen and Dr. Jan van Dalen devoted their effort on the methodology check. Prof. Leo Kroon and Dr. Chris de Blois provided their feedback and polished my texts.

In Chapter 4, Prof. Henk van Zuylen suggested implementing the Bayesian inference for demand estimation. I set up the Bayesian framework for OD estimation, together with Dr. Roelof Kuik. Two distributions of errors, normal distributions and log-normal distributions, are proposed. After deriving the posterior demand based on these two distributions, I have conducted a case study with the Markov Chain Monte Carlo simulation, Gibbs sampling and Metropolis-Hasting sampling. A part of this chapter has been published. I was the first author,

together with Dr. Roelof Kuik and Prof. Henk van Zuylen. Prof. Leo Kroon, Dr. Jan van Dalen and Dr. Chris de Blois gave comments on this chapter and polished my texts.

In Chapter 5, I came up with the idea of demand forecasting in the Bayesian framework. Dr. Roelof Kuik and I proposed the innovative concept of the origin destination tuple, which was supported by Prof. Henk van Zuylen. This proposed concept connects the travel activity model and the transportation planning. I have implemented the state-space model associated with the auto-regressive model, the Kalman filtering and the multi-process model to forecast the demand. To my knowledge, there is no paper to apply the multi-process model to forecast demand. Dr. Roelof Kuik and Dr. Jan van Dalen devoted their efforts on the methodology checking and the case-study adjustment. The initial version of this chapter has been published. I was the first author, together with Dr. Roelof Kuik and Prof. Henk van Zuylen. Prof. Leo Kroon and Dr. Chris de Blois gave comments on this chapter and polished my texts.

Chapter 2

Literature Review

2.1 Introduction

There are two ways to define freight: one is based on freight vehicles and the other is based on the commodities transported. Consequently, determining a freight OD matrix has two approaches: one is the vehicle-trip-based approach and the other is the commodity-based approach. The vehicle-trip-based approach focuses on the movements of freight vehicles. The idea is to use data captured on the roads, mainly loop detector data, to estimate a freight truck OD matrix. The commodity-based approach, in contrast, focuses on the flow of transported cargo. Due to the inability to model empty trips, where the complexity of transportation activities and the confidential nature of cargo information are involved, the application of this approach is limited.

Additionally, the use of multiple data sources for traffic OD estimation, especially Automatic Vehicle Identification (AVI) data, has emerged in recent years. In principle, AVI data makes it possible to distinguish trucks from other vehicles and vehicles can be traced from one AVI location to another. Furthermore, the path flow data obtained from AVI may increase estimation accuracy substantially (Dixon and Rilett, 2002; Zhou and Mahmassani, 2006).

In this chapter, the classical problem of OD matrix estimation is demonstrated. Papers with the vehicle-trip-based approach are reviewed from both the modeling and the solution point of view. Two general estimation methods, point estimation and distribution estimation, are discussed. Following this, the use of freight flow data from multiple data sources is demonstrated.

2.2 Linear Relation between Origin Destination Matrix and Flow

The general relationship between the observed flow and the origin-destination (OD) demand makes use of route proportions. Denote the flow at link a as V_a, the OD demand from origin i to destination j as T^{ij}, the route proportion at link a from origin i to destination j as A_a^{ij} (in some cases, the route proportion is dependent on the flow.), and the error term as E_a. The so-called flow-demand equation is formulated as follows.

$$V_a = \sum_{ij} A_a^{ij} T^{ij} + E_a \tag{2.1}$$

Using matrix notation, flows are represented as a column vector with length n, $V^\top = [V_1, \ldots, V_n]$. The OD demand T is a column vector with length $m = \alpha \cdot \beta$, where α is the number of origins and β is the number of destinations. The demand with a format of a column vector can be obtained by stacking the columns of an OD matrix like Table 1.2, $T^\top = [(T^1)^\top, \ldots, (T^\alpha)^\top]^\top$. The route proportion as A is with a matrix of $n \times m$, $A = [A_a^{ij}]_{n \times m}$. The error term is a column

vector with length n, $E^\top = [E_1, \ldots, E_n]$. The relation is expressed as follows.

$$V = A \times T + E \tag{2.2}$$

Traffic assignment is applied to determine the route proportions A. There are four main approaches to determine these proportions for trip generation, illustrated in Table 2.1.

Table 2.1: Classification of Static Traffic Assignment

		Congestion Effect Modeled	
		NO	YES
Random Utility Route Choice Modeled	NO	All-or-Nothing	Deterministic Equilibrium
	YES	Stochastic	Stochastic Equilibrium

First, the All-or-Nothing rule applies the shortest path method without considering congestion. The shortest path is one of the key technologies used in distributed route guidance systems. It is concerned with finding the shortest path from a specific origin to a specific destination in a given network, while minimizing the total distance, time or cost associated with the path (Deng and Tong, 2011). Well-known algorithms, including Bellman's dynamic programming algorithm (Bellman, 1956), Dijkstra's algorithm (Dijkstra, 1959) and Bellman-Ford's successive algorithm (Bellman, 1956; Ford and Fulkerson, 1962), are referred to as standard shortest path algorithms. For navigation systems, which can track and trace the trajectories of vehicles, the A-star algorithm (Kumar and Kumar, 2011) is more suitable. This algorithm improves Dijkstra's approach by maintaining a heuristic estimate of how close a given node is to the best route. This is provided by calculating the Euclidian distance to the destination at every node (Jenkins, 2007). Bell (2009) introduces the hyper-star algorithm as a multi-path A-star algorithm for risk averse vehicle navigation. The hyper-star algorithm takes travel time reliability into account, and can deliver all paths that may be optimal. It overcomes the shortcomings of the A-star algorithm which offers only one path.

Second, in the deterministic equilibrium assignment a distribution of trips over multiple routes may arise if it is assumed that the traveler chooses the shortest route. The key assumption underlying the user equilibrium (UE) assignment model is that each traveler has perfect information concerning the attributes of the network and each traveler chooses a route that minimizes travel time or travel costs, such that all travelers between the same origin and destination have the same travel time or cost (Wardrop, 1952). Equilibrium methods take account of the volume-dependency of travel times and result in the calculation of link flows and travel times. Equilibrium flow algorithms require iterating back and forth between assigning flows and calculating travel times (Sheffi, 1985). A variational inequality formulation is usually applied to solve the network user equilibrium problem (Friesz et al., 1993; Wu et al., 1994; Ran and Boyce, 1994).

Third, stochastic assignment distributes the trips between two zones over several routes connecting those zones, making assumptions concerning route choice behavior. This approach is mainly useful for the analysis of traffic during non-congested periods. The definition of stochastic assignment is that all travelers choose their perceived shortest path from origin to destination without taking the effects of congestion into account. The stochastic assignment model assumes that the value of the generalized travel time that a traveler attaches to a route, known as travel time, follows a probability distribution. The average of this probability distribution generally equals the generalized objective travel time. The standard deviation of the probability distribution is a measure for the behavioral differences between travelers.

Fourth, the stochastic user equilibrium assignment (SUE) or Wardrop (1952) principle combines the properties of the stochastic assignment and the deterministic user equilibrium assignment (Daganzo and Sheffi, 1977). This model is based on the assumption that travelers have imperfect information about network paths and vary in their perceptions of the network attributes. In the road network, there are many alternative routes that can be used to travel from a single origin zone to a single destination zone. Often, trips from various points within an origin zone to various points in a destination zone will use entirely different major roads to make the trip. Also, individuals will judge each alternative in a different way. These mechanisms cause many paths between origin and destination to be used, even if link costs are assumed to be independent of link flows. SUE leads to an equilibrium in which no traveler can improve his perceived travel time by unilaterally changing routes (Wardrop, 1952).

In this thesis, a static route choice model without congestion effect is applied, as Viti and Corman (2012). The contributions focus on the demand estimation.

Furthermore, the error term in Equation (2.2) may be from different sources, such as observational devices and traffic simulation. The first type of error is labeled as observation error or measurement error (Zhou and Mahmassani, 2006), which results from disruptions of the devices, for instance. The second is from the route choice mapping. The inconsistency of the shortest paths in the traffic assignment and the real travel behavior in the road network leads to a difference between the observed flow and the expected flow.

Due to the fact that the dimension of observed flows defined by the number of links n is much less than the dimension of an unobserved OD matrix m in Equation (2.2), estimating an OD matrix based on link flows is an under-specified problem (Van Zuylen and Willumsen, 1980). This under-specification increases the uncertainty of getting an accurate OD matrix.

2.3 Origin Destination Matrix Estimation Model with Traffic Data

The estimation of the traffic OD matrix has been subject to research for about three decades. There are two main approaches to model the OD matrix estimation: point estimation and

distribution estimation. In statistics, point estimation involves the use of sample data to infer a single value that serves as a best guess or a best estimate of an unknown, fixed population parameter. Distribution estimation aims to predict the random variables. Point estimation can be done with maximum likelihood, least squares and maximization of entropy or minimization of information methods. Distribution estimation is done with Bayesian inference and Kalman filtering.

Other methods that have drawn attention are variational inequality (Nie and Zhang, 2008), gradient approximation (Frederix et al., 2011) and Thompson estimation (Zhang et al., 2010). Most of the research uses loop detector data, while some of the most recent research combines this with AVI data, such as camera data or Bluetooth data. Table 2.2 presents an overview of the literature.

Table 2.2: Summary of OD Estimation Literature

		Loop	Loop & AVI	Loop & Survey
Point Estimation Method	Maximum Likelihood	Spiess (1987) Cascetta and Nguyen (1988) Nihan and Davis (1989) Watling and Maher (1992)		Watling (1994)
	Generalized Least Squares	Li and Moor (2002)	Asakura et al. (2000) Zhou and Mahmassani (2006)	Cascetta (1984)
	Information Method	Van Zuylen and Willumsen (1980) Van Zuylen (1981)		
Distribution Estimation Method	Bayesian Inference	Hazelton (2008)	Van Der Zijpp (1997)	
	Kalman Filtering	Chang and Wu (1994) Ashok and Ben-Akiva (2000) Lin and Chang (2007) Zhou and Mahmassani (2007)	Dixon and Rilett (2002) Barceló et al. (2010)	

2.3.1 Point Estimation Methods

Point estimation is used to determine a single value which serves as a best estimate of an unknown population parameter. Maximum likelihood and generalized least squares are the two common methods to arrive at point estimates. Information and entropy in information theory (Brillouin, 1956; Cover and Thomas, 2006) are measures of the average uncertainty in a random variable (Cover and Thomas, 2006). The optimum value is obtained by minimizing the uncertainty.

Maximum Likelihood

The maximum likelihood determines values of the model parameters that have the highest likelihood to give the observed data. This method was applied mainly in early studies of OD matrix estimation by Spiess (1987), Cascetta and Nguyen (1988), Nihan and Davis (1989), Watling and Maher (1992), and Watling (1994).

Spiess (1987) applies maximum likelihood in a convex programming problem, in which the elements of an OD matrix are assumed to be outcomes of flow observations with a Poisson distribution with an unknown mean. He ignores the connectivity and topology of the network and defines his likelihood function solely in terms of a set of independent observations on each OD pair (Hazelton, 2000). Nihan and Davis (1989) estimate an intersection OD matrix by minimizing the error between observed and predicted exiting counts. Watling (1994) applies maximum likelihood to a partial registration plate survey, in which all possible combinations of the observed data are considered. At that time, since the maximum likelihood cannot be obtained analytically, alternative numerical techniques were applied.

Generalized Least Squares

The least squares method minimizes the sum of the squared deviations between a prior OD matrix and an estimated OD matrix, based on the observed flows. It became popular in the eighties and the beginning of the nineties, and has been applied by many researchers: Cascetta (1984), Carey and Revelli (1986), Cascetta and Nguyen (1988), Bell (1991), Bierlaire and Toint (1995), Yang (1995). Estimators based on least squares have the advantage of being relatively easy to solve mathematically. Especially for larger problems, such as the simultaneous estimation of OD matrices for several time steps in large networks, this methodology gives a feasible solution. Additionally, the least squares method gives the same results as maximum likelihood, if normal distributions of the OD demand variables are assumed.

Information or Entropy based Method

Van Zuylen and Willumsen (1980) and Van Zuylen (1981) gave an approach to generate the most likely OD matrix based on maximization of the entropy of the trip matrix or minimization of the information with respect to a prior OD matrix. Assuming that there are no errors in the observations of the link flows and independence among the flows in the highway sections, they formulate the equality of the assigned and the observed traffic flows on the links of the network. Willumsen (1984) extends the entropy approach with a scaling factor of the flow observations to estimate an OD matrix.

2.3.2 Distribution Estimation Methods

Distribution estimation methods take the stochastic nature of freight demand into account. The parameters of the distributions represent the features of the sample data. A Bayesian inference method is a typical distribution-based approach. In general, Bayesian inference methods can deal with all kinds of distributions.

Bayesian inference

Bayesian inference updates a prior OD distribution based on the flow observations to generate

a posterior OD matrix. This approach reduces the overall uncertainty of the estimates by producing posterior distributions for the parameters as well as predictive distributions for future OD flows (Perrakis et al., 2011). The approach began gaining popularity in the middle of the nineties, due to a paper by Tebaldi and West (1998), although Maher (1983) had already introduced this method to estimate OD matrices. Later, many researchers contributed to this method, such as Hazelton (2000), Li (2005), Sun et al. (2006), Castillo et al. (2008a), Hazelton (2010) and Perrakis et al. (2011).

Maher (1983) assumes a multivariate normal distribution of a prior OD matrix and considers normally distributed errors in the observations. This makes the computations very fast. Li (2005) applies a Bayesian method to deal with the under-specification problem. He states that a Bayesian analysis provides a research framework by specifying prior demand that amounts to introducing extra information based on accumulated knowledge. He uses an expectation maximization algorithm to overcome the problem of an analytically intractable likelihood. In addition, Bayesian Networks are applied by Castillo et al. (2008a) to represent the relation between OD demand variables, where linear relations of each layer in Bayesian Networks are built up. In order to deal with empirical distributions, Tebaldi and West (1998), Hazelton (2010) and Perrakis et al. (2011) propose a Markov Chain Monte Carlo (MCMC) simulation in Bayesian inference methods to estimate the OD matrix.

Kalman Filtering

Kalman filtering is widely used to adapt model parameters in a rolling horizon to the measured characteristics of the modeled reality. This method usually considers a state space and induces observable values. The relationships for the dynamics of the states and how the states induce observations may include errors. These errors are usually assumed to have normal distributions making computations easy and efficient. Chang and Wu (1994), Ashok and Ben-Akiva (2000), Dixon and Rilett (2002), Zhou and Mahmassani (2007), and Barceló et al. (2010) use Kalman filtering to estimate and predict dynamic OD matrices.

Zhou and Mahmassani (2007) present a structural state-space model to systematically incorporate regular demand pattern information, structural deviations and random fluctuations. By considering demand deviations from the prior estimate of the regular pattern as a time-varying process with a smooth trend, a polynomial trend filter is developed to capture possible structural deviations in the real-time demand. An optimal adaptive procedure is proposed based on a Kalman filtering framework, to capture day-to-day demand evolution, and to update the prior demand pattern estimates using new real-time estimates and observations obtained every day.

2.3.3 Discussion

Compared with point estimation methods, Bayesian inference treats parameters as random variables, instead of single realizations. Bayesian inference considers prior information and

determines the posterior distribution of the parameters, conditioned on the observed data. The core characteristic of Bayesian inference is the updating of prior beliefs to obtain posterior belief (Greenberg, 2008).

Maximum likelihood, least squares, Bayesian inference and Kalman filtering show great similarity in their basic mechanisms if normal distributions are assumed for the errors. Kalman filtering and Bayesian methods are popular for OD estimation and prediction due to the recursive method of updating estimates based on new (recent) data. Kalman filtering, where normality dominates distributions, is a special case of recursive Bayesian estimation (Koopman et al., 2012).

2.4 Capturing Freight Flow by Multiple Data Sources

The general linear relation between flows and demand is well known. This flow-demand equation is connected by the route proportion. Multiple data sources bring different ways to capture flows including link flows and path flows. The route proportion of the path flows is discussed, which is different from the one of the link flows.

2.4.1 Link Flows Obtained from Loop Detectors and Weigh-in-Motion Equipment

Loop detectors record the total traffic flow on a link during a certain time interval. They can record passing vehicles and register characteristics, such as weight and length, which may be used to estimate the share of trucks among the total number of vehicles. If we denote the observed truck flow over loop detector l_x for a given time period as V_{l_x} and the link proportion that is the share of trucks passing detector x when traveling from origin i to destination j as $A_{l_x}^{ij}$, then the relationship between the observed truck flow and truck OD trips is:

$$V_{l_x} = \sum_{ij} A_{l_x}^{ij} \cdot T^{ij} + \varepsilon_{l_x} \tag{2.3}$$

The error term ε_{l_x} could be assumed with zero mean and a certain variance. Weigh-in-Motion equipment is installed in a limited number, only nine locations in the Netherlands, for instance. Thus, estimating OD matrices only based on the truck flow data from WiM is inaccurate.

2.4.2 Path Flows Obtained from Cameras and Bluetooth Scanners

Cameras have the advantage of capturing the vehicle identification: trucks can be recognized when passing cameras during their trips. These identified continuous recordings recognized by

the set of cameras on the highways reduce the uncertainty in the OD demand. The set of cameras which identify the paths is denoted as $c_x...c_y$. The relationship between path flow $W_{c_x...c_y}$ from cameras and OD trips T^{ij} is:

$$W_{c_x...c_y} = \sum_{ij} A^{ij}_{c_x...c_y} \cdot T^{ij} + \varepsilon_{c_x...c_y} \tag{2.4}$$

where the route proportion $A^{ij}_{c_x...c_y}$ is the share of trucks passing a set of cameras $c_x...c_y$ when traveling from origin i to destination j.

Bluetooth scanners can identify devices on-board of vehicles using their 16 digit MAC addresses. The method to obtain the trip flow from Bluetooth scanners is similar to cameras. $W_{b_x...b_y}$ denotes the path flows from a set of Bluetooth scanners $b_x...b_y$ and $A^{ij}_{b_x...b_y}$ is the share of trucks passing a set of Bluetooth scanners when traveling from origin i to destination j:

$$W_{b_x...b_y} = \sum_{ij} A^{ij}_{b_x...b_y} \cdot T^{ij} + \varepsilon_{b_x...b_y} \tag{2.5}$$

Regarding the characteristics of Bluetooth data, there are at least three disadvantages when compared with cameras. First, incomplete flow data may arise when no Bluetooth devices are on board of a freight truck. Vehicles without Bluetooth devices will not be visible for Bluetooth scanners. This incomplete information could be captured by the error term in Equation (2.5). Second, vehicles may carry multiple Bluetooth devices that lead to redundantly captured information. Current Bluetooth systems are able to filter multiple signals. Third, Bluetooth devices cannot be directly related to the type of vehicle, implying that trucks cannot be distinguished from the traffic flow. A solution may be to estimate the proportions of Bluetooth recordings coming from trucks and other vehicles. This could be done by looking at camera data at locations where both cameras and Bluetooth scanners are available. The estimated proportion is assumed to apply to all the links.

2.4.3 Route Proportion associated with Path Flows and Demand

Link flows usually cannot determine the exact OD demand, because the rank of the route proportion matrix A is not full in most of the cases, with many ones in each row. These ones make it difficult to estimate OD demand. When vehicles can be identified, we can do a little better by processing the counts and therefore obtain more detailed information about OD volumes. The path flows from the identification function cannot exactly represent the OD pairs, if the devices are not located at all origins and all destinations. The combined use of identification devices can reduce the number of ones in each row of the route proportion matrix, which decreases the uncertainty in the OD demand.

To illustrate the concept of the route proportions in the case of link flows and path flows, a small artificial road network with four OD pairs and three links is presented in Figure 2.1. On each link, a loop detector and a camera are installed.

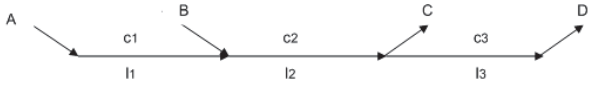


Figure 2.1: Small Network for Explaining the Route Proportion

The OD pairs are summarised as $T = \begin{pmatrix} T^{AC} & T^{AD} & T^{BC} & T^{BD} \end{pmatrix}^\top$. Loop detectors are located on the three links. So the link flows can be represented as $V_l = \begin{pmatrix} V_{l_1} & V_{l_2} & V_{l_3} \end{pmatrix}^\top$. Following the linear relation $V = A \cdot T$, the route proportion matrix A is defined as:

$$\begin{pmatrix} V_{l_1} \\ V_{l_2} \\ V_{l_3} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} T^{AC} \\ T^{AD} \\ T^{BC} \\ T^{BD} \end{pmatrix} \tag{2.6}$$

It shows, for instance, that trucks traveling from B to C and trucks going from B to D are captured in the flow recorded at loop detector l_2. The equation also shows that trips T^{ij} cannot be uniquely identified based on the three flow recordings V_{l_1}, V_{l_2}, and V_{l_3}, which illustrates the underspecification problem.

Cameras do allow one to identify the trips in this network, by combining the recordings from different devices. Given the three available cameras, seven possible combinations exist: c_1, c_2, c_3, c_1c_2, c_1c_3, c_2c_3, and $c_1c_2c_3$. The recorded flows of these combinations are denoted as W_{c1}, W_{c2}, W_{c3}, W_{c1c2}, W_{c1c3}, W_{c2c3}, and W_{c1c2c3}, where W_{c1c2}, for instance, is the number of trucks recorded by cameras c_1 and c_2. The flow-demand equation for these combined camera flows is:

$$\begin{pmatrix} W_{c1} \\ W_{c2} \\ W_{c3} \\ W_{c1c2} \\ W_{c2c3} \\ W_{c1c3} \\ W_{c1c2c3} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} T^{AC} \\ T^{AD} \\ T^{BC} \\ T^{BD} \end{pmatrix} \tag{2.7}$$

Clearly, not all combined camera flows contribute to identifying OD trips: no trucks are recorded solely by camera c_1 or camera c_3, nor by the combination of cameras c_1 and c_3. Thus, three pieces of combined flow information, W_{c1c2}, W_{c2c3}, and W_{c1c2c3}, suffice to uniquely identify the OD trips in this example. Due to the specific location of camera 2 and the identification function of cameras, camera $c2$ is able to capture the demand from B to C, associated with W_{c2}.

Comparing the route proportion matrix for link flows in Equation (2.6) and the route proportion matrix for path flows in Equation (2.7), there are more ones in the route proportion matrix for link flows than for path flows. Path flows thus reduce the uncertainty to estimate OD demand.

Path flows and link flows are clearly not unrelated. In fact, the route proportions for path flows can be obtained from those of the link flows. Denoting the route proportion of each device as D, the route proportion of single loop detector l_1 or single camera c_1 without identification function is denoted as $D_1 = (1 \quad 1 \quad 0 \quad 0)$, which is the same as the first row from the route proportion matrix A in Equation (2.6). The same principle holds for the other devices: $D_2 = (1 \quad 1 \quad 1 \quad 1)$ and $D_3 = (0 \quad 1 \quad 0 \quad 1)$. D_1, D_2 and D_3 are the rows of the route proportion matrix in Equation (2.6).

The route proportion of path flows based on the route proportion of link flows is obtained through element-wise matrix multiplication, the so-called Hadamard product (Shao et al., 2014), represented by $\circ$. If a camera c is not triggered, then the associated vector of route proportions may be defined as $\overline{D}_c = \iota - D_c$, where ι is a vector of ones of appropriate length. For instance, $\overline{D}_{c_3} = \iota - D_{c_3} = (1\ 1\ 1\ 1) - (0\ 1\ 0\ 1) = (1\ 0\ 1\ 0)$. The element-wise multiplication of the appropriate route proportion vectors gives the vector associated with a particular path. For instance, the triggered camera 1 and camera 2 are associated with the route mapping $A_{c1c2\bar{c}_3}$:

$$A_{c_1c_2\bar{c}_3} = D_{c_1} \circ D_{c_2} \circ \overline{D}_{c_3} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}^\top = (1 \quad 0 \quad 0 \quad 0).$$

The same principle is applied to other route proportion of valid path flows:

$$A_{\bar{c}_1c_2c_3} = \overline{D}_{c_1} \circ D_{c_2} \circ D_{c_3} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}^\top = (0 \quad 0 \quad 0 \quad 1)$$

$$A_{c_1c_2c_3} = D_{c_1} \circ D_{c_2} \circ D_{c_3} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}^\top = (0 \quad 1 \quad 0 \quad 0)$$

$$A_{\bar{c}_1c_2\bar{c}_3} = \overline{D}_{c_1} \circ D_{c_2} \circ \overline{D}_{c_3} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}^\top \circ \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}^\top = (0 \quad 0 \quad 1 \quad 0)$$

These four route proportions are the same as the route proportion matrix in Equation (2.7). If the set of cameras cannot offer the valid path flow, there are zeros in the corresponding route proportion. For instance, if the set of combined cameras consist of only one camera c_1, the associated mapping $A_{c_1\overline{c}_2\overline{c}_3}$ has:

$$A_{c_1\overline{c}_2\overline{c}_3} = D_{c_1} \circ \overline{D}_{c_2} \circ \overline{D}_{c_3} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}^\top \circ \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}^\top \circ \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}^\top = (0 \quad 0 \quad 0 \quad 0).$$

In general, having n cameras installed, there exist $2^n - 1$ possible combined camera recordings, for which the associated mappings can be derived as described.

2.5 Multiple Data Sources for Origin Destination Estimation

With the development of technology, different data collection devices have been introduced. These devices capture different characteristics of transport data. Automatic Vehicle Identification (AVI) data has attracted research interest for estimating OD matrices. AVI is used to collect OD information about vehicle movements between detector stations or key locations in the transportation network. The advantage of exploiting AVI data for extracting information on networks is its extreme flexibility and possibility to cover a large part of links with a relatively small share of the total demand chaptured (Viti and Corman, 2012).

Zhou and Mahmassani (2006) estimate OD matrices by extracting the link-to-link information from AVI counts in fixed locations of AVI, without considering the optimal locations of the AVI devices and without estimating the penetration rates of AVI devices. They develop a joint estimation method based on the ordinary least-squares model, taking errors into account, such as model assumption errors, sensor errors, sampling errors, aggregation errors and dynamic traffic assignment estimation errors. They argue that it is advantageous to locate AVI detectors in major OD demand zones with large traffic attraction or production in order to capture the essential OD distribution pattern in the network. Furthermore, Zhou and List (2010) discuss the selection of the AVI locations and the maximization of the expected information gain for OD matrix estimation, considering several important error sources, such as measurement errors and assignment errors. A scenario-based, stochastic optimization procedure and a beam search algorithm were developed to find suboptimal point and point-to-point sensor locations subject to budget constraints.

Others, like Dixon and Rilett (2002) and Asakura et al. (2000) addressed the added value of AVI data for OD estimation, using generalized least squares and Kalman filter approaches and

least squares, respectively. Dixon and Rilett (2002) addressed the issue of market penetration rates using AVI and counts data.

2.6 Conclusion

In this chapter, the methods for estimating OD matrices have been reviewed. There are two approaches: point estimation and distribution estimation. Maximum likelihood, least squares and information methods belong to the point estimation approaches; Bayesian inference and Kalman filtering are distribution estimation approaches. The main advantage of distribution estimation is that it takes the stochastic features of flows and demand into account. Bayesian inference is able to update the prior information to obtain posterior belief.

In addition, link flows and path flows are introduced, associated with the linear relation between flows and demand. The route mapping of loop detectors has been developed to link flows and demand. Based on the route proportion of link flows, the route proportion of path flows is proposed through element-wise matrix multiplication, the so-called Hadamard product. In such a way, a mapping with the combined use of cameras can be obtained.

Building on the literature review, the following chapters further develop the methodology for estimating OD demand and demonstrate the use of multiple data sources.

Chapter 3

Kullback-Leibler Divergence Method for Freight Truck OD Estimation

3.1 Introduction

Because demand management is so important in the field of transportation, it helps if decision makers are able to get insight into travel behavior in the road network. Usually, demand is unobservable, while we can only capture flow data in the network. A proper estimation method is required to obtain an accurate OD demand. Nowadays, different types of devices are installed in the network to capture the flow data. These new data sources may help to get high-quality demand data.

In this chapter, information methods are applied to estimate a freight truck OD matrix. Van Zuylen and Willumsen (1980) and Van Zuylen (1981) proposed an information minimization approach. In their approaches, a Stirling's approximation is applied which requires large flow data. In order to elaborate the theoretical background of their papers and relax the approximation method, the relative entropy or Kullback-Leibler divergence approach, another information method, is described. Mathematically, the information minimization method and the Kullback-Leibler divergence approach appear to have a strong connection. In addition, Van Zuylen and Willumsen (1980) applied only loop detector data, while we take multiple data sources into account, such as cameras in addition to loop detectors. The path flow data from cameras for instance can improve the quality of the estimated OD matrix, which is shown by a case study.

Research questions arise such as what are the connections and differences between the information minimization approach and the Kullback-Leibler divergence approach; whether the Kullback-Leibler divergence approach is a proper method to estimate demand; and what is the benefit of using multiple data sources to estimate OD demand. Based on a case of the A15 region in the Netherlands, we illustrate how different data capturing technologies are actually implemented, and demonstrate the benefit of these multiple data sources for estimating a freight truck OD matrix.

3.2 Review of the Information Minimization Method

In an early study, Van Zuylen and Willumsen (1980) reached an estimate of an OD matrix through information minimization and entropy maximization. The under-specification issue associated with OD matrix estimation arises from the fact that the information available in the counts on the links is insufficient to determine a complete OD matrix. Choosing a trip matrix that adds as little flow information as possible seems reasonable (Van Zuylen and Willumsen, 1980). The work of Van Zuylen and Willumsen (1980) on estimating the most likely OD matrix starts from Brillouin's information measure in information theory (Brillouin, 1956).

Van Zuylen and Willumsen (1980) measure the difference between a prior OD matrix and an estimated OD matrix. The information I_a (Brillouin, 1956) contained in V_a vehicle counts at

a link a, with n_a^{ij} vehicles traveling from origin i to destination j via a, and a prior probability q_a^{ij} of a vehicle traveling from i to j via a is expressed as:

$$I_a = -\ln\left(V_a! \prod_{ij} \frac{(q_a^{ij})^{n_a^{ij}}}{n_a^{ij}!}\right) \tag{3.1}$$

The total number of vehicles travelling from i to j via link a, n_a^{ij}, can be written as a proportion A_a^{ij} of the total number of vehicles going from i to j. V_a is the summation of n_a^{ij} over ij.

$$n_a^{ij} = T^{ij} A_a^{ij} \tag{3.2}$$

$$V_a = \sum_{i,j} n_a^{ij} = \sum_{i,j} T^{ij} A_a^{ij} \tag{3.3}$$

The q_a^{ij} in Equation (3.1) is a prior probability of a vehicle traveling via link a from origin i to destination j. This probability is normalized over all vehicle counts on link a from the prior OD matrix t^{ij}.

$$q_a^{ij} = \frac{t^{ij} A_a^{ij}}{S_a} \tag{3.4}$$

where

$$S_a = \sum_{ij} t^{ij} A_a^{ij}$$

In view of Equations (3.2) and (3.4), the information on link a can be expressed as a function of the actual demand, as follows:

$$I_a(T) = -\ln\left(V_a! \prod_{ij} \left(\frac{t^{ij} A_a^{ij}}{S_a}\right)^{T^{ij} A_a^{ij}} / \prod_{ij} (T^{ij} A_a^{ij})!\right) \tag{3.5}$$

Considering the complexity of calculation, Stirling's approximation is applied by Van Zuylen and Willumsen (1980) to approximate the factorial part, $\ln X! \cong X \ln X - X$. Thus, the information on a link can be formulated as in Equation (3.6).

$$I_a(T) = -\sum_{ij} T^{ij} A_a^{ij} \ln\left(\frac{T^{ij} S_a}{V_a t^{ij}}\right) \tag{3.6}$$

Note that the application of Stirling's approximation is applied to the situation where the factorial part is a large number (Nemes, 2010). It means that Van Zuylen and Willumsen (1980) assume that the link flows are always large. For the case where there are small link flows, such as night time, this approximation may lead to less accuracy.

Further, assuming that link flows in the highway sections are independent, we have the total information I over the network as:

$$I(T) = -\sum_{a}\sum_{ij} T^{ij} A_a^{ij} \ln\left(\frac{T^{ij} S_a}{V_a t^{ij}}\right) \tag{3.7}$$

subject to

$$V_a = \sum_{i,j} T^{ij} A_a^{ij} \tag{3.8}$$

This independence actually ignores the network structure. Summing the information about all vehicles on each link of the network is not proper. Usually, the network is fixed with a certain number of links, while the number of vehicles on each link varies case by case. This will be further addressed in the next section.

Since the expression of information $I(T)$ is concave in the variable T for both the objective and constraint parts, Lagrangian relaxation is applied. Lagrangian relaxation uses a Lagrangian factor to account for the equality constraints between the observed and allocated traffic flows on the links. A new function $\varphi(T, \lambda)$, where $T = (T^{ij})_{ij}$ and $\lambda = (\lambda_a)_a$ is set up with variables T and λ.

$$\max_{T,\lambda} \varphi(T, \lambda) = \sum_{a}\sum_{ij} T^{ij} A_a^{ij} \ln\frac{T^{ij} S_a}{V_a t^{ij}} + \sum_{a} \lambda_a \left(\sum_{ij} T^{ij} A_a^{ij} - V_a\right) \tag{3.9}$$

By partially differentiating the Lagrangian over T^{ij} and λ_a, respectively, Van Zuylen and Willumsen (1980) arrive at the estimated OD trips by updating a prior OD matrix with multipliers X_a. Since Equation (3.9) is convex either in T^{ij} with a fixed λ_a or in λ_a with a fixed T^{ij}, it has only one maximum value.

$$T^{ij} = t^{ij} \prod_{a} X_a^{A_a^{ij}/g^{ij}} \tag{3.10}$$

$$X_a = \frac{V_a}{S_a} e^{-(1+\lambda_a)}$$

$$g^{ij} = \sum_{a} A_a^{ij}$$

$$\sum_{ij} T^{ij} A_a^{ij} - V_a = 0 \tag{3.11}$$

The idea expressed in Equation (3.10) is that the prior trips t^{ij} are updated through the multipliers X_a and the route proportion A_a^{ij} divided by g^{ij}, the number of times that a trip between i and j is detected. The multipliers involve the Lagrangian factor λ_a, and the ratio of observed and prior expected link counts V_a/S_a. Substituting Equation (3.10) into (3.11), the estimated demand T^{ij} can be obtained in principle. But Equation (3.10) is in the form of a

posynomial (Duffin et al., 1967), where the power of a matrix is a matrix as well. It is impossible to solve the formulas analytically. The approach proposed in Van Zuylen and Willumsen (1980) often does not converge.

In order to relax the assumption of the Stirling's approximation and to have a proper approach to model the information measure in the network, the Kullback-Leibler divergence method is proposed in the next section.

3.3 Kullback-Leibler Divergence to Estimate OD Demand

The Kullback-Leibler divergence method (Kullback and Leibler, 1951) is also called information measure or relative entropy method in probability theory and in information theory. The Kullback-Leibler divergence (Cover and Thomas, 2006) is a non-symmetric measure of the difference between two probability distributions P and Q. Typically, P represents the true distribution of data, observations, or a precisely calculated theoretical distribution, while the measure Q represents a prior, theory, or approximation of P.

The original idea of connecting OD trips and link flows is done through an experiment: randomly selecting one vehicle and assigning the vehicle to the road network. The probability of the vehicles having the OD pair (i, j) is T^{ij}/T, where T^{ij} is the number of trucks with OD pair (i, j) and T is the total number of trucks in all OD pairs. Given counts on a link a, the probability of a truck having OD pair (i, j) is the conditional probability, denoted as $p(ij|a)$, which is proportional to the joint probability of OD pair (i, j) and link a, $p(ij, a)$. The joint probability is the product of the route proportion $p(a|ij)$ (or p_a^{ij}) and the probability of an OD trip at (i, j), $p(i, j)$. The concept indicates that the demand follows a multinomial distribution.

$$p(ij, a) = p(a|ij)p(ij) = A_a^{ij} \cdot \frac{T^{ij}}{T} \tag{3.12}$$

The conditional probability $p(ij|a)$ is proportional to $p(ij, a)$ and is given as:

$$p(ij|a) = \frac{p(ij, a)}{p(a)} = \frac{p(ij, a)}{\sum_{i'j'} p(i'j', a)} = \frac{p(a|ij)p(ij)}{\sum_{i'j'} p(a|i'j')p(i'j')} = \frac{T^{ij}A_a^{ij}}{\sum_{i'j'} T^{i'j'}A_a^{i'j'}} \tag{3.13}$$

Thus, the probability of vehicles from OD matrix T_{ij} contributing to the flow on link a is expressed as

$$P(ij|a) = \frac{T^{ij}A_a^{ij}}{V_a} \tag{3.14}$$

Denoting the probability that trucks from OD pair ij in the prior OD matrix t^{ij} appear on link a as $Q(ij|a)$, we have

$$Q(ij|a) = \frac{t^{ij}A_a^{ij}}{\sum_{i'j'} t^{i'j'}A_a^{i'j'}} \tag{3.15}$$

Thus, the divergence is represented by the expected number of extra counts required to get the probability of the real OD matrix T^{ij} when using an assignment probability based on a prior OD matrix t^{ij}. The Kullback-Leibler divergence of the discrete probability distribution of each vehicle at link a, denoted as $D_{KL,a}$, is expressed in Equation (3.16). It measures the difference between $P(ij|a)$ and $Q(ij|a)$, given the prior OD matrix t^{ij} and the estimated OD matrix T^{ij} of a single truck in link a.

$$D_{KL,a}(T) = \sum_{ij} \ln\left(\frac{P(ij|a)}{Q(ij|a)}\right) P(ij|a) = \sum_{ij} \frac{T^{ij}A_a^{ij}}{V_a} \ln\left(\frac{T^{ij}S_a}{t^{ij}V_a}\right) \tag{3.16}$$

Assuming an independent network structure without correlations among link flows, the Kullback-Leibler divergence of each vehicle over the road network is simply the summation of $D_{KL,a}(T)$ over all links:

$$D_{KL}(T) = \sum_a D_{KL,a}(T) = \sum_a \sum_{ij} \frac{T^{ij}A_a^{ij}}{V_a} \ln\left(\frac{T^{ij}S_a}{t^{ij}V_a}\right) \tag{3.17}$$

subject to

$$V_a = \sum_{ij} T^{ij}A_a^{ij}$$

To find the maxima and minima of $D_{KL}(T)$ subject to the equality constraints, Lagrange relaxation is applied. Denoting the Lagrange multipliers for each link a as λ_a, the new objective function is presented in Equation (3.18), with two variables T^{ij} and λ_a.

$$\max_{T,\lambda} F(T,\lambda) = \sum_a \sum_{ij} \frac{T^{ij}A_a^{ij}}{V_a} \ln\left(\frac{T^{ij}S_a}{t^{ij}V_a}\right) + \sum_a \lambda_a \left(\sum_{ij} T^{ij}A_a^{ij} - V_a\right) \tag{3.18}$$

Differentiating $F(T,\lambda)$, where $T = (T^{ij})_{ij}$, over the variable T^{ij}, we have the expression of the estimated OD matrix T^{ij} in Equation (3.19).

$$\begin{aligned}
\frac{\partial F(T^{ij},\lambda_a)}{\partial T^{ij}} &= \sum_a \frac{A_a^{ij}}{V_a} \ln\left(\frac{T^{ij}S_a}{t^{ij}V_a}\right) + \sum_a \frac{T^{ij}A_a^{ij}}{V_a} \cdot \frac{1}{T^{ij}} + \sum_a \lambda_a A_a^{ij} \\
&= \sum_a \frac{A_a^{ij}}{V_a} \ln\left(\frac{T^{ij}S_a}{t^{ij}V_a}\right) + \sum_a \frac{A_a^{ij}}{V_a}(1 + \lambda_a V_a) \\
&= 0
\end{aligned}$$

$$\ln\left[\prod_a \left(\frac{T^{ij}S_a}{t^{ij}V_a}\right)^{A_a^{ij}/V_a}\right] = -\sum_a \frac{A_a^{ij}}{V_a}(1+\lambda_a V_a)$$

$$\left(\frac{T^{ij}}{t^{ij}}\right)^{\sum_a A_a^{ij}/V_a} = \prod_a \left(\frac{V_a}{S_a}\right)^{A_a^{ij}/V_a} \exp[-\sum_a \frac{A_a^{ij}}{V_a}(1+\lambda_a V_a)]$$

Hence,

$$T^{ij} = t^{ij}\prod_a X_a^{A_a^{ij}/(V_a g^{ij})} \tag{3.19}$$

$$X_a = \frac{V_a}{S_a}e^{-(1+\lambda_a V_a)}$$

$$g^{ij} = \sum_a \frac{A_a^{ij}}{V_a}$$

Equation (3.19) gives the optimal T^{ij} as a function of λ_a. Closed form expressions of both T^{ij} and λ_a cannot be easily obtained. In Section 3.5 a Genetic Algorithm is proposed to find numerical approximations of the optimal values.

3.4 Connection between Information Minimization and Kullback-Leibler Divergence

The information minimization method applied to the OD estimation starts with a physics interpretation that vehicles travel on the link a in a road network where the probability of vehicles traveling from i to j is given by q_a^{ij} as in Equation (3.4). Based on the information minimization method, the information for all vehicles on link a is expressed as $I_a(T)$. This link-based information is aggregated to the network level through Stirling's approximation. Summing the information about all vehicles of the network is not suitable. In order to have relatively stable information measures, it could be suitable to add up the average information of vehicles per link. In such a way, it allows to represent the network structure. Denoting the information of each vehicle in a link as $I_a^{'}$, Equation (3.6) can be updated in Equation (3.20).

$$I_a^{'}(T) = \frac{I_a(T)}{V_a} = -\sum_{ij}\frac{1}{V_a}T^{ij}A_a^{ij}\ln\left(\frac{T^{ij}S_a}{V_a t^{ij}}\right) \tag{3.20}$$

Mathematically, the expression for information at each link I_a in Equation (3.6) used by Van Zuylen and Willumsen (1980) has a strong connection with the expression of Kullback-Leibler divergence at each link D_{KL}^a in Equation (3.16):

$$V_a \cdot D_{KL,a}(T) = -I_a(T). \tag{3.21}$$

Consequently, the expressions for the estimated demand T^{ij} from the information method in Equation (3.10) and from the Kullback-Leibler divergence in Equation (3.19) are quite similar. There are two main differences. One is that the Lagrangian factor λ_a in Equation (3.19) is multiplied with a factor V_a. It can be simplified if we assume a new Lagrangian multiplier $\lambda' = \lambda_a V_a$. The other is that the mapping A_a^{ij} is scaled by the flow V_a in Equation (3.19). Comparing the exponents of $A_a^{ij} / \sum_a A_a^{ij}$ in Equation (3.10) and the scaled exponent of $\frac{A_a^{ij}}{V_a} / \sum_a \frac{A_a^{ij}}{V_a}$ in Equation (3.19), the difference between these two approaches is small when the deviations of the flows varying among links are limited. The difference could be large when the deviations of the flows significantly vary among links.

3.5 Kullback-Leibler Divergence Method with Multiple Data Sources

Different data sources have distinct characteristics. Here, we consider the link flow data from loop detectors and the path flow data from cameras. The information for freight truck OD estimation is contained in a set of freight flow observations. The observation set consists of information from independent link flows. Due to independence, the set could be extended to have both link flow data and path flow data, generated from multiple data sources besides loop detectors, such as cameras and Bluetooth. Starting from the Lagrangian expression of link flow information V_{l_x} from loop detector l_x on this link, and path flow information $W_{c_x...c_y}$ indicated by several cameras $c_x...c_y$, we get the following equation:

$$
\begin{aligned}
&\varphi(T^{ij}, \lambda_{l_x}, \lambda_{c_x...c_y}) \\
=&\sum_{l_x} \frac{A_{l_x}^{ij} T^{ij}}{V_{l_x}} \cdot \ln \frac{T^{ij} S_{l_x}}{V_{l_x} t^{ij}} + \sum_{l_x} \lambda_{l_x} (\sum_{ij} T^{ij} A_{l_x}^{ij} - V_{l_x}) \\
&+ \sum_{c_x...c_y} \frac{A_{c_x...c_y}^{ij} T^{ij}}{W_{c_x...c_y}} \cdot \ln \frac{T^{ij} S_{c_x...c_y}}{W_{c_x...c_y} t^{ij}} + \sum_{c_x...c_y} \lambda_{c_x...c_y} (\sum_{ij} T^{ij} A_{c_x...c_y}^{ij} - W_{c_x...c_y})
\end{aligned}
$$

where,

$$S_{l_x} = \sum_{ij} t^{ij} A_{l_x}^{ij}$$

$$S_{c_x...c_y} = \sum_{ij} t^{ij} A_{c_x...c_y}^{ij}$$

From partial differentiation over T^{ij}, we have

$$\begin{aligned}
&\frac{\partial\varphi(T^{ij},\lambda_{l_x},\lambda_{c_x\ldots c_y})}{\partial T^{ij}}\\
=&\sum_{l_x}\frac{A_{l_x}^{ij}}{V_{l_x}}\cdot\ln\left(\frac{T^{ij}S_{l_x}}{V_{l_x}t^{ij}}\right)+\sum_{l_x}\frac{A_{l_x}^{ij}}{V_{l_x}}+\sum_{l_x}\lambda_{l_x}A_{l_x}^{ij}\\
&+\sum_{c_x\ldots c_y}\frac{A_{c_x\ldots c_y}^{ij}}{W_{c_x\ldots c_y}}\cdot\ln\left(\frac{T^{ij}S_{c_x\ldots c_y}}{W_{c_x\ldots c_y}T^{ij}}\right)+\sum_{c_x\ldots c_y}\frac{A_{c_x\ldots c_y}^{ij}}{W_{c_x\ldots c_y}}+\sum_{c_x\ldots c_y}\lambda_{c_x\ldots c_y}A_{c_x\ldots c_y}^{ij}\\
=&0
\end{aligned}$$

Thus,

$$\begin{aligned}
&\ln\left[\prod_{l_x}\left(\frac{T^{ij}S_{l_x}}{V_{l_x}t^{ij}}\right)^{A_{l_x}^{ij}/V_{l_x}}\prod_{c_x\ldots c_y}\left(\frac{T^{ij}S_{c_x\ldots c_y}}{W_{c_x\ldots c_y}t^{ij}}\right)^{A_{c_x\ldots c_y}^{ij}/W_{c_x\ldots c_y}}\right]\\
=&-\sum_{l_x}(1+\lambda_{l_x}V_{l_x})\frac{A_{l_x}^{ij}}{V_{l_x}}-\sum_{c_x\ldots c_y}(1+\lambda_{c_x\ldots c_y}W_{c_x\ldots c_y})\frac{A_{c_x\ldots c_y}^{ij}}{W_{c_x\ldots c_y}}
\end{aligned}$$

Thus,

$$\begin{aligned}
&\left(\frac{T^{ij}}{t^{ij}}\right)^{\sum_{l_x}A_{l_x}^{ij}/V_{l_x}+\sum_{c_x\ldots c_y}A_{c_x\ldots c_y}^{ij}/W_{c_x\ldots c_y}}\\
=&\prod_{l_x}\left(\frac{V_{l_x}}{S_{l_x}}\right)^{A_{l_x}^{ij}/V_{l_x}}\prod_{c_x\ldots c_y}\left(\frac{W_{c_x\ldots c_y}}{S_{c_x\ldots c_y}}\right)^{A_{c_x\ldots c_y}^{ij}/W_{c_x\ldots c_y}}\cdot\\
&\exp\left[-\sum_{l_x}(1+\lambda_{l_x}V_{l_x})\frac{A_{l_x}^{ij}}{V_{l_x}}\right]\exp\left[-\sum_{c_x\ldots c_y}(1+\lambda_{c_x\ldots c_y}W_{c_x\ldots c_y})\frac{A_{c_x\ldots c_y}^{ij}}{W_{c_x\ldots c_y}}\right]
\end{aligned}$$

Hence, the estimated truck ODs T^{ij} is found as follows:

$$T^{ij}(\lambda_{l_x},\lambda_{c_x\ldots c_y})=t^{ij}X_0^{ij}\prod_{l_x}(X_{l_x}^{ij})^{A_{l_x}^{ij}/V_{l_x}g^{ij}}\prod_{c_x\ldots c_y}(X_{c_x\ldots c_y}^{ij})^{A_{c_x\ldots c_y}^{ij}/W_{c_x\ldots c_y}g^{ij}} \tag{3.22}$$

where,

$$X_0^{ij}=\prod_{l_x}\left(\frac{V_{l_x}}{S_{l_x}}\right)^{A_{l_x}^{ij}/V_{l_x}g^{ij}}\prod_{c_x\ldots c_y}\left(\frac{W_{c_x\ldots c_y}}{S_{c_x\ldots c_y}}\right)^{A_{c_x\ldots c_y}^{ij}/W_{c_x\ldots c_y}g^{ij}}$$

$$X_{l_x}^{ij}=e^{-(1+\lambda_{l_x}V_{l_x})}$$

$$X_{c_x\ldots c_y}^{ij}=e^{-(1+\lambda_{c_x\ldots c_y}W_{c_x\ldots c_y})}$$

$$g^{ij} = \sum_{l_x} \frac{A_{l_x}^{ij}}{V_{l_x}} + \sum_{c_x \ldots c_y} \frac{A_{c_x \ldots c_y}^{ij}}{W_{c_x \ldots c_y}}$$

In addition, by partial differentiation over λ_{l_x} and $\lambda_{c_x \ldots c_y}$, we have

$$\frac{\partial \varphi(T^{ij}, \lambda_{l_x}, \lambda_{c_x \ldots c_y})}{\partial \lambda_{l_x}} = \sum_{ij} T^{ij} A_{l_x}^{ij} - V_{l_x} = 0 \tag{3.23}$$

$$\frac{\partial \varphi(T^{ij}, \lambda_{l_x}, \lambda_{c_x \ldots c_y})}{\partial \lambda_{c_x \ldots c_y}} = \sum_{ij} T^{ij} A_{c_x \ldots c_y}^{ij} - W_{c_x \ldots c_y} = 0 \tag{3.24}$$

3.6 Genetic Algorithm to Find OD Demand

Combining Equations (3.22), (3.23) and (3.24), the OD matrix T^{ij} cannot be solved analytically as indicated in Section 3.2. Additionally, there are several issues during the OD estimation.

First, the underspecification problem that the number of equations is less than the number of variables leads to multiple solutions for the OD matrix.

Second, either measurement errors or assignment ($A_{l_x}^{ij}$ and $A_{c_x \ldots c_y}^{ij}$) errors may exist, which violate Equations (3.23) and (3.24). Involving errors in the observations means bringing uncertainty to the measurements. The more different devices, the less consistent the observation data.

Third, there may exist inconsistencies of different data sources and inconsistencies of in-flows and out-flows. The inconsistency of in-flows and out-flows of a node in the road network can be eliminated using fuzzy reasoning (Kikuchi et al., 2000) or a likelihood estimator (Van Zuylen and Branston, 1982). The first approach assumes that each observation is a member of a fuzzy (uncertain) set. The volume that is consistent and fits as well as possible to the observed data (maximum membership to the set) is used (Kikuchi et al., 2000). The other approach assumes that traffic counts have a certain probability distribution and estimates the consistent traffic volumes as the most likely volumes (Van Zuylen and Branston, 1982).

In order to deal with the three issues mentioned and to have at least one solution from the feasible solution space, the deviations between the estimated flow and the observed flow are minimized as illustrated in the following equations. Substituting the expression of T^{ij} into the minimization functions, the Lagrangian multipliers of two data sources, loops and cameras, can be obtained.

$$\min_{\lambda_{l_x}} \sum_{l_x} (V_{l_x} - \sum_{ij} A_{l_x}^{ij} \cdot T^{ij}(\lambda_{l_x}, \lambda_{c_x \ldots c_y}))^2 \tag{3.25}$$

$$\min_{\lambda_{c_x \ldots c_y}} \sum_{c_x \ldots c_y} (W_{c_x \ldots c_y} - \sum_{ij} A_{c_x \ldots c_y}^{ij} \cdot T^{ij}(\lambda_{l_x}, \lambda_{c_x \ldots c_y}))^2 \tag{3.26}$$

Since the expression of $T^{ij}(\lambda_{l_x}, \lambda_{c_x \ldots c_y})$ in Equation (3.22) is in the form of a posynomial (Duffin et al., 1967), it is hard to solve the optimization problems by non-linear programming.

A heuristic approach is an option. There are three main characteristics of heuristic approaches: easy to learn, easy to implement, and generally efficient. But they do not guarantee finding an optimal solution. Heuristic approaches can be applied to many general problems, because they do not rely on rigorous mathematical characteristics. Among the heuristic approaches, such as simulated annealing, tabu search, genetic algorithm and ant colonies, we chose one of the general methods to solve Equation (3.25) and (3.26), which is a Genetic Algorithm (GA).

A Genetic Algorithm is a global search heuristic, a particular class of evolutionary algorithm that uses techniques inspired by evolutional biology such as inheritance, mutation, selection and crossover. The general procedure starts with a set of solutions called population. Then GA evaluates the fitness of each individual in the population. All such things can be handled as weighted components of the objective function, making it easy to adapt GA to the particular requirements of a wide range of possible objectives. However, GA also brings criticisms. First, it is hard for GA to handle repeated fitness function evaluations for complex problems (Mitchell et al., 1997), which is often the most prohibitive and limiting issue of artificial evolutionary algorithms. Finding the optimal solution to complex high dimensional problems often requires very expensive fitness function evaluations. In some real world problems, a single function evaluation may require several hours to several days to complete the simulation. Second, the stop criterion in GA is not clear. In many problems, GA may have a tendency to converge towards local optima or even arbitrary points rather than the global optimum of the problem (Mitchell et al., 1997).

3.7 Case Study of the A15 Motorway

A case study numerically demonstrates the use of multiple data sources for estimating demand based on the Kullback-Leibler divergence together with a Genetic Algorithm. To show the added value of path flows, six scenarios of different combinations of detectors are designed. A test of the estimated demand being sensitive to the prior information involved in the Kullback-Leibler divergence method is carried out.

3.7.1 General Settings

This case study considers part of the A15 motorway in the Netherlands, which connects the Port of Rotterdam to Germany. Statistics Netherlands has freight truck demand data for this area aggregated at the level of three-digital postcodes around the A15 motorway. The areas with the three-digital postcodes are illustrated in Figure 3.1 using the software MapPoint.

Part of the A15 motorway, from Hoogvliet (east) to Havens (west) within the three-digital zones 307 and 308, the same section as the test bed in Ma et al. (2010), is selected for further study. This area is quite nearby the port and has a stable geography, which is suitable to be

Figure 3.1: Three-Digit Post Code Areas around the A15 Motorway

used as a test bed. There are seven highway sections, four on-ramps (in1, in2, in3 and in6) and four off-ramps (out4, out5, out7 and out8), illustrated in Figure 3.2.

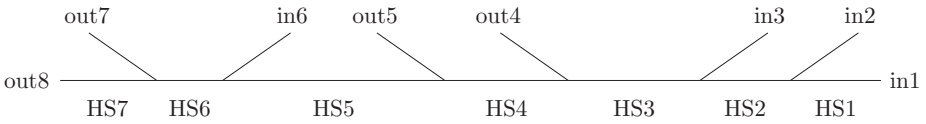


Figure 3.2: Part of A15 motorway from Hoogvliet to Havens

A prior OD matrix is derived from the demand data for zones 307 and 308 in Figure 3.1 from Statistics Netherlands. The data refer to the morning peak of March 11th, 2008. Each three-digit postcode area includes several on-ramps and off-ramps as origins and destinations. The prior OD data have been obtained by splitting the demand equally to the on-ramps and off-ramps within each area. This leads to an approximate prior OD matrix as shown in Table 3.1.

Table 3.1: Prior OD Matrix from Statistics Netherlands

	out4	out5	out7	out8
in1	2014	1869	1696	1236
in2	1989	1877	1699	1236
in3	1200	1862	1588	1236
in6	-	-	1696	1236

In addition, we take an OD matrix on this motorway from previous work of Ma et al. (2010) as the ground truth matrix shown in Table 3.2. This ground truth matrix is used to evaluate the estimated demand.

The link flows from loop detectors and path flows from cameras are generated based on the ground truth demand, taking the error terms in Equation (2.3) and Equation (2.4) into account. These error terms are assumed to follow normal distributions with zero mean. The path flows are assumed to be more accurate than the link flows, so the variance of the link flows is set to be 50 and the variance of the path flows is set to be 1.

Table 3.2: Ground Truth OD Matrix

	out4	out5	out7	out8
in1	1709	1527	1401	1569
in2	2295	2569	2105	2357
in3	3549	3275	3638	3163
in6	-	-	1576	1362

3.7.2 Estimation Accuracy among Six Scenarios with Kullback-Leibler Divergence Method

The proposed Kullback-Leibler divergence method is applied together with the settings of GA. To consider the influence of the different combinations of detectors, six scenarios are designed with three randomly selected locations. The detectors are added up step by step.

1. only loops on highway sections 3, 4 and 6;
2. full coverage of loops;
3. only cameras on highway sections 3, 4 and 6;
4. full coverage of cameras;
5. full coverage of loops plus cameras on highway sections 3, 4 and 6;
6. full coverage of both cameras and loops.

Table 3.3 gives an overview of the relationship among six scenarios. The horizontal axis represents the number of loop devices from non loops to the full coverage of loops. The vertical axis is the number of cameras. Scenarios 2, 5 and 6 take the full coverage of loops as a base to add on cameras. There is no such scenario between scenario 4 and scenario 6 that has full coverage of cameras plus three loops. Full coverage of cameras can identify demand properly. Whatever the number of loops is added up may not have a big influence.

Table 3.3: Six Scenarios

Cameras				
Full	sce 4	-	sce 6	
HS3,4,6	sce 3	-	sce 5	
Non	-	sce 1	sce 2	
	Non	HS3,4,6	Full	Loops

To evaluate the demand estimation accuracy among the six scenarios, the average deviation between the ground truth demand and the estimated demand in different combinations of detectors is taken as the criterion. The results are demonstrated in Table 3.4. The second column in the table is the prior demand from Statistics Netherlands; the third column is the ground

truth demand as a benchmark; the column with |%| is the absolute ratio between the estimated demand and the ground truth demand. The last row gives the average relative differences over the 14 OD combinations. Table 3.4 illustrates that the scenario with only three loops on the highway has the largest average deviation, 40%, between the ground truth demand and the estimated demand. When more loops are added to cover the whole network, the average deviation reduces to 15.44%. If cameras are installed on the highway sections 3, 4 and 6, together with a full coverage of loops, the average deviation slightly decreases to 12.44%. The reason of the insignificantly decreased deviations from scenario 2 to scenario 5 could be that the limited number of cameras, offering the path flow, cannot cover the rest of the freedom from the link flow data. Furthermore, in the scenario with full coverage of cameras and loops, the average deviation is as low as expected, only 0.32%.

Due to the fact that cameras measure the path flows, allocating three cameras on highway sections 3, 4 and 6 gives better demand information, leading to an average deviation of 35.21%, than 40.00% from the situation with three loops. The scenario with full coverage of cameras has the lowest deviation, only 0.04%, which is even lower than that from the scenario with the full coverage of both cameras and loops. It can be argued from the fact that the flow data from loop detectors has a much higher variance which stretches the estimated result in the wrong direction.

Table 3.4: Average Deviations between Ground Truth Demand and Estimated Demand in Different Combinations of Detectors

OD	Prior Demand	Ground Truth	Loops 346	|%|	Full Loops	|%|	Cameras 346	|%|	Full Cameras	|%|	Full Loops and Cameras 346	|%|	Full Loops and Full Cameras	|%|
1 – 4	2014	1709	2923.47	71.06	1337.31	21.75	2923.65	71.07	1708.99	0.05	1465.39	14.25	1707.99	0.17
1 – 5	1869	1527	2867.31	87.77	1443.73	5.45	2456.55	60.87	1527.05	0.05	1426.09	6.61	1524.47	0.34
1 – 7	1696	1401	2537.07	81.09	1836.00	31.05	2777.50	98.25	1401.06	0.06	1749.35	24.86	1410.59	1.19
1 – 8	1236	1569	1848.95	17.84	1589.85	1.33	2024.17	29.01	1569.07	0.06	1565.36	0.23	1563.23	1.13
2 – 4	1989	2295	2887.18	25.80	2362.98	2.96	2887.36	25.81	2295.10	0.04	2523.02	9.94	2295.80	0.10
2 – 5	1877	2569	2879.59	12.09	2247.38	12.52	2467.07	3.97	2568.97	0.03	2198.90	14.41	2570.23	0.11
2 – 7	1699	2105	2541.56	20.74	2559.50	21.59	2782.41	32.18	2105.00	0.04	2452.45	16.51	2101.76	0.28
2 – 8	1236	2357	1848.95	21.55	2154.94	8.57	2024.17	14.12	2357.02	0.03	2150.07	8.78	2357.93	0.28
3 – 4	1200	3549	1741.89	50.92	3852.42	8.55	1741.99	50.92	3549.04	0.02	3564.77	0.44	3549.40	0.06
3 – 5	1862	3275	2856.57	12.78	3679.15	12.34	2447.35	25.27	3274.91	0.02	3746.14	14.39	3275.44	0.08
3 – 7	1588	3638	2375.52	34.70	3254.83	10.53	2600.63	28.51	3638.01	0.02	3306.04	9.12	3636.19	0.09
3 – 8	1236	3163	1848.95	41.54	2838.57	10.26	2024.17	36.00	3162.95	0.03	3009.54	4.85	3164.03	0.12
6 – 7	1696	1576	2412.22	53.06	1069.79	32.12	1699.53	7.84	1575.98	0.05	1212.20	23.08	1574.73	0.21
6 – 8	1236	1362	1757.96	29.07	1867.67	37.13	1238.57	9.06	1361.94	0.06	1725.78	26.71	1363.72	0.34
Average Deviation				40.00		15.44		35.21		0.04		12.44		0.32

Table 3.5: Average Deviations with Unit Prior OD Demand

OD	Prior Demand	Ground Truth	Loops 346	|%|	Full Loops	|%|	Cameras 346	|%|	Full Cameras	|%|	Full Loops and Cameras 346	|%|	Full Loops and Full Cameras	|%|
1 – 4	1	1709	2517.67	47.32	1163.04	31.95	2517.70	47.32	1709.01	0.05	1261.20	26.20	0.13	0.13
1 – 5	1	1527	2629.46	72.20	1465.86	4.00	2456.96	60.90	1527.03	0.05	1471.26	3.65	0.33	0.33
1 – 7	1	1401	2285.93	63.16	1789.60	27.74	2372.15	69.32	1400.95	0.06	1733.76	23.75	0.91	0.91
1 – 8	1	1569	2285.93	45.69	1787.54	13.93	2372.15	51.19	1569.05	0.05	1739.77	10.88	0.87	0.87
2 – 4	1	2295	2517.67	9.70	2126.65	7.34	2517.70	9.70	2294.99	0.04	2228.19	2.91	0.07	0.07
2 – 5	1	2569	2629.46	2.35	2367.70	7.84	2456.96	4.36	2568.99	0.03	2343.11	8.79	0.09	0.09
2 – 7	1	2105	2285.93	8.60	2483.30	17.97	2372.15	12.69	2104.95	0.03	2430.00	15.44	0.23	0.23
2 – 8	1	2357	2285.93	3.02	2348.18	0.37	2372.15	0.64	2356.98	0.04	2324.76	1.37	0.21	0.21
3 – 4	1	3549	2517.67	29.06	4263.42	20.13	2517.70	29.06	3548.89	0.02	4063.59	14.50	0.03	0.03
3 – 5	1	3275	2629.46	19.71	3537.32	8.01	2456.96	24.98	3274.89	0.02	3556.67	8.60	0.05	0.05
3 – 7	1	3638	2285.93	37.17	3071.70	15.57	2372.15	34.80	3638.06	0.02	3144.27	13.57	0.07	0.07
3 – 8	1	3163	2285.93	27.73	2752.65	12.97	2372.15	25.00	3163.03	0.03	2860.45	9.57	0.09	0.09
6 – 7	1	1576	1727.66	9.62	1375.44	12.73	1469.01	6.79	1576.05	0.05	1411.96	10.41	0.14	0.14
6 – 8	1	1362	1727.66	26.85	1562.58	14.73	1469.01	7.86	1362.02	0.05	1525.99	12.04	0.24	0.24
Average Deviation				28.73		13.95		27.47		0.04		11.55		0.25
Function Value				211.52		14.16		370.34		356.10		109.80		14552.00

3.7.3 Sensitivity to Prior Demand

In order to test the sensitivity of the average deviation between the estimated demand and the ground truth demand with respect to the prior demand (Antoniou et al., 2015), there are three artificially designed combinations of OD trips besides the prior demand from Statistics Netherlands. The first alternative prior is the unit prior OD demand in Table 3.6. The second alternative prior is with large differences among the prior OD trips in Table 3.7, where we gave some prior demand as 1, while others as 1000. The third one is with relatively small differences among the prior OD trips in Table 3.8, where we gave some prior demand as 100, while others as 1000. The variances of the link flows from loops and the path flows from cameras are assumed to be the same.

Table 3.6: Alternative Prior Demand 1

	out4	out5	out7	out8
in1	1	1	1	1
in2	1	1	1	1
in3	1	1	1	1
in6	-	-	1	1

Table 3.7: Alternative Prior Demand 2

	out4	out5	out7	out8
in1	1	1000	1	1000
in2	1	1000	1	1000
in3	1	1000	1	1000
in6	-	-	1	1000

Table 3.8: Alternative Prior Demand 3

	out4	out5	out7	out8
in1	100	1000	100	1000
in2	100	1000	100	1000
in3	100	1000	100	1000
in6	-	-	100	1000

Further, Table 3.5, Table 3.9 and Table 3.10 illustrate the average deviations under the three alternative prior demands. These tables show that adding detectors to the basis of loops (three loops, the full coverage of loops, the full coverage of loops and three cameras, and the full coverage of both loops and cameras), can reduce the average deviation. With the basis of cameras, the average deviation in the scenario of three cameras is not larger than in the scenario of the full coverage of both loops and cameras, except for the prior demand with large differences among the OD trips in Table 3.9. This could arise from the wrong pattern of the

prior demand. Even with many flow observations, this bad prior demand worsens the estimated results. Whatever the prior OD trips are, the scenario with the full coverage of cameras always offers very low average deviations, 0.04%, between the estimated OD trips and the ground truth. In addition, the function values from the Genetic Algorithm in Table 3.9 with large differences among the OD trips are quite large. It may take more generations to reach convergence.

The function values of the objective functions, Equations (3.25) and (3.26), in the tables represent the deviation between the flow observations and the assigned flow from the estimated demand. In the scenario of the full coverage of loops and cameras, the function value is largest among the scenarios. It may come from the fact that the larger the number of detectors, the more errors add up together.

In a nutshell, information methods take measures of the differences between a true distribution and a prior distribution. Thus, the quality of the prior distribution is essential. Unrepresentative prior demand results in bad estimated demand as demonstrated in this section.

Table 3.9: Average Deviations with Large Differences among the Prior Demand

OD	Prior Demand	Ground Truth	Loops 346	|%|	Full Loops	|%|	Cameras 346	|%|	Full Cameras	|%|	Full Loops and Cameras 346	|%|	Full Loops and Full Cameras	|%|
1 – 4	1	1709	2505.10	47.58	30.94	98.19	2517.68	47.32	1709.04	0.05	126.83	92.58	16.14	99.06
1 – 5	1000	1527	3302.08	116.25	2585.17	69.30	2457.02	60.90	1526.96	0.05	2283.58	49.55	2502.33	63.87
1 – 7	1	1401	3.90	99.72	59.57	95.75	4.74	99.66	1401.01	0.05	8.43	99.40	6.54	99.53
1 – 8	1000	1569	3901.62	148.67	3519.14	124.29	4739.61	202.08	1569.05	0.05	3784.20	141.19	2548.14	62.41
2 – 4	1	2295	2505.10	10.15	242.97	89.41	2517.68	9.70	2294.95	0.03	819.95	64.27	36.42	98.41
2 – 5	1000	2569	3302.08	28.54	4171.17	62.37	2457.02	4.36	2568.91	0.03	3391.58	32.02	4020.56	56.50
2 – 7	1	2105	3.90	99.81	169.57	91.94	4.74	99.77	2104.92	0.04	15.13	99.28	11.29	99.46
2 – 8	1000	2357	3901.62	65.53	4722.06	100.34	4739.61	101.09	2357.11	0.03	5099.13	116.34	3706.53	57.26
3 – 4	1	3549	2505.10	29.41	5680.61	101.21	2517.68	29.06	3549.04	0.02	6783.00	91.12	158.66	95.53
3 – 5	1000	3275	3302.08	0.83	2732.16	33.19	2457.02	24.98	3275.02	0.02	2226.95	32.00	4722.03	44.18
3 – 7	1	3638	3.90	99.89	448.74	88.52	4.74	99.87	3637.79	0.02	19.65	99.46	23.78	99.35
3 – 8	1000	3163	3901.62	23.35	4017.20	29.70	4739.61	49.85	3163.01	0.03	4422.66	39.82	4781.79	51.18
6 – 7	1	1576	5.45	99.65	1711.71	138.26	2.94	99.81	1576.12	0.05	86.78	99,96	70.43	95.53
6 – 8	1000	1362	5449.02	300.08	26.78	98.03	2935.08	115.50	1361.92	0.06	596.33	56,22	1715.11	25.93
Average Deviation				83.53		87.18		74.57		0.04		79.51		74.87
Function Value				142,651.40		34,450,374.12		17.28		7.02		47,139,007.75		811.36

Table 3.10: Average Deviations with Reduced Differences among the Prior Demand

OD	Prior Demand	Ground Truth	Loops 346	|%|	Full Loops	|%|	Cameras 346	|%|	Full Cameras	|%|	Full Loops and Cameras 346	|%|	Full Loops and Full Cameras	|%|
1 – 4	100	1709	2517.67	47.32	550.91	67.76	2517.64	47.32	1708.82	0.05	757.52	55.67	1714.57	0.56
1 – 5	1000	1527	3200.34	109.58	2057.83	34.76	2457.01	60.90	1527.06	0.05	1944.06	27.31	1534.64	0.98
1 – 7	100	1401	363.73	74.04	1118.39	20.17	431.30	69.21	1401.00	0.06	1167.28	16.68	1332.49	5.12
1 – 8	1000	1569	3637.25	131.82	2478.89	57.99	4313.01	174.89	1569.04	0.05	2337.22	48.96	1619.47	3.28
2 – 4	100	2295	2517.67	9.70	1495.00	34.86	2517.64	9.70	2294.99	0.03	1824.37	20.51	2292.75	0.22
2 – 5	1000	2569	3200.34	24.58	2883.30	12.23	2457.01	4.36	2569.00	0.03	2675.03	4.13	2566.53	0.24
2 – 7	100	2105	363.73	82.72	1920.96	8.74	431.30	79.51	2105.03	0.04	1946.85	7.51	2124.95	1.13
2 – 8	1000	2357	3637.25	54.32	3026.72	28.41	4313.01	82.99	2356.83	0.03	2879.64	22.17	2346.83	0.57
3 – 4	100	3549	2517.67	29.06	5507.10	55.17	2517.64	29.06	3548.99	0.02	4971.13	40.07	3546.82	0.12
3 – 5	1000	3275	3200.34	2.28	2430.28	25.79	2457.01	24.98	3274.91	0.02	2751.87	15.97	3272.25	0.17
3 – 7	100	3638	363.73	90.00	2833.41	22.12	431.30	88.14	3637.98	0.02	2945.19	19.04	3650.76	0.42
3 – 8	1000	3163	3637.25	14.99	2854.26	9.76	4313.01	36.36	3163.05	0.02	2956.81	6.52	3156.94	0.30
6 – 7	100	1576	469.82	70.19	2847.00	80.65	267.09	83.05	1576.08	0.05	2660.51	68.81	1588.50	0.83
6 – 8	1000	1362	4698.16	244.95	91.48	93.28	2670.90	96.10	1361.99	0.06	277.40	79.63	1341.15	1.56
Average Deviation				70.40		39.41		63.33		0.04		30.93		1.11
Function Value				18.83		13.97		16.16		7.42		14.17		37085.22

3.7.4 Summary of the Case Study

In summary, this case study first shows that path flow data from cameras play a significant role in estimating the freight truck OD matrix. The situation with full coverage of cameras has the lowest average deviation between the ground truth demand and the estimated demand, as 0.04%. Although the final scenario with the full coverage of both cameras and loops exceeds the scenario with full coverage of cameras as well, the average deviation is not as low as the scenario with only full cameras. It is because loops introduce extra errors. Second, the influence of the prior OD matrix in the Kullback-Leibler divergence method to the estimated OD matrix is significant. Information methods take measures of the differences between a true distribution and a prior distribution. Bad prior demand has a rather strong impact on the estimation. Even the very accurate observations in the road network may not entirely lead to a low average deviation between the ground truth demand and the estimated demand. Third, a Genetic Algorithm is one option of a heuristic method to estimate freight truck demand, but it is time consuming, around 20 minutes for each scenario with a laptop of 4GB RAM and 32-bit Operating System.

3.8 Numerical Comparison between Information Minimization and Kullback-Leibler Divergence

The Kullback-Leibler divergence and the information minimization method differ with respect to the weighing of flows. In addition, the Stirling's approximation applied in the information minimization is making the assumption that the flow is large. It means that this method may not be suitable for the situation that the flow is small, during the night for instance. In order to test the proper use of the Stirling's approximation and further to compare the information minimization method and the Kullback-Leibler divergence, a test with small flows is designed. The A15 motorway is applied as the road network. We assume that the ground truth demand is 5 for all OD pairs. With this demand, the link flows and the path flows are generated with variances of 0.5 and 0.01, respectively. The prior demand is designed as 1 for each OD pair.

With these settings, the average deviations of the estimated demand and the ground truth demand based on both methods are shown in Table 3.11. It demonstrates that the Kullback-Leibler divergence offers smaller deviations than the information minimization for all the situations except the full-cameras situation where they are equal. Therefore, the Kullback-Leibler divergence method is able to better deal with the situation of small demand than the information minimization method.

Table 3.11: Average Deviations of Estimated Demand and Ground Truth from Information Method and Kullback-Leibler Divergence Method (%)

	Loops 346	Full Loops	Cameras 346	Full Cameras	Loop and Cameras 346	Full Loop and Full Cameras
Information Method	3.61	2.15	1.35	0.66	4.00	1.44
Kullback-Leibler divergence	2.29	1.16	0.14	0.66	1.46	1.39

3.9 Conclusion

In this chapter, we reexamine the information minimization method from Van Zuylen and Willumsen (1980) for OD estimation, and introduce the Kullback-Leibler divergence method on top of the information minimization method. Theoretically, these two methods have strong connections: information is mathematically the negative value of a product between a scale of flow counts and the Kullback-Leibler divergence. The expressions of the estimated demand from both methods can reach similar results if the flow data is large. For the situation where the flow data is small, the test demonstrates that the Kullback-Leibler divergence method has lower average deviations between the estimated demand and the ground truth demand. It can be argued from the fact that Stirling's approximation applied in the information minimization method has the underlying assumption that the flow should be large. Therefore, the Kullback-Leibler divergence method can be treated as a generalized approach of the information minimization method, although the Kullback-Leibler divergence method still belongs to the point estimation methods discussed in Chapter 2, where a single value serves as a best estimate of an unknown population parameter, and has no error involved.

In addition, comparing with previous research illustrated in Table 2.2, three types of data are integrated in the Kullback-Leibler divergence method, link flow from loop detectors, path flow from cameras, and prior demand from Statistics Netherlands. The effectiveness of the combination of these data sources has been demonstrated. The A15 case study shows that the more cameras are installed, the fewer uncertainties are involved, and the more accurate the estimated OD matrix is. Additionally, since information methods take measures of the differences between a true distribution and a prior distribution, the estimated OD matrix is sensitive to the prior demand. But if there are accurate observed flow data in the road network, like the path flow data from cameras, the impact of the prior OD matrix is small. The average deviation between the estimated demand and the ground truth demand is always 0.04% in the A15 case study, whatever the prior OD matrix is. Thus, the high quality of prior demand results in good estimated demand.

Chapter 4

Hierarchical Bayesian Networks for Freight Truck OD Estimation

[1]Ma,Y., Kuik, R. and Zuylen, H.J. van. Freight Origin Destination Estimation based on Multiple Data Sources (2012). IEEE on Intelligent Transportation Systems, USA.

4.1 Introduction

Road traffic has stochastic characteristics. This is also the case for freight traffic. Both the demand and the flows of freight vehicles in a network during a certain time period are non-deterministic. In order to properly represent the truck demand and flows, stochastic methods need to be applied. Bayesian inference is one approach used to account for the stochastic nature of phenomena. In contrast with traditional estimation methods, parameters in Bayesian inference are treated as random variables, instead of as single deterministic realizations. Typically, Bayesian inference updates the probability density function of prior belief based on the likelihood of the evidence to obtain posterior belief (Greenberg, 2008). This method takes the posterior distribution of the unknown parameters, conditioned on the observed data.

In this chapter, a hierarchical Bayesian network (Castillo et al., 2008a), a particular Bayesian inference approach, is used to estimate a freight truck OD matrix. Castillo et al. (2008a) develop the framework of Bayesian networks to estimate demand, assuming normal distributions both for the prior demand and for the flows, given demand. They assume a single type of data source: loop detectors. Moreover, they assume flow data without measurement errors. In this thesis, models with both normal distributions and log-normal distributions are proposed. The closed form of the demand estimation from normal distributions leads to a fast calculation, while log-normal distributions require a simulation approach to get the estimated demand. In addition, compared with the information minimization method in Chapter 3, hierarchical Bayesian networks allow errors being taken into account. Multiple data sources are investigated to improve the accuracy of the estimation. Examples of data sources are link flow data from loop detectors and path flow data from cameras.

The sensor location problem has received more and more attention over the last two decades. Its main objective is to determine optimal sensor locations in a transportation network to estimate OD demand, such as Bianco et al. (2001), Gentili and Mirchandani (2005), Castillo et al. (2008b), Hu et al. (2009), Larsson et al. (2010), Zhou and List (2010), and Fei and Mahmassani (2011). Among them, Bianco et al. (2001), Gentili and Mirchandani (2005), Hu et al. (2009), and Larsson et al. (2010) make the assumptions of a static traffic assignment and no errors involved. However, Castillo et al. (2008b), Zhou and List (2010), and Fei and Mahmassani (2011) take errors such as measurement errors, estimation errors and simulation errors into account. Especially, Zhou and List (2010) applied a least mean square OD estimator, analysed the covariance matrix of estimated demand, and gave advice about sensor locations. Furthermore, a few studies have been conducted to locate cameras or path-flow detectors for sensor location problems. Zhou and List (2010) focused on locating a limited number of point-flow detectors and path-flow detectors in a network to update an estimated OD trip table. Viti and Corman (2012) describe several rules for sensor locations, such as the OD-coverage rule guaranteeing that all OD pairs are observed at least for a small portion, the maximum flow fraction rule where the portion of flow measured belonging to that OD pair with respect to all other OD pairs

measured by that sensor is maximized, the maximum flow intercepting rule, and the maximal net OD flow captured rule. Combining several elements mentioned in the literature, this chapter intends to address sensor coverage in a framework of hierarchical Bayesian networks, aiming to demonstrate the randomness reduction in demand estimation and static traffic allocation after adding multiple data source detectors, such as loops and cameras. Prior error, measurement error and estimation error are considered during the analysis. Due to the assumption of the known static route proportion, simulation error is not applied in the thesis.

The research questions are whether hierarchical Bayesian networks in the respective situations of normal distributions and log-normal distributions are able to achieve estimation of freight truck demand, whether there is any connection between the models based on these two distributions, how multiple data sources can be combined to estimate freight truck demand, and how the sensor coverage influences the demand estimation.

4.2 Literature Review

There are two main approaches to model the OD matrix estimation: point estimation and distribution estimation. In statistics, point estimation involves the use of sample data to infer about a single value that serves as a best guess or a best estimate of an unknown, fixed population parameter. Distribution estimation aims to predict the random variables. Point estimation can be done with maximum likelihood, least squares and maximization of entropy or minimization of information methods. Distribution estimation is done with Bayesian inference and Kalman filtering.

4.2.1 Point Estimation Methods

Point estimation is used to determine a single value which serves as a best estimate of an unknown population parameter. Maximum likelihood and generalized least squares are the two common methods to arrive at point estimates. Information and entropy in information theory (Brillouin, 1956; Cover and Thomas, 2006) are measures of the average uncertainty in a random variable (Cover and Thomas, 2006). The optimum value is obtained by minimizing the uncertainty.

Maximum Likelihood

The maximum likelihood determines values of the model parameters that have the highest likelihood to give the observed data. This method was applied mainly in early studies of OD matrix estimation by Spiess (1987), Cascetta and Nguyen (1988), Nihan and Davis (1989), Watling and Maher (1992), and Watling (1994).

Spiess (1987) applies maximum likelihood in a convex programming problem, in which the elements of an OD matrix are assumed to be outcomes of flow observations with a Poisson

distribution with an unknown mean. He ignores the connectivity and topology of the network and defines his likelihood function solely in terms of a set of independent observations on each OD pair (Hazelton, 2000). Nihan and Davis (1989) estimate an intersection OD matrix by minimizing the error between observed and predicted exiting counts. Watling (1994) applies maximum likelihood to a partial registration plate survey, in which all possible combinations of the observed data are considered. At that time, since the maximum likelihood cannot be obtained analytically, alternative numerical techniques were applied.

Generalized Least Squares

The least squares minimize the sum of squared deviations between a prior OD matrix and an estimated OD matrix, based on the observed flows. It became popular in the eighties and the beginning of the nineties, and has been applied by many researchers: Cascetta (1984), Carey and Revelli (1986), Cascetta and Nguyen (1988), Bell (1991), Bierlaire and Toint (1995), Yang (1995). Estimators based on least squares have the advantage of being relatively easy to solve mathematically. Especially for larger problems, such as the simultaneous estimation of OD matrices for several time steps in large networks, this methodology gives a feasible solution. Additionally, the least squares method gives the same results as maximum likelihood, if normal distributions of the OD demand variables are assumed.

Information or Entropy based Method

Van Zuylen and Willumsen (1980) and Van Zuylen (1981) gave an approach to generate the most likely OD matrix based on maximization of the entropy of the trip matrix or minimization of the information with respect to a prior OD matrix. Assuming that there are no errors in the observations of the link flows and independence among the flows in the highway sections, they formulate the equality of the assigned and the observed traffic flows on the links of the network. Willumsen (1984) extends the entropy approach with a scaling factor of the flow observations to estimate an OD matrix.

4.2.2 Distribution Estimation Methods

Distribution estimation methods take the stochastic nature of freight demand into account. The parameters of the distributions represent the features of sample data. A Bayesian inference method is a typical distribution-based approach. In general, Bayesian inference methods can deal with all kinds of distributions.

Bayesian inference

Bayesian inference updates a prior OD distribution based on the flow observations to generate a posterior OD matrix. This approach reduces the overall uncertainty of the estimates by producing posterior distributions for the parameters as well as predictive distributions for future

OD flows (Perrakis et al., 2011). The approach began gaining popularity in the middle of the nineties, due to a paper by Tebaldi and West (1998), although Maher (1983) had already introduced this method to estimate OD matrices. Later, many researchers contributed to this method, such as Hazelton (2000), Li (2005), Sun et al. (2006), Castillo et al. (2008a), Hazelton (2010) and Perrakis et al. (2011).

Maher (1983) assumes a multivariate normal distribution of a prior OD matrix and considers normally distributed errors in observations. This makes the computations very fast. Li (2005) applies a Bayesian method to deal with the under-specification problem. He states that a Bayesian analysis provides a research framework by specifying prior demand that amounts to introducing extra information based on accumulated knowledge. He uses an expectation maximization algorithm to overcome the problem of an analytically intractable likelihood. In addition, Bayesian Networks are applied by Castillo et al. (2008a) to represent the relation between OD demand variables, where linear relations of each layer in Bayesian Networks are built up. In order to deal with empirical distributions, Tebaldi and West (1998), Hazelton (2010) and Perrakis et al. (2011) propose a Markov Chain Monte Carlo (MCMC) simulation in Bayesian inference methods to estimate the OD matrix.

Kalman Filtering

Kalman filtering is widely used to adapt model parameters in a rolling horizon to the measured characteristics of the modeled reality. This method usually considers a state space and induces observable values. The relationships for the dynamics of the states and how the states induce observations may include errors. These errors are usually assumed to have normal distributions making computations easy and efficient. Chang and Wu (1994), Ashok and Ben-Akiva (2000), Dixon and Rilett (2002), Zhou and Mahmassani (2007), and Barceló et al. (2010) use Kalman filtering to estimate and predict dynamic OD matrices.

Zhou and Mahmassani (2007) present a structural state-space model to systematically incorporate regular demand pattern information, structural deviations and random fluctuations. By considering demand deviations from the prior estimate of the regular pattern as a time-varying process with a smooth trend, a polynomial trend filter is developed to capture possible structural deviations in the real-time demand. An optimal adaptive procedure is proposed based on a Kalman filtering framework, to capture day-to-day demand evolution, and to update the prior demand pattern estimates using new real-time estimates and observations obtained every day.

4.2.3 Discussion

Compared with point estimation methods, Bayesian inference treats parameters as random variables, instead of single realizations. Bayesian inference considers prior information and determines the posterior distribution of the parameters, conditioned on the observed data. The

core characteristic of Bayesian inference is the updating of prior beliefs to obtain posterior belief (Greenberg, 2008).

Maximum likelihood, least squares, Bayesian inference and Kalman filtering show great similarity in their basic mechanisms if normal distributions are assumed for the errors. Kalman filtering and Bayesian methods are popular for OD estimation and prediction due to the recursive method of updating estimates based on new (recent) data. Kalman filtering, where normality dominates distributions, is a special case of recursive Bayesian estimation (Koopman et al., 2012).

4.3 Methodological Framework of Hierarchical Bayesian Networks

Bayesian inference starts with modeling a stochastic process by means of a prior distribution of parameters $\Theta \sim f(\Theta)$, where $f(\cdot)$ is a probability density function. A sampling distribution of the data is involved given parameters: $y \sim f(y|\theta)$, where $f(\cdot|\theta)$ is a probability density function of y given outcomes θ of Θ. From the prior distribution and the sampling distribution, the joint distribution of parameters and data is $f(\theta, y) = f(\theta)f(y|\theta)$. The posterior distribution is obtained as follows:

$$f(\theta|y) = \frac{f(\theta, y)}{f(y)} = \frac{f(y|\theta)f(\theta)}{f(y)} \propto f(\theta)f(y|\theta) \tag{4.1}$$

where $f(y) = \int_\theta f(\theta, y)d\theta$ is the marginal distribution of the data, and $f(y|\theta)/f(y)$ is called the likelihood ratio. Since $f(y)$ does not depend on the parameter θ, the posterior distribution is proportional to the joint distribution.

For instance, supposing that θ is the parameter of transport demand which needs to be estimated and y is the flow observation data, the probability of transport demand given observed flows, $f(Demand|Flow)$, follows Equation (4.1) as

$$\begin{aligned} f(Demand|Flow) &= \frac{f(Demand, Flow)}{f(Flow)} \\ &= \frac{f(Flow|Demand)f(Demand)}{f(Flow)} \\ &\propto f(Demand)f(Flow|Demand). \end{aligned}$$

Hierarchical Bayesian networks belong to the field of Bayesian inference. They make use of a network structure to represent the relationships between variables. Hierarchical Bayesian networks (HBNs), also called belief networks, are probabilistic models that represent a set of stochastic variables and their conditional dependencies via a directed acyclic graph (Greenberg, 2008). HBNs consist of a set of nodes, each of which represents a variable, and a set of directed arcs connecting the nodes. A node where a directed arc starts is called a parent, and a node

where a directed arc ends is called a child. A conditional distribution can be represented by the variable of each child node given its parents. Hierarchical Bayesian networks are acyclic graphs. The simplest Bayesian Network is a graph with only two nodes and one arc, where θ is a parameter and y is data, see Figure 4.1.

Figure 4.1: Minimal Bayesian Network

In this section, hierarchical Bayesian networks are applied to estimate freight truck demand, taking as inputs a prior distribution of the OD matrix and a sampling distribution that specifies how flow data arise given a particular OD matrix.

In the following, we explain how to determine the posterior OD matrix. There are two proposed approaches to obtain the posterior density. One is to assume that all the density functions have normal distributions, as Castillo et al. (2008a) did. Johnson and Wichern (2002) argue that for the sampling distributions, once the magnitudes of the flows are large enough, the density functions of the sampling distributions can be assumed to follow normal distributions. Assuming normal distributions, an analytical approach can be used. However, if the symmetric shape of the normal distribution under-represents the probability of large flows, a log-normal distribution may be applied. As an analytical approach is not feasible in this case, Markov Chain Monte Carlo simulation is applied to estimate the unknown parameters of OD demand.

In hierarchical Bayesian networks, a prior distribution on an OD of freight trucks is updated taking freight truck flow observations as evidence. The prior beliefs of the available historical demand data are updated with the likelihood ratio of the conditional density function of observed flows given freight truck demand and the density function of observed flows. We extend the method of Castillo et al. (2008a), by considering three kinds of errors: prior error, estimate error and measurement error; and also by applying multiple data sources to increase the estimation accuracy of the OD matrix. Three steps are involved: 1) specification of the hierarchical Bayesian networks; 2) derivation of the joint distribution in the hierarchical Bayesian networks; 3) derivation of the posterior distribution in the hierarchical Bayesian networks.

Step 1: Specification of the Hierarchical Bayesian Networks

The elements to construct a hierarchical Bayesian network (Rossi et al., 2005) are the prior freight truck OD parameters, the freight truck OD matrix, and the observed freight flow. These are represented by three layers in Figure 4.2. The available information consists of the historical OD information, which could be the monthly demand as a scalar or a historical OD matrix, and the observed freight truck flows in the road network.

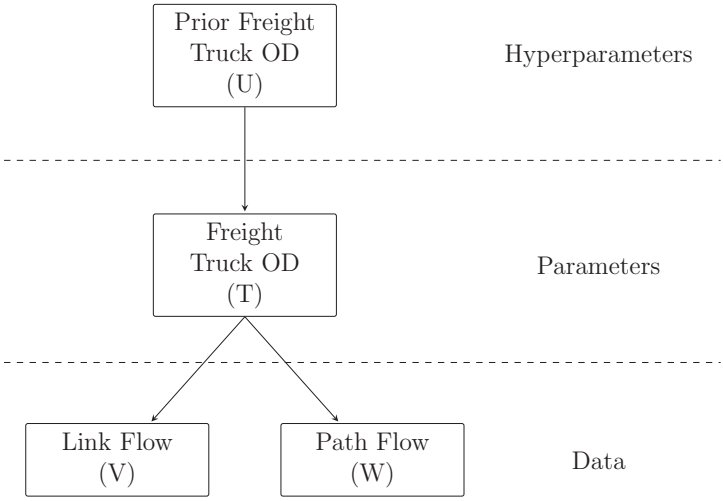


Figure 4.2: Bayesian Networks for Freight OD Estimation

The hyperparameters consist of the stochastic prior OD distribution parameter U. There are two different situations in which to consider this prior OD information, since the prior information comes from different sources in the format of either a matrix or a scalar. One situation is that the prior OD variables between origin i and destination j pairs, denoted as U^{ij}, with the same dimension as the number of origins and destinations as the freight truck demand, T^{ij}. The other situation is that the prior OD data is the total demand of freight trucks represented by a scalar, such as a monthly freight truck demand from a statistics office, and denoting the total prior OD matrix as U. In order to separate the total prior OD volume into the matrix U^{ij} with the same dimension as the estimated OD matrix T^{ij}, a constant weight of the prior OD parameter x^{ij} is introduced to represent the proportion of OD pair ij:

$$U^{ij} = x^{ij}U \tag{4.2}$$

$$\sum_{ij} x^{ij} = 1$$

U is a random variable following a normal distribution $U \sim N(\mu, \Sigma_U)$. In other words, μ is the mean of U, with the relation of $U = \mu + \varepsilon_U$. ε_U is the prior error which follows a normal distribution $\varepsilon_U \sim N(0, \Sigma_U)$. In this way, U^{ij} is also a random variable with mean equal to x^{ij} times the expected value of U, $U^{ij} \sim N(x^{ij}\mu, x^{ij}\Sigma_U)$.

The middle parameter layer identifies the freight truck OD matrix, which is an unobserved hidden variable. A linear relation is assumed between OD demand T^{ij} and prior demand U^{ij}:

$$T^{ij} = U^{ij} + E^{ij} = x^{ij}U + E^{ij} \tag{4.3}$$

where the estimation error $E^{ij} \sim N(0, \Sigma_T^{ij})$. Note that the estimation error E^{ij} indicates the freedom for U^{ij}, although U^{ij} in Equation (4.2) is perfectly correlated with U^{kl}. Equation (4.3) generalizes the prior information in Castillo et al. (2008a), where the total mean flow is taken as the prior knowledge. It also brings stochasticity to the prior OD information in Van Zuylen and Willumsen (1980).

The bottom data layer in Figure 4.2 consists of data about the flows of freight trucks, which are observed by different detectors. These detectors, such as loops, cameras and Bluetooth scanners, generate two types of information: link-based truck flows, V, and path-based truck flows, W, as mentioned in Section 2.5. The linear relation, as below, between the observed flows and the freight truck demand is given by the route proportions, A, taking measurement errors into account. The route proportions in the linear relation are assumed to be deterministic.

$$V_m = \Sigma_{ij} A_m^{ij} T^{ij} + E_m \tag{4.4}$$

Here m is the indicator of multiple devices. E_m is the measurement error with covariance D.

Note that the mean prior μ, the proportion of prior demand x^{ij}, all the variances, and the route proportion A are given.

Step 2: Derivation of the Joint Distribution in Hierarchical Bayesian Networks

A hierarchical model is built up through a sequence of two or more conditional distributions that specify the prior information (Rossi et al., 2005). Each arrow in Figure 4.2 represents a conditional distribution. U and T represent the parameters which are unobserved random variables, and V and W reflect the data. The joint distribution of the three layers in Figure 4.2 is:

$$f(U, T, V, W) = f(U) f(T|U) f(V|T) f(W|T) \tag{4.5}$$

where $V|T$ and $W|T$ are called the sampling distribution, i.e. the distribution of the observed data given an OD matrix.

Step 3: Derivation of the Posterior Distribution in Hierarchical Bayesian Networks

The posterior probability of the parameters of the freight truck OD matrix given observed data $f(T, U|V, W)$ is proportional to the joint distribution of the OD matrix and the observed data, which is the product of the density function of the data given certain parameters, $f(V|T)$ and $f(W|T)$, and the density function of these parameters:

$$f(T, U|V, W) = \prod_{l,c} \frac{f(V_l|T) f(W_c|T) f(T|U) f(U)}{f(V, W)} \tag{4.6}$$

where

$$f(V,W) = \int_U \int_T \prod_{l,c} f(V_l|T) f(W_c|T) f(T|U) f(U) dU dT$$

The link flow observations V and the path flow observations W are assumed to be conditionally independent given T. Note that the stochastic prior OD, U, as a hyperparameter is updated, together with the joint posterior of the freight truck OD matrix T, given all the flows in the road network as parameters:

$$f(T,U|V,W) = \prod_{l,c} \frac{f(V_l|T) f(W_c|T) f(T|U) f(U)}{f(V,W)} \tag{4.7}$$

4.4 Posterior Demand Estimation

In this section, two approaches to estimate the posterior demand distribution are presented. One is an analytical approach associated with the assumption of normal distributions, and the other is a simulation approach associated with a log-normal distribution of errors.

4.4.1 Analytical Approach of Posterior Estimation with Normal Distributions

To obtain the density function of $f(T,U|V,W)$, the right hand side of Equation (4.7) should be analyzed. If the density functions in the right hand side of Equation (4.7) are assumed to be normal, then the posterior density function on the left hand side of the equation follows a normal distribution as well. Mean and variance characterize a normal distribution.

Basic Multivariate Normal Distributions for Demand Estimation

From Koopman et al. (2012) and Rossi et al. (2005), suppose that estimated demand T and observed flow V are joint normally distributed random variables with the following expectation and covariance matrix:

$$\mathbb{E}\begin{pmatrix} T \\ V \end{pmatrix} = \begin{pmatrix} \bar{T} \\ \bar{V} \end{pmatrix}$$

$$COV\begin{pmatrix} T \\ V \end{pmatrix} = \begin{pmatrix} \Sigma_{TT} & \Sigma_{TV} \\ \Sigma_{TV}^{\top} & \Sigma_{VV} \end{pmatrix}$$

Here Σ_{VV} is assumed to be a nonsingular matrix. Shao et al. (2014) mentioned that most of the conventional OD demand estimation models mainly make use of the first-order statistical property (i.e. the mean) of the hourly traffic counts. The second-order property (say the covariance) of the count data is usually ignored. Here, we consider the both statistical properties.

Then, the conditional distribution of T given V is normally distributed with a mean vector

$$\mathbb{E}(T|V) = \bar{T} + \Sigma_{TV}\Sigma_{VV}^{-1}(V - \bar{V}), \tag{4.8}$$

and a variance-covariance matrix

$$COV(T|V) = \Sigma_{TT} - \Sigma_{TV}\Sigma_{VV}^{-1}\Sigma_{TV}^{\top}. \tag{4.9}$$

Denoting the covariance matrix of the prior demand as C and the covariance of the flow data as D that has been referred to in Equation (4.4) where D is the covariance of E_m, the covariance matrices of Σ_{TV} and Σ_{VV} can be expressed as follows:

$$\Sigma_{TV} = COV(T, V) = COV[T, (AT + E)] = COV(T, T)A^{\top} = CA^{\top}$$

$$\Sigma_{VV} = COV[(AT + E), (AT + E)] = A \cdot COV(T, T) \cdot A^{\top} + D = ACA^{\top} + D$$

Note that the relation $CA^{\top}(ACA^{\top} + D)^{-1} = (A^{\top}D^{-1}A + C^{-1})^{-1}A^{\top}D^{-1}$ helps to transfer the computations from the OD dimensions to the flow measurement dimensions based on the Sherman-Morrison formula (Sherman and Morrison, 1950). The proof is as follows.

Proof.

$$\begin{aligned}
&A^{\top}D^{-1}ACA^{\top} + A^{\top} = A^{\top}D^{-1}ACA^{\top} + A^{\top} \\
\Longleftrightarrow &(A^{\top}D^{-1}A + C^{-1})CA^{\top} = A^{\top}D^{-1}(ACA^{\top} + D) \\
\Longleftrightarrow &CA^{\top}(ACA^{\top} + D)^{-1} = (A^{\top}D^{-1}A + C^{-1})^{-1}A^{\top}D^{-1}
\end{aligned}$$

□

Combining Equations (4.7) and (4.8) the expectation and covariance of the trips conditional on the flow information are obtained as:

$$\begin{aligned}
\mathbb{E}(T|V) &= \bar{T} + CA^{\top}(ACA^{\top} + D)^{-1}(V - \bar{V}) \\
&= \bar{T} + (A^{\top}D^{-1}A + C^{-1})^{-1}A^{\top}D^{-1}(V - \bar{V})
\end{aligned} \tag{4.10}$$

$$\begin{aligned}
COV(T|V) &= C - CA^{\top}(ACA^{\top} + D)^{-1}AC^{\top} \\
&= C - (A^{\top}D^{-1}A + C^{-1})^{-1}A^{\top}D^{-1}AC^{\top}
\end{aligned} \tag{4.11}$$

The covariance matrix of the posterior demand in Equation (4.11) is related to the structure of the sensor network presented by the route proportions A, the variances of the prior demand C, and the variances of the observed flows D, but independent of the flow data V. The flow data play an essential role in the conditional expectation of the posterior demand in Equation (4.10).

Expressions based on Multiple Data Sources

The basic conditional distribution of estimation demand given observed flow has been represented in the previous session. The general expressions of the conditional expectation in Equation (4.10) and the covariance of the demand given the flow data in Equation (4.11), may hide any relationships among the freight truck demand T, the prior demand U, the link flow data V, and the path flow data W.

The expressions of estimated demand considering multiple data sources and prior demand are discussed, assuming a fixed x^{ij} as X in Equation (4.3), and denoting $\bar{T}$, $\bar{U}$, $\bar{V}$, and $\bar{W}$ as the corresponding means. We combine demand parameters as $\begin{pmatrix} T & U \end{pmatrix}^\top$, and combine loop flow vector and camera flow vector as $\begin{pmatrix} V & W \end{pmatrix}^\top$. V_m is defined as a vector element in $\begin{pmatrix} V & W \end{pmatrix}^\top$, where m is an index of multiple data sources, including loops and cameras. The route proportion matrix is defined as $\begin{pmatrix} A_l & 0 \\ A_c & 0 \end{pmatrix}$, where zeros serve U. A_l is a matrix with a number of rows equal to the number of loop detectors and with a number of columns equal to the number of ODs, and A_c is a matrix with a number of rows equal to the number of camera combinations and with a number of columns equal to the number of ODs. The notation, A_m, is introduced for a row vector in the matrix of $\begin{pmatrix} A_l & 0 \\ A_c & 0 \end{pmatrix}$. The error vector is defined as $\begin{pmatrix} E_l \\ E_c \end{pmatrix}$, where E_l is a column vector with the error for each loop and E_c is a column vector with the error for each camera combination. E_m is introduced as a vector element in the error vector. The covariance matrix of the prior demand is defined as C, and the covariance matrix of the flow data as $\begin{pmatrix} \Sigma_l & 0 \\ 0 & \Sigma_c \end{pmatrix}$ which simplifies the correlations between link flows and path flows. Σ_l is a matrix with the covariance of each loop flow in the diagonal, and Σ_c is a matrix with covariance of each camera combination in the diagonal. D_m is introduced to represent a matrix element in the covariance matrix of the flow data. The expected value and covariance of the freight truck demand T and prior demand U are obtained as follows, with the covariance matrix of the prior demand $C = \begin{pmatrix} X\Sigma_U X^\top & X\Sigma_U \\ \Sigma_U X^\top & \Sigma_U \end{pmatrix}$:

$$
\begin{aligned}
\mathbb{E}\left[\begin{pmatrix} T \\ U \end{pmatrix} \middle| \begin{pmatrix} V \\ W \end{pmatrix}\right] &= \begin{pmatrix} \bar{T} \\ \bar{U} \end{pmatrix} + \begin{pmatrix} (X\Sigma_U X^\top + \Sigma_T)A_l^\top & (X\Sigma_U X^\top + \Sigma_T)A_c^\top \\ \Sigma_U X^\top A_l^\top & \Sigma_U X^\top A_c^\top \end{pmatrix} \cdot \\
&\quad \begin{pmatrix} A_l(X\Sigma_U X^\top + \Sigma_T)A_l^\top + \Sigma_l & A_l(X\Sigma_U X^\top + \Sigma_T)A_c^\top \\ Ac(X\Sigma_U X^\top + \Sigma_T)A_l^\top & A_c(X\Sigma_U X^\top + \Sigma_T)A_c^\top + \Sigma_c \end{pmatrix}^{-1} \cdot \\
&\quad \left[\begin{pmatrix} V \\ W \end{pmatrix} - \begin{pmatrix} \bar{V} \\ \bar{W} \end{pmatrix}\right] \\
&= \begin{pmatrix} A_l^\top \Sigma_l^{-1} A_l + A_c^\top \Sigma_c^{-1} Ac + \Sigma_T^{-1} & -X\Sigma_T^{-1} \\ -\Sigma_T^{-1} X^\top & X\Sigma_T^{-1} X^\top + \Sigma_U^{-1} \end{pmatrix}^{-1} \cdot \\
&\quad \begin{pmatrix} A_l^\top \Sigma_l^{-1} V + A_c^\top \Sigma_c^{-1} W \\ \Sigma_U^{-1} \bar{U} \end{pmatrix}
\end{aligned}
$$

$$
\begin{aligned}
COV\left[\begin{pmatrix} T \\ U \end{pmatrix} \middle| \begin{pmatrix} V \\ W \end{pmatrix}\right] &= \begin{pmatrix} X\Sigma_U X^\top + \Sigma_T & X\Sigma_U \\ \Sigma_U X^\top & \Sigma_U \end{pmatrix} - \\
&\quad \begin{pmatrix} (X\Sigma_U X^\top + \Sigma_T)A_l^\top & (X\Sigma_U X^\top + \Sigma_T)A_c^\top \\ \Sigma_U X^\top A_l^\top & \Sigma_U X^\top A_c^\top \end{pmatrix} \cdot \\
&\quad \begin{pmatrix} A_l(X\Sigma_U X^\top + \Sigma_T)A_l^\top + \Sigma_l & A_l(X\Sigma_U X^\top + \Sigma_T)A_c^\top \\ A_c(X\Sigma_U X^\top + \Sigma_T)A_l^\top & A_c(X\Sigma_U X^\top + \Sigma_T)A_c^\top + \Sigma_c \end{pmatrix}^{-1} \cdot \\
&\quad \begin{pmatrix} A_l(X\Sigma_U X^\top + \Sigma_T) & A_l X\Sigma_U \\ A_c(X\Sigma_U X^\top + \Sigma_T) & A_c X\Sigma_U \end{pmatrix}
\end{aligned}
$$

In detail, the conditional expectation of the posterior OD matrix is related to the flow data V and W, the structure of the sensors A_l and A_c, the variance of the prior OD matrix Σ_U, and the variance of the errors including Σ_T, Σ_l and Σ_c. The covariance of the posterior OD matrix is independent on the observed flows V and W, but dependent on the choices made upfront, such as the demand portion X, the structure of the sensor network A_l and A_c, the variances of the prior OD matrix Σ_U, and the variances of the error terms $\Sigma_T, \Sigma_l, \Sigma_c$.

In conclusion, normal distributions give an analytical way to obtain the mean and covariance matrix of the posterior probability of the estimated demand in Equation (4.7). In this way, a fast computation in numerical studies is facilitated.

4.4.2 Simulation Approach of Posterior Estimation with Log-Normal Distributions

The assumed normal distributions are convenient to find and update parameters. However, they do not plausibly describe the underlying traffic process in situations with relatively high probabilities of large traffic demand and flow. The symmetry of the assumed normal distribution implies that the probabilities of overshoots and undershoots of the mean have equal values. If the demands and link counts are not too small, the boundary at zero will be largely irrelevant (Shao et al., 2014). It means that the log-normal distribution excluding zero is not an issue. In addition, the flow errors with normal distributions imply that the errors are independent on devices. Errors are additive to the flows. In most situations, the flow errors are dependent on devices. Different devices have different error percentages. Errors are proportional to the flows. In order to cope with the skewness of large flows and take the multiplicative errors into account, a log-normal distribution is proposed.

Below the models derived from Figure 4.2 are different from the ones in Section 4.2.1. First, the stochastic prior demand U has the following relationship with the known OD demand μ. The error term, E_{apri}, is normally distributed with a zero mean and covariance Σ_U. The term $\exp(E_{apri})$ is the error of the prior demand, called prior error.

$$U = \mu \cdot \exp(E_{apri}) \tag{4.12}$$

where $E_{apri} \sim N(0, \Sigma_U)$.

In the parameter layer, the relation between the demand T^{ij} and the prior demand U is expressed with an error term $\exp(E^{ij})$.

$$T^{ij} = x^{ij}U \cdot \exp(E^{ij}) \tag{4.13}$$

where, $E^{ij} \sim N(0, \Sigma_T^{ij})$ and the prior demand U and the error E^{ij} are independently distributed. The error term adjusts the prior demand $x^{ij}U$ for each OD pair ij to obtain the demand T^{ij}. Since the prior demand U is log-normally distributed, and the exponent of Equation (4.13) is always positive, the expression guarantees positive values of the estimated demand T^{ij}. Equation (4.13) can also be formulated as:

$$\ln T^{ij} = \ln(x^{ij}U) + E^{ij} = \ln x^{ij} + \ln U + E^{ij} \tag{4.14}$$

In the data layer, the flow V is related to the route proportion A, the demand T and the random measurement error E_m. The exponent of measurement errors $\exp(E_m)$ in Equation (4.15) represents the uncertain scales of the observed flow and the estimated flow from each detector.

$$V_m = (\sum_{ij} A_m^{ij} \cdot T^{ij}) \cdot \exp(E_m) \tag{4.15}$$

where, $E_m \sim N(0, D_m)$, D_m is a diagonal matrix element in $\begin{pmatrix} \Sigma_l & 0 \\ 0 & \Sigma_c \end{pmatrix}$ (note that we assume that Σ_l and Σ_c are diagonal), V_m is a vector element in $\begin{pmatrix} V \\ W \end{pmatrix}$, and A_m is a row in $\begin{pmatrix} A_l \\ A_c \end{pmatrix}$.

The measurement equation can be written as:

$$\ln V_m = \ln \left(\sum_{ij} A_m^{ij} \cdot T^{ij} \right) + E_m \tag{4.16}$$

Notice that the use of a log-normal distribution brings multiplicative errors, such as in Equations (4.12), (4.13) and (4.15), instead of additive errors as in the case of a normal distribution. As Caroll (2006) points out, much attention has been paid to additive measurement error models. Much less work has been done for multiplicative error models. The multiplicative errors in this study represent the scales of variables and cover the large variability of demand and flows.

Further, the posterior probability density function is expressed as follows.

$$f(T, U|V) = \prod_m \frac{f(V_m|T) f(T|U) f(U)}{f(V)}$$

where,

$$f(V) = \int_T \int_U \prod_m f(V_m|T) f(T|U) f(U) dU dT$$

Probability Density Functions

The prior demand U in Equation (4.12) follows a log-normal distribution $U \sim LN(\ln \mu, \Sigma_U)$. The parameter $\ln \mu$ is the mean of the prior demand distribution on the log scale and will be referred to as log-mean below. Similarly, the parameter Σ_U is the standard deviation of the distribution on the log scale. This quantity will be referred to as log-std below. These values are the parameters of the associated normal distribution. $T|U$ follows a log-normal distribution with log-mean $\ln(x^{ij}U)$ and log-std Σ_T^{ij}, $T|U \sim LN(\ln(x^{ij}U), \Sigma_T^{ij})$. $V|T$ follows a log-normal distribution with $V|T \sim LN(\ln(\sum_{ij} A_m^{ij} \cdot T^{ij}), D_m)$, where D_m is a covariance of flow.

Markov Chain Monte Carlo simulation

Since the log-mean of $V|T$, $\ln(\sum_{ij} A_m^{ij} \cdot T^{ij})$, is the logarithm over summations, the integral over the high dimensional term $f(V)$ is hard to compute explicitly, and an analytical solution is not available. Instead, the Markov Chain Monte Carlo simulation method is applied, which is based on sampling of the probability density functions. The differences between numerical integration

(Rossi et al., 2005) and Monte Carlo Integration are presented in Table 4.1, which motivates us to apply the Markov Chain Monte Carlo simulation method.

Table 4.1: Comparison of Numerical Integration and Monte Carlo Integration

	Numerical Integration	**Monte Carlo Integration**
Convergence speed	Fast	Slow
Convergence speed depends on dimension of the integral	Yes. Speed decreases for higher dimensional integrals.	No.
Computational burden	Increasing exponentially with the dimension of the integral	May not increase exponentially with the dimension of the integral
Dimension of Integral	Low-dimensional integrals	High-dimensional integrals

Among Markov Chain Monte Carlo simulation methods, the Gibbs sampler is one solution to the problem of numerical integration. Gibbs sampling is commonly used as a means of statistical inference, especially Bayesian inference, and when the model is built up from hierarchies of relatively standard distributions (Rossi et al., 2005). The idea of the Gibbs sampler starts at a point of a prior demand U. Draw $U|T,V$ from the density function of $f(U|T,V)$, and then draw $T|U,V$ from the density function of $f(T|U,V)$. Repeat as long as desired to obtain $f(T,U|V)$. The first few generated samples, the so called burn-in period, are discarded. The way to judge the convergence is to plot the posterior and to check whether the draws as stationary.

The $U|T,V$ is drawn from the density function $f(U|T,V)$, which is proportional to the product of $f(T|U)$ and $f(U)$. Assume $\Sigma_T^{ij} = (\sigma_T^{ij})^2 I$ where I is an identity matrix with the dimensions of $|OD| \times |OD|$ and $\Sigma_U = (\sigma_U)^2$.

$$
\begin{aligned}
& f(U|T,V) \\
\propto & f(T|U) f(U) \\
\propto & \prod_{ij} \frac{1}{\sigma_T^{ij} T^{ij} \sqrt{2\pi}} \exp\left[-\frac{(\ln T^{ij} - \ln(x^{ij} U))^2}{2(\sigma_U)^2}\right] \frac{1}{\sigma_U U \sqrt{2\pi}} \exp\left[-\frac{(\ln U - \ln \mu)^2}{2(\sigma_T^{ij})^2}\right] \\
\propto & \prod_{ij} \frac{1}{U} \exp\left[-\frac{(\ln T^{ij} - \ln x^{ij} - \ln U)^2}{2(\sigma_T^{ij})^2} - \frac{(\ln U - \ln \mu)^2}{2(\sigma_U)^2}\right]
\end{aligned}
$$

Based on the rule of completing the square, this density function of $U|T,V$ follows a log-normal distribution, with the parameters log-mean a and log-std b, where

$$
a = \frac{1}{\sum_{ij}(1/(\sigma_T^{ij})^2) + 1/(\sigma_U)^2} \left[\sum_{ij} \frac{1}{(\sigma_T^{ij})^2}(\ln T^{ij} - \ln x^{ij}) + \frac{1}{(\sigma_U)^2} \ln \mu\right]
$$

$$b = \frac{1}{\sum_{ij}(1/(\sigma_T^{ij})^2) + 1/(\sigma_U)^2}$$

Then we draw $T|U, V_m$ from the density function of $f(T|U, V_m)$, which is proportional to the product of $f(V_m|T)$ and $f(T|U)$.

$$\begin{aligned} & f(T|U, V_m) \\ \propto & f(V_m|T)f(T|U) \\ \propto & \exp\left[-\frac{(\ln V_m - \ln \sum_{ij} A_m^{ij} T^{ij})^2}{2D_m}\right] \times \prod_{ij} \exp\left[-\frac{(\ln T^{ij} - \ln(x^{ij}U))^2}{2(\sigma_T^{ij})^2}\right] \end{aligned} \tag{4.17}$$

Equation (4.17) can be simulated with general-purpose tools including the Metropolis class of algorithms to produce Markov Chain samples (Rossi et al., 2005). Among the general tools, the Metropolis-Hastings algorithm is a Markov Chain Monte Carlo method for obtaining a sequence of random samples from a probability distribution for which direct sampling is hard (Chib and Greenberg, 1995). This sequence can be used to compute an integral. The advantage of this approach is that it permits the sampling from a distribution for which the normalizing constant is not available.

The basic idea of the Metropolis-Hastings algorithm is as follows. It starts with an arbitrary starting point and searches for a next step. It is based on a proposed density function, from which random values are sampled. This movement is accepted when the sample from the proposed density function is not the same as the initial starting point. This movement is regarded as a trend to the target density function. Then iterations are carried out until the samples become stable (Rossi et al., 2005). Generally speaking, the approach used to draw the joint probability of $f(T, U|V)$ is through Gibbs sampling nested by the Metropolis-Hastings algorithm.

4.4.3 Summing up

In this section, hierarchical Bayesian networks have been introduced to obtain the posterior OD matrix through updating the prior OD matrix with the flow observations in the road network. Two approaches were investigated. One is the analytical approach based on the assumption of normally distributed flows and demand associated with additive errors. The other is based on log-normal distributions of flows and demand and makes use of Markov Chain Monte Carlo simulation associated with multiplicative errors. The computations involving the first method are fast because of the closed form solution. The second modeling approach is more realistic, but the computation process is time consuming.

4.5 Evaluation Criteria

The proposed method is evaluated based on the deviations between the conditional expectation of demand and the ground truth demand, and on the differences between the predicted flows and the observed flows. For the case with log-normal distribution, model complexity is additionally used to examine the power of data to estimate parameters. For the case with normal distributions, sensor coverage is discussed to demonstrate the randomness reduction in demand estimation and static traffic assignment after adding multiple data source detectors.

4.5.1 Demand Estimation Accuracy

Using the hierarchical Bayesian networks, the conditional expectation and variance of the estimated demand in Equations (4.10) and (4.11) can be obtained. Further, the difference between the conditional expectation $\mathbb{E}(T|V)$ and the ground truth demand T is further specified. This difference, as the estimation error Δ between the conditional expectation and the ground truth demand can be derived by the following equations.

$$\begin{aligned}
\Delta &= \mathbb{E}(T|V) - T \\
&= \bar{T} + CA^\top(ACA^\top + D)^{-1}(V - \bar{V}) - T \\
&= \bar{T} + CA^\top(ACA^\top + D)^{-1}(A \cdot T + E - \bar{V}) - T \\
&= [CA^\top(ACA^\top + D)^{-1}A - I]T + CA^\top(ACA^\top + D)^{-1}E - CA^\top(ACA^\top + D)^{-1}\bar{V} + \bar{T} \\
&= L_1 T + L_2 E + L_3
\end{aligned}$$

where,

$$L_1 = CA^\top(ACA^\top + D)^{-1}A - I,$$

$$L_2 = CA^\top(ACA^\top + D)^{-1},$$

$$L_3 = -CA^\top(ACA^\top + D)^{-1}\bar{V} + \bar{T} = -L_1\bar{T}.$$

The expectation of the estimation error is zero with a general property.

$$\mathbb{E}(\Delta) = \mathbb{E}(\mathbb{E}(T|V)) - \mathbb{E}(T) = \mathbb{E}(T) - \mathbb{E}(T) = 0 \tag{4.18}$$

The covariance of the difference $COV(\Delta, \Delta)$ depends on the route proportions A, the variance of the prior demand C, and the variance of the flow data D, but is independent of the flow observations V.

$$
\begin{aligned}
COV(\Delta,\Delta) &= \mathbb{E}(\Delta\Delta^\top) \\
&= \mathbb{E}[(L_1\bar{T} + L_2E + L_3)(L_1\bar{T} + L_2E + L_3)^\top] \\
&= L_1\mathbb{E}(\bar{T}\cdot\bar{T}^\top)L_1^\top + L_2DL_2^\top + L_1\bar{T}L_3^\top + L_3\bar{T}L_1^\top + L_3L_3^\top \\
&= L_1(C + \bar{T}\cdot\bar{T}^\top)L_1^\top + L_2DL_2^\top + L_1\bar{T}L_3^\top + L_3\bar{T}L_1^\top + L_3L_3^\top \\
&= L_1CL_1^\top + L_2DL_2^\top
\end{aligned}
$$

4.5.2 Flow Prediction Accuracy

Considering the fact that the flow observations in the network are unique measurements, they can be used to benchmark the quality of the estimated demand. Thus, the posterior estimated demand is reassigned to the network, obtaining the predicted flows. The deviations between the predicted flows and the observed flows are called flow prediction errors, denoted as $\diamondsuit$.

$$
\begin{aligned}
\diamondsuit &= A\cdot\mathbb{E}(T|V) - V \\
&= A\cdot[\bar{T} + CA^\top(ACA^\top + D)^{-1}(V - \bar{V})] - V \\
&= A\cdot\bar{T} + Q(A\cdot T + E - \bar{V}) - AT - E \\
&= (QA - A)\cdot T + (Q - I)E + A\cdot\bar{T} - Q\cdot\bar{V}
\end{aligned}
$$

where, $Q = ACA^\top(ACA^\top + D)^{-1}$.

In the following, the expectation and the covariance matrix of the flow prediction error are determined. The expectation is zero, which means that the proposed model can find the correct mean value of the flow. The covariance matrix depends on the sensor locations A, the covariance matrix of the prior demand C, and the covariance matrix of the flow observations D.

For convenience, we define

$$M_1 = QA - A = ACA^\top(ACA^\top + D)^{-1}A - A = M_2A,$$

$$M_2 = Q - I = ACA^\top(ACA^\top + D)^{-1} - I,$$

$$M_3 = A\cdot\bar{T} - Q\cdot\bar{V} = A\cdot\bar{T} - ACA^\top(ACA^\top + D)^{-1}\cdot\bar{V}.$$

As a result, we have

$$\mathbb{E}(\diamondsuit) = \mathbb{E}[A(\mathbb{E}(T|V) - V)] = 0$$

$$COV(\diamondsuit,\diamondsuit) = M_1CM_1^\top + M_2DM_2^\top = M_2ACA^\top M_2 + M_2DM_2^\top$$

4.5.3 Model Complexity

Model complexity in Van der Linde (2012) is quantified as the power of data for parameters, how hard it is to learn parameters from data, how sensitive parameters are to observations, or how large the estimation variance is. The model complexity is also called the general degrees of freedom of a modelling procedure (Ye, 1998), which is defined as:

$$COV(T) - \mathbb{E}(COV(T|V)) = COV(\mathbb{E}(T|V)) \tag{4.19}$$

This equation follows the Law of Total Variance (Weiss et al., 2006), where $\mathbb{E}(COV(T|V))$ is called the unexplained part of the covariance of T, and $COV(\mathbb{E}(T|V))$ is called the explained part. If there is no measured flow data V, $COV(\mathbb{E}(T|V))$ is the covariance of the prior demand. The expectation of the conditional variance $\mathbb{E}(COV(T|V))$ will be large, but the variance of the conditional expectation $COV(\mathbb{E}(T|V))$ will be small since its value is the covariance of the prior demand. Adding detectors to measure flow leads to a decrease of the value of $\mathbb{E}(COV(T|V))$, and an increase of the value of $COV(\mathbb{E}(T|V))$. In other words, the more detectors involved, the higher the model complexity.

4.5.4 Sensor Coverage for the Case with Normal Distributions

Sensor coverage is represented by the route proportion A, connecting demand and flows from different detectors. The impact of the different combinations of sensors to the demand estimation is analysed by means of the route proportion. The conditional expectation in Equation (4.10) gains from flow observations. Thus, the smaller $(A^\top D^{-1} A + C^{-1})^{-1}$ in the case with normal distributions, the more the gain. $A^\top A$ can be analyzed since the covariance matrix of the prior demand C and the covariance matrix of the flow observations D are fixed. In general, the larger the norm of $A^\top A$ is, the smaller $(A^\top D^{-1} A + C^{-1})^{-1}$ is. Thus, adding detectors helps to increase the value of $A^\top A$, which is proved as follows.

Proof. Suppose $A = \begin{pmatrix} A_0 \\ A_1 \end{pmatrix}$, then

$$A^\top A = \begin{pmatrix} A_0^\top & A_1^\top \end{pmatrix} \begin{pmatrix} A_0 \\ A_1 \end{pmatrix} = A_0^\top A_0 + A_1^\top A_1 \geq A_0^\top A_0 \geq 0 \qquad \square$$

As long as D and D_0 are positive definite matrices, $A^\top D^{-1} A + C^{-1} \geq A_0^\top D_0^{-1} A_0 + C^{-1}$.

Actually, this approach to represent sensor coverage fits into the three rules of Viti and Corman (2012). First is the maximum flow fraction rule (so called F1), which says that the sensors should be installed so that the portion of flow measured belonging to that OD pair with respect to all other OD pairs measured by that sensor is maximized. Thus priority must be given to locations containing the largest information distributed over fewer OD pairs. Our

approach proves that adding detectors helps to increase the value gain presented by $A^\top A$ (A is mapping), which is aligned with this rule. The second rule of maximal OD demand fraction rule (F3) and the third rule of maximal net OD flow captured rule (F4) are analogous to the first rule. Comparing with these rules, our approach provides another aspect of the value gain from the mapping $A^\top A$ to view the sensor coverage problem.

4.6 Application of Normal Distributions

The road network of part of the A15 motorway between Hoogvliet and Havens (from locations 43.1km to 49.9km) in the Netherlands serves as a case study. Its structure is illustrated in Figure 4.3. The available information consists of an OD demand based on survey data obtained from Statistics Netherlands. In order to test the methodology and represent the stochastic behavior of the model, a ground truth OD matrix, link flow data, and path flow data are generated. The methods are tested with both normal distributions and log-normal distributions.

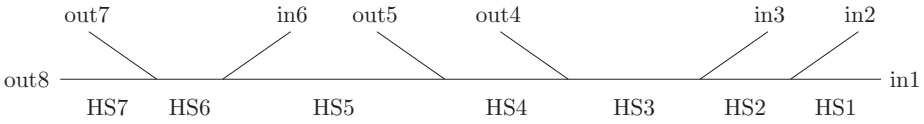


Figure 4.3: Part of A15 motorway from Hoogvliet to Havens

Seven scenarios are employed to illustrate the usage and effectiveness of the flow data from the different combinations of loops and cameras. These scenarios are defined in a hierarchical way, in the sense that a base scenario with loops or cameras is stepwise extended with more loops or cameras. Specifically, the first three scenarios consist of three loops on highway section 3, 4 and 6, which is extended with loops on all highway sections, and cameras on highway section 3, 4 and 6. The next three scenarios consist of cameras on highway section 3, 4 and 6, which is extended with cameras on all highway sections, and loops on highway section 3, 4 and 6. Last scenario consists of loops and cameras on all sections. In detail,

1. loops on highway section 3, 4 and 6;
2. loops on highway section 3, 4 and 6; plus loops on highway section 1, 2 and 5;
3. loops on highway section 3, 4 and 6; plus loops on highway section 1, 2 and 5; plus cameras on highway section 3, 4 and 6;
4. cameras on highway section 3, 4 and 6;
5. cameras on highway section 3, 4 and 6; plus cameras on highway section 1, 2 and 5;

6. cameras on highway section 3, 4 and 6; plus cameras on highway section 1, 2 and 5; plus loops on highway section 3, 4 and 6;

7. both loops and cameras on all highway sections.

4.6.1 Data Generation with Normal Distributions

The process of data generation is illustrated in Figure 4.4. For the model with normal distributions, we take the demand data from Statistics Netherlands as the expected value μ of the prior demand U, $\mu = 22,434$. The prior variance Σ_U is 200. x^{ij} is taken as 1/14 (14 OD pairs). To generate the ground truth OD matrix T^{ij} as step (1) in Figure 4.4, a multivariate normal distributed error term E^{ij} is introduced according to Equation (4.3), where $E^{ij} \sim N(0, 20)$. Based on the generated demand T as the ground truth demand, the link flow data and the path data are obtained with certain values of variances, as step (2) in Figure 4.4. Both variances of the link flow from loops Σ_l and the path flow from cameras Σ_c to be 1. Note that although the estimation error is independent on the flow data as mentioned before, the purpose here is to generate data for further analysis.

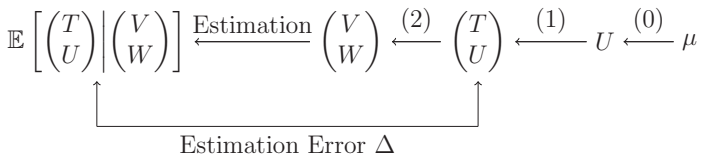


Figure 4.4: Process of Generating Data

4.6.2 Demand Estimation Accuracy

Once the data have been generated, the model should be able to re-estimate the generated ground truth demand T. The expectation of the difference between the conditional expectation and the ground truth demand in Equations (4.18) is equal to zero, which means that the proposed model can reach the conditional expectation of the OD demand given the observations. Thus, the expectation of the difference is not influenced by the mean of the prior demand or other parameters of the assumed distributions.

However, the covariance of the estimation error varies among the scenarios due to the size of route proportion A, flow covariance D, prior demand covariance C and flow data V. We demonstrate the features of the covariance matrices among the scenarios using their trace. The larger the trace of the covariance matrices, the more randomness is involved. Table 4.2 demonstrates the trace of the covariance matrices of the estimation errors in the scenarios. There is 90.21%

reduction of the trace of the demand estimation error covariance, from the situation with three loops to the situation with full coverage of cameras and loops.

Table 4.2: Trace of the Covariance Matrices of the Estimation Errors in Each Scenario

Loops 346	Full Loops	Full Loops +Cameras346	Cameras 346	Full Cameras	Full Cameras + Loops346	Full Cameras + Full Loops
2695.70	1811.58	1809.71	2474.10	268.60	266.33	263.87

4.6.3 Flow Prediction Accuracy

Traces are used to represent the features of the covariance matrix of the flow prediction errors in Table 4.3. There is around 61.09% reduction of the covariance of the flow prediction errors from the worst case with three loops installed to the best case with full coverage of cameras and loops. Taking the respective basis of loops and cameras, adding equipment always decreases the eigenvectors of the covariance matrices of flow prediction, increasing the certainty of prediction. For example, 24.01 for the three loops situation, 18.69 for the situation of full coverage of loops, 18.68 for the situation of full coverage of loops plus three cameras, and 9.39 for full coverage of both loops and cameras. The reason is that the measurement errors of the detectors cancel out when more equipment is installed.

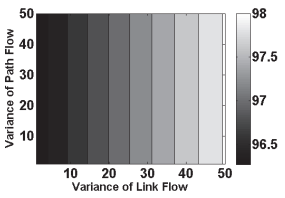
Table 4.3: Trace of Covariance Matrices of Flow Prediction Errors in Each Scenario

Loops 346	Full Loops	Full Loops +Cameras346	Cameras 346	Full Cameras	Full Cameras + Loops346	Full Cameras + Full Loops
24.01	18.69	18.68	22.67	9.42	9.41	9.39

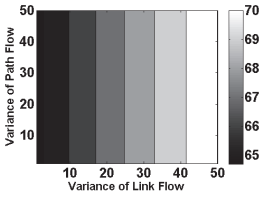
4.6.4 Variances

Since flow variances from loops and cameras are assumed, insight into the relation between these two variances is obtained. Assuming the variance of the prior demand Σ_U and the demand variance Σ_T are fixed, Figure 4.5 illustrates the mean eigenvalues along both the variance of the link flow and the variance of the path flow. Figures 4.5 (a), (b), (c) and (g) take the loops as the basis where the number of loops is increasing step by step, and Figures 4.5 (d), (e), (f) and (g) take cameras as the basis where the number of cameras is increasing step by step. In general, the lower the variances of the detectors are, the smaller the mean values of the eigenvalues. Adding detectors always helps to reduce the randomness. With the full coverage of both loops and cameras, the mean values are the smallest, ranging from 0 to 22.

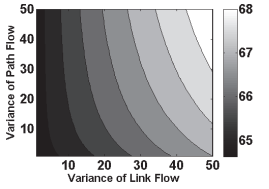
Additionally, Figure 4.5 (g) is not symmetric, even if both cameras and loops have full coverage of the network. This is because the path flow data from the cameras is able to identify



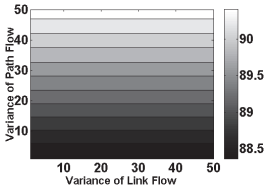
(a) Loops 346



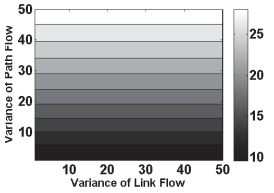
(b) Full loops



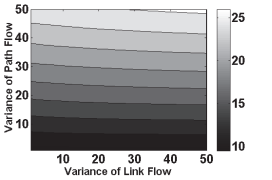
(c) Full Loops + Cameras 346



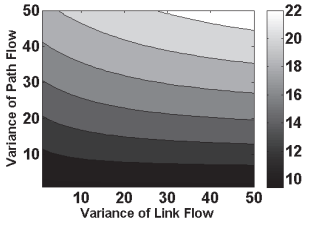
(d) Cameras 346



(e) Full Cameras



(f) Full Cameras + Loops 346



(g) Full Loops and Full Cameras

Figure 4.5: Mean of the Eigenvalue of the Covariance Matrix from Estimation Errors (different colors) along the Variances of Link Flow and Path Flow

the OD matrix exactly. Fixing a variance of the path flow, no matter what the variance of the link flow changes to, the mean eigenvalues are stable. Comparing Figures 4.5 (f) and (g), the stability of the mean eigenvalues from the path flow in Figure 4.5 (f) is better, although the mean values are a bit larger. It may be due to the fewer loops that bring lower link variances. Since three cameras bring more certainty to the eigenvalues, the slope in Figure 4.5 (c) with the scenario of full coverage of loops and three cameras, is larger than the one in Figure 4.5 (f) with the scenario of full coverage of cameras and three loops.

4.6.5 Sensor Coverage

Table 4.4 numerically shows the trend of increasing eigenvalues of $A^\top A$ when adding devices as illustrated in section 4.5. As the number of loop detectors installed is increasing step by step, the mean value of the eigenvalue of $A^\top A$ increases from 2.0714 in the scenario of only three loops installed, to 3.6429 in the case of full coverage of loops, to 4.6429 in the case of full coverage of both cameras and loops. Although the mean eigenvalues in both cases of full coverage of loops plus three cameras and full coverage of loops plus full coverage of cameras are the same, the median of the eigenvalues in the last scenario is higher, namely 1.3383. As the number of loop detectors installed is increased, the eigenvectors increase from 1 in the three-camera situation to 4.6429 in the case of full coverage of both cameras and loops.

Table 4.4: Eigenvalues of $A^\top A$ in Each Scenario (Each Column has the same dimension as T)

	Loops 346	Full Loops	Full Loops +Cameras346	Cameras 346	Full Cameras	Full Cameras + Loops346	Full Cameras + Full Loops
Minimum	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.0000	0.0000	0.0000	1.0000	1.0000	1.0000
	0.0000	0.6766	1.2733	0.0000	1.0000	1.0000	1.6766
	0.0000	1.0322	1.6461	0.0000	1.0000	1.0000	2.0322
	0.0000	1.3395	2.8012	0.0000	1.0000	1.0000	2.3395
	0.0000	1.7380	3.4151	2.0000	1.0000	1.0000	2.7380
	1.2075	2.8868	4.4782	3.0000	1.0000	2.2075	3.8868
	3.7146	6.7016	9.4937	3.0000	1.0000	4.7146	7.7016
Maximum	24.0779	36.6254	41.8924	6.0000	1.0000	25.0779	37.6254
Mean	2.0714	3.6429	4.6929	1.0000	1.0000	3.0714	4.6429
Median	0.0000	0.3383	0.6366	0.0000	1.0000	1.0000	1.3383

Compared to the sensor location papers, the insight of randomness reduction in demand estimation is gained from sensor coverage, represented by the positive mapping $A^\top A$ for the case with normal distributions. Adding detectors leads to an increase of $A^\top A$ and a reduction of the estimation covariance. Other researches about sensor coverage, such as Bianco et al. (2001), Zhou and List (2010) and Fei and Mahmassani (2011), focus on the sensor location problem which minimizes the budgets and maximizes the sensor usage. We extend the use of sensor coverage, obtaining an insight into reducing randomness when adding detectors and argues why adding detectors is able to reduce randomness based on $A^\top A$.

4.7 Application of Log-Normal Distributions

The same road network is used as in the application of the normal distributions. From data generation to results evaluation, the framework is illustrated in Figure 4.6. The Metropolis-Hastings sampling is applied to draw $T|U,V$ with only one sample with the fixed flow data. Then $U|T$ can be easily drawn from the log-normal distribution. Based on Gibbs sampling, the joint density function of the posterior demand $T,U|T$ can be achieved.

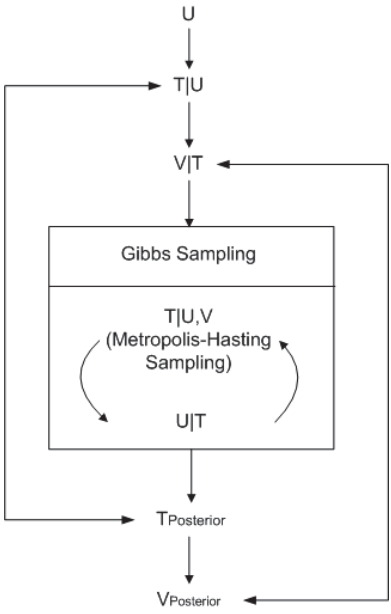


Figure 4.6: Framework to Evaluate the Model in the Situation of the Log-Normal Distributions

Three aspects of the results are evaluated: demand estimation accuracy, flow prediction accuracy, and the sensitivity of demand to flows. The demand estimation accuracy and the flow prediction accuracy are applied to demonstrate whether the flow data associated with hierarchical Bayesian networks can lead the demand to the correct values. In addition, the sensitivity of the demand parameter to observation data is evaluated by means of the model complexity, which demonstrates how well the flow data controls the demand estimation.

4.7.1 Data Generation with Log-normal Distributions

Data are generated for the stochastic prior total demand U, the stochastic ground truth demand T, and the stochastic flow data V. The prior demand of part of the A15 motorway from Statistics

Netherlands follows a log-normal distribution $U \sim LN(\ln \mu, \Sigma_U)$, where the log-std of the prior demand Σ_U is assumed to be 0.02. This small value is chosen because of the magnitude reduction of the logarithm $\ln \mu$. For instance, $\ln(2000) = 7.60$. Taking Σ_U as 0.02 is reasonable. The ground truth demand T is generated based on Equation (4.13) with an assumed variance of 0.02. Next the stochastic link flows and path flows are generated with a log-variance 0.0001 in the log-normal distribution.

4.7.2 Estimation and Prediction Accuracy

The demand expectation with the log-normal distribution is zero. The focus is on estimation covariance. The trace of the covariance matrices of the demand estimation errors and the trace of the covariance matrices of the flow prediction errors are taken to measure accuracy. Table 4.5 demonstrates the trend of the absolute mean values of demand estimation errors in seven combinations of detectors. The trace of the covariance matrices in the situations associated with full coverage of cameras is much smaller than in other situations. This is due to the identity mapping in the case of full coverage of cameras. In general, adding devices always helps to reduce the absolute mean values of the demand estimation errors. From the worst case with three loops installed to the best case with full coverage of cameras and loops, there is 77.38% trace reduction of the demand estimation error covariance.

Table 4.5: Trace of the Covariance Matrix of the Estimation Errors in Each Scenario with Log-Normal Distributions

Loops 346	Full Loops	Full Loops +Cameras346	Cameras 346	Full Cameras	Full Cameras + Loops346	Full Cameras + Full Loops
2,807,782.22	1,936,089.47	1,851,283.94	2,662,375.40	650,865.44	645,199.22	635,095.64

In addition, the posterior demand T is reassigned to the network. The absolute deviation between the reassigned flow and the measured flow is the absolute value of the flow prediction error, illustrated in Table 4.6. The traces demonstrate the same trend as Table 4.5. Adding detectors helps to reduce errors. From the worst case with three loops installed to the best case with full coverage of cameras and loops, there is 75.80% trace reduction of flow prediction error covariance.

Table 4.6: Trace of the Covariance Matrices of the Flow Prediction Errors in Each Scenario with Log-Normal Distributions

Loops 346	Full Loops	Full Loops +Cameras346	Cameras 346	Full Cameras	Full Cameras + Loops346	Full Cameras + Full Loops
8,544.65	8,333.98	8,289.99	22,323.80	2,342.85	2,261.45	2,067.74

Technically, the Bayesian Network Method with a log-normal distribution is quite time consuming to carry out the computation. With 1000 draws of flows and 60,000 iterations in the Gibbs sampler, it usually takes approximately five hours for the software of Matlab to get convergence on a laptop with 4GB RAM and 32-bit Operating System.

4.7.3 Model Complexity

The model complexity criterion is applied to demonstrate how hard it is to derive parameters from data. $\mathbb{E}(COV(T|V))$, the unexplained part in the Law of Total Variance, should get smaller. $COV(\mathbb{E}(T|V))$, the explained part in the Law of Total Variance, should get larger. The case study of the A15 motorway is used mainly to illustrate the effects of multiple data sources. Adding devices means having stronger data observations.

To better represent the results, both explained and unexplained parts are normalized by the covariance of the prior demand $COV(T)$. Since the demand is multivariate with high-dimensionality, presenting the results requires an index with good numerical stability. The matrix trace is such a criterion to demonstrate the characteristics of a matrix. Table 4.7 and Table 4.8 demonstrate respectively the trace of the normalized explained part $COV(\mathbb{E}(T|V))$ and the normalized unexplained part $\mathbb{E}(COV(T|V))$. The values in the tables add up to almost one in each scenario. Table 4.7 shows that adding observations increases the value of model complexity. From scenario 1 (three loops) to scenario 2 (full coverage of loops), the model complexity doubles from 0.1749 to 0.3880. This means that the parameters are very sensitive to perturbations of the flow data, since there is a higher rank of the mapping matrix in scenario 2 than in scenario 1. Scenarios with full coverage of cameras have high values of model complexity, up to 0.73. The reason is that full coverage of cameras is able to capture the demand explicitly. The mapping matrix of full coverage of cameras has a full rank. In addition, the model complexity in the last three scenarios stays almost at the same values, although it still increases a little according to the number of loops involved.

Table 4.7: Trace of the Normalized Explained Part $COV(\mathbb{E}(T|V))$ in Scenarios

Sce 1 Loops 346	Sce 2 Full Loops	Sce 3 Full Loops +Cameras346	Sce 4 Cameras 346	Sce 5 Full Cameras	Sce 6 Full Cameras + Loops346	Sce 7 Full Cameras + Full Loops
0.1749	0.3880	0.4147	0.2107	0.7305	0.7312	0.7316

The findings from Table 4.5 and Table 4.8 are consistent. Both demand estimation error and model complexity are able to evaluate the results.

Table 4.8: Trace of the Normalized Unexplained Part $\mathbb{E}(COV(T|V))$ in Scenarios

Sce 1 Loops 346	Sce 2 Full Loops	Sce 3 Full Loops +Cameras346	Sce 4 Cameras 346	Sce 5 Full Cameras	Sce 6 Full Cameras + Loops346	Sce 7 Full Cameras + Full Loops
0.8339	0.6135	0.5935	0.8024	0.2816	0.2870	0.2813

4.8 Conclusion

Hierarchical Bayesian networks take the stochastic features of the freight system into account. To obtain the posterior demand associated with the mean and the covariance, the prior demand is updated by different types of flow data in the road network, taking errors into account. Shao et al. (2014) mentioned that most of the conventional OD demand estimation models mainly make use of the first-order statistical property (i.e. the mean) of the traffic counts, but the second-order property (i.e. the covariance) of the count data is usually ignored. Here, we consider both statistical properties and provide the detailed mathematical derivation. Contributing to the research body of Bayesian inference (Maher, 1983; Tebaldi and West, 1998; Hazelton, 2000; Castillo et al., 2008a; Perrakis et al., 2011), where normal distributions dominate the quantities, we propose both normal distributions and log-normal distributions, taking the means and the covariances into account.

If flow errors are independent of devices, errors are additive to flows; then a normal distribution can be applied, which allows one to adopt an analytical approach and to quickly obtain the posterior demand. If flow errors are considered to be dependent on devices, such as different devices having different percentage errors, errors are proportional or multiplicative to flows. Then a log-normal distribution is applied. So far, we have not seen relevant papers applying the log-normal distribution to OD estimation. A log-normal distribution gives a plausible description of a right skewed flow distribution, in which large flows have a relatively high probability of occurrence. If the demands and link counts are not too small, the boundary at zero will be largely irrelevant (Shao et al., 2014). In this case, an analytical solution is not possible. Instead, Markov-Chain Mont-Carlo simulation is applied with Gibbs sampling, nested by Metropolis-Hastings sampling to arrive at estimates. The computation time associated with the log-normal distributions is longer than in the case of the normal distributions because of the sampling simulation.

In addition, there are three types of errors considered: prior error, estimation error and observation error. The prior error represents the error in the prior data from surveys for instance. The estimation error is the error generated when estimating demand parameters. The observation error, also called measurement error, could result from disruptions of the devices. Usually, errors modeled in researches, such as Zhou and Mahmassani (2006), are additive. The assumption of additive errors is that whatever the error generating problems are, the errors stay the same. In our model of log-normal distributions, multiplicative errors are proposed, which

actually relaxes the underlying assumption. These multiplicative errors represent the proportion of the deviation to the baseline, for instance, the percentage failure of each device for observation error. Since the two situations have different approaches to generate inputs, it is infeasible to compare the estimation results from the two situations. An extension could compare the two situations with a third data set.

Furthermore, in the framework of hierarchical Bayesian networks, three types of data are integrated, prior demand data, link flow from loop detectors, and path flow from cameras. The researches have considered one or two types of data sources, illustrated in Table 2.2. Most of the research is based on loop detectors (Van Zuylen and Willumsen, 1980; Spiess, 1987; Watling, 1994; Ashok and Ben-Akiva, 2000; Hazelton, 2008). Zhou and Mahmassani (2006), Van Der Zijpp (1997) and Dixon and Rilett (2002) take both loop detectors and cameras into account, but use different methods, such as least squares. Cascetta (1984) uses the survey data together with loop detectors. He takes the objective function as minimizing the estimated demand and survey demand, with the assumption that survey demand has a good quality. In this chapter, hierarchical Bayesian networks allow to update survey demand as prior demand based on flow data, whatever the survey quality is. The prior demand associated with a relatively large variance has little impact on the estimated demand. The impact has been dominated by the flow observations.

The effectiveness of path flow data from cameras is addressed, combined with link flow data from loop detectors. The case study of the A15 motorway in the Netherlands demonstrates that, with the assumption of a normal distribution, the trace of the covariance matrix benchmarks the added value of extra detectors to the estimation performance. Path flows obtained from cameras reduce the randomness significantly, around 90.21% trace reduction of the estimated demand error covariance. There is 61.09% trace reduction of flow prediction error covariance from the worst case with three loops installed to the best case with full coverage of cameras and loops. In the log-normal case, these two values are 77.38% and 75.80%, respectively. Thus, path flow data plays an essential role to estimate accurate demand. The results from the model complexity are consistent with the demand estimation errors. The model complexity could be applied as a new evaluation criterion.

Last but not least, the insight of randomness reduction in demand estimation is gained from sensor coverage by $A^\top D^{-1} A + C^{-1}$, a part of the expectation of the estimated demand conditional on the flow data for the case with normal distributions. Antoniou et al. (2015) indicates that the performance of OD estimation methods depends highly upon the structure and the properties of the supply/demand interaction - that is of the assignment map A. We have proved that the positive mapping $A^\top A$ can represent $A^\top D^{-1} A + C^{-1}$ since the covariance matrix of the prior demand C and the covariance matrix of the flow observations D are fixed. Adding detectors leads to an increase of $A^\top A$ and a reduction of the estimation covariance. Our approach also fits into three rules in Viti and Corman (2012): the maximum flow fraction

rule, the maximal OD demand fraction rule and the maximal net OD flow captured rule, where the priority must be given to locations containing the largest information distributed over fewer OD pairs. Comparing with these rules, we provide another aspect of the value gain from the mapping $A^\top A$ to view the sensor coverage problem. Additionally, other researches such as Bianco et al. (2001), Zhou and List (2010) and Fei and Mahmassani (2011), focus on the sensor location problem which minimizes the budgets and maximizes the sensor usage. Especially, Zhou and List (2010) applied a least mean square OD destination estimator, analysed the covariance matrix of estimated demand, and gave advice about sensor locations. A few studies have been conducted to locate cameras or path-flow detectors for sensor location problems. Zhou and List (2010) focused on locating a limited number of point-flow detectors and path-flow detectors in a network to update an estimated OD trip table. Compared with these papers, our approach addresses a sensor coverage problem in a framework of hierarchical Bayesian networks, obtaining an insight into reducing randomness when adding detectors and arguing why adding detectors is able to reduce randomness based on $A^\top A$.

Chapter 5

Day-to-Day Origin Destination Tuple Estimation and Forecasting

[1]Ma,Y., Kuik, R. and Zuylen, H.J. van. (2013) Day-to-day Origin Destination Tuple Estimation And Prediction with Hierarchical Bayesian Networks Using Multiple Data Sources, Journal of Transportation Research Record, 2342(1) p51-61

5.1 Introduction

In the previous chapter, we developed a hierarchical Bayesian network model to estimate origin destination pairs in a road network. This model allowed us to estimate the parameters of the trip demand equation associated with the network using data from two different data sources, loop detectors and camera recordings, while taking prior information about OD pairs into account. Compared with more traditional approaches, such as information methods (Van Zuylen and Willumsen, 1980) and general least squares (Cascetta, 1984; Bell, 1991; Bierlaire and Toint, 1995), this model contributes to the literature on modeling the stochastic nature of traffic demand.

In the present chapter, we extend the Bayesian network model in two directions: the incorporation of forecasting flow one day ahead, and the consideration of origin-destination tuples. Both extensions entail the exploitation of systematic demand patterns in time for the forecasting of next day's demand. Forecasts of future traffic demand support traffic management in taking measures to alleviate congestion. Statistical forecasting approaches, such as time series analysis (West and Harrison, 1997), have been applied to transportation demand prediction based on one previous time period. For instance, Ashok and Ben-Akiva (2002) applied autoregressive models to the recursive estimation and prediction of transport demand. Zhou and Mahmassani (2007) developed a polynomial trend filter to capture possible structural deviations in the forecasted demand, by considering demand deviations from the a priori estimate of the regular pattern as a time-varying process with smooth trend. Considering multiple previous time periods, a multi-process model is applied in this chapter. This model, consisting of a mixture of dynamic linear relations, assigns probabilities to each designed scenario of weighted previous demands.

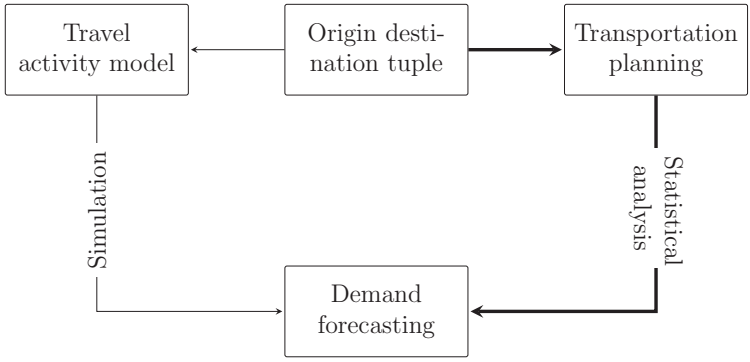


Figure 5.1: ODT Connects the Travel Activity Model and Transportation Planning (the bold line is the applied approach)

In addition, origin destination tuples (ODTs) introduced in Chapter 1 consider combinations or chains of links in a network, as opposed to individual links, when estimating the parameters

of the trip matrix. The notion of origin destination tuples stems from travel activity-based research considering a behavioral perspective of traffic demand in a road network (Kitamura, 1996; Bowman and Ben-Akiva, 2001; Chorus et al., 2008). Travel activity-based research assumes that people plan ahead and choose attributes of each trip, including mode, destination, and departure time, while considering a chain of links in the network, instead of separate links. Jones et al. (1990) provide a comprehensive definition of activity analysis: it is a framework in which traveling is analyzed as daily or as multi-day patterns of behavior, related to and derived from differences in life style and activity participation among the population. They take into account the fact that travelers have travel plans as a trip chain, such as from home to work and back (HWH) or a work tour with at least one additional stop for another activity (HWH+). In addition, travel activity-based models integrate household activities, land use distributions, regional demographics and transportation networks in an explicitly time-dependent fashion (McNally, 1996). In order to understand the individual choice behavior, researchers mostly represent it as a discrete choice model (Bowman and Ben-Akiva, 2001; Chorus et al., 2008), and microsimulation (McNally, 1996). The disaggregate discrete choice activity schedule presented by Bowman and Ben-Akiva (2001) is specified and estimated from the available daily survey and service data at the transportation system level. They generate time and mode specific trip matrices for forecasting. The model is designed to capture interactions among individual decisions throughout a 24 hours day by explicitly representing tours and their interrelationships in an activity pattern. Wang and Cheng (2001) develop a spatial-temporal data model to support activity based transport demand modeling in a GIS environment, identifying spatial and temporal opportunities for activity participation. Activity patterns are conceptualized as a sequence of staying at or travelling between activity locations. The activity based research enriches the trip generation in the transportation planning process. But the activity based model does not touch upon the road network associated with the observed data, which is essential for the transportation planning and this thesis. In addition, survey data (Stavins, 1999; Kroes and Sheldon, 1988) is the main information source supporting the activity based research. Although surveys may demonstrate some trip chains of travelers, the sparseness of survey data limits to represent travel activities over time. Consequently, it is hard to estimate demand from patterns of behavior and it is hard to use survey data of OD information as input for traffic assignment. Hence, the scope of behavioral research is normally limited to the demand side, ignoring the road network.

Importantly, as indicated in Kitamura (1996), how many trips in an activity based demand model are made depends on how the visits to different places are sequenced and combined into trip chains. Activity-based demand forecasting should be based on a model of activity engagement, and then forecast the number of trips, given a set of activities to be pursued. In this sense, ODTs can bridge the gap between transportation planning and travel activity-based research, considering the demand pattern and the road network.

The forecasting approaches in the activity based model rely on macro-simulation of mobility and activity participation (e.g., number of trips, total travel distance), and on micro-simulation of replicating the decision mechanisms underlying activity engagement. The underlying forecasting methods are Monte Carlo simulation, as McNally (1996) and Kitamura et al. (2000) did. Since this study contributes to the transportation planning side, the demand forecasting approach will be aligned with statistical forecasting approaches applied in transporting planning, as illustrated by the bold line in Figure 5.1.

Actually, ODTs bring extra challenges to the estimation and forecasting of demand. Firstly, additional demand parameters need to be estimated on top of OD pairs. This further aggravates the problem of under-specification (Van Zuylen and Willumsen, 1980), where the number of ODTs including OD pairs is much larger than the number of links. Secondly, one reason for the ignorance of the trip chain in transportation planning could be the anonymous loop detector data, which cannot identify vehicles. ODTs cannot be estimated and forecasted accurately without identification detectors.

The network studied in Chapter 4 investigates the use of information from different sources employing artificially constructed information scenarios. In practice, the existence of identification monitoring systems, such as automated number plate recognition (ANPR) cameras and Bluetooth scanners, allow partial observation of the trajectories connecting links that make up OD pairs. The question arises how these OD tuples can be structurally embedded in the demand estimation.

In the following section, the methodology is presented. The case study in Section 5.3 illustrates this method. Section 5.4 finalizes the chapter with discussions.

5.2 Methodology

In this section, the hierarchical Bayesian network model is applied to obtain the posterior forecasted origin destination tuples. Considering the stochastic nature of the model and the computational efficiency, Kalman filtering with normally distributed error terms is used to get the mean and variance of the ODTs.

5.2.1 Hierarchical Bayesian Networks to Forecast Origin Destination Tuples

Hierarchical Bayesian networks represent probabilistic dependencies between variables in a directed acyclic graph. Each node of the graph is regarded as a random variable and is connected by its conditional probability given the value of its parents in the graph (Gyftodimos and Flach, 2002). Figure 5.2 illustrates the diagram of a hierarchical Bayesian network for forecasting origin destination tuples with three layers: hyper-parameters, parameters and data layers. In

the hyper-parameter layer, the prior ODTs data are located. Information about these priors is obtained from surveys. The variables corresponding to ODT volumes are in the parameter layer. The observations in the data layer are link flows and path flows.

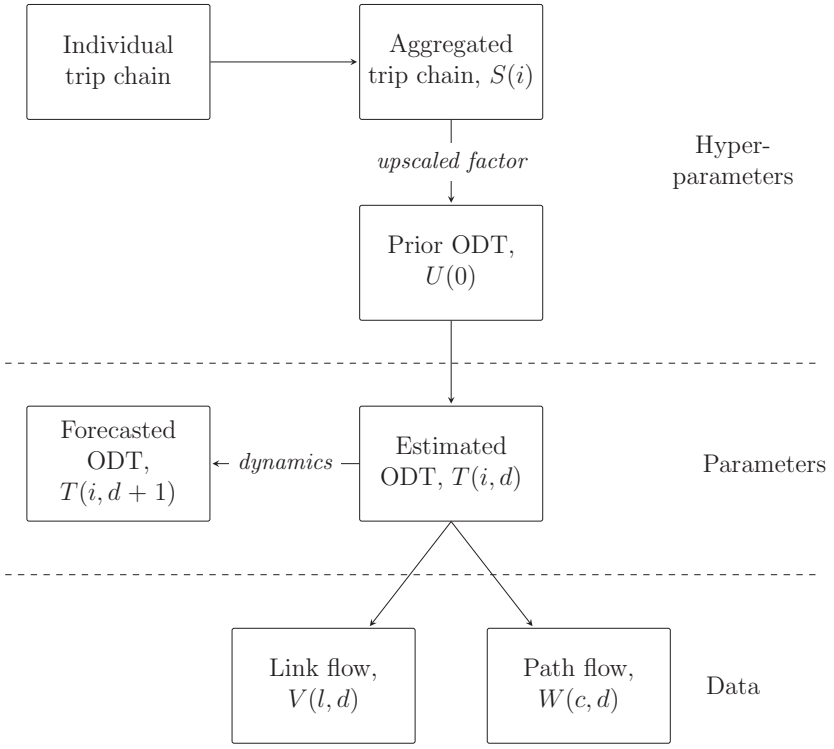


Figure 5.2: Hierarchical Bayesian Networks for Forecasting OD Tuples

Hyper-Parameter Level

The setup of the hyper-parameter level is based on individual trip chains, which are derived from survey data. These individual activity trip chains are related to the scheduled time and the activity locations. With the scheduled time, travelers determine the departure time; and with the activity locations, they decide about the travel patterns. Here, we assume that travel modes are either cars or trucks (Note that in the activity-based research mentioned in Chapter 1, scheduled time, activity locations, and travel modes are three elements.). Aggregating vehicles with the same ODT pattern gives the sample demand of ODT i, denoted as S_i. A deterministic growth factor is introduced, denoted as β_i to scale up this sample demand to the population level of the prior ODT U_0, a scalar representing prior demand for a network.

$$U_0 = \sum_i \beta_i \cdot S_i \tag{5.1}$$

U_0 is a random variable following a normal distribution $U_0 \sim N(\sum_i \beta_i \cdot S_i, \Sigma_{U_0})$. Introducing the constant weight of the prior demand parameter, x_i, similar to the weight in Equation (4.2), the separated prior ODT i is presented as:

$$U_i = x_i \cdot U_0 \tag{5.2}$$

Parameter Level

The demand variables at the parameter level are the factors to be estimated and predicted. The ODT denoted as $T_{i,d}$ is a parameter in this layer. West and Harrison (1997) treat the prior U_0 as an input of the parameter $T_{i,d}$ when $d = 0$. The connection between $T_{i,0}$ and U_0 is as follows.

$$T_{i,0} = U_i + E_i = x_i \cdot U_0 + E_i \tag{5.3}$$

There U_0 and E_i are assumed to be independent and the estimation error is described as $E_i \sim N(0, \Sigma_{T,i})$.

A forecasted ODT for the next day, $T_{i,d+1}$, is obtained through an auto-regressive model of past demands. The weights α, which are considered as unknown, indicate the share of the past demand for forecasting the future demand. The applied model of the dynamics is a multi-process model (West and Harrison, 1997). Here we treat the weights as deterministic in the demand dynamics model. In Equation (5.4), C_i is a constant:

$$T_{i,d+1} = C_i + \sum_{z=0}^{x-1} \alpha_{z+1} T_{i,d-z} + \varepsilon_{i,d+1} \tag{5.4}$$

Demand on different days is conveniently summarised into vectors including the demand from day d to day $d-(x-1)$, $\vec{T}_{i,d} = (T_{i,d}, T_{i,d-1}, \ldots, T_{i,d-(x-1)})^\top$, where x is the length of the recursive process. The first component of next day's error vector $\vec{\varepsilon}_{i,d+1}$ is assumed to have a normal distribution with zero mean and variance $\sigma_{i,d+1}$. Equation (5.4) can be written in matrix notation as:

$$\vec{T}_{i,d+1} = \vec{C}_i + B \cdot \vec{T}_{i,d} + \begin{pmatrix} \varepsilon_{i,d+1} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5.5}$$

where,

$$B = \begin{pmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{x-1} & \alpha_x \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}$$

and $\vec{C_i} = \iota \cdot C_i$, where $\iota = (1, 0, ..., 0)^\top$.

The model can produce the demand at day $d + k$ by recursively applying Equation (5.5). Analytically, the expression is derived as follows:

$$\vec{T}_{i,d+k} = (\sum_{z=1}^{k} B^{k-z}) \cdot \vec{C_i} + B^k \cdot \vec{T}_{i,d} + \sum_{z=1}^{k} B^{k-z} \begin{pmatrix} \varepsilon_{i,d+z} \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{5.6}$$

The long-term mean demand has the following relation with the constant $\vec{C_i}$, assuming that $1 - B$ is positive definite $(1 - B > 0)$:

$$\overline{\vec{T_i}} = (1 - B)^{-1} \vec{C_i} \tag{5.7}$$

Data Level with Multiple Data Sources

For the measurement model at the data level, two types of flow data are generated from different devices. First, the link flow data represents the traffic counts on links during a certain time period. Second, the path flow data refers to the traffic counts of vehicles that pass a particular path with multiple links. The path flow data indicates origin-destination information, which may reduce the uncertainty due to the under-specification of the estimation problem.

Loop detectors measuring link flows include anonymous counts. In principle, they cannot distinguish trip chains of vehicles. Denoting the flow observation on link l on day d as $V_{l,d}$, the relation between observed flows and ODTs is linked with the route proportion $A_{l,d,i}$.

$$V_{l,d} = \sum_i A_{l,d,i} T_{i,d} + \zeta_{l,d} \tag{5.8}$$

Path flows are generated by devices that identify vehicles. For instance, cameras track the trajectories of vehicles, which provides information about the traveling routes and even ODT information. Denoting a path as c, the path flow as $W_{c,d}$ and the route proportion as $A_{c,d,i}$, the linear relation between path flow and ODT is expressed in Equation (5.9):

$$W_{c,d} = \sum_i A_{c,d,i} T_{i,d} + \xi_{c,d} \tag{5.9}$$

The error $\xi_{c,d}$ is assumed to be independent from the error $\zeta_{l,d}$ in Equation (5.8).

5.2.2 Posterior Estimation Method with Normal Distribution

After having the hierarchical Bayesian model with relations among the layers, the estimation and forecasting of the posterior ODT is carried out. We assume normal distributions for the errors, so that an analytical approach is effective. The linear relations in the hierarchical Bayesian networks method fit with the linear state-space models. According Koopman et al. (2012), we have the state-space model as follows:

$$\vec{T}_{d+1,i} = \vec{C}_i + B_d\vec{T}_{d,i} + \vec{\varepsilon}_{d,i}, \qquad \vec{\varepsilon}_{d,i} \sim N(0, \vec{Q}_d) \tag{5.10}$$

$$\vec{V}_d = \vec{A}_d\vec{T}_d + \vec{\zeta}_d, \qquad \vec{\zeta}_d \sim N(0, \vec{H}_d) \tag{5.11}$$

Equation (5.11) is called the observation equation. Equation (5.10) is called the state equation, which represents a first order vector autoregressive model, the Markovian nature of which accounts for many properties of the state-space model (Koopman et al., 2012).

In order to update the estimate of the OD demand, a new observation $\vec{V}_d$ is brought in (Koopman et al., 2012). This procedure is called filtering. Kalman filtering is applicable to recursive Bayesian inference when all error terms have a multivariate normal distribution. It operates recursively on streams of noisy input data to produce a statistical estimate of the underlying system state.

Kalman filtering involves both forecasting and updating processes (West and Harrison, 1997). The advantage of the Kalman filtering is to decrease the computational complexity. The whole procedure is illustrated in Figure 5.3 with four steps. The first step is that the observed flows up to day d, $\vec{V}(d)$, are used to estimate the distribution of the demand at day d. This estimated demand at day d is denoted as $\vec{T}_{d|d}$, and serves as prior for day $d+1$. Through dynamic forecasting, the demand at day $d+1$ is forecasted in step 2. In step 3, the demand $\vec{T}_{d+1|d}$ forecasts the flow data at day $d+1$ as $\vec{f}_{d+1|d}$ based on Equations (5.8) and (5.9). $\vec{f}$ is the forecasted flow from the forecasted demand, and differs from the observed flow $\vec{V}(d)$. Once the observed flow at the forecasted day $d+1$ is obtained, there is likely a deviation between the forecasted flow $\vec{f}_{d+1|d}$ and the observed flow $\vec{V}_{d+1}$ in step 4. The deviation and the prior forecasted state at day $d+1$ are used to get the posterior demand, $\vec{T}_{d+1|d+1}$.

Actually, the whole process of updating ODT has the same mechanism as forecasting the flow directly. Since Kalman filtering follows the principle of Markov chains that the next state only depends on the previous state, the updating in each iteration only requires data from the previous period. However, the direct forecasting requires data about the entire history. Thus, the computation time of updating the ODTs in the Kalman filtering framework is much less than the time to update the flows directly.

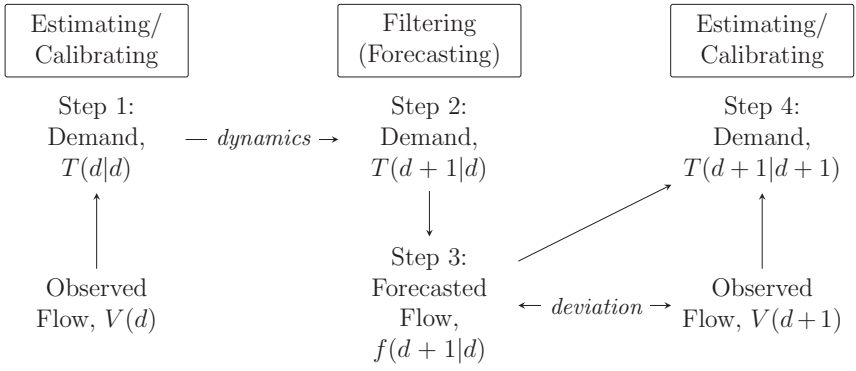


Figure 5.3: Kalman Filter for ODT Estimation and Forecastion at Day Level

Here specifically, the four steps of the Kalman filter updating process (West and Harrison, 1997) are as follows, illustrated also in Figure 5.3.

Step 1: Initializing the posterior at previous days d with a normal distribution, having mean $\vec{m}_d$ and covariance $\vec{\sigma}_d$.

$$\vec{T}_d \sim N(\vec{m}_d, \vec{\sigma}_d)$$

Step 2: Computing the forecasted ODT with mean $\vec{a}_{d+1}$ and covariance $\vec{R}_{d+1}$.

$$\vec{T}_{d+1}|\vec{V}_d \sim N(\vec{a}_{d+1}, \vec{R}_{d+1})$$

where,

$$\vec{a}_{d+1} = B_{d+1}\vec{m}_d$$

$$\vec{R}_{d+1} = B_{d+1}\vec{\sigma}_d B_{d+1}^{\top} + \vec{Q}_{d+1}$$

Step 3: One-step forecast of flow data with mean $\vec{f}_{d+1}$ and covariance $\vec{Q}_{d+1}$.

$$\vec{V}_{d+1}|\vec{V}_d \sim N(\vec{f}_{d+1}, \vec{Q}_{d+1})$$

where,

$$\vec{f}_{d+1} = \vec{A}_{d+1}\vec{a}_{d+1}$$

$$\vec{Q}_{d+1} = \vec{A}_{d+1}^{\top}\vec{R}_{d+1}\vec{A}_{d+1} + \vec{H}_{d+1}$$

Step 4: Posterior at day $d+1$ updating mean $\vec{m}_{d+1}$ and covariance $\vec{\sigma}_{d+1}$.

$$\vec{T}_{d+1}|\vec{W}_{d+1} \sim N(\vec{m}_{d+1}, \vec{\sigma}_{d+1})$$

$$\vec{m}_{d+1} = \vec{a}_{d+1} + \vec{X}_{d+1}\vec{e}_{d+1}$$

$$\vec{\sigma}_{d+1} = \vec{R}_{d+1} - \vec{X}_{d+1}\vec{Q}_{d+1}\vec{X}_{d+1}^{\top}$$

where,

$$\vec{X}_{d+1} = \vec{R}_{d+1}\vec{A}_{d+1}^{\top}\vec{Q}_{d+1}^{-1}$$
$$\vec{e}_{d+1} = \vec{V}_{d+1} - \vec{f}_{d+1} \tag{5.12}$$

5.2.3 Approach to Handle the Evolution Parameters in the Multi-Process Model

The evolution parameters α weigh the demand in past time periods to arrive at a new value. The class of models consisting of a mixture of dynamic linear models is referred to as multi process models, which were originally introduced into the statistics literature by West and Harrison (1997).

The vector of evolution parameters $\vec{\alpha}$ in the multi-process model are constant over time but unknown. We assume to have a prior probability distribution on a finite set Λ of possible values for α. This set of possible evolution parameters, $\Lambda = \{\alpha\}$, is assumed discrete, so as to make the calculation feasible. A realization of this parameter is denoted as $\alpha \in \Lambda$. Given a vector of weight values on the elements of Λ, the model in Equation (5.5) can be analyzed with normal distributions to produce sequences of prior, posterior and forecasted distributions of ODTs that are sequentially estimated over time as the flow observations are processed (West and Harrison, 1997). The means and variances of the distributions all depend on the specific weight values $p(\alpha)$ under consideration. The inference about the ODTs at day $d+1$ is based on the density of the demand given all the available flow observations over historical time, $p(T_{d+1}|\alpha, V_d, V_{d-1}, \ldots)$.

In order to have the density function of the weights α, given all the observed flow data $p(\alpha|V_d, V_{d-1}, \ldots)$, we start with an initial prior density $p(\alpha|V_0)$. Information is sequentially processed to provide inference via posterior $p(\alpha|V_d, V_{d-1}, \ldots)$. This is sequentially updated using Bayes' theorem as per Equation (5.13):

$$p(\alpha|V_d, V_{d-1}, \ldots) \propto p(\alpha|V_{d-1}, V_{d-2}, \ldots)p(V_d|\alpha, V_{d-1}, V_{d-2}, \ldots) \tag{5.13}$$

5.3 Case Study on the A15 Motorway

The multi-process model is tested in a real network of part of the A15 motorway (between entry 17 and exit 15 from east to west) in the Netherlands. There are seven highway sections, four on-ramps as origins and four off-ramps as destinations. Loop detectors are installed on each highway section. Cameras are on highway sections 3, 4, and 6.

Prior demand information is obtained from Statistics Netherlands. The travel survey of individual trip chains in this area is obtained from Statistics Netherlands, including the departure

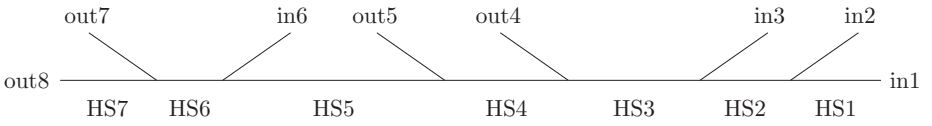


Figure 5.4: Part of A15 motorway from Hoogvliet to Havens

time, travel patterns and so on. The travel pattern information is used to understand the types of trip chains of these travelers within one day. Besides the 14 OD pairs in this network, two types of trip chains are assumed to be present in the survey to run the further experiments: one is in3-out4 and then in6-out7; the other is in3-out5 and then in6-out7. These two types of trip chains are the introduced Origin Destination Tuples. The 14 OD pairs in this case are the simplest case of ODTs. Route proportion A is extended with the ODTs. For instance, considering cameras on highway sections 3 and 6, ODT in3-out4∼in6-out7 is actually part of OD in3-out4 and OD in6-out7. Thus, the route proportion A is as follows:

$$\begin{pmatrix} W_{c_3} \\ W_{c_6} \\ W_{c_3c_6} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} T^{34} \\ T^{67} \\ T^{34,67} \end{pmatrix}$$

The expectation of the prior demand is available, after up-scaling the demand of the trip chains from the sampled data S_i. The variance of the prior demand (Σ_U) is for each ODT. In the case study, we randomly generate the four-day demands as prior demands, assuming the ODT patterns are similar everyday. The evolution variance is assumed as 1 (ε_{d+1}) (Zhou and Mahmassani, 2007). With these 1000-day demand data, the traffic flow of loop detectors and cameras are derived by the linear relations in Equations (5.8) and (5.9). The measurement errors (ζ_{d+1}) of flow data are normal random variables with mean zero and variances 50 and 10 for loops and cameras, respectively. The ratio of the evolution variance of the demand and the measurement variance of the loop flow is 0.02 (1/50), which means that the most recent real-time estimate receives a relatively small weight (Zhou and Mahmassani, 2007). The ratio for cameras is 0.1, which implies that the camera flow data actually plays a role with less randomness and higher accuracy. We also manipulated these variances. The findings are in the discussion.

The constant C_i, the conditional mean of the series, has been set according to its long-term relation with expected demand under stationarity, which has been set equal to prior demand U_i

$$C_i = (1 - \sum_{z=0}^{x-1} \alpha_{z-1}) U_i \tag{5.14}$$

Three scenarios ($\vec{\alpha}$) with weights for demand in the previous four days are considered: non-stationarity, strong auto-correlation and weak auto-correlation. In each case, a similar and a

distinct pattern of weights is applied. Table 5.1 gives an overview. The non-stationary model is with $\sum\alpha = 1$, $C = 0$. For the auto-regressive stationary model with $\sum\alpha < 1$, both a strong auto-correlation model and a weak auto-correlation model are considered. The strong auto-correlation model has $\sum\alpha = 0.8$, and the weak auto-correlation model has $\sum\alpha = 0.2$. Each α in the weak auto-correlation model is one fourth of the corresponding α in the strong auto-correlation model. The prior weights for each scenario are 10%, 40% and 50%, respectively. These weights can be interpreted as the trust levels of these scenarios. For instance, people who believe that scenario 1 is almost impossible to happen, associate it with a probability of 10%.

Table 5.1: Designed Scenarios (α)

Models ($\sum\alpha$)	Cases	Scenario 1	Scenario 2	Scenario 3
Non-Stationarity (1)	Distinct	$(0, 0.1, 0, 0.9)$	$(0.8, 0.1, 0, 0.1)$	$(0.25, 0.25, 0.25, 0.25)$
	Similar	$(0.8, 0.05, 0, 0.15)$	$(0.8, 0.1, 0, 0.1)$	$(0.8, 0.1, 0.1, 0)$
Strong AC (0.8)	Distinct	$(0, 0.08, 0, 0.72)$	$(0.64, 0.08, 0, 0.08)$	$(0.2, 0.2, 0.2, 0.2)$
	Similar	$(0.64, 0.04, 0, 0.12)$	$(0.64, 0.08, 0, 0.08)$	$(0.64, 0.08, 0.08, 0)$
Weak AC (0.2)	Distinct	$(0, 0.02, 0, 0.18)$	$(0.16, 0.02, 0, 0.02)$	$(0.05, 0.05, 0.05, 0.05)$
	Similar	$(0.16, 0.01, 0, 0.03)$	$(0.16, 0.02, 0, 0.02)$	$(0.16, 0.02, 0.02, 0)$

To demonstrate the behavior of the model, scenario 2 is selected as the true scenario to generate demand and flow data. The results should show that scenario 2 has the highest probability. In reality, the true scenario is unknown and unlikely included in the designed scenarios. The results from another scenario (0.1, 0.1, 0.7, 0.1), different from those in Table 5.1, are demonstrated as well. The experiments of the stationary weights within and beyond the designed scenarios are carried out. In addition, the forecast accuracy from different types of detectors is checked by the mean absolute deviation ratio (MADR), defined as follows:

$$MADR = \frac{|ForecastedDemand - TrueDemand|}{TrueDemand} \times 100\% \tag{5.15}$$

5.3.1 Experiment One: Non-Stationary Weights within the Designed Scenarios

To demonstrate the proposed methodology and better fit the assumed distributions, we generate the day-to-day demand of 1000 days with parameters α in Equation (5.5) as scenario 2 (0.8, 0.1, 0, 0.1) under the non-stationary model with $C = 0$ in Equation (5.4). This setting of scenario 2 means that the future demand shares 10% of the demand in the second and fourth previous days, respectively, 80% from the first days and no share from the third day. The initial values of U_0 are obtained from Statistics Netherlands. The initial demand covariance is assumed to be one. The demand volumes are proposed for further tests, as around 2000, 1000, 500, 100, 20. Since the standard deviation is 1, the ratios between demand and standard deviation are 2000, 1000, 500, 100, 20.

Identification of Scenarios

After running the model, scenario 2 wins the highest probability 100%. The convergence is influenced by the ratio between the demand volume and the standard deviation, because the exponential part of the normal density function is sensitive to the ratio. If the ratio is large, the exponential part goes to one very fast. The convergence steps for both the distinct case and the similar case are demonstrated in Table 5.2 for different ratios between demand volume and standard deviation. The model appears to easier identify the right scenario in the distinct case than in the similar case. It takes more steps for the similar case to converge, that is 12 steps for the ratio as 100, and 1000 steps for the ratio as 20. The probability in the similar case with the ratio 20 is plotted in Figure 5.5, demonstrating that the probability in scenario 2 converges to 100% in around 100 steps.

Table 5.2: Convergence Steps to 100% Probability for Distinct Case and Similar Case with Non-Stationarity

	Demand Volumes/Standard Deviation				
	2000	1000	500	100	20
Distinct Case	2	2	2	2	8
Similar Case	2	2	4	12	1000

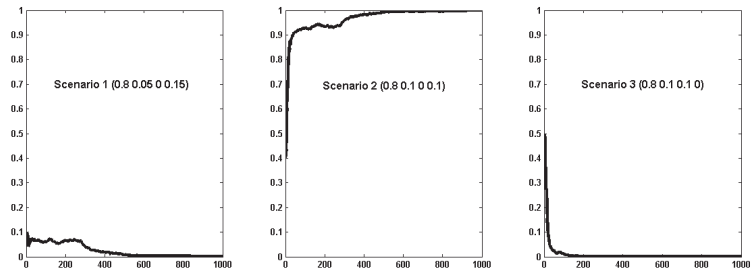


Figure 5.5: Convergence of Weights p in the Similar Case where the Ratio between Demand Volume and Standard Deviation is 20

Forecasted Performance from Different Detectors

To demonstrate the combined use of loop detectors and cameras, the mean absolute deviation ratios (MADR) between the true demand and the forecasted demand are introduced. These MADR values are computed based on the last 200 days of demand after convergence. The true demand is generated with scenario 2. The absolute deviation ratios for both cameras and loops, and only loops are presented in Table 5.3. The table shows that the MADR from the combined

use of cameras and loops, 0.25%, is lower than the ratios from loops, 0.46%. Especially, the last two ODTs gain significantly from cameras. The ratios from loops are 0.55% and 2.10% for the last two ODTs, while 0.13% and 0.20% from cameras plus loops. Comparing the ratios among ODTs under certain detectors, some are higher or lower than others. This may arise from the sensor locations, mapping the observed flows to the demand.

Table 5.3: Mean Absolute Deviation Ratios (MADR) between True (generated from Scenario 2) and Forecasted Demand with Non-Stationarity (|%|)

	Distinct Scenarios		Similar Scenarios	
	Cameras and Loops	Loops	Cameras and Loops	Loops
ODT 1-4	0.22	0.20	0.19	0.24
ODT 1-5	0.32	0.37	0.28	0.51
ODT 1-7	0.02	0.36	0.04	0.08
ODT 1-8	0.29	0.09	0.35	0.09
ODT 2-4	0.41	0.19	0.46	0.52
ODT 2-5	0.03	0.10	0.03	0.12
ODT 2-7	0.06	0.15	0.00	0.36
ODT 2-8	0.24	0.00	0.18	0.17
ODT 3-4	1.22	2.28	1.16	0.88
ODT 3-5	0.06	0.19	0.06	0.15
ODT 3-7	0.05	0.30	0.05	0.10
ODT 3-8	0.09	0.15	0.09	0.06
ODT 6-7	0.06	0.28	0.07	1.04
ODT 6-8	0.52	0.02	0.48	0.20
ODT 3-4∼6-7	0.13	0.55	1.24	2.07
ODT 3-5∼6-7	0.20	2.10	0.20	0.34
MADR (|%|)	0.25	0.46	0.31	0.43

For the similar case, where the true demand is generated from scenario 2, Table 5.3 shows that MADR for the combined use of cameras and loops, 0.31%, is a little bit lower than the ratios from loops, 0.43%. As expected that trip chains can only been identified by cameras, the last two real ODTs have lower ratios when using both cameras and loops, 1.24% and 0.20%, than using only loops, 2.07% and 0.34%, respectively.

There is no significant difference in MADR between the distinct scenarios and the similar scenarios. It means that the model successfully forecasts the demand to a correct scenario. The only difference between the distinct scenarios and the similar scenarios appears to be the convergence steps in Table 5.2. In addition, the MADRs for the situations with and without cameras have no significant difference either. This may arise from the demand being generated from a scenario which is part of the designed scenarios.

5.3.2 Experiment Two: Non-Stationary Weights beyond the Designed Scenarios

In reality, people do not know the true weights for the demand forecast, which are most likely beyond the designed scenarios. Keeping the model settings as in the first experiment, we generate 1000-day demand data with an extra scenario of (0.1, 0.1, 0.7, 0.1). People may believe that the third previous day has 0.7 proportion to forecast demand. The other days contribute 0.1, each.

Identification of Scenarios

After running the model, the weights for scenario 3 converge to unity and for the rest to zero, in both situations with and without cameras, in both distinct and similar scenarios. It means that scenario 3 is identified as the most likely to forecast the demand. The underlying reason is that the Euclidean distance between scenario 3 and the true scenario is the shortest, 0.52. The Euclidean distances from the other two scenarios are 1.07 and 0.99, respectively. Even though no designed scenario is true, scenario 3 with the unit weight probability is able to lead to the ratios with low differences between the true demand with the scenario of (0.1, 0.1, 0.7, 0.1) and the forecasted demand.

Forecasted Performance from Different Detectors

Regarding MADR in the situations with and without cameras, the ratios go to 0.33% in the situation with cameras and 1.34% without cameras, as illustrated in Table 5.4. Comparing the absolute deviations with and without cameras in this table, especially for the ODTs which can be identified by cameras, the forecasted demand can achieve almost the true demand with an absolute deviation ratio, 0.22% for ODT 3-5∼6-7; while there is a relatively large ratio in the situation with only loops, 3.85% from ODT 3-5∼6-7 for instance. The large deviations of MADRs come from the fact that ODTs always rely on cameras. The similar scenarios appear to the same situation in Table 5.4.

Comparing the MADRs of the two experiments, in the case of non-stationary weights within the designed scenarios, even if cameras are able to identify the vehicles, the path flow data cannot achieve significantly more accurate results than only link flow data. In the other situation of the true scenario beyond the three designed scenarios, which is more realistic, the path flow data provide much more accurate forecasts than the loop data. Path flow data from cameras play a significant role to identify the true scenario in estimating the ODTs.

Additionally, the non-stationary dynamics for demand forecasting may generate a drift in the forecasted flow data, illustrated in Figure 5.6, although the drift is not significant. The figure plots the measured flow data with the largest variance, and the forecasted data from three scenarios. The lines of three scenarios are too close to separate. This drift may arise from

Table 5.4: Mean Absolute Deviation Ratios (MADR) between True (generated with Another Scenario) and Forecasted Demand with Non-Stationarity (|%|)

	Distinct Scenarios		Similar Scenarios	
	Cameras and Loops	Loops	Cameras and Loops	Loops
ODT 1-4	0.09	1.15	0.25	0.75
ODT 1-5	0.38	0.84	0.11	0.98
ODT 1-7	0.05	0.90	0.11	0.12
ODT 1-8	0.13	0.22	0.46	0.49
ODT 2-4	0.58	0.45	0.45	0.99
ODT 2-5	0.06	0.27	0.14	0.04
ODT 2-7	0.00	0.59	0.17	0.21
ODT 2-8	0.43	0.46	0.20	0.85
ODT 3-4	1.16	1.97	1.14	1.35
ODT 3-5	0.07	0.79	0.07	0.57
ODT 3-7	0.05	0.59	0.05	0.58
ODT 3-8	0.14	0.21	0.11	0.00
ODT 6-7	0.06	0.72	0.06	0.29
ODT 6-8	0.55	1.06	0.56	1.28
ODT 3-4∼6-7	1.37	7.30	1.38	9.36
ODT 3-5∼6-7	0.22	3.85	0.22	2.88
MADR (|%|)	0.33	1.34	0.34	1.30

the model feature of random walk. The drift may not be realistic in a transport situation, as it would suggest that the flow data at the same location increases day by day.

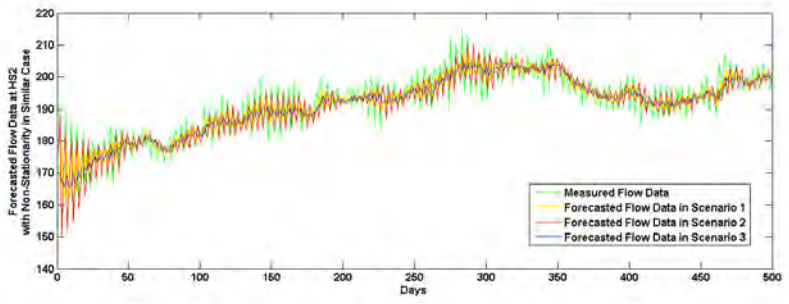


Figure 5.6: One Day Ahead Forecasted Flow Data at HS2 with Non-Stationarity in Similar Case

5.3.3 Dynamics under Strong and Weak Stationarity

The auto-regressive model has stationary weights. We define the strong auto-correlation as $\sum \alpha = 0.8$, and the weak auto-correlation as $\sum \alpha = 0.2$. The constant C in the auto-regressive model is predefined in Equation (5.14).

Identification of Scenarios

The auto-regressive models under strong and weak auto-correlation can find the true scenario 2, when the data is generated from scenario 2. The convergence steps are not as sensitive to the ratios between the demand volumes and the standard deviations, for both the strong and the weak auto-correlation. Table 5.5, from the strong auto-correlation, does not show any gain of reducing the demand volume. Comparing the convergence steps of the distinct case and the similar case, shows that the distinct case has fast and stable convergence. The model can easily identify the correct scenario, independent of the ratio between the demand volumes and to the standard deviations. In the similar case, it takes more steps to converge, and identify the correct scenario. Due to the presence of demand volume in the exponent of the normal distribution, small ratios require more convergence steps. When the true scenario is beyond the designed scenarios, scenario 3 gains the unit probability.

Table 5.5: Convergence Steps to 100% Probability for Distinct Case and Similar Case with Strong Auto-Correlation ($\sum \alpha = 0.8$)

	Demand Volumes/Standard Deviation				
	2000	1000	500	100	20
Distinct Case	6	8	10	14	7
Similar Case	26	39	108	108	106

Forecast Performances for Different Detectors

The forecast performances in both the similar case and the distinct case are quite the same. So the following discussion is based on the similar case, since it is comparatively difficult to find good results. The forecasted flow data based on the strong auto-correlation model is plotted in Figure 5.7, taking the HS2 for instance. The figure includes the measured flow data and the forecasted data in the three scenarios. These lines are too close to separate. But there is no drift in the forecasted flow data as in Figure 5.6. Figure 5.7 shows stable patterns after 180 days, especially for scenario 2. Since the forecasted demand converges to the mean demand of $\overline{T}$ in Equation (5.7), the forecasted flow data will eventually converge to a certain value. The MADR between true demand and forecasted demand are all around $1 * e^{-5}$, irrespective the use of detectors and the scenario cases.

For the weak auto-correlation model ($\sum \alpha = 0.2$), the forecasted flow data from three scenarios are also quite close to each other, because the α in the weak auto-correlation model is one fourth (0.2/0.8) of the α in the strong auto-correlation model. The forecasted flow in the weak auto-correlation model converges faster than the strong auto-correlation model, around 20 steps and 180 steps, respectively. The MADR between true demand and forecasted demand in the

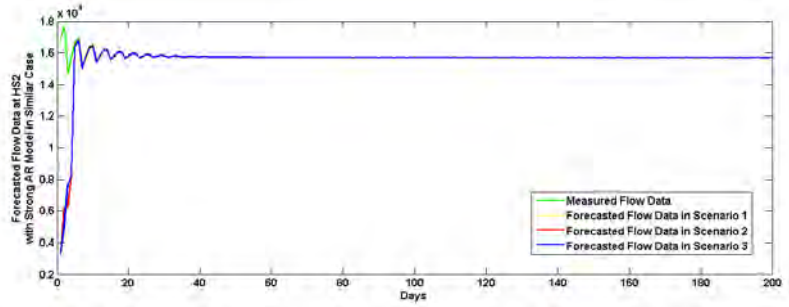


Figure 5.7: One Day Ahead Forecasted Flow Data at HS2 with Strong Auto-Correlation Model ($\sum \alpha = 0.8$)

weak auto-correlation model are all around $1 * e^{-5}$ as well, independent of the use of detectors and the scenario cases.

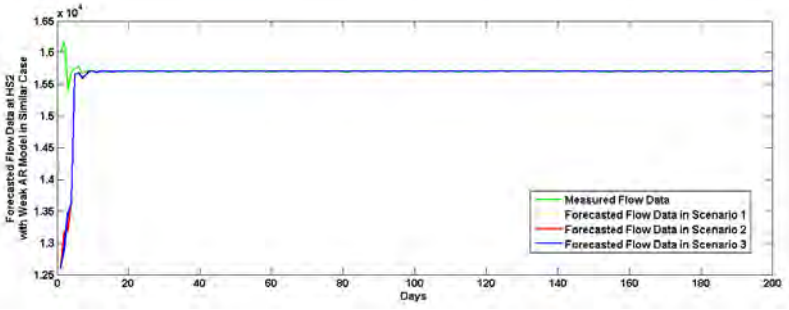


Figure 5.8: One Day Ahead Forecasted Flow Data at HS2 with Weak Auto-Regressive Model ($\sum \alpha = 0.2$)

In summary, the dynamic model with stationarity much easier leads to accurate forecasting results than the model with non-stationarity. The advantages of cameras show only in the non-stationary situation, if the true scenario is beyond the designed scenarios.

5.4 Conclusion

In this chapter, origin destination tuples (ODT) have been introduced and examined. ODTs consider combinations or chains of links in a network, as opposed to individual links, when estimating the parameters of the trip matrix. Conceptually, ODTs bridge the gap between a travel activity model that ignores the road network, and transportation planning where vehicles are anonymous. Demand forecasting has been researched from the perspective of both travel

activity modeling and transport planning. Thus, ODTs forecasting based on statistical analysis explores the research of demand forecasting.

There are three main contributions in this chapter. Firstly, we propose the concept of origin destination tuple as a sequential dependence of the OD matrices, based on the identification function of cameras. This concept bridges transportation modeling, which considers only OD pairs assigned anonymously to the road network, and activity-based model research that focuses on travel choice behavior with discrete choice models (Bowman and Ben-Akiva, 2001; Chorus et al., 2008). The activity based research enriches the trip generation in the transportation planning process. But the activity based model does not touch upon the road network associated with the observed data, which is essential for transportation planning and this thesis. ODTs take care of both travel behavior in the road network and trip chains as one of the travel choices.

Secondly, ODTs bring extra uncertainties, on top of OD pairs, to the under-specification problem (Van Zuylen and Willumsen, 1980) associated with the estimation and forecasting of ODs. Taking advantage of monitoring systems that are able to identify trip chains of vehicles, the path flows from identification devices such as cameras, decrease the uncertainties due to the OD tuples, especially in the situation of non-stationary weights beyond the designed scenarios. Comparing the mean absolute deviation ratios (MADR) of two experiments in the non-stationary weight, when the true scenario is within the designed scenarios, even if cameras are able to identify the vehicles, path flow data cannot achieve significantly more accurate results than only link flow data. In the other situation of the true scenario beyond the three designed scenarios, which is more realistic, the path flow data provide a much more accurate MADR than only the loop data. When the dynamic model is stationary, the forecasting results are so good that the influences of the path flow data from cameras to find out the true scenario cannot be identified. The MADRs are all around $1 * e^{-5}$ with and without cameras. Thus, the advantages of cameras show only in the non-stationary situation, if the true scenario is beyond the designed scenarios.

Lastly, a hierarchical Bayesian network with a multi-process model is suitable to forecast demand. To our knowledge, there is no paper in this field published on this method. The dynamics include the non-stationary model, the strong auto-correlation model and the weak auto-correlation model. The multi-process model in the Baysian framework has the advantages of taking the demand in several previous days into account instead of just-one-day before demand (Ashok and Ben-Akiva, 2000; Zhou and Mahmassani, 2007), giving each pre-defined scenario of combinations of the previous days demand a prior probability, and coming up with posterior probabilities for each scenario. Usually, one scenario, which has the shortest Euclidean distance with the true scenario, gains a unit probability, and the rest has zero probability. In the case of having the same designed scenarios, they will share the unit probability with the same proportions of the prior probabilities. With this multi-process model, people's experience is taken into account in the pre-defined scenario. The observed data together with errors update the prior probability and end up with the most likely scenario. The multi-process model with the

non-stationary model has its own weakness as well. Firstly, the probability converges within two steps, when the ratio between demand volume and standard deviation is more than 100. This results from the large numbers for the power of the exponent in the normal distribution. As long as the demand volume gets smaller, convergence is getting smooth as indicated in Figure 5.5. Secondly, the model sometimes takes a large number of iterations to converge, as was also indicated in West and Harrison (1997), say 2000 days. This depends on the combination of the variances of observed data and the scenarios. Lastly, the non-stationary model of demand forecasting may generate a drift in the forecasted flow data, arising from the model feature of random walk. This drift may not be realistic in the transport situation that the flow data at the same location increases day by day. The multi-process model with a stationary-dynamics model has none of these issues. The dynamic model with stationarity much easier leads to accurate forecasting results than the model with non-stationarity.

Chapter 6

Conclusions and Recommendations

Traditionally, official traffic and transportation statistics are to a large extent based on data obtained through (online) surveys, secondary data sources and administrative databases. With the advent of large scale electronic transactions, telecommunication and widespread use of sensors, opportunities arise to base statistics on captured data. Advanced monitoring systems for these captured data are installed in various types of infrastructures. Particularly in the road network, data from loop detectors and cameras offer abundant and detailed transport and traffic information. In the Netherlands, the national data warehouse is responsible for storing these captured data, and provides the data to data users. Consistent with the original purpose of the detected data, these data are applied to research in the field of transportation planning. Meanwhile, these advanced monitoring data enrich the data resources and data collection methods applied by statistics offices.

Transportation demand management based on monitoring data is an example of a link between transportation planning and official statistics. It is particularly interesting for a statistics office to know the annual number of freight vehicles travelling from specific origins to specific destinations. This overlaps with the interest of transportation planning, in which demand data is used to understand the transportation behavior in the road network. Currently, the statistics office gets information about origins and destinations from questionnaires sent to logistics companies, while transportation planning estimates and predicts the demand based on observations about the road network. Statistics office and transportation planning have different perspectives of the same issue. To combine these two perspectives, this thesis applies the Kullback-Leibler method and the hierarchical Bayesian network model, taking survey data as prior information and monitoring data as evidence.

The main findings of the thesis are summarised in the next section. In section 6.2, conclusions are drawn from the use of multiple data sources. Following that, the evaluation of demand estimation and forecasting methods is summarized in section 6.3. In section 6.4, recommendations of data usage are addressed for Statistics Netherlands. Future research finalizes this chapter.

6.1 Main Achievements in Each Chapter

In Chapter 1, various types of transportation data are introduced, besides survey data from statistics offices. Based on the characteristics of the different data sources, transportation data are grouped into three categories: (i) point data, such as loop detectors and Weigh-in-Motion (WiM), providing counts information; (ii) point-to-point data or path data, such as cameras, Bluetooth and GSM, giving path data including travel time and OD information; (iii) and route data, such as GPS having counts, travel time, OD information and vehicle trajectory. To better capture the transport behavior, these data can be used jointly. For instance, due to location restrictions, the Weigh-in-Motion systems in the Dutch highway network mainly provide the

transport cargo weight data for each passed vehicle. The system cannot present the network information, although cameras are involved in the WiM systems. In contrast, loop detectors cannot identify trucks, but they do give a general impression of truck flows and traffic flows. The flows from loop detectors are representative for the network. Assuming that the truck types in the network are similar to the truck types in the WiM locations, we can have the average cargo weight per truck per hour.

In Chapter 2, demand estimation methods have been reviewed. There are two approaches: point estimation and distribution estimation. Maximum likelihood, least squares and information methods belong to the point estimation approaches; Bayesian inference and Kalman filtering are the distribution estimation approaches. The main advantage of distribution estimation is that the stochastic features of flows and demand are taken into account. Bayesian inference updates the prior information to obtain a posterior belief. Additionally, to obtain the route proportion with the combined use of cameras, the Hadamard product, via element-wise matrix multiplication, is proposed based on the route proportion of link flows.

In Chapter 3, the Kullback-Leibler divergence model is proposed to estimate OD demand, as an alternative to the information minimization method (Van Zuylen and Willumsen, 1980). The Kullback-Leibler divergence model relaxes Stirling's approximation in the information minimization method, and can be regarded as a generalization of the information minimization method. Compared with previous research considering one or two types of data sources (Van Zuylen and Willumsen, 1980; Watling, 1994; Ashok and Ben-Akiva, 2000; Van Der Zijpp, 1997; Zhou and Mahmassani, 2006), we take three data sources: link flow from loop detectors, path flow from cameras, and prior demand from Statistics Netherlands. The effectiveness of path flow data from cameras has been demonstrated in the A15 case study. Additionally, a high quality of the prior demand is essential in the Kullback-Leibler divergence model, if flow data can not perfectly estimate demand. As long as the flow data is sufficiently accurate, taking path flow for instance, the demand estimation error remains 0.04% in the A15 case study, independent of the prior demand.

In Chapter 4, the hierarchical Bayesian networks are introduced to estimate OD demand, taking the stochastic features of the freight system into account. The prior demand in the framework of hierarchical Bayesian networks associated with a relatively large variance, called uninformative prior (Tibshirani, 1989), appears to have little impact on the estimated demand. The impact has been dominated by the flow observations. In this sense, hierarchical Bayesian networks have more advantages than the information methods in Chapter 3. Additionally, hierarchical Bayesian networks can include various types of distributions. Two distributions of errors are proposed in this chapter: the normal distributions for the additive errors and the log-normal distributions for the multiplicative errors. To calculate or approximate the posterior density function of the demand, there are two approaches. One is an analytical approach applied to the normal distributions, getting the mean and covariance of the posterior demand; and the

other is a simulation approach associated with Gibbs sampling nested by the Metropolis-Hastings sampling applied for the situation with the log-normal distributions. The case study of the A15 motorway in the Netherlands demonstrates, for the normal distribution case, that there is 90.21% trace reduction of the estimated demand error covariance and 61.09% trace reduction of flow prediction error covariance, from the worst case with three loops installed to the best case with full coverage of cameras and loops. For the log-normal distribution case, these two values are 77.38% and 75.08%, respectively. Thus, hierarchical Bayesian networks associated with both distributions have been shown to be able to achieve significant trace reductions. Furthermore, three types of errors are considered: prior errors, estimation errors and observation errors. The prior error represents the error in the prior data from survey, for instance. The estimation error is the error generated when estimating demand parameters. The observation error, also called measurement error, results from disruptions of the devices. Usually, errors modeled in research, such as Zhou and Mahmassani (2006), are additive. The assumption of additive errors is that, whatever the error generated, the errors stay the same. In our model of log-normal distributions, multiplicative errors are proposed, which actually relaxes the underlying assumption. These multiplicative errors represent the proportion of the deviation to the baseline, for instance, the percentage failure of each device for observation errors. The multiplicative errors associated with log-normal distributions could better represent the errors to each device, and may be suitable for the situation of a small demand with a large variance.

Chapter 5 contains three main contributions. Firstly, we propose the concept of origin destination tuple as a sequential dependence of the OD matrices, based on the identification function of cameras. This concept bridges transportation modeling, which considers only the OD pairs assigned anonymously to the road network, and activity-based model research that focuses on travel choice behavior with discrete choice models (Bowman and Ben-Akiva, 2001; Chorus et al., 2008). The activity based research enriches the trip generation in the transportation planning process. But the activity based model does not touch upon the road network associated with the observed data, which is essential for transportation planning and this thesis. ODTs take care of both travel behavior in the road network and trip chains as one of the travel choices. Secondly, ODTs bring extra uncertainties, on top of OD pairs, to the under-specification problem (Van Zuylen and Willumsen, 1980) associated with the estimation and forecasting of ODs. Taking advantage of monitoring systems that are able to identify trip chains of vehicles, the path flows from identification devices such as cameras decrease the uncertainties due to the OD tuples, especially in the situation of non-stationary weights of previous demand in the dynamic model beyond the designed scenarios. Comparing the mean absolute deviation ratios (MADR) of two experiments with the non-stationary weights, when the true scenario is within the designed scenarios, even if cameras are able to identify the vehicles, path flow data cannot achieve significantly more accurate results than only link flow data. In the other situation where the true scenario is beyond the three designed scenarios, which is more realistic, the path flow

data provide much more accurate MADR than only the loop data. But the path flow from cameras does not play a significant role to identify the true scenario when the dynamic model is stationary. The MADRs are all around $1 * e^{-5}$ with and without cameras. Thus, the advantages of cameras only show in the non-stationary situation, if the true scenario is beyond the designed scenarios. Lastly, the hierarchical Bayesian network with a multi-process model is suitable to forecast demand. To our knowledge, there is no paper in this field published on this method. The dynamics include the non-stationary model, the strong auto-correlation model and the weak auto-correlation model. The multi-process model in the Baysian framework has the advantage of taking the demand in several previous days into account instead of just-one-day before demand (Ashok and Ben-Akiva, 2000; Zhou and Mahmassani, 2007), giving a prior probability to each pre-defined scenario of a combination of the previous days demand, and coming up with posterior probabilities for each scenario. Usually, one scenario, which is the shortest Euclidean distance from the true scenario, gains a unit probability, and the rest have zero probability. If they have the same designed scenarios, they will share the unit probability with the same proportions as the prior probabilities. With this multi-process model, people's experience is taken into account in the pre-defined scenario. The observed data together with errors updates the prior probability and end up with the most likely scenario.

6.2 Use of Multiple Data Sources

Data can be collected in various ways. Traditional surveys consist of rich information with many variables. Taking surveys for logistics companies, for instance, there are data about departure time, arrival time, location, travel mode, cargo weight, and so on. Capturing data via surveys is not limited to locations. Location information can be easily obtained in questionnaires. But surveys require man-power and time to process data. Nowadays the on-line surveys by Statistics Netherlands achieve a response rate around 80%, but questionnaires may be incomplete. Compared with surveys, advanced monitoring systems capture data automatically during the whole day. The data capturing error is much lower than the error from surveys. However, the advanced monitoring system may provide limited types of data. Loop detectors can provide link flow data, time mean speed and lengths of vehicles. Cameras with an identification function can capture more information, but the data is limited to number plates and path flow data. There are only eighteen Weigh-in-Motion (WiM) detectors installed in the Dutch road network, obtaining the weight of vehicles. All these monitoring systems provide only specific information at specific locations. Advanced monitoring systems can never replace surveys. Combining survey data and monitoring data extends and enriches the current data collection methods. Surveys collect data with various variables and are not bound to a particular location, but may be incomplete. Advanced monitoring systems collect data continuously at specific locations, but have few variables.

Integrating these data sources allows one to extend the use of transport data and to obtain more insight into transport behavior. Due to the location restrictions, the WiM systems in the Dutch highway network mainly provide the transport cargo weight data for each passed vehicles. The systems cannot present network information, although cameras are involved in the WiM systems. In contrast, loop detectors cannot identify trucks, but they do give a general impression of truck flows and traffic flows. The flows from loop detectors are representative for the network. Assuming the truck types in the network are similar to the truck types on WiM locations, joint information from loop detectors and WiM provides the average cargo weight per truck by hours of the day and days of the week, and also gives the ratio of weight and truck flow. In Figure 1.2, the ratio of weight and truck flow is much higher during the period from 20:00 to 05:00 am the next day than for the rest of the time. The patterns are a consequence of the fact that during daytime, a range of commercial vehicles, including vans, light trucks and lorries, travel on the roads, while during the evening hours and night time, there are mainly heavy lorries with full loads, such as international transport from and to the Port of Rotterdam. The occupancy rate of trucks on the A15 highway from the port area to the hinterland is high during the early times of the day (00:00-07:00) and then gradually lowers as the network presence of passenger cars increases.

Among these monitoring systems, different data capturing methods provide different types of data in the road network. Two types of captured traffic flow observations can be distinguished. One is the link flow observed from loop detectors and Weigh-in-Motion. The other is path flow obtained with cameras and Bluetooth scanners. Link flow data only captures anonymous data, while path flow data identifies vehicles. Path flow data considerably increase the accuracy of demand estimation based on loop detectors. In addition, the identification function of path flows allows to bring a concept of origin destination tuple. Traditionally, vehicles traveling along an OD pair to fulfill demand are assumed to be homogeneous and the trips of vehicles are taken as independent. In reality, vehicle trips are inter-related. One vehicle may contribute to time dependent OD matrices several times a day, according to its schedule or travel plans. Origin destination tuples, defined as a set of OD pairs, take travel behavior into account and quantitatively represent trip chains from the demand aspect. With the help of cameras and Bluetooth scanners, the origin destination tuples of the identified freight trucks can be measured. The innovative concept of origin destination tuples could be obtained with survey data, but hardly with link flow observations. The combination of survey data and path flow data demonstrates the trip chains of freight trucks.

6.3 Evaluation of Demand Estimation and Forecasting Methods

To estimate demand, information methods, including the information minimization and the Kullback-Leibler method, and the hierarchical Bayesian networks have been applied. Hierarchical Bayesian networks with a multi-process model have been investigated to forecast demand.

6.3.1 Demand Estimation Methods

Two main methods have been applied to estimate the truck demand for infrastructure usage. One is the Kullback-Leibler divergence method, and the other is the hierarchical Bayesian network method. The Kullback-Leibler divergence method is an information method. Compared with the information minimization approach (Van Zuylen and Willumsen, 1980), the Kullback-Leibler divergence method better represents the underlying concept of information minimization for estimating the OD demand and relaxes the assumption implied by Stirling's approximation. Stirling's approximation is feasible when flow data are large. When the flow data are small, less than 10, the Kullback-Leibler divergence method has lower average deviations between the estimated demand and the ground truth demand than the information minimization method. The Kullback-Leibler divergence model is regarded as a generalized approach of the information minimization method.

Hierarchical Bayesian networks take the stochastic nature of freight truck demand and freight truck flows into account. This is more realistic than the assumption that demand and flow are deterministic. Using flow observations in the road network, the prior information is updated to yield the posterior demand. The demand information from official statistics is taken as the prior information. Two situations have been considered. If flow errors are independent of devices, errors are additive to flows; then a normal distribution is applied, which allows one to adopt an analytical approach and to quickly obtain the posterior demand. If flow errors are considered to be dependent of devices, such as different devices having different error percentages, errors are proportional or multiplicative to flows; then a log-normal distribution is applied. A log-normal distribution gives a plausible description of a right skewed flow distribution, in which large flows have a relatively high probability of occurrence. In this case, an analytical solution is not possible. Instead, Markov-chain Mont-Carlo simulation is applied with Gibbs sampling, nested by Metropolis-Hastings sampling to arrive at estimates.

A comparison of the hierarchical Bayesian networks and the Kullback-Leibler divergence method for freight truck demand estimation yields several insights. Firstly, the framework of the hierarchical Bayesian networks is flexible. It can be extended as necessary with additional variables; while the Kullback-Leibler divergence method measures the probable divergence of two distributions.

Secondly, the Kullback-Leibler divergence method considers a discrete probability distribution, which is characterized by a probability mass function that a truck from an OD trip contributes to the flow on a link. The Bayesian method takes the stochastic nature of truck behavior into account, which involves continuous probability distributions of demand and flows. The probability density functions are used to represent the distributions, rather than having deterministic values, as assumed in the information methods. In addition, prior errors, estimation errors and observation errors are explicitly considered in the hierarchical Bayesian networks, making the approach more realistic; while information methods including the Kullback-Leibler divergence method usually do not take stochastic errors into account.

Thirdly, the Kullback-Leibler divergence method and the hierarchical Bayesian network method have different mechanisms to estimate truck demand. Hierarchical Bayesian networks provide the posterior demand via updating a prior demand based on flow observations in the road network. This takes the network structure, measurement errors, assignment errors and the correlations among flows and demand explicitly into account; while the Kullback-Leibler divergence method, assuming independent link flows, minimizes the divergence between the probability of link flows from the prior demand and from the estimated demand. Thus, hierarchical Bayesian networks have realistic settings with less assumptions. In hierarchical Bayesian networks, the quality of the prior information appears to have limited relevance for estimating freight truck demand. The results of estimated demand rely mainly on the flow observations in the road network, the variation of the flow data, and the variation of the prior demand. The Kullback-Leibler divergence method treats the weights of prior demand and posterior demand as equal, where prior information plays a significant role. A wrong chosen prior demand results in inaccurately estimated demand. Thus, the under-specification problem is covered by the variance of prior data in hierarchical Bayesian networks. Updated by the observed data, the impact of the under-specification problem can be limited.

Lastly, a Bayesian method produces more accurate results than an information method. The hierarchical Bayesian networks associated with normal distributions is able to get the zero mean deviation between the estimated demand and the ground truth demand, while the heuristic for the Kullback-Leibler divergence method associated with a heuristic approach has no guarantee to reach optimal values.

6.3.2 Demand Forecasting Method

The stochastic nature of traffic demand and traffic flows is also considered in the context of demand forecasting. In order to perform day-to-day forecasting, we investigate a multi-process model, in the context of hierarchical Bayesian networks. A hierarchical Bayesian network with a multi-process model is suitable to forecast demand. To our knowledge, there is no paper in this field published on this method. This model has its own advantages of taking the demand in several previous days into account instead of just-one-day before (Ashok and Ben-Akiva, 2000;

Zhou and Mahmassani, 2007). Each pre-defined scenario of the combination of the previous days demand has a prior probability, and the model comes up with posterior probabilities for each scenario. Usually, one of the scenarios gains a unit probability, and the rest has zero probability. In the case of the same scenarios, they will share the unit probability with the same proportions of the prior probabilities. In such a way, people's experience in the pre-defined scenario and the observed data including errors are involved in the multi-process model.

The multi-process model has its own weakness as well. Firstly, the probability converges within two steps, too fast to be believed, when the ratio between demand volume and standard deviation is more than 100. This results from the large numbers in the exponent in the normal distribution. When scaling down the demand, convergence is getting smooth as is indicated in Figure 5.5. Secondly, the model sometimes takes a large number of iterations to converge, as is also indicated in West and Harrison (1997), like 2000 days. It depends on the combination of the variances of observed data and scenarios. Lastly, the auto-regressive model for the demand forecasting may generate a drifting in the forecasted flow data, illustrated in Figure 5.6. It may not be realistic in the transport situation that the flow data at the same location increases day by day. An autoregressive model with a constant is applied then. This constant is able to stabilize the forecasting process, avoiding the drifting. The absolute deviation ratios between true demand and forecasted demand are almost zeros, independent of the use of detectors and the scenarios.

6.4 Recommendations for Statistics Netherlands

This thesis introduces different kinds of advanced monitoring data, such as loop detectors, cameras, Weigh-in-Motion (WiM), and GSM. Regarding the characteristics of these data, loop detector data and Weigh-in-Motion data are point data, providing counts information; cameras, bluetooth and GSM give path data, including travel time and OD information; and GPS represents route data having counts, travel time, OD information and vehicle trajectory. Based on these modern information sources, Statistics Netherlands can publish statistics on, for example, traffic flow and traffic density. Since the time intervals of the captured monitoring data are short, Statistics Netherlands could publish short term data or near real-time information, in addition to the annual, quarterly or monthly reports.

An automated data capturing approach may reduce the administrative burden for transport companies, but can never replace the traditional surveys. The surveys for logistics companies, for instance, provide information about departure time, arrival time, location, travel mode, cargo weight and so on. Advanced monitoring system is mostly restricted to few variables. Loop detectors, for example, limit to provide link flow data and length categories of vehicles. The advanced monitoring systems thus merely enrich the traditional data capturing approach. Combining survey data and monitoring data extends the data collection methods, as surveys collect data with

various variables and advanced monitoring systems collect data continuously. This opens the opportunity for Statistics Netherlands to publish new variables, such as transportation demand as the number of incoming and outgoing trucks per day per municipality to be published each month. Transportation demand can be partially obtained from surveys, but monitoring data in the network allow to estimate transportation demand on a daily basis. Combining different demand sources, Statistics Netherlands can have transportation demand statistics. In 2015, Statistics Netherlands published an article (van der Sangen, 2015), demonstrating the monthly traffic intensity on Dutch motorways over the period 2011-2014. It is a great step for Statistics Netherlands to have published statistics based on minute data from more than 20 thousand loop detectors.

Weigh-in-Motion (WiM) is a particular data source that Statistics Netherlands might exploit. Weigh-in-Motion data include the variables vehicle type, weight, speed, number of axles, and number plate. Based on these data, Statistics Netherlands could report on the frequency, average speed, and average weight of each type of vehicles. Derived from Weigh-in-Motion data, the indicators of dynamic weight capacity utilization and transportation impact cost are introduced and validated in Ma et al. (2012). These two novel indices can be used for the quantitative evaluation of the transportation performance from multiple perspectives, including the performance of logistics companies and the performance of the road network. Dynamic weight capacity utilization draws on the notion of momentum, comparing the observed and ideal situation of cargo weight and traffic speed. Transportation impact cost fuses weight, transportation value and traffic conditions to present the transportation inefficiency. Based on these indicators, Statistics Netherlands could generate the corresponding statistics, which may give insight into the road performance from a logistics point of view and the transportation efficiency of logistics companies. Meanwhile, there are some restrictions of capacity utilization based on WiM data. The WiM data does not reveal the transported cargo (or volume) capacity utilization. This has two consequences. First, truck flows cannot be translated into commodity flows and corresponding statistics. Second, though weight capacity utilization may be measured, it will be difficult, if possible at all, to measure volume capacity utilization.

Additionally, from the data processing point of view, the hierarchical Bayesian networks method enriches the methodologies applied by Statistics Netherlands. It supports Statistics Netherlands to have short term publications by forecasting short term variables.

Furthermore, the large amount of advanced monitoring data may require visualization tools, a large capacity IT infrastructure and data scientists. For the further development of official statistics, this is a point of particular interest for Statistics Netherlands.

6.5 Future Research

There are several extensions to the research in this thesis. Firstly, the real flow data in a larger road network can be used to extend the case study. Based on the real flow data, the hierarchical Bayesian networks with normal distributions and log-normal distributions can be compared. The parameters in the model, such as variances of flows and variances of prior demand, are fixed. This assumption can be relaxed using the data-driven Bayesian network approach based on captured data. Consequently, standard deviations of normal/log-normal distributions will follow certain distributions. In addition, dynamic OD demand estimation (Antoniou et al., 2015) with stochastic route choice could be considered (Ashok and Ben-Akiva, 2000; Dixon and Rilett, 2002; Zhou and Mahmassani, 2006; Barceló et al., 2010), which relaxes the assumption of static assignment in this thesis. The framework of hierarchical Bayesian networks with another dimension of time is feasible, as shown in Chapter 5. This extension is useful for real time transportation planning and for having short term statistics.

Secondly, for the research on origin destination tuples, in this thesis it is assumed that the ODT patterns are known. This assumption can be relaxed, extending to the case that the ODT patterns are being updated over time with real time capture data such as GPS data. Path prediction (Yavaş et al., 2005; Krumm and Horvitz, 2006; Morzy, 2006, 2007; Jeung et al., 2008) associated with pattern-based prediction methods can be applied. Pattern-based prediction methods consider the movement of each object to be independent. Yavaş et al. (2005) propose a method to predict future paths based on mobility patterns. Jeung et al. (2008) propose a hybrid prediction model which combines motion functions and mobility patterns.

Thirdly, short term forecasting statistics, such as demand forecasting combined with weather forecasting, could be generated. It may provide useful information for travellers. During the process, a hierarchical Bayesian network method associated with large capacity IT infrastructures should be implemented. In addition, surveys from Statistics Netherlands, secondary data sources and internet have lots of text information or non-numerical data, such as comments and descriptions of freight products. Those unstructured data contain rich information as well. Text analytics could be applied to extract the information from sentences. One simple example is that text analytics counts the records of words in text and provides the correlations among words. In such a way, Statistics Netherlands can capture extra information.

Fourthly, based on such rich information, many research topics can be carried out. Besides demand estimation and forecasting in this thesis, three researches have been conducted: time slot allocation (Ma et al., 2010), traffic density estimation (Ma et al., 2011), and dynamic weight capacity utilization (Ma et al., 2012). These three topics align with the concept of using multiple data sources. Time slot allocation provides a strategy to reduce congestion via early or late departures of travellers. The travellers are paid a certain incentive, represented as value of time. One practical question arises: how to ensure that these travellers follow the agreement? Identification devices, such as cameras or Bluetooth, can be applied. In addition,

dynamic weight capacity utilization combines the use of WiM systems and loop detectors. For the logistics companies, transport volume, besides weight, is another essential element. How to automatically measure the volume utilization is quite challenging. Translating weight to volume requires product information.

Lastly, the concept of big data nowadays has been discussed a lot. From a pure data point of view, big data is related to three high V's. They are high volume, high velocity, and high variety. My research touched upon various transport sensor data, such as loop detectors, cameras, Bluetooth, and Weigh-in-Motion. Usually this data has several gigabytes or even terabytes, which can be considered as high volume. Since all these sensors capture data every minute or even every second, the data is updated very fast with high velocity. Together with the variables from survey data, transport sensor data is provided in several formats, which gives high variety. Together with traditional data processing, nowadays big data is also concerned with IT infrastructures such as the cloud and data warehouses, and visualizations. The research described in this thesis is not related to these two aspects; it mainly contributes to data science in terms of methodology. Data science bridges the gap between traditional IT and business, providing the solutions to the business based on data processing methods. Furthermore, from a data structure point of view, there are structured data or numerical data, and unstructured data or text data. My thesis mainly contributes to structured data analytics. As an extension, content analytics could be applied. For instance, drivers may send messages saying that they are departing and which routes they will go. This is all real time and accurate information, which may help to improve the accuracy of demand estimation. The text mining methods usually include entropy analysis and Bayesian networks. These methods are considered in this thesis.

Summary

Traffic and transportation statistics are mainly published as aggregated data, traditionally based on surveys or secondary data sources like public registers and companies' administrations. Nowadays, advanced monitoring systems are installed in the road network, offering more abundant and detailed transport information than surveys and secondary data sources. Usually, these rich data are applied to the research in the field of transportation planning. It gives an opportunity to national statistics offices to update their databases and apply new methods to generate statistics. Transportation demand estimation and prediction are taken as examples. Quantitative information on transportation demand is important for national and regional policy makers who want to know the number of freight vehicles traveling from origins to destinations. Traditionally, they extract this information largely from the national statistics offices. Moreover, transportation research needs the demand data to understand transportation behavior in the road network, such as congestion and pollution. Usually, statistics offices get the demand information from questionnaires sent to logistics companies, and transportation researchers estimate the demand based on the observations in the road network.

The contributions of this dissertation are as follows. First, this dissertation considers fusing different data sources. Flow observations are distinct as two types. One is the link flow observed from loop detectors and Weigh-in-Motion, for instance; the other is the path flow from cameras and Bluetooth scanners which can identify vehicles. The path flow can significantly increase the accuracy of the estimated demand, since it reduces the uncertainty of the match between the flow observation and the demand.

In addition, an innovative concept of origin destination tuples is proposed and validated. Since cameras and Bluetooth scanners can identify vehicles, they bring the opportunity to get an insight into the trip chains of freight trucks. The trip chains are quantitatively represented from the demand aspect as origin destination tuples, sets of origin destination pairs. The innovative concept of origin destination tuples can be found in the survey data, but hardly from the link flow observations. The combination of survey data and the path flow demonstrates the trip chains of freight trucks.

Furthermore, the Kullback-Leibler divergence method is proposed to generalize the information minimization method and to relax the assumption from Stirling's approximation. The hierarchical Bayesian method takes the stochastic nature of demand and flows into account.

This is more realistic than the deterministic values of demand and flow. The demand information from Statistics Netherlands is taken as the prior information with a certain distribution. Using updates from the flow observations in the road network as evidence, the posterior demand is obtained for two situations. The first situation assumes that the errors follow a normal distribution. This assumption leads to an analytical approach, which can be applied to quickly obtain the posterior demand. As the symmetric shape of the normal distribution may underrepresent the probability of a large flow, the other situation of the log-normal distribution is also discussed. In this case, an analytical approach is no longer feasible. When the Markov Chain Mont Carlo simulation is applied with Gibbs sampling, nested by Metropolis-Hastings sampling, the computations to reach an equilibrium take lots of time, which makes the method practically infeasible.

The last chapter describes how the hierarchical Bayesian network combined with a multi-process model is applied to forecast demand. To our knowledge, there is no paper in this field published on this method. The multi-process model in the Bayesian framework has the advantage of taking the demand of several previous days into account instead of just-one-day before demand, giving each pre-defined scenario of a combination of the previous days demand a prior probability, and coming up with posterior probabilities for each scenario. Usually, one scenario, which has the shortest Euclidean distance with the true scenario, gains a unit probability, and the rest has zero probability. With this multi-process model, people's experience is taken into account in the pre-defined scenario.

Nederlandse Samenvatting (Summary in Dutch)

Verkeer en vervoersstatistieken worden voornamelijk gepubliceerd als geaggregeerde gegevens die in het verleden werden verkregen met enquêtes en secundaire databronnen zoals overheidsregisters en bedrijfsadministraties. De geavanceerde monitoring systemen die tegenwoordig geïnstalleerd zijn in het wegennet bieden meer en gedetailleerdere informatie over verkeer en vervoer dan enquêtes en secundaire databronnen. Gewoonlijk worden deze rijke data gebruikt voor onderzoek op het gebied van transportplanning. Ook voor statistische bureaus, zoals het Centraal Bureau voor de Statistiek in Nederland, bieden ze een uitgelezen mogelijkheid om de bestaande databronnen verder uit te breiden en aan te vullen, en om nieuwe methoden voor het maken van statistieken te ontwikkelen. Een voorbeeld is de toepassing van deze gegevens voor het schatten en voorspellen van de vervoersvraag. Kwantitatieve informatie over de vervoersvraag is van belang voor nationale en regionale beleidsmakers die inzicht willen hebben in het aantal vrachtwagens met hun oorsprong en bestemming. Deze informatie ontlenen zij van oudsher grotendeels aan de statistische bureaus. Bovendien zijn schattingen van de vervoersvraag belangrijk voor transportonderzoek dat zich richt op analyse van verplaatsingsgedrag in het wegennet voor inzicht in, onder andere, congestie en vervuiling. Statistische bureaus krijgen hun informatie over de vraag meestal op basis van vragenlijsten verzonden naar logistieke bedrijven, terwijl transportonderzoeken de vraag schatten op basis van waarneming van het wegennet.

De bijdragen van deze dissertatie zijn de volgende. Allereerst, behandelt het proefschrift het combineren van vervoersgegevens uit verschillende bronnen. Er zijn twee soorten waarnemingen van de vervoerstroom in het wegennet. De ene is de linkstroom die wordt waargenomen met lusdetectoren in het wegdek en Weigh-in-Motion installaties; en de andere is de padstroom die wordt waargenomen met camera's en Bluetooth scanners, die voertuigen kunnen identificeren. Het gebruik van informatie over de padstroom kan tot een aanzienlijke verhoging van de nauwkeurigheid van de geschatte vraag leiden, omdat het de onzekerheid over de fit tussen de stroomwaarneming en de vraag vermindert.

Daarnaast introduceert en valideert het proefschrift een innovatief concept van Herkomst-Bestemming-Tupels. Aangezien camera's en Bluetooth scanners voertuigen kunnen identificeren, bieden ze de mogelijkheid om inzicht te geven in de routeketens van vrachtwagens. Deze rou-

teketens worden kwantitatief uitgewerkt als Herkomst-Bestemming-Tupels, verzamelingen van herkomsten en bestemmingen. Het innovatieve concept van Herkomst-Bestemming-Tupels kan worden uitgewerkt met enquêtegegevens, maar nauwelijks door een koppeling van stroomwaarnemingen. De combinatie van de enquêtegegevens en de padstroom maakt de routeketens van vrachtwagens zichtbaar.

Verder presenteert het proefschrift de Kullback-Leibler methode als generalisatie van de informatieminimaliseringsmethode, waarmee de aanname van Stirling's formule wordt omzeild. Toepassing van de hiërarchische Bayesiaanse methode houdt rekening met de stochastische aard van de vervoersvraag en -stromen. Deze aanpak is realistischer dan de deterministische modellering van de vraag in bestaande informatieminimaliseringsmethoden. De vraaginformatie gepubliceerd in officiële statistieken wordt gebruikt als priorinformatie met een bepaalde verdeling. De posterior vraaginformatie wordt verkregen door het updaten met waargenomen stroomgegevens over het wegennet. Hierbij worden twee situaties onderscheiden. De eerste situatie is gebaseerd op de veronderstelling dat fouten normaal verdeeld zijn. Deze veronderstelling leidt tot een analytische oplossing die snel tot de berekening van de posterior vraag leidt. Een nadeel van de veronderstelde normale verdeling is dat de symmetrische vorm een ondervertegenwoordiging van de kans op grote vervoersstromen impliceert. De tweede situatie is daarom gebaseerd op de veronderstelling dat fouten niet-symmetrisch, in het bijzonder log-normaal, verdeeld zijn. In dit geval is een analytische oplossing niet langer binnen bereik en wordt gebruikgemaakt van Markov Chain Monte Carlo simulatie in combinatie met Gibbs sampling, genest in Metropolis-Hastings sampling. Hoewel deze niet-symmetrische modellering een meer plausibele weergave van vervoersstromen nastreeft, is de benodigde rekentijd voor het bereiken van een evenwicht zodanig dat de praktische inzetbaarheid van de methode op dit moment beperkt is.

De laatste bijdrage van het proefschrift, van belang vanuit het oogpunt van multidisciplinariteit, betreft een toepassing van een hierarchisch Bayesiaans netwerk met een multi-proces model op het voorspellen van de vraag. Voor zover ons bekend, zijn er geen studies die dit eerder op deze manier hebben gedaan. Het gebruik van een multi-procesmodel in een Bayesiaans kader heeft het voordeel dat de vraag op verschillende voorafgaande dagen in beschouwing wordt genomen in plaats van alleen de vraag op de direct voorafgaande dag. Hierbij worden prior kansen op vooraf gedefinieerde scenarios met wegingen van de vraag op voorgaande dagen bijgesteld tot posterior kansen op deze scenarios. Meestal krijgt het scenario met de kortste Euclidische afstand tot het werkelijke scenario een posterior kans gelijk aan een, en de andere scenarios een kans gelijk aan nul. Deze vooraf gedefinieerde scenarios maken het mogelijk dat de ervaringen van mensen worden meegenomen in het multi-proces model.

Bibliography

Afandizaden, Z., Yadi, H., 2006. Estimation of freight od matrix using waybill data and traffic counts in iran roads. Iranian Journal of Science and Technology Transaction B-Engineering 30 (B1), 129–144.

Antoniou, C., Barceló, J., Breen, M., Bullejos, M., Casas, J., Cipriani, E., Ciuffo, B., Djukic, T., Hoogendoorn, S., Marzano, V., et al., 2015. Towards a generic benchmarking platform for origin–destination flows estimation/updating algorithms: Design, demonstration and validation. Transportation Research Part C: Emerging Technologies .

Asakura, Y., Hato, E., Kashiwadani, M., 2000. Origin-destination matrices estimation model using automatic vehicle identification data and its application to the han-shin expressway network. Transportation 27 (4), 419–438.

Ashok, K., Ben-Akiva, M., 2000. Alternative approaches for real-time estimation and prediction of time-dependent origin–destination flows. Transportation Science 34 (1), 21–36.

Ashok, K., Ben-Akiva, M. E., 2002. Estimation and prediction of time-dependent origin-destination flows with a stochastic mapping to path flows and link flows. Transportation Science 36 (2), 184–198.

Barceló, J., Montero, L., Marquès, L., Carmona, C., 2010. Travel time forecasting and dynamic origin-destination estimation for freeways based on bluetooth traffic monitoring. Transportation Research Record: Journal of the Transportation Research Board 2175 (-1), 19–27.

Bell, M., 1991. The estimation of origin-destination matrices by constrained generalised least squares. Transportation Research Part B: Methodological 25 (1), 13–22.

Bell, M., 2009. Hyperstar: A multi-path astar algorithm for risk averse vehicle navigation. Transportation Research Part B: Methodological 43 (1), 97–107.

Bellman, R., 1956. On a routing problem. Notes 16 (1).

Bianco, L., Confessore, G., Reverberi, P., 2001. A network based model for traffic sensor location with implications on o/d matrix estimates. Transportation Science 35 (1), 50–60.

Bierlaire, M., Toint, P., 1995. Meuse: An origin-destination matrix estimator that exploits structure. Transportation Research Part B: Methodological 29 (1), 47–60.

Bowman, J., Ben-Akiva, M., 2001. Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice 35 (1), 1–28.

Brillouin, L., 1956. Science and information theory. Dover Publications.

Carey, M., Revelli, R., 1986. Constrained estimation of direct demand functions and trip matrices. Transportation science 20 (3).

Caroll, R. J., 2006. Covariance analysis in generalized linear measurement error models. Statistics in Medicine 8 (9), 1075–1093.

Cascetta, E., 1984. Estimation of trip matrices from traffic counts and survey data: A generalized least squares estimator. Transportation Research Part B: Methodological 18 (4), 289–299.

Cascetta, E., Nguyen, S., 1988. A unified framework for estimating or updating origin/destination matrices from traffic counts. Transportation Research Part B: Methodological 22 (6), 437–455.

Castillo, E., Menéndez, J., Sánchez-Cambronero, S., 2008a. Predicting traffic flow using bayesian networks. Transportation Research Part B: Methodological 42 (5), 482–509.

Castillo, E., Menéndez, J. M., Sánchez-Cambronero, S., 2008b. Traffic estimation and optimal counting location without path enumeration using bayesian networks. Computer-Aided Civil and Infrastructure Engineering 23 (3), 189–207.

Chang, G., Wu, J., 1994. Recursive estimation of time-varying origin-destination flows from traffic counts in freeway corridors. Transportation Research Part B: Methodological 28 (2), 141–160.

Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. The American Statistician 49 (4), 327–335.

Chorus, C., Arentze, T., Timmermans, H., 2008. A random regret-minimization model of travel choice. Transportation Research Part B: Methodological 42 (1), 1–18.

Cover, T., Thomas, J., 2006. Elements of information theory. Wiley-interscience.

Daganzo, C., Sheffi, Y., 1977. On stochastic models of traffic assignment. Transportation Science 11 (3), 253–274.

Deng, Y., Tong, H., 2011. Dynamic shortest path algorithm in stochastic traffic networks using pso based on fluid neural network. Journal of Intelligent Learning Systems and Applications 3 (1), 11–16.

Dijkstra, E., 1959. A note on two problems in connexion with graphs. Numerische mathematik 1 (1), 269–271.

Dixon, M., Rilett, L., 2002. Real-time od estimation using automatic vehicle identification and traffic count data. Computer-Aided Civil and Infrastructure Engineering 17 (1), 7–21.

Duffin, R., Peterson, E., Zener, C., 1967. Geometric programming: theory and application. Wiley New York.

Fei, X., Mahmassani, H. S., 2011. Structural analysis of near-optimal sensor locations for a stochastic large-scale network. Transportation Research Part C: Emerging Technologies 19 (3), 440–453.

Ford, L., Fulkerson, D. R., 1962. Flows in networks. Vol. 1962. Princeton Princeton University Press.

Frederix, R., Viti, F., Corthout, R., Tampère, C., 2011. New gradient approximation method for dynamic origin-destination matrix estimation on congested networks. Transportation Research Record: Journal of the Transportation Research Board 2263 (-1), 19–25.

Friesz, T., Bernstein, D., Smith, T., Tobin, R., Wie, B., 1993. A variational inequality formulation of the dynamic network user equilibrium problem. Operations Research 41 (1), 179–191.

Gentili, M., Mirchandani, P. B., 2005. Locating active sensors on traffic networks. Annals of Operations Research 136 (1), 229–257.

Greenberg, E., 2008. Introduction to Bayesian Econometrics.

Gyftodimos, E., Flach, P., 2002. Hierarchical bayesian networks: A probabilistic reasoning model for structured domains. In: Proceedings of the ICML-2002 Workshop on Development of Representations. University of New South Wales. pp. 23–30.

Haghani, A., Hamedi, M., Sadabadi, K. F., Young, S., Tarnoff, P., 2010. Freeway travel time ground truth data collection using bluetooth sensors. In: Transportation Research Board 89th Annual Meeting. Transportation Research Board, Washington, DC.

Hazelton, M., 2000. Estimation of origin-destination matrices from link flows on uncongested networks. Transportation Research Part B: Methodological 34 (7), 549–566.

Hazelton, M., 2008. Statistical inference for time varying origin–destination matrices. Transportation Research Part B: Methodological 42 (6), 542–552.

Hazelton, M., 2010. Bayesian inference for network-based models with a linear inverse structure. Transportation Research Part B: Methodological 44 (5), 674–685.

Hu, S.-R., Peeta, S., Chu, C.-H., 2009. Identification of vehicle sensor locations for link-based network traffic applications. Transportation Research Part B: Methodological 43 (8), 873–894.

Jenkins, M., 2007. Introduction to route calculation. NAVTEQ-Network for developers 1.

Jeung, H., Liu, Q., Shen, H. T., Zhou, X., 2008. A hybrid prediction model for moving objects. In: Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. IEEE, pp. 70–79.

Johnson, R. A., Wichern, D. W., 2002. Applied multivariate statistical analysis. Vol. 5. Prentice hall Upper Saddle River, NJ.

Jones, P., Koppelman, F., Orfeuil, J., 1990. Activity analysis: State-of-the-art and future directions. Developments in dynamic and activity-based approaches to travel analysis , 34–55.

Kikuchi, S., Miljkovic, D., van Zuylen, H., 2000. Examination of methods that adjust observed traffic volumes on a network. Transportation Research Record: Journal of the Transportation Research Board 1717 (-1), 109–119.

Kitamura, R., 1996. Applications of models of activity behavior for activity based demand forecasting. In: Activity-Based Travel Forecasting Conference, New Orleans, Louisiana.

Kitamura, R., Chen, C., Pendyala, R. M., Narayanan, R., 2000. Micro-simulation of daily activity-travel patterns for travel demand forecasting. Transportation 27 (1), 25–51.

Koopman, S., et al., 2012. Time series analysis by state space methods. Vol. 38. OUP Oxford.

Kroes, E., Sheldon, R., 1988. Stated preference methods: an introduction. Journal of Transport Economics and Policy , 11–25.

Krumm, J., Horvitz, E., 2006. Predestination: Inferring destinations from partial trajectories. In: UbiComp 2006: Ubiquitous Computing. Springer, pp. 243–260.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22 (1), 79–86.

Kumar, S., Kumar, S., 2011. Automated taxi/cab system using a* algorithm. In: Proceedings of the International Conference on Advances in Computing and Artificial Intelligence. ACM, pp. 13–20.

Larsson, T., Lundgren, J. T., Peterson, A., 2010. Allocation of link flow detectors for origin-destination matrix estimation comparative study. Computer-Aided Civil and Infrastructure Engineering 25 (2), 116–131.

Li, B., 2005. Bayesian inference for origin-destination matrices of transport networks using the em algorithm. Technometrics 47 (4), 399–408.

Li, B., 2009. Markov models for bayesian analysis about transit route origin-destination matrices. Transportation Research Part B: Methodological 43 (3), 301–310.

Li, B., Moor, B. D., 2002. Dynamic identification of origin–destination matrices in the presence of incomplete observations. Transportation Research Part B: Methodological 36 (1), 37–57.

Lin, P.-W., Chang, G.-L., 2007. A generalized model and solution algorithm for estimation of the dynamic freeway origin–destination matrix. Transportation Research Part B: Methodological 41 (5), 554–572.

Ma, Y., Van Dalen, J., De Blois, C., Kroon, L., 2011. Estimation of dynamic traffic densities for official statistics. Transportation Research Record: Journal of the Transportation Research Board 2256 (-1), 104–111.

Ma, Y., van Dalen, J., Zuidwijk, R., de Blois, C., 2012. Dynamic weight capacity utilization and efficiency in freight transport: An application of weigh-in-motion data. In: Transportation Research Board 91st Annual Meeting. No. 12-3427.

Ma, Y., van Zuylen, H., Chen, Y., van Dalen, J., 2010. Allocating departure time slots to optimize dynamic network capacity. Transportation Research Record: Journal of the Transportation Research Board 2197 (-1), 98–106.

Maher, M., 1983. Inferences on trip matrices from observations on link volumes: a bayesian statistical approach. Transportation Research Part B: Methodological 17 (6), 435–447.

McNally, M., 1996. An activity-based microsimulation model for travel demand forecasting .

Mitchell, T., et al., 1997. Machine learning.

Morzy, M., 2006. Prediction of moving object location based on frequent trajectories. In: Computer and Information Sciences–ISCIS 2006. Springer, pp. 583–592.

Morzy, M., 2007. Mining frequent trajectories of moving objects for location prediction. In: Machine Learning and Data Mining in Pattern Recognition. Springer, pp. 667–680.

Nemes, G., 2010. On the coefficients of the asymptotic expansion of n! Journal of Integer Sequences 13 (2), 3.

Nie, Y., Zhang, H., 2008. A variational inequality formulation for inferring dynamic origin–destination travel demands. Transportation Research Part B: Methodological 42 (7), 635–662.

Nihan, N., Davis, G., 1989. Application of prediction-error minimization and maximum likelihood to estimate intersection od matrices from traffic counts. Transportation Science 23 (2), 77–90.

NMmagazine, 2016. Demogelijkheden van bluetooth en gsm data .

Perrakis, K., Karlis, D., Cools, M., Janssens, D., Vanhoof, K., Wets, G., 2011. A bayesian approach for modeling origin-destination matrices. Transportation Research Part A: Policy and Practice .

Quayle, S., Koonce, P., 2010. Arterial performance measures using mac readers portland experience. North American Travel Monitoring report. htm .

Quayle, S., Koonce, P., DePencier, D., Bullock, D., 2010. Arterial performance measures with media access control readers. Transportation Research Record: Journal of the Transportation Research Board 2192 (-1), 185–193.

Raju, P., 2003. Fundamentals of gps. Satellite Remote Sensing and GIS Applications in Agricultural Meteorology , 121.

Ran, B., Boyce, D., 1994. Dynamic urban transportation network models: theory and implications for intelligent vehicle-highway systems. No. 417.

Rossi, P., Allenby, G., McCulloch, R., 2005. Bayesian statistics and marketing.

Schneider, M., Linauer, M., Hainitz, N., Koller, H., 2009. Traveller information service based on real-time toll data in austria. Intelligent Transport Systems, IET 3 (2), 124–137.

Shan, J., Li, X., 2008. Estimating the highway freight origin-destination matrix from multisource data based on fuzzy programming theory. In: Modelling, Simulation and Optimization, 2008. WMSO'08. International Workshop on. IEEE, pp. 204–209.

Shao, H., Lam, W. H., Sumalee, A., Chen, A., Hazelton, M. L., 2014. Estimation of mean and covariance of peak hour origin-destination demands from day-to-day traffic counts. Transportation Research Part B: Methodological 68, 52–75.

Sheffi, Y., 1985. Urban transportation networks: Equilibrium analysis with mathematical programming methods .

Sherman, J., Morrison, W. J., 1950. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. The Annals of Mathematical Statistics 21 (1), 124–127.

Spiess, H., 1987. A maximum likelihood model for estimating origin-destination matrices. Transportation Research Part B: Methodological 21 (5), 395–412.

Stanley, J., Hensher, D., Loader, C., 2009. Road transport and climate change: stepping off the greenhouse gas. Transportation Research Part A: Policy and Practice .

Stavins, R., 1999. The costs of carbon sequestration: a revealed-preference approach. The American Economic Review 89 (4), 994–1009.

Sun, S., Zhang, C., Yu, G., 2006. A bayesian network approach to traffic flow forecasting. Intelligent Transportation Systems, IEEE Transactions on 7 (1), 124–132.

Tebaldi, C., West, M., 1998. Bayesian inference on network traffic using link count data. Journal of the American Statistical Association , 557–573.

Tibshirani, R., 1989. Noninformative priors for one parameter of many. Biometrika 76 (3), 604–608.

Van der Linde, A., 2012. A bayesian view of model complexity. Statistica Neerlandica 66 (3), 253–271.

van der Sangen, M., 2015. A frist for statistsics netherlands: lauching statistiscs based on big data. In: http://www.cbs.nl/NR/rdonlyres/4E3C7500-03EB-4C54-8A0A-753C017165F2/0/afirstforlaunchingstatisticsbasedonbigdata.pdf.

Van Der Zijpp, N., 1997. Dynamic origin-destination matrix estimation from traffic counts and automated vehicle identification data. Transportation Research Record: Journal of the Transportation Research Board 1607 (-1), 87–94.

Van Zuylen, H., 1981. Some improvements in the estimation of an od matrix from traffic counts. In: Proceedings of the 8th international symposium on transportation and traffic theory. Toronto, Canada, University of Toronto Press, Toronto.

Van Zuylen, H., Branston, D., 1982. Consistent link flow estimation from counts. Transportation Research Part B: Methodological 16 (6), 473–476.

Van Zuylen, H., Willumsen, L., 1980. The most likely trip matrix estimated from traffic counts. Transportation Research Part B: Methodological 14 (3), 281–293.

Viti, F., Corman, F., 2012. A noval approach to the sensor location problem for measuring the observed network flow variability. In: Proceedings of the 5th international symposium of traffic network reliability. Hong Kong, China.

Wang, D., Cheng, T., 2001. A spatio-temporal data model for activity-based transport demand modelling. International Journal of Geographical Information Science 15 (6), 561–585.

Wardrop, J., 1952. Some theoretical aspects of road traffic research. In: Inst Civil Engineers Proc London/UK/.

Watling, D., 1994. Maximum likelihood estimation of an origin-destination matrix from a partial registration plate survey. Transportation Research Part B: Methodological 28 (4), 289–314.

Watling, D., Maher, M., 1992. A statistical procedure for estimating a mean origin-destination matrix from a partial registration plate survey. Transportation Research Part B: Methodological 26 (3), 171–193.

Weiss, N. A., Holmes, P. T., Hardy, M., 2006. A course in probability. Pearson Addison Wesley.

West, M., Harrison, J., 1997. Bayesian forecasting and dynamic models. Springer Verlag.

Willumsen, L., 1984. Estimating time-dependent trip matrices from traffic counts. In: Ninth International Symposium on Transportation and Traffic Theory, VNU Science Press. pp. 397–411.

Wu, J., Florian, M., Marcotte, P., 1994. Transit equilibrium assignment: a model and solution algorithms. Transportation Science 28 (3), 193–203.

Yang, H., 1995. Heuristic algorithms for the bilevel origin-destination matrix estimation problem. Transportation Research Part B: Methodological 29 (4), 231–242.

Yavaş, G., Katsaros, D., Ulusoy, Ö., Manolopoulos, Y., 2005. A data mining approach for location prediction in mobile environments. Data & Knowledge Engineering 54 (2), 121–146.

Ye, J., 1998. On measuring and correcting the effects of data mining and model selection. Journal of the American Statistical Association 93 (441), 120–131.

Zhang, Y., Qin, X., Dong, S., Ran, B., 2010. Daily od matrix estimation using cellular probe data. In: Transportation Research Board 89th Annual Meeting. No. 10-2472.

Zhou, X., List, G., 2010. An information-theoretic sensor location model for traffic origin-destination demand estimation applications. Transportation Science 44 (2), 254–273.

Zhou, X., Mahmassani, H., 2006. Dynamic origin-destination demand estimation using automatic vehicle identification data. Intelligent Transportation Systems, IEEE Transactions on 7 (1), 105–114.

Zhou, X., Mahmassani, H., 2007. A structural state space model for real-time traffic origin–destination demand estimation and prediction in a day-to-day learning framework. Transportation Research Part B: Methodological 41 (8), 823–840.

About the author

Yinyi Ma was born in 1984 in Nanjing, China. She holds a bachelor degree of Transportation Engineering in China. From 2007 to 2008, she studied the Transportation Infrastructure and Logistics major at Delft University of Technology, the Netherlands, under the supervision of Prof. dr. Henk van Zuylen. In 2009, she joined the PhD program in Erasmus University, the Netherlands. Her project is funded by Statistics Netherlands (CBS). Her research interests include: transportation demand management, travel demand forecasting, transportation performance evaluation and econometrics modeling. Three of her research papers have been published in the Journal of Transportation Research Record. Some other research findings have been presented in international conferences including the Transportation Research Board, INFORMS, the World Conference on Transportation Research, and IEEE Conference on Intelligent Transportation Systems. During her study period in the Netherlands, she got a fellowship of the International Road Federation, USA. She also won a bronze medal in the Young European Arena of Research, Future Visions of Transport. In 2012, she spent four months in the Transportation Center of Northwestern University as a visiting scholar, under the supervision of Prof. Hani Mahmassani.

Author portfolio

Journal Publications

Y. Ma, H. van Zuylen, Y. Chen, J. van Dalen
"Allocating Departure Time Slots to Optimize Dynamic Network Capacity",
Journal of Transportation Research Record, 2 (2197) (2011), pp 98-106

Y.Ma, J. van Dalen, C. de Blois, L. Kroon
"Estimation of Dynamic Traffic Density for Official Statistics based on Combined Use of GPS and Loop Detector Data",
Journal of Transportation Research Record 2256 (2012), pp 104-111

Y. Ma, R. Kuik, H. van Zuylen
"Day-to-Day Origin Destination Tuple Estimation and Prediction with Hierarchical Bayesian Networks Using Multiple Data Sources",
Journal of Transportation Research Record 2343 (2013), pp 51-61

Conference Presentations

2009: **Transportation Research Board**, Washington D.C., USA
2009: **Advanced Forum on Transportation of China**, Beijing, China
2010: **Transportation Research Board**, Washington D.C., USA
2010: **World Conference on Transportation Research**, Lisbon, Portugal
2011: **Transportation Research Board**, Washington D.C., USA
2012: **Transportation Research Board**, Washington D.C., USA
2012: **Institute for Operations Research and the Management Science**, Phoenix, USA
2012: **IEEE on Intelligent Transforation Systems**, Anchorage, USA
2013: **Transportation Research Board**, Washington D.C., USA

Awards

2009: **International Road Federation Fellowship** Washington D.C., USA
2010: **Bronze Medal in the European Arena of Research** Brussels, Belgium
2012: **Visiting Scholar in Northwestern University** Evanston, USA

Selected Skills and Languages

Computer skills:	**R, Matlab, SPSS(statistics, modeler, text), SQL**
Simulation tools:	**eViews, Dynasmart, ArcGIS, VISSIM, eM-Plant, Arena**
Languages:	**Chinese, English, Dutch**

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

ERIM PH.D. SERIES RESEARCH IN MANAGEMENT

The ERIM PhD Series contains PhD dissertations in the field of Research in Management defended at Erasmus University Rotterdam and supervised by senior researchers affiliated to the Erasmus Research Institute of Management (ERIM). All dissertations in the ERIM PhD Series are available in full text through the ERIM Electronic Series Portal: http://repub.eur.nl/pub. ERIM is the joint research institute of the Rotterdam School of Management (RSM) and the Erasmus School of Economics at the Erasmus University Rotterdam (EUR).

DISSERTATIONS LAST FIVE YEARS

Abbink, E.J., Crew Management in Passenger Rail Transport, Promotor(s): Prof.dr. L.G. Kroon & Prof.dr. A.P.M. Wagelmans, EPS-2014-325-LIS, http://repub.eur.nl/ pub/76927

Acar, O.A., *Crowdsourcing for Innovation: Unpacking Motivational, Knowledge and Relational Mechanisms of Innovative Behavior in Crowdsourcing Platforms*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2014-321-LIS, http://repub.eur.nl/pub/76076

Akin Ates, M., *Purchasing and Supply Management at the Purchase Category Level: strategy, structure and performance*, Promotor(s): Prof.dr. J.Y.F. Wynstra & Dr. E.M. van Raaij, EPS-2014-300-LIS, http://repub.eur.nl/pub/50283

Akpinar, E., *Consumer Information Sharing*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2013-297-MKT, http://repub.eur.nl/pub/50140

Alexander, L., *People, Politics, and Innovation: A Process Perspective*, Promotor(s): Prof.dr. H.G. Barkema & Prof.dr. D.L. van Knippenberg, EPS-2014-331-S&E, http: //repub.eur.nl/pub/77209

Almeida e Santos Nogueira, R.J. de, *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. J.M.C. Sousa, EPS-2014-310-LIS, http://repub.eur.nl/pub/51560

Bannouh, K., *Measuring and Forecasting Financial Market Volatility using High-frequency Data*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2013-273-F&A, http://repub.eur.nl/pub/38240

Ben-Menahem, S.M., *Strategic Timing and Proactiveness of Organizations*, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2013-278-S&E, http://repub.eur.nl/pub/39128

Benning, T.M., *A Consumer Perspective on Flexibility in Health Care: Priority Access Pricing and Customized Care*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2011-241-MKT, http://repub.eur.nl/pub/23670

Benschop, N, *Biases in Project Escalation: Names, frames & construal levels*, Promotors: Prof.dr. K.I.M. Rhode, Prof.dr. H.R. Commandeur, Prof.dr. M.Keil & Dr. A.L.P. Nuijten, EPS-2015-375-S&E, hdl.handle.net/1765/79408

Berg, W.E. van den, *Understanding Salesforce Behavior using Genetic Association Studies*, Promotor(s): Prof.dr. W.J.M.I. Verbeke, EPS-2014-311-MKT, http://repub.eur.nl/pub/51440

Betancourt, N.E., *Typical Atypicality: Formal and Informal Institutional Conformity, Deviance, and Dynamics*, Promotor(s): Prof.dr. B. Krug, EPS-2012-262-ORG, http://repub.eur.nl/pub/32345

Bliek, R. de, *Empirical Studies on the Economic Impact of Trust*, Promotor(s): Prof.dr. J. Veenman & Prof.dr. Ph.H.B.F. Franses, EPS-2015-324-ORG, http://repub.eur.nl/pub/78159

Blitz, D.C., *Benchmarking Benchmarks*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. W.F.C. Verschoor, EPS-2011-225-F&A, http://repub.eur.nl/pub/22624

Boons, M., *Working Together Alone in the Online Crowd: The Effects of Social Motivationsand Individual Knowledge Backgrounds on the Participation and Performance of Members of Online Crowdsourcing Platforms*, Promotor(s): Prof.dr. H.G. Barkema & Dr. D.A. Stam, EPS-2014-306-S&E, http://repub.eur.nl/pub/50711

Brazys, J., *Aggregated Marcoeconomic News and Price Discovery*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-351-F&A, http://repub.eur.nl/pub/78243

Burger, M.J., *Structure and Cooptition in Urban Networks*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.R. Commandeur, EPS-2011-243-ORG, http://repub.eur.nl/pub/26178

Byington, E., *Exploring Coworker Relationships: Antecedents and Dimensions of Interpersonal Fit,Coworker Satisfaction, and Relational Models*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-292-ORG, http://repub.eur.nl/pub/41508

Camacho, N.M., *Health and Marketing: Essays on Physician and Patient Decision- Making*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-237-MKT, http://repub.eur.nl/pub/23604

Cancurtaran, P., *Essays on Accelerated Product Development*, Promotor(s): Prof.dr. F. Langerak & Prof.dr.ir. G.H. van Bruggen, EPS-2014-317-MKT, http://repub.eur.nl/pub/76074

Caron, E.A.M., *Explanation of Exceptional Values in Multi-dimensional Business Databases*, Promotor(s): Prof.dr.ir. H.A.M. Daniels & Prof.dr. G.W.J. Hendrikse, EPS-2013-296-LIS, http://repub.eur.nl/pub/50005

Carvalho, L. de, *Knowledge Locations in Cities: Emergence and Development Dynamics*, Promotor(s): Prof.dr. L. Berg, EPS-2013-274-S&E, http://repub.eur.nl/pub/38449

Cox, R.H.G.M., *To Own, To Finance, and To Insure - Residential Real Estate Revealed*, Promotor(s): Prof.dr. D. Brounen, EPS-2013-290-F&A, http://repub.eur.nl/pub/40964

Deichmann, D., *Idea Management: Perspectives from Leadership, Learning, and Network Theory*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2012-255-ORG, http://repub.eur.nl/pub/31174

Deng, W., *Social Capital and Diversification of Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2015-341-ORG, http://repub.eur.nl/pub/77449

Desmet, P.T.M., *In Money we Trust? Trust Repair and the Psychology of Financial Compensations*, Promotor(s): Prof.dr. D. de Cremer, EPS-2011-232-ORG, http://repub.eur.nl/pub/23268

Dollevoet, T.A.B., *Delay Management and Dispatching in Railways*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-272-LIS, http://repub.eur.nl/pub/38241

Doorn, S. van, *Managing Entrepreneurial Orientation*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-258- STR, http://repub.eur.nl/pub/32166

Douwens-Zonneveld, M.G., *Animal Spirits and Extreme Confidence: No Guts, No Glory?* Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2012-257-F&A, http://repub.eur.nl/pub/31914

Duca, E., *The Impact of Investor Demand on Security Offerings*, Promotor(s): Prof.dr. A. de Jong, EPS-2011-240-F&A, http://repub.eur.nl/pub/26041

Duyvesteyn, J.G. *Empirical Studies on Sovereign Fixed Income Markets*, Promotor(s): Prof.dr P.Verwijmeren & Prof.dr. M.P.E. Martens, EPS-2015-361-F&A, hdl.handle.net/1765/79033

Duursema, H., *Strategic Leadership: Moving Beyond the Leader-Follower Dyad*, Promotor(s): Prof.dr. R.J.M. van Tulder, EPS-2013-279-ORG, http://repub.eur.nl/pub/39129

Eck, N.J. van, *Methodological Advances in Bibliometric Mapping of Science*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-247-LIS, http://repub.eur.nl/pub/26509

Elemes, A, *Studies on Determinants and Consequences of Financial Reporting Quality*, Promotor: Prof.dr. E.Peek, EPS-2015-354-F&A, http://hdl.handle.net/1765/79037

Ellen, S. ter, *Measurement, Dynamics, and Implications of Heterogeneous Beliefs in Financial Markets*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2015-343-F&A, http://repub.eur.nl/pub/78191

Eskenazi, P.I., *The Accountable Animal*, Promotor(s): Prof.dr. F.G.H. Hartmann, EPS-2015-355-F&A, http://repub.eur.nl/pub/78300

Essen, M. van, *An Institution-Based View of Ownership*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. G.M.H. Mertens, EPS-2011-226-ORG, http://repub.eur.nl/pub/22643

Evangelidis, I., *Preference Construction under Prominence*, Promotor(s): Prof.dr. S.M.J. van Osselaer, EPS-2015-340-MKT, http://repub.eur.nl/pub/78202

Faber, N., *Structuring Warehouse Management*, Promotor(s): Prof.dr. MB.M. de Koster, Prof.dr. Ale Smidts, EPS-2015-336-LIS, http://repub.eur.nl/pub/78603

Fernald, K., *The Waves of Biotechnological Innovation in Medicine: Interfirm Cooperation Effects and a Venture Capital Perspective*, Promotor(s): Prof.dr. E.Claassen, Prof.dr. H.P.G.Pennings & Prof.dr. H.R. Commandeur, EPS-2015-371-S&E, http://hdl.handle.net/1765/79120

Fourne, S.P., *Managing Organizational Tensions: A Multi-Level Perspective on Exploration, Exploitation and Ambidexterity*, Promotor(s): Prof.dr. J.J.P. Jansen & Prof.dr.S.J. Magala, EPS-2014-318-S&E, http://repub.eur.nl/pub/76075

Gharehgozli, A.H., *Developing New Methods for Efficient Container Stacking Operations*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2012-269-LIS, http://repub.eur.nl/pub/37779

Gils, S. van, *Morality in Interactions: On the Display of Moral Behavior by Leaders and Employees*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-270-ORG, http://repub.eur.nl/pub/38027

Ginkel-Bieshaar, M.N.G. van, *The Impact of Abstract versus Concrete Product Communications on Consumer Decision-making Processes*, Promotor(s): Prof.dr.ir. B.G.C.Dellaert, EPS-2012-256-MKT, http://repub.eur.nl/pub/31913

Gkougkousi, X., *Empirical Studies in Financial Accounting*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. E. Peek, EPS-2012-264-F&A, http://repub.eur.nl/pub/37170

Glorie, K.M., *Clearing Barter Exchange Markets: Kidney Exchange and Beyond*, Promotor(s): Prof.dr. A.P.M. Wagelmans & Prof.dr. J.J. van de Klundert, EPS-2014-329-LIS, http://repub.eur.nl/pub/77183

Hekimoglu, M., *Spare Parts Management of Aging Capital Products*, Promotor: Prof.dr.ir. R. Dekker, EPS-2015-368-LIS, http://hdl.handle.net/1765/79092

Heij, C.V., *Innovating beyond Technology. Studies on how management innovation, cocreation and business model innovation contribute to firm's (innovation) performance*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-370-STR, http://repub.eur.nl/pub/78651

Heyde Fernandes, D. von der, *The Functions and Dysfunctions of Reminders, Promotor(s): Prof.dr. S.M.J. van Osselaer*, EPS-2013-295-MKT, http://repub.eur.nl/pub/41514

Heyden, M.L.M., *Essays on Upper Echelons & Strategic Renewal: A Multilevel Contingency Approach*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch & Prof.dr. H.W. Volberda, EPS-2012-259-STR, http://repub.eur.nl/pub/32167

Hoever, I.J., *Diversity and Creativity*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2012-267-ORG, http://repub.eur.nl/pub/37392

Hogenboom, A.C., *Sentiment Analysis of Text Guided by Semantics and Structure*, Promotor(s):Prof.dr.ir. U.Kaymak & Prof.dr. F.M.G. de Jong, EPS-2015-369-LIS, http://hdl.handle.net/1765/79034

Hogenboom, F.P., *Automated Detection of Financial Events in News Text*, Promotor(s): Prof.dr.ir. U. Kaymak & Prof.dr. F.M.G. de Jong, EPS-2014-326-LIS, http://repub.eur.nl/pub/77237

Hollen, R.M.A., *Exploratory Studies into Strategies to Enhance Innovation-Driven International Competitiveness in a Port Context: Toward Ambidextrous Ports*, Promotor(s) Prof.dr.ing. F.A.J. Van Den Bosch & Prof.dr. H.W.Volberda, EPS-2015-372-S&E, hdl.handle.net/1765/78881

Hoogendoorn, B., *Social Entrepreneurship in the Modern Economy: Warm Glow, Cold Feet*, Promotor(s): Prof.dr. H.P.G. Pennings & Prof.dr. A.R. Thurik, EPS-2011-246-STR, http://repub.eur.nl/pub/26447

Hoogervorst, N., *On The Psychology of Displaying Ethical Leadership: A Behavioral Ethics Approach*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2011- 244-ORG, http://repub.eur.nl/pub/26228

Hout, D.H. van, *Measuring Meaningful Differences: Sensory Testing Based Decision Making in an Industrial Context; Applications of Signal Detection Theory and Thurstonian Modelling*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. G.B. Dijksterhuis, EPS- 2014-304-MKT, http://repub.eur.nl/pub/50387

Houwelingen, G.G. van, *Something To Rely On*, Promotor(s): Prof.dr. D. de Cremer & Prof.dr. M.H. van Dijke, EPS-2014-335-ORG, http://repub.eur.nl/pub/77320

Hurk, E. van der, *Passengers, Information, and Disruptions*, Promotor(s): Prof.dr. L.G. Kroon & Prof.mr.dr. P.H.M. Vervest, EPS-2015-345-LIS, http://repub.eur.nl/pub/78275

Hytonen, K.A., *Context Effects in Valuation, Judgment and Choice*: A Neuroscientific Approach, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2011-252-MKT, http://repub.eur.nl/pub/30668

Iseger, P. den, *Fourier and Laplace Transform Inversion with Applications in Finance*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2014-322-LIS, http://repub.eur.nl/pub/76954

Jaarsveld, W.L. van, *Maintenance Centered Service Parts Inventory Control*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-288-LIS, http://repub.eur.nl/pub/39933

Jalil, M.N., *Customer Information Driven After Sales Service Management: Lessons from Spare Parts Logistics*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-222-LIS, http://repub.eur.nl/pub/22156

Kappe, E.R., *The Effectiveness of Pharmaceutical Marketing*, Promotor(s): Prof.dr. S. Stremersch, EPS-2011-239-MKT, http://repub.eur.nl/pub/23610

Karreman, B., *Financial Services and Emerging Markets*, Promotor(s): Prof.dr. G.A. van der Knaap & Prof.dr. H.P.G. Pennings, EPS-2011-223-ORG, http://repub.eur.nl/pub/22280

Khanagha, S., *Dynamic Capabilities for Managing Emerging Technologies*, Promotor(s): Prof.dr. H.W. Volberda, EPS-2014-339-S&E, http://repub.eur.nl/pub/77319

Kil, J., *Acquisitions Through a Behavioral and Real Options Lens*, Promotor(s): Prof.dr. H.T.J. Smit, EPS-2013-298-F&A, http://repub.eur.nl/pub/50142

Klooster, E. van 't, *Travel to Learn: the Influence of Cultural Distance on Competence Development in Educational Travel*, Promotor(s): Prof.dr. F.M. Go & Prof.dr. P.J. van Baalen, EPS-2014-312-MKT, http://repub.eur.nl/pub/51462

Koendjbiharie, S.R., *The Information-Based View on Business Network Performance: Revealing the Performance of Interorganizational Networks*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.mr.dr. P.H.M. Vervest, EPS-2014-315-LIS, http://repub.eur.nl/pub/51751

Koning, M., *The Financial Reporting Environment: The Role of the Media, Regulators and Auditors*, Promotor(s): Prof.dr. G.M.H. Mertens & Prof.dr. P.G.J. Roosenboom, EPS-2014-330-F&A, http://repub.eur.nl/pub/77154

Konter, D.J., *Crossing Borders with HRM: An Inquiry of the Influence of Contextual Differences in the Adoption and Effectiveness of HRM*, Promotor(s): Prof.dr. J. Paauwe & Dr. L.H. Hoeksema, EPS-2014-305-ORG, http://repub.eur.nl/pub/50388

Korkmaz, E., *Bridging Models and Business: Understanding Heterogeneity in Hidden Drivers of Customer Purchase Behavior*, Promotor(s): Prof.dr. S.L. van de Velde & Prof.dr. D. Fok, EPS-2014-316-LIS, http://repub.eur.nl/pub/76008

Kroezen, J.J., *The Renewal of Mature Industries: An Examination of the Revival of the Dutch Beer Brewing Industry*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2014- 333-S&E, http://repub.eur.nl/pub/77042

Kysucky, V., *Access to Finance in a Cros-Country Context*, Promotor(s): Prof.dr. L. Norden, EPS-2015-350-F&A, http://repub.eur.nl/pub/78225

Lam, K.Y., *Reliability and Rankings*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-230-MKT, http://repub.eur.nl/pub/22977

Lander, M.W., *Profits or Professionalism? On Designing Professional Service Firms*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2012-253-ORG, http://repub.eur.nl/pub/30682

Langhe, B. de, *Contingencies: Learning Numerical and Emotional Associations in an Uncertain World*, Promotor(s): Prof.dr.ir. B. Wierenga & Prof.dr. S.M.J. van Osselaer, EPS-2011-236-MKT, http://repub.eur.nl/pub/23504

Legault-Tremblay, P.O., *Corporate Governance During Market Transition: Heterogeneous responses to Institution Tensions in China*, Promotor: Prof.dr. B. Krug, EPS-2015-362-ORG, http://repub.eur.nl/pub/78649

Lenoir, A.S. *Are You Talking to Me? Addressing Consumers in a Globalised World*, Promotor(s) Prof.dr. S. Puntoni & Prof.dr. S.M.J. van Osselaer, EPS-2015-363-MKT, , http://hdl.handle.net/1765/79036

Leunissen, J.M., *All Apologies: On the Willingness of Perpetrators to Apologize*, Promotor(s): Prof.dr. D. de Cremer & Dr. M. van Dijke, EPS-2014-301-ORG, http://repub.eur.nl/pub/50318

Li, D., *Supply Chain Contracting for After-sales Service and Product Support*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2015-347-LIS, http://repub.eur.nl/pub/78526

Li, Z., *Irrationality: What, Why and How*, Promotor(s): Prof.dr. H. Bleichrodt, Prof.dr. P.P. Wakker, & Prof.dr. K.I.M. Rohde, EPS-2014-338-MKT, http://repub.eur.nl/pub/77205

Liang, Q.X., *Governance, CEO Identity, and Quality Provision of Farmer Cooperatives*, Promotor(s): Prof.dr. G.W.J. Hendrikse, EPS-2013-281-ORG, http://repub.eur.nl/pub/39253

Liket, K., *Why 'Doing Good' is not Good Enough: Essays on Social Impact Measurement*, Promotor(s): Prof.dr. H.R. Commandeur & Dr. K.E.H. Maas, EPS-2014-307-STR, http://repub.eur.nl/pub/51130

Loos, M.J.H.M. van der, *Molecular Genetics and Hormones: New Frontiers in Entrepreneurship Research*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2013-287-S&E, http://repub.eur.nl/pub/40081

Lovric, M., *Behavioral Finance and Agent-Based Artificial Markets*, Promotor(s): Prof.dr. J. Spronk & Prof.dr.ir. U. Kaymak, EPS-2011-229-F&A, http://repub.eur.nl/pub/22814

Lu, Y., *Data-Driven Decision Making in Auction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck & Prof.dr. W. Ketter, EPS-2014-314-LIS, http://repub.eur.nl/pub/51543

Manders, B., *Implementation and Impact of ISO 9001*, Promotor(s): Prof.dr. K. Blind, EPS-2014-337-LIS, http://repub.eur.nl/pub/77412

Markwat, T.D., *Extreme Dependence in Asset Markets Around the Globe*, Promotor(s): Prof.dr. D.J.C. van Dijk, EPS-2011-227-F&A, http://repub.eur.nl/pub/22744

Mees, H., *Changing Fortunes: How China's Boom Caused the Financial Crisis*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2012-266-MKT, http://repub.eur.nl/pub/34930

Mell, J.N., *Connecting Minds: On The Role of Metaknowledge in Knowledge Coordination*, Promotor: Prof.dr.D.L. van Knippenberg, EPS-2015-359-ORG, http://hdl.handle.net/1765/78951

Meuer, J., *Configurations of Inter-firm Relations in Management Innovation: A Study in China's Biopharmaceutical Industry*, Promotor(s): Prof.dr. B. Krug, EPS-2011-228-ORG, http://repub.eur.nl/pub/22745

Micheli, M.R., *Business Model Innovation: A Journey across Managers' Attention and Inter-Organizational Networks*, Promotor(s): Prof.dr. J.J.P. Jansen, EPS-2015-344-S&E, http://repub.eur.nl/pub/78241

Mihalache, O.R., *Stimulating Firm Innovativeness: Probing the Interrelations between Managerial and Organizational Determinants*, Promotor(s): Prof.dr. J.J.P. Jansen, Prof.dr.ing. F.A.J. van den Bosch, & Prof.dr. H.W. Volberda, EPS-2012-260-S&E, http://repub.eur.nl/pub/32343

Milea, V., *News Analytics for Financial Decision Support*, Promotor(s): Prof.dr.ir. U. Kaymak, EPS-2013-275-LIS, http://repub.eur.nl/pub/38673

Naumovska, I., *Socially Situated Financial Markets: A Neo-Behavioral Perspective on Firms, Investors and Practices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. A. de Jong, EPS-2014-319-S&E, http://repub.eur.nl/pub/76084

Nielsen, L.K., *Rolling Stock Rescheduling in Passenger Railways: Applications in short term planning and in disruption management*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2011-224-LIS, http://repub.eur.nl/pub/22444

Nuijten, A.L.P., *Deaf Effect for Risk Warnings: A Causal Examination applied to Information Systems Projects*, Promotor(s): Prof.dr. G.J. van der Pijl, Prof.dr. H.R. Commandeur & Prof.dr. M. Keil, EPS-2012-263-S&E, http://repub.eur.nl/pub/34928

Osadchiy, S.E., *The Dynamics of Formal Organization: Essays on bureaucracy and formal rules*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens, EPS-2011-231-ORG, http://repub.eur.nl/pub/23250

Ozdemir, M.N., *Project-level Governance, Monetary Incentives, and Performance in Strategic R&D Alliances*, Promotor(s): Prof.dr.ir. J.C.M. van den Ende, EPS-2011-235-LIS, http://repub.eur.nl/pub/23550

Peers, Y., *Econometric Advances in Diffusion Models*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-251-MKT, http://repub.eur.nl/pub/30586

Peters, M., *Machine Learning Algorithms for Smart Electricity Markets*, Promotor(s): Prof.dr. W. Ketter, EPS-2014-332-LIS, http://repub.eur.nl/pub/77413

Porck, J., *No Team is an Island: An Integrative View of Strategic Consensus between Groups*, Promotor(s): Prof.dr. P.J.F. Groenen & Prof.dr. D.L. van Knippenberg, EPS-2013-299-ORG, http://repub.eur.nl/pub/50141

Porras Prado, M., *The Long and Short Side of Real Estate, Real Estate Stocks, and Equity*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2012-254-F&A, http://repub.eur.nl/pub/30848

Poruthiyil, P.V., *Steering Through: How organizations negotiate permanent uncertainty and unresolvable choices*, Promotor(s): Prof.dr. P.P.M.A.R. Heugens & Prof.dr. S.J. Magala, EPS-2011-245-ORG, http://repub.eur.nl/pub/26392

Pourakbar, M., *End-of-Life Inventory Decisions of Service Parts*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2011-249-LIS, http://repub.eur.nl/pub/30584

Pronker, E.S., *Innovation Paradox in Vaccine Target Selection*, Promotor(s): Prof.dr. H.J.H.M. Claassen & Prof.dr. H.R. Commandeur, EPS-2013-282-S&E, http://repub.eur.nl/pub/39654

Protzner, S. *Mind the gap between demand and supply: A behavioral perspective on demand forecasting*, Promotor(s): Prof.dr. S.L. van de Velde & Dr. L. Rook, EPS-2015-364-LIS, http://repub.eur.nl/pub/79355

Pruijssers, J.K., *An Organizational Perspective on Auditor Conduct*, Promotor(s): Prof.dr. J. van Oosterhout & Prof.dr. P.P.M.A.R. Heugens, EPS-2015-342-S&E, http://repub.eur.nl/pub/78192

Retel Helmrich, M.J., *Green Lot-Sizing*, Promotor(s): Prof.dr. A.P.M. Wagelmans, EPS-2013-291-LIS, http://repub.eur.nl/pub/41330

Rietveld, N., *Essays on the Intersection of Economics and Biology*, Promotor(s): Prof.dr. A.R. Thurik, Prof.dr. Ph.D. Koellinger, Prof.dr. P.J.F. Groenen, & Prof.dr. A. Hofman, EPS-2014-320-S&E, http://repub.eur.nl/pub/76907

Rijsenbilt, J.A., *CEO Narcissism: Measurement and Impact*, Promotor(s): Prof.dr. A.G.Z. Kemna & Prof.dr. H.R. Commandeur, EPS-2011-238-STR, http://repub.eur.nl/pub/23554

Rosch, D. *Market Efficiency and Liquidity*, Promotor: Prof.dr. M.A. van Dijk, EPS-2015-353-F&A, http://hdl.handle.net/1765/79121

Roza-van Vuren, M.W., *The Relationship between Offshoring Strategies and Firm* Performance: Impact of innovation, absorptive capacity and firm size, Promotor(s): Prof.dr. H.W. Volberda & Prof.dr.ing. F.A.J. van den Bosch, EPS-2011-214-STR, http://repub.eur.nl/pub/22155

Rubbaniy, G., *Investment Behaviour of Institutional Investors*, Promotor(s): Prof.dr. W.F.C. Verschoor, EPS-2013-284-F&A, http://repub.eur.nl/pub/40068

Schoonees, P. *Methods for Modelling Response Styles*, Promotor: Prof.dr P.J.F. Groenen, EPS-2015-348-MKT, http://repub.eur.nl/pub/79327

Shahzad, K., *Credit Rating Agencies, Financial Regulations and the Capital Markets*, Promotor(s): Prof.dr. G.M.H. Mertens, EPS-2013-283-F&A, http://repub.eur.nl/pub/39655

Sousa, M.J.C. de, *Servant Leadership to the Test: New Perspectives and Insights*, Promotor(s): Prof.dr. D.L. van Knippenberg & Dr. D. van Dierendonck, EPS-2014-313-ORG, http://repub.eur.nl/pub/51537

Spliet, R., *Vehicle Routing with Uncertain Demand*, Promotor(s): Prof.dr.ir. R. Dekker, EPS-2013-293-LIS, http://repub.eur.nl/pub/41513

Staadt, J.L., *Leading Public Housing Organisation in a Problematic Situation: A Critical Soft Systems Methodology Approach*, Promotor(s): Prof.dr. S.J. Magala, EPS-2014-308-ORG, http://repub.eur.nl/pub/50712

Stallen, M., *Social Context Effects on Decision-Making: A Neurobiological Approach*, Promotor(s): Prof.dr.ir. A. Smidts, EPS-2013-285-MKT, http://repub.eur.nl/pub/39931

Tarakci, M., *Behavioral Strategy: Strategic Consensus, Power and Networks*, Promotor(s): Prof.dr. D.L. van Knippenberg & Prof.dr. P.J.F. Groenen, EPS-2013-280-ORG, http://repub.eur.nl/pub/39130

Teixeira de Vasconcelos, M., *Agency Costs, Firm Value, and Corporate Investment*, Promotor(s): Prof.dr. P.G.J. Roosenboom, EPS-2012-265-F&A, http://repub.eur.nl/pub/37265

Troster, C., *Nationality Heterogeneity and Interpersonal Relationships at Work*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2011-233-ORG, http://repub.eur.nl/pub/23298

Tsekouras, D., *No Pain No Gain: The Beneficial Role of Consumer Effort in Decision- Making*, Promotor(s): Prof.dr.ir. B.G.C. Dellaert, EPS-2012-268-MKT, http://repub.eur.nl/pub/37542

Tuijl, E. van, *Upgrading across Organisational and Geographical Configurations*, Promotor(s): Prof.dr. L. van den Berg, EPS-2015-349-S&E, http://repub.eur.nl/pub/78224

Tuncdogan, A., *Decision Making and Behavioral Strategy: The Role of Regulatory Focus in Corporate Innovation Processes*, Promotor(s): Prof.dr.ing. F.A.J. van den Bosch, Prof.dr. H.W. Volberda, & Prof.dr. T.J.M. Mom, EPS-2014-334-S&E, http://repub.eur.nl/pub/76978

Uijl, S. den, *The Emergence of De-facto Standards*, Promotor(s): Prof.dr. K. Blind, EPS-2014-328-LIS, http://repub.eur.nl/pub/77382

Vagias, D., *Liquidity, Investors and International Capital Markets*, Promotor(s): Prof.dr. M.A. van Dijk, EPS-2013-294-F&A, http://repub.eur.nl/pub/41511

Veelenturf, L.P., *Disruption Management in Passenger Railways: Models for Timetable, Rolling Stock and Crew Rescheduling*, Promotor(s): Prof.dr. L.G. Kroon, EPS-2014-327-LIS, http://repub.eur.nl/pub/77155

Venus, M., *Demystifying Visionary Leadership: In search of the essence of effective vision communication*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013-289- ORG, http://repub.eur.nl/pub/40079

Vermeer, W., *Propagation in Networks:The impact of information processing at the actor level on system-wide propagation dynamics*, Promotor: Prof.mr.dr. P.H.M.Vervest, EPS-2015-373-LIS, http://repub.eur.nl/pub/79325

Visser, V.A., *Leader Affect and Leadership Effectiveness:How leader affective displays influence follower outcomes*, Promotor(s): Prof.dr. D.L. van Knippenberg, EPS-2013- 286-ORG, http://repub.eur.nl/pub/40076

Vlam, A.J., *Customer First? The Relationship between Advisors and Consumers of Financial Products*, Promotor(s): Prof.dr. Ph.H.B.F. Franses, EPS-2011-250-MKT, http://repub.eur.nl/pub/30585

Waltman, L., *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*, Promotor(s): Prof.dr.ir. R. Dekker & Prof.dr.ir. U. Kaymak, EPS-2011-248- LIS, http://repub.eur.nl/pub/26564

Wang, T., *Essays in Banking and Corporate Finance*, Promotor(s): Prof.dr. L. Norden & Prof.dr. P.G.J. Roosenboom, EPS-2015-352-F&A, http://repub.eur.nl/pub/78301

Wang, Y., *Information Content of Mutual Fund Portfolio Disclosure*, Promotor(s): Prof.dr. M.J.C.M. Verbeek, EPS-2011-242-F&A, http://repub.eur.nl/pub/26066

Wang, Y., *Corporate Reputation Management: Reaching Out to Financial Stakeholders*, Promotor(s): Prof.dr. C.B.M. van Riel, EPS-2013-271-ORG, http://repub.eur.nl/pub/38675

Weenen, T.C., *On the Origin and Development of the Medical Nutrition Industry*, Promotor(s): Prof.dr. H.R. Commandeur & Prof.dr. H.J.H.M. Claassen, EPS-2014-309-S&E, http://repub.eur.nl/pub/51134

Wolfswinkel, M., *Corporate Governance, Firm Risk and Shareholder Value*, Promotor(s): Prof.dr. A. de Jong, EPS-2013-277-F&A, http://repub.eur.nl/pub/39127

Yang, S., *Information Aggregation Efficiency of Prediction Markets*, Promotor(s): Prof.dr.ir. H.W.G.M. van Heck, EPS-2014-323-LIS, http://repub.eur.nl/pub/77184

Zaerpour, N., *Efficient Management of Compact Storage Systems*, Promotor(s): Prof.dr.ir. M.B.M. de Koster, EPS-2013-276-LIS, http://repub.eur.nl/pub/38766

Zhang, D., *Essays in Executive Compensation*, Promotor(s): Prof.dr. I. Dittmann, EPS-2012-261-F&A, http://repub.eur.nl/pub/32344

Zwan, P.W. van der, *The Entrepreneurial Process: An International Analysis of Entry and Exit*, Promotor(s): Prof.dr. A.R. Thurik & Prof.dr. P.J.F. Groenen, EPS-2011-234-ORG, http://repub.eur.nl/pub/23422

Erasmus Research Institute of Management - ERiM

THE USE OF ADVANCED TRANSPORTATION MONITORING DATA FOR OFFICIAL STATISTICS

Traffic and transportation statistics are mainly published as aggregated data, which are traditionally obtained through surveys and secondary data sources like public registers and companies' administrations. Nowadays, advanced monitoring systems are installed in the road network, offering more abundant and detailed transport information than surveys and secondary data sources. Usually, these rich data are applied to the research in the field of transportation planning. But they also provide opportunities to national statistics offices to update their databases and apply new methods to generate statistics. Transportation demand estimation and prediction are taken as examples. Quantitative information on transportation demand is important for national and regional policy makers who want to know the number of freight vehicles traveling from origins to destinations. Traditionally, they extract this information largely from the national statistics offices. Transportation research needs the demand data to understand transportation behaviour in the road network, such as congestion and pollution.

In the thesis, information methods and hierarchal Bayesian networks are used to demonstrate the approaches to estimate transportation demand. To forecast transportation demand, the hierarchical Bayesian network associated with the multi-process model is applied and tested. Additionally, an innovative concept of origin destination tuple is introduced. Origin destination tuples are able to represent trip chain observations obtained with cameras or Bluetooth scanners.

Design & layout: B&T Ontwerp en advies (www.b-en-t.nl) Print: Haveka (www.haveka.nl)

ERiM

The Erasmus Research Institute of Management (ERIM) is the Research School (Onderzoekschool) in the field of management of the Erasmus University Rotterdam. The founding participants of ERIM are the Rotterdam School of Management (RSM), and the Erasmus School of Economics (ESE). ERIM was founded in 1999 and is officially accredited by the Royal Netherlands Academy of Arts and Sciences (KNAW). The research undertaken by ERIM is focused on the management of the firm in its environment, its intra- and interfirm relations, and its business processes in their interdependent connections.

The objective of ERIM is to carry out first rate research in management, and to offer an advanced doctoral programme in Research in Management. Within ERIM, over three hundred senior researchers and PhD candidates are active in the different research programmes. From a variety of academic backgrounds and expertises, the ERIM community is united in striving for excellence and working at the forefront of creating new business knowledge.

ERIM PhD Series

Research in Management

Erasmus Research Institute of Management - ERiM
Rotterdam School of Management (RSM)
Erasmus School of Economics (ESE)
Erasmus University Rotterdam (EUR)
P.O. Box 1738, 3000 DR Rotterdam,
The Netherlands

Tel. +31 10 408 11 82
Fax +31 10 408 96 40
E-mail info@erim.eur.nl
Internet www.erim.eur.nl