DYNAMIC ORIGIN-DESTINATION MATRIX ESTIMATION ON MOTORWAY NETWORKS

This thesis is a result of a Ph.D. study performed at the Transportation Planning and Traffic Engineering Section of the Faculty of Civil Engineering of Delft University of Technology from 1991 to 1995

DYNAMIC ORIGIN-DESTINATION MATRIX ESTIMATION ON MOTORWAY NETWORKS

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus,
Prof. ir. K.F.Wakker
in het openbaar te verdedigen
ten overstaan van een commissie,
door het College van Dekanen aangewezen,
op dinsdag 7 mei 1996 te 16.00 uur

door

Nanne Jacob VAN DER ZIJPP

wiskundig ingenieur

geboren te Dacca, Pakistan

Dit proefschrift is goedgekeurd door de promotoren:

Prof.dr.ir. R. Hamerslag
Prof.dr. G.J. Olsder

Samenstelling promotiecommissie:

Rector Magnificus	-	voorzitter
Prof.dr.ir. R. Hamerslag	-	TU Delft, promotor
Prof.dr. G.J. Olsder	-	TU Delft, promotor
Prof.dr.ir. P.H.L. Bovy	-	TU Delft
Prof.dr. R.M. Cooke	-	TU Delft
Prof.dr.ir. J.J.M. Evers	-	TU Delft
Prof.dr.ir. A.W. Heemink	-	TU Delft
Dr.ir. E.C. van Berkum	-	Goudappel Coffeng

Zijpp, N.J. van der

Dynamic Origin-Destination Matrix Estimation on Motorway Networks
/ Nanne van der Zijpp - Delft
Universiteitsdrukkerij Delft - Ill. - With index, ref. - With summary in Dutch
ISBN: 90 - 9009499 - 7
NUGI: 849
Subject headings: dynamic origin destination matrix estimation, dynamic traffic management / traffic models

Distributed by:

Delft University of Technology
Faculty of Civil Engineering
Transportation Planning and Traffic Engineering Section
P.O.Box 5048
2600 GA Delft
The Netherlands

Printed in The Netherlands

Table of Contents

Appendices

Preface

This report is based on doctoral research into dynamic origin-destination matrix estimation that was carried out at Delft University, in the transportation planning and traffic engineering section, between 1991 and 1995. I would like to thank Rudi Hamerslag, professor of transport models, who supervised the research during all these years, and Geert Jan Olsder, professor of mathematical systems theory, for his supervision during the last stages.

I feel privileged for the facilities that TU Delft has offered me. My stay at Delft University was financed from various contributions and contracts. I would like to acknowledge the 'Beek stimuleringsfonds' and the 'Cornelis Lelystichting' for their financial support, and would like to thank AKZO and the EC DRIVE office for granting contracts to TU Delft that helped to pay for my appointment. Furthermore, I thank NWO, Shell, and the 'Universiteitsfonds', for their travel related contributions.

During the years, my research has been influenced by many people. I would like to thank Ben Immers for his concern with my research in its early stages, and professor Hobeika for inviting me to the Virginia Tech Center for Transportation Research. I thank all the members of this center for making this not only a useful, but also a very pleasant stay. I would like to thank the students I cooperated with: Tammo Hoeksema, Namita Arora, Jeroen Rijsdijk, my colleagues at TU Delft, in particular Hein Botma, Erik de Romph, Jim Stada, Lori Tavasszy, and Marcel Westerman, and my present colleagues at the Center for Transport Studies in London. The discussions we have had, have inspired me a lot, and have helped to shape my ideas on transport research.

Special thanks also to professor Bovy, for his constructive criticism on the earlier versions of this report, to Sylvia Wamsteker for taking care of a variety of administrative and organizational tasks, and to David Crosta for proof-reading the main text of this thesis.

Most of all, however, I owe to Lucia van Velsen. Her continuous support enabled me to complete this thesis.

Rotterdam, April 1996 Nanne van der Zijpp

Overview of Notation

indices

i	entry
j	exit
r	origin
s	destination
m	number of entries
n	number of exits
l	number of link volume observations
h	number of license plate reader locations
k	link volume observation
a,b	license plate reader locations
t	departure period
p	period
$x(i,j)$	mapping of index i,j into vector, $x(i,j)=(i\text{-}1)n+j$

variables

$\tilde{q}(t)$	Vector (length m) of idealized entry volumes in period t
$q(t)$	Vector (length m) of entry volume observations, in period t
$\tilde{y}(t)$	Vector (length l) of idealized link volumes in period t
$y(t)$	Vector (length l) of link volume observations in period t
$f_{ij}(t)$	Number of trips from entry i to exit j, departing in period t
$f(t)$	EE flow vector (length mn). This vector contains the elements of the flow matrix, $f_{ij}(t)$, row by row, $f_{x(i,j)}(t)=f_{ij}(t)$ $i=1,\ldots m,\ \ j=1,\ldots n$
$e_{ab}(t)$	trajectory count
$e(t)$, $\theta(r,s)$	Vector of trajectory counts, location of element $e_{rs}(t)$ in the vector $e(t)$.
$g_{rs}^{ij}(t)$	Trajectory count contribution, the number of trips with departure period t that contribute to both EE flow $f_{ij}(t)$ and trajectory count $e_{rs}(t)$.
$g(t)$, $\phi(r,s,i,j)$	Vector of trajectory count contributions, location of $g_{rs}^{ij}(t)$ in this vector.
α_k	recognition rate at site k.
p_{rs}^{ij}	Trajectory count contribution probability; the probability that a trip in flow $f_{ij}(t)$ contributes to $e_{rs}(t)$.
p	Vector of trajectory count contribution probabilities.
$y^{\mathrm{H}}(t)$	Combined observation vector, $y^{\mathrm{H}\prime}(t)=[y'(t)\ e'(t)]'$
$b(t)$	Split vector (length mn). This vector contains the elements of the split-probability matrix, $b_{ij}(t)$, row by row, i.e. $b_{x(i,j)}(t)=b_{ij}(t)$ $i=1,\ldots m,\ \ j=1,\ldots n$

$\overline{b}(t)$	Estimated split vector (length *mn*)
$x(t)$	Displacement of split vector relative to historical value
$w(t)$	Drift variable (length *mn*)
$u(t)$	Systematic component in variation of split proportions (length *mn*)
$v(t)$	Observation error on link volume observations (length *l*)
$y^{H}(t)$	Combined observations in period *t*
$Y^{H}(t)$	Set of combined observation up to and including period *t*

matrices

t	Assignment map τ_{ijk}=1, if flow $f_{ij}(t)$ contributes to observation $y_k(t)$, and zero otherwise.
k	Path-license plate reader incidence map. κ_{ijr}=1 if route *i-j* uses license plate reader *r* and zero otherwise, *r*=1,2,…*h*
U	Link-flow incidence matrix (height *mn*, width *l*). $U_{x(i,j),k} = \tau_{ijk}$ $i=1,\ldots m,\ \ j=1,\ldots n,\ \ k=1,\ldots l$
$\tilde{H}(t)$	Idealized measurement matrix, a matrix with the same dimensions as *U*, but with the idealized entry volumes as its non-zero elements, i.e.: $\tilde{H}_{x(i,j),k}(t) = \tau_{ijk}\,\tilde{q}_i(t)$ $i=1,\ldots m,\ \ j=1,\ldots n,\ \ k=1,\ldots l$
$H(t)$	A matrix equivalent to $\tilde{H}(t)$, but with the idealized entry-volumes replaced with the observed volumes.
Q_t	Variance covariance matrix of the drift variable (size *mn*×*mn*)
R_t	Variance covariance matrix of the observation error (size *l*×*l*)
K_t	Kalman Gain matrix

operators

t(.)	truncation operator
r(.)	reflection operator

constants

0	a vector of zeros
1	a vector of ones
e_i	the *i*th unit vector
I	the identity matrix
$\mathrm{p}^{m,n}$	repeating column matrix of size *mn*×*n*, containing *n* columns of lengths *m*. The nonzero elements of $\mathrm{p}^{m,n}$ are defined with: $\mathrm{p}^{m,n}{}_{x(i,j),i}=1$ $i=1,2,\ldots m,\ j=12,\ldots n$

notational conventions

′	matrix transpose
$\mathbf{0}\leq x\leq\mathbf{1}$	the inequality applies to *all* elements of x
p(.)	probability density
P(.)	probability distribution
x~P(.)	x is a random variable with distribution P(.)
P(x|y)~	the conditional distribution of x is equivalent to:
MVN[μ,Σ]	multivariate normal distribution with mean μ, and variance covariance matrix Σ
MVN[μ,Σ]$\Big\|_{x}$	the value of the distribution in point x
Poisson[λ]	poisson distribution with parameter λ
TMVN[μ,Σ]	truncated multivariate normal distribution with parameters μ, and Σ
RMVN[μ,Σ]	reflected multivariate normal distribution with parameters μ, and Σ
MLNM[n,$p_1\ldots p_m$]	multinomial distribution with parameters n and $p_1\ldots p_m$
$\underset{x_1,\, x_2\ldots x_n}{\operatorname{argmax}} \mathrm{J}(x_1,\, x_2\ldots x_n)$	the arguments x_1,$x_2\ldots x_n$ that maximize the function J
max J(.)	the maximum of the function J

1. Introduction to the Problem

1.1 Background

The ascendancy of automotive technology has led to the development of a flourishing traffic engineering discipline. Authorities turn to traffic engineers for advice on issues related to planning and management of traffic. The *analysis* of traffic always plays an important role in approaching these issues. Therefore much research is directed towards fundamental issues, such as estimating the future number of travellers, their departure times and routes and determining their travel delays.

Variables that determine the state of a traffic system, change over time. However, traditional traffic analysis involves no time differentiation other than distinguishing between peak and off-peak periods. In traffic engineering, the term *static* is used to denote a methodology in which only time-aggregated variables and their mutual relations are considered, while the term *dynamic* is used to denote a methodology in which the development of the variables in time is of central importance.

The lack of practical applications of dynamic traffic analysis can be explained in part by a lack of appropriate theoretical development. Traffic is a result of human behaviour. This behaviour is only partly understood and this is a cause of large inherent uncertainty. Another factor that discourages dynamic analysis is the lack of data. Reliable, detailed, time-differentiated traffic data are needed to test hypotheses and to apply dynamic models in practice. In general, traffic data that satisfy these demands are not available. Finally, the importance for practical applications of this kind of analysis may not have been fully recognized.

In the last decade, various factors have contributed to placing dynamic traffic analysis high on the agenda of transport research. These factors are connected with both *operational* and *planning* aspects.

Operational tasks include the control of individual intersections, the *coordinated* control of multiple intersections, information provision to travellers, route guidance, access control to motorways, and the like. As for supporting operational tasks with dynamic traffic analysis, an increasing awareness is growing that models that have the ability to forecast traffic conditions can contribute to the efficiency, safety and reliability of traffic systems. Also the increasing availability of facilities for automated data collection and the increasing cost of recurrent and non-recurrent congestion have contributed to this.

The term Advanced Transport Telematics (ATT) is now used for applications in the area of control of traffic and information provision to travellers. Research in the field of ATT is supported by large government programs such as the ‘Dedicated Road Infrastructure and Vehicle safety in Europe' (DRIVE) programme of the European Commission, the Intelligent Transportation Systems (ITS) programme in the USA, and the Vehicle Information and Communication System (VICS) program in Japan. These programmes function as umbrellas under which transport research can be classified.

Although most research is directed towards operational issues, *strategic* issues such as the planning of road infrastructure and travel demand management can also benefit from dynamic traffic analysis. The introduction of the time dimension in traffic analysis can lead to better answers to questions that a policy maker would be interested in, such as the impact of a proposed change in infrastructure on location, time of occurrence and extent of recurrent congestion. In particular, questions related to the environmental impact of traffic are difficult to answer on the basis of aggregated data only, as models that estimate noise production, exhaust emissions and smog need to be supplied with detailed characteristics of traffic flows. It is expected that in the near future new legislation in the USA and Europe will lead to extra requirements to the planning process of road traffic infrastructure, which will further stimulate the use of dynamic traffic analysis.

Dynamic traffic analysis encompasses a wide area of interrelated traffic phenomena. These phenomena may be traveller decisions such as trip frequency, destination choice, mode choice, route choice and departure time choice, but also the interaction between supply (of infrastructure) and demand (for travel). Insight in the latter category of phenomena is needed to estimate link capacities, speed density relationships, queuing, etc.

1.2 Objective of the Study

The objective of this study is the estimation of time varying *travel demand* on small road networks such as motorway corridors. This is a specific subproblem within the framework of dynamic traffic analysis. Estimates of travel demand are summarized in Entry-Exit (EE) matrices containing a number of trips for every combination of entry and exit. Likewise, time varying travel demand is summarized in a *dynamic* EE-matrix. A dynamic EE-matrix is a series of EE-matrices ordered with respect to trip departure time. For this purpose the time-axis is divided into intervals of which a typical length would be ten minutes. The elements of a dynamic EE-matrix are denoted by $f_{ij}(t)$, where i represents the entry, j represents the exit and t represents the departing period. For convenience of notation EE-matrices are rearranged in EE-flow vectors, denoted by $f(t)$.

EE-flows give rise to various categories of observations, such as traffic counts and survey data. These observations can be used to estimate the EE-flows. This thesis concentrates on the use of observations that can be collected in an automated manner. A typical example of such observations are traffic counts, which in many instances are collected routinely by road authorities using induction loops. Another example is the observation of individual vehicle trajectories using an *Automated Vehicle Identification* (AVI) technique such as license plate recognition based on image processing. The latter category of observations is referred to as *trajectory counts* (an exact definition will be given in a later chapter).

Let the vector of traffic counts and trajectory counts that relate to $f(t)$ be denoted by $y(t)$ and $e(t)$ respectively, and define the *combined observation* $y^{\mathsf{H}}(t)$ by:

$$y^{\mathsf{H}}(t) = \begin{bmatrix} y\,(t) \\ e\,(t) \end{bmatrix} \tag{1.1}$$

As errors are involved in the observation process, $y^{\mathsf{H}}(t)$ is not only a function of $f(t)$, but also

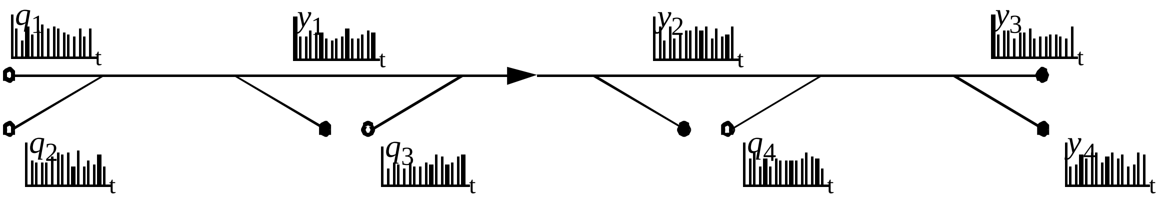


Figure 1.1: Example network. For all entries and a subset of the other links, time series of observations are available. Observations of entry volumes are denoted by $q_i(t),i=1,2,\ldots4, t=1,2,\ldots$, other traffic counts are denoted by $y_k(t),k=1,2,\ldots4$. Each EE-pair is connected via one route.

of a random vector $\varepsilon(t)$ that accounts for observation errors:

$$y^H(t)=y^H(f(t),\varepsilon(t)) \tag{1.2}$$

Let $Y^H(t)$ denote the set of all observations until and including period t, i.e. $Y^H(t)=\{y^H(1), y^H(2),\ldots y^H(t)\}$, then the problem considered in this thesis is estimating the vector $f(t)$ on the basis of available observations, i.e. the definition of an estimator $\bar{f}(Y^H(t))$ for $f(t)$.

The first part of the thesis deals with the simplified case, where the observation vector consists of traffic counts only. The extension to the case of combined observations is given in chapter 7.

1.3 Assumptions and Limitations

This section constrains the problem area by making a number of assumptions about the road network that is considered, the traffic flow characteristics on this network, and the availability and properties of the observations.

Network properties

Any network considered in this thesis is embedded in a *surrounding network*, and therefore is referred to as a *subnetwork*. A typical example of such a subnetwork is a motorway corridor. The union between the subnetwork and the surrounding network is referred to as the *full network*. The events on the surrounding network are assumed to be beyond our observation.

An important characteristic of the problem is that the subnetwork Entry-Exit flows that are to be estimated correspond with parts of Origin-Destination (OD) flows on the full network, and therefore depend on travel decisions taken in view of traffic conditions on the full network. Travel decisions may relate to departure time, destination, mode, route and the like.

For the subnetwork being considered, it is assumed that for each EE-pair only one connecting path exists. The subnetwork may hence be represented by a directed tree (see figure 1.1). This enables the definition of an *assignment map* t, with:

$$\begin{gathered}\tau_{ijk}=1, \text{ if the path from entry } i \text{ to exit } j \text{ traverses link } k\\ \tau_{ijk}=0, \text{ otherwise}\\ i=1,2,\ldots m,\ j=1,2,\ldots n,\ k=1,2,\ldots l\end{gathered} \tag{1.3}$$

Moving time coordinate system

Throughout the thesis the presence of a *moving time coordinate system* (MTCS) will be assumed. The idea behind a moving time coordinate system is that, if travel times can not be neglected, boundaries between consecutive periods are not given by fixed points on the time axis, but by time space trajectories, see figure 1.2. If these trajectories are chosen in an appropriate manner then the majority of vehicles complete their trip through the study area in one time zone, i.e. in figure 1.2, their trajectories do not cross boundaries of periods. Vehicles not satisfying these conditions give rise to *assignment errors*. The relative significance of assignment errors can be made arbitrarily small by increasing the length of the sampling periods (at the cost of the number of observations). Also specifying the travel delays more accurately reduces the assignment error. However, the development of new theory or tools for the estimation of travel times is not within the scope of the present thesis.

Formally, the construction of an MTCS may be thought of as follows:

- Divide the time axis into intervals, for example by using a regular grid. Let τ_i denote the boundaries between interval i-1 and interval i, i=1,2,….
- Choose an arbitrary location in the network, for example the most upstream node. Refer to this location as the *reference location*.
- Define the moving time coordinate system by the mapping $\{\mathbb{R}^+,P\}\rightarrow\mathbb{N}^+$

$$t^{\mathrm{MTCS}}(\tau,p)=\max\{\ i\mid\tau_i+\tau^{\mathrm{delay}}(\tau,p)<\tau,i\in\mathbb{N}^+\}$$

$$\tau\in\mathbb{R}^+,\ \ p\in P \tag{1.4}$$

where τ denotes the instant on the continuous time axis, $\tau^{\mathrm{delay}}(\tau,p)$ denotes the median of the distribution of travel delays on the path from the reference location to point p encountered by motorists who depart at the reference location at instant τ-$\tau^{\mathrm{delay}}(\tau,p)$, and P is the set of points in the network for which one wishes to define the moving time coordinate system.

Observations

The set of subnetwork links is divided into three disjunct categories: *entry links*, *exit links*, and a third category referred to as *internal links*. It is assumed that all entry flows and a part of the other link flows are observed. The traffic counts on the entry links are denoted by $q_i(t)$, i=1,2,…m, t=1,2,…, while counts on internal and exit links are denoted by $y_k(t)$, k=1,2,…l.

A distinction is made between the *idealized link flows* and their corresponding traffic counts. Idealized link flows are marked with the symbol ‘ ˜ ’ and are defined by sums of EE-flows with a given departure period, i.e.:

$$\tilde{q}_i(t)\equiv\sum_{j=1}^{n}f_{ij}(t) \tag{1.5}$$

and:

$$\tilde{y}_k(t)\equiv\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}(t)\,\tau_{ijk} \tag{1.6}$$

where $\tilde{q}_i(t)$ are idealized entry flows and $\tilde{y}_k(t)$ are idealized internal link flows or exit flows.

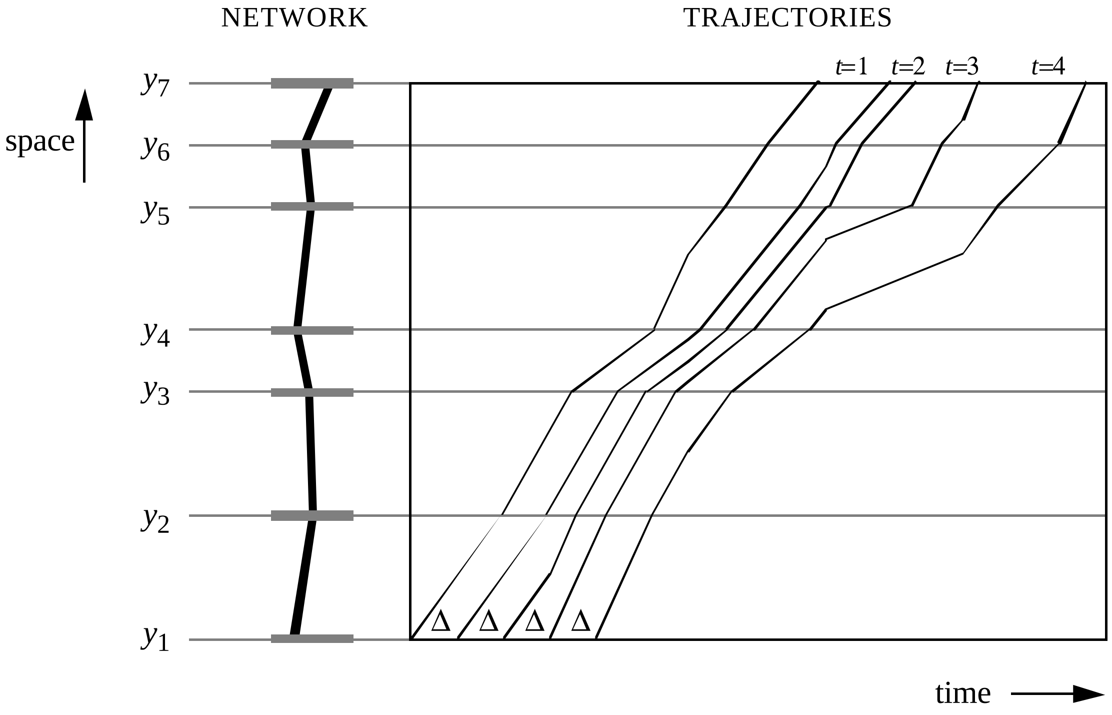


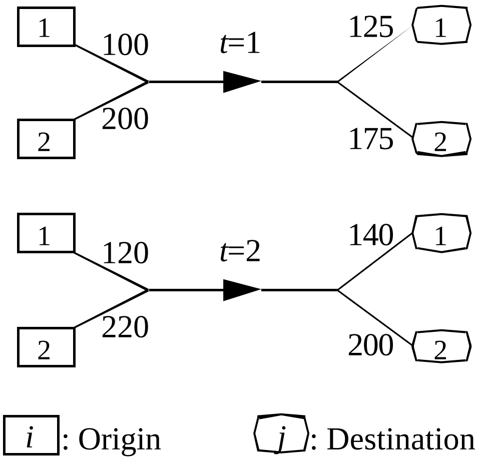
Figure 1.2: Moving time coordinate system. The time space trajectories divide the consecutive periods.

In the absence of observation errors or assignment errors, idealized flows and observed flows would be equal. The idealized link flows play an important role in specifying a motorway EE travel demand model in chapter 2.

Assumption of slowly varying split probabilities

In general, the EE flows can not be solved from a single set of traffic counts. For example in the light of equations (1.5) and (1.6), the flows on the links of the network shown in figure 1.1 correspond to 13 linear combinations of EE-flows. Due to the conservation of flows at the 5 internal nodes of the network, only 8 of these observations are linearly independent. On the other hand the topology of the network allows for 13 nonzero EE-flows. Therefore an infinite number of EE-matrices match the observations, and additional assumptions are needed to obtain a problem with a unique solution. This set of assumptions will be referred to as the *EE-travel demand model.* The combination of a model with a procedure to estimate its unknown parameters will be referred to as an EE-*estimation method*.

After a short discussion of alternative modelling assumptions at the beginning of chapter 2, this thesis concentrates on a class of EE-estimation methods based on the *assumption of slowly varying split probabilities*. Methods in this class will be referred to as *split ratio methods*. An introduction to this class of methods is given by *Cremer* in *Papageorgiou (1991)*, pp. 310-315. When split ratio methods were first introduced (*Cremer and Keller, 1981*) *split proportions*, denoted by $b_{ij}(t)$, and defined as the proportion of vehicles entering at entry i destined for exit j, were central to these methods. The idea behind split ratio methods is illustrated in figure 1.3. By assuming that the splits b_{11} and b_{21} remain constant during two periods, two



t=1:
$$100\, b_{11} + 200\, b_{21} = 125$$

t=2:
$$120\, b_{11} + 220\, b_{21} = 140$$

$\boldsymbol{b_{11}}$**=.25** $\boldsymbol{b_{12}}$**=.75**

Figure 1.3: Application of the split ratio principle: assuming the split proportions are constant during two periods, results in two independent linear equations with two unknowns; from these equations the split proportions can be solved.

independent linear equations can be set up from which b_{11} and b_{21} can be solved. Figure 1.3 also shows that a group of split proportions related to a common exit can be solved separately from the other split proportions.

Elaborating on *Van Der Zijpp and Hamerslag (1993)*, in this thesis we will be concerned with split probabilities rather than split proportions. This distinction is only small, but occasionally leads to new insight, such as a derivation of spatial correlations between traffic counts in chapter 5.

<u>*Estimating future entry flows*</u>

After the split probabilities have been estimated, the estimate for the flow vector is obtained by multiplying the observed entry flows with the estimated split probabilities, i.e.:

$$\bar{f}_{ij}(t) = q_i(t)\bar{b}_{ij}(t\text{-}1) \qquad (1.7)$$

where $\bar{b}_{ij}(t\text{-}1)$ is the most up to date estimate of $b_{ij}(t\text{-}1)$ at the end of period t-1. This implies that the prediction horizon for each EE-pair equals the traveltime for the connecting route, as the entry flows must be available before an estimate of a future flow can be made. If a longer prediction horizon is needed, the entry flows need to be extrapolated, for example using historical patterns. This is likely to result in loss of accuracy. However, empirical data show that little day to day variation exists in the entry flow patterns, see e.g. *Van Der Zijpp (1993)*. A method to estimate future entry flows based on historical patterns and a scaling algorithm was described in *De Romph (1994)*. This issue will not be further addressed in this thesis.

1.4 Outline of the thesis

The problem that was described in the present chapter is in its heart an underspecified problem: one set of link flows may correspond to many different EE-matrices. Hence, it is not possible to uniquely identify an EE-matrix from observations that are derived from link flows, or in the dynamic context, to uniquely identify a sequence of EE-matrices from time series of such observations.

To eliminate the underspecification, a model referred to as the motorway model will be specified in chapter 2. On a number of points, usage of this model refines a number of earlier proposed dynamic approaches. For example in the motorway model, the split *probabilities* supersede the split *proportions* as the unknown parameters. This implies that a part of the variation in EE flows that was formerly subscribed to variation in the unknown parameters, is now statistically described by a random selection process that is inherent to individual motorists making uncoordinated travel decisions. To expand the practical applicability of dynamic EE-estimation methods, the commonly used restrictions that the entry flows are exactly known and traffic counts do not involve internal links are relaxed by explicitly taking errors in entry volume observations into account and allowing for the use of internal link counts. Furthermore, due to the moving time coordinate system that was introduced in the present chapter, the motorway model is applicable to larger networks, as the assumption of small travel *times* is relaxed into the assumption of small travel time *dispersion*.

Taking the motorway model as point of departure a second step in the research involves deriving estimators for its parameters, deriving system properties following from the model, and investigating how additional sources of information may be used to improve the estimates.

Chapter 3 contains a literature review of dynamic EE-estimation methods. To enable the use of the traditional methods as a reference for alternative ones, extensions of the existing methods that can process internal link flows are described, leaving the original versions that use only entry and exit flows as a special case. In chapter 4, a new estimation method is proposed. This method will be based on Bayesian updating and is aimed at overcoming a number of problems that were found with the traditional methods. The method will compute a probability distribution rather than a point estimate for the unknown split probabilities. The derivation of point estimates from this distribution will be referred to as postprocessing, and a number of alternative routines for this purpose will be proposed.

System properties implied by or applicable to the motorway model are considered in chapters 5 and 6. Chapter 5 exploits the relations described in the motorway model to derive spatial correlations between traffic counts. Knowledge of these properties combined with usage of proper statistical tools is expected to lead to better estimates of the split probabilities and better founded judgements about the reliability of these estimates. Chapter 6 puts the assumption of slowly varying split probabilities to the test through an analysis of tolltickets that were collected by road authorities in France. This analysis also reveals that there is very little day to day variation in the split proportions, implying that historic information is highly relevant for the estimation of split probabilities. A way is proposed to utilize this historic information.

In the future new technologies will make it possible to trace individual vehicles in an automated manner either by installing Automated Vehicle Identification (AVI) equipment at multiple locations or by letting vehicles transmit their trajectories. In chapter 7 it is investigated how data arising from these technologies may be used in combination with induction loop data. The aim is the development of a method that uses a mixture of historic data, traffic counts, automated license plate surveys, and maybe data obtained from a group of vehicles equipped as probe vehicles.

In the last part of the thesis, many of the theoretical findings have been tested in two series of experiments. The first series of experiments involves synthetic EE-flows and traffic counts that are generated according to the specifications of the motorway model (chapter 8). The second series of experiments involves traffic counts on the Amsterdam beltway that were collected during one month (chapter 9).

A separate series of appendices deals with technical issues, like the details of the minimization algorithms that are needed to implement the EE-estimation methods (appendix A), an approximation of the mean and the variance of the special class of statistical distributions that was introduced in chapter 4 (appendix B), notational conventions and mathematical preliminaries needed for chapter 5 (appendix C). An overview of the notation used in the thesis is found prior to the present chapter.

2. Modelling Motorway EE-Travel Demand

2.1 Introduction

As has been illustrated in chapter 1, the EE flows can not be solved from a single set of traffic counts. Therefore additional assumptions are needed to obtain a problem with a unique solution. This set of assumptions will be referred to as the *EE-travel demand model*. Such a model may be based on various principles, such as closeness to a prior matrix, compliance with a static model, maximization of the number of micro states or minimization of the total travel time in a system, see e.g. *van Zuylen and Willumsen (1980)*, *Maher (1983)*, *Cascetta and Nguyen (1988)*, *Hamerslag and Immers (1988)*, and *Bell (1991b)*. A distinction is made between models that describe momentary interrelations between EE-flows (static models) and models that describe how EE-flows develop in time (dynamic models). This chapter discusses models in both categories, with an emphasis on a dynamic model to which the assumption of slowly varying split probabilities is central, referred to as the *motorway model*. In various ways, usage of this model represents an elaboration of approaches known from literature, such as proposed by *Cremer and Keller (1981, 1984, 1987)*, *Keller and Ploss (1987)* and *Nihan and Davis (1987, 1989)*. Later chapters will discuss properties that follow from the motorway model and estimators for its parameters.

2.2 Static models

EE-travel demand models that describe momentary interrelations between EE-flows may help a transport planner in determining travel demand at an aggregate level, and may even reflect relationships that remain valid over a long planning horizon. Nevertheless these models are considered to be of limited significance for the dynamic EE-estimation problem as they require a certain level of aggregation to be sufficiently plausible.

Having said this it should be noted that due to lack of data, usage of static models often is the only option. For example if only for a few locations in the network, time series of traffic counts are available while one has time-aggregated counts for a larger number locations, a pragmatic approach is to compute a time-aggregated matrix by calibrating a static model, and to distribute this matrix over multiple time slices proportional to the patterns observed in the time-differentiated counts, see *De Romph et. al. (1994)*.

Apart from aggregation level requirements, another concern when applying a static model to a subnetwork such as a motorway corridor is that the derivation of many static models explicitly relies on the definition of a trip as being a movement from origin to destination, usually in relation to human activities such as work, leisure or shopping. For this reason the application of such a model to a subnetwork would not be theoretically justified.

An exception is a model that predicts the most likely trip matrix, given a set of traffic counts, to be the one that maximises entropy (*Van Zuylen and Willumsen, 1980*). Examples of dynamic or at least time-dependent approaches using this assumption are *Willumsen (1984)* and *VanAerde et. al. (1993).*

Another option is to experiment with a model for which validity for a full network implies validity for subnetworks (*Van Der Zijpp and De Romph 1994, 1995*). A model that was shown to have such properties is the well known gravity model with exponential deterrence function. This model defines the following relation:

$$f_{ij} = a_i b_j \exp(-\beta c_{ij}) \qquad (2.1)$$

with:

a_i	entry production ability
b_j	exit attraction ability
c_{ij}	generalized travel costs
β	parameter in deterrence function

2.2.1 Dynamic models

Applying dynamic models to the EE-estimation problem implies taking the time-varying nature of EE-flows into account. On the one hand this makes the estimation of EE-travel demand more complex, as compared to the static case, a larger number of unknown parameters need to be estimated. On the other hand, in combination with dynamic models one may use time-series of observations rather than time-aggregated observations.

The dynamic models that are considered in the present context impose certain continuity requirements on the EE-flows or variables related to EE-flows. Such modelling assumptions help to resolve the under-specification of the dynamic EE-estimation problem, as now multiple sets of observations become relevant for the EE-flows in a given period.

As an example consider the assumption of constant split ratios discussed in chapter 1. For the example network shown in figure 1.3 this assumption helps to solve the split proportions and hence the EE-flows. A similar line of reasoning can be applied to dynamic EE-estimation at different levels. This is illustrated in figure 2.1. In this example the gravity model (2.1) is referred to with arrow **A**. One can use the assumption that the *model parameters* a_i and b_j are slowly varying in time (arrow **B**) (see *Van Der Zijpp and De Romph 1994, 1995*), or one can use the assumption that the *matrix cells* are slowly changing (arrow **C**) (see *Ashok and Ben-Akiva, 1993*). Finally one can use the assumption that the *split parameters* are slowly changing (arrow **D**) (see e.g. *Cremer and Keller, 1981*).

Since the matrix of split proportions is implied by the matrix of EE-cells, assuming slowly varying EE-matrix cells also implies slowly varying EE-split matrix cells. The opposite of this is not true, as variation in entry flows may result in variation in EE-flows, while the EE-splits remain constant. Hence **C** represents a stronger assumption than **D**. For a similar reason **B** represents a stronger assumption than **C**.

2.3 Predictive capabilities versus model fit

A model can be considered as a mathematical simplification of the real world. In the previous sections various underlying ideas of models have been discussed. In the present context, calibration of a ‘hidden’ model, or estimation of its parameters, are used to select the best of

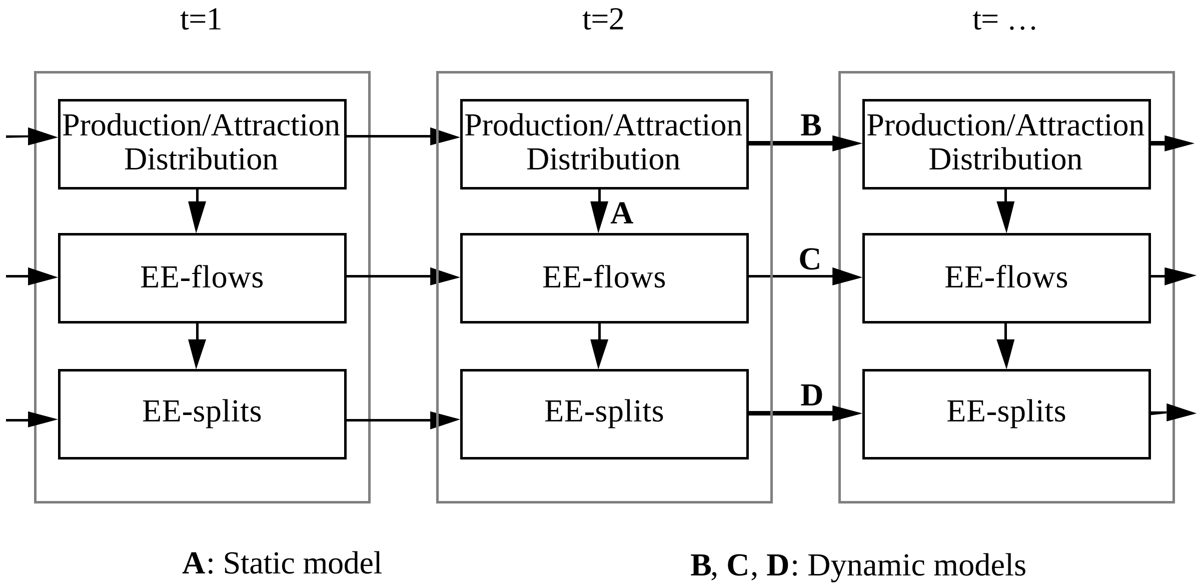


Figure 2.1: Structure of time-dependent model

many possible states in view of a given set of observations. An ever recurring issue when specifying a model is its level of complexity. On the one hand, there should be enough data to estimate the parameters in the model, i.e. the model should be identifiable. On the other hand the model should not oversimplify, i.e. it should not overrule observations that represent valuable information. The following example was taken from *Van Der Zijpp and Heydecker (1996)*, and discusses the issue of model selection.

<u>Example</u>

Consider a transport system for which the travel demand T is to be estimated. Travel demand is derived from human activities such as work, shopping or leisure activities and is influenced by a large number of factors. Suppose all factors that influence travel demand are summarized in a state vector s. This vector then implies the size of the travel demand vector $T(s)$. Observations are denoted as $o(s)$, and are assumed to represent sums of elements of the vector $T(s)$.

The present example is concerned with selecting the best model from a range of alternatives. In this context a model is thought of as a set of relationships to be satisfied by the EE-flow vector. In order to define ‘best’ in a more precise manner, define $\hat{T}^k(o)$ as the estimate corresponding to the kth model, and define the *expected model prediction error* for model k as:

$$\mathrm{E}_s[\ d(T(s),\hat{T}^k(o))\ ] \qquad (2.2)$$

The estimation procedure is to be made precise later, for the time being one can think of $\hat{T}^k(o)$ as a least squares estimate. The function $d(.,.)$ denotes some distance measure, e.g. L^2. Measure (2.2) is defined in terms of an unknown functional form $T(s)$ and a probability distribution of s, and hence does not give any direct guidelines in selecting a model.

In order to bound this measure to a minimum and a maximum, the *specification gap* is def-

defined as:

$$\mathrm{E}_s[\ d(T(s),T^{k}*(s))\ ] \tag{2.3}$$

where $T^k*(s)$ denotes the projection of $T(s)$ on the space of model predictions of model i, see figure 2.2. This gap provides a lower bound for (2.2). The more a model restricts the space of permitted solutions, the larger is this gap. The size of this gap does not depend on the quality or amount of data that is represented by $o(s)$.

Furthermore, define the *estimation gap* as:

$$\mathrm{E}_s[\ d(\hat{T}^{k}(o),T^{k}*(s))\ ] \tag{2.4}$$

This represents the expected difference between the *optimal model prediction* $T^k*(s)$ and the *actual model prediction* $\hat{T}^k(o)$. The estimation gap decreases if extra (well posed) constraints are added, either in the form of extra observations, or in the form of extra modelling assumptions. As an example of this mechanism, consider at the one extreme a ‘fixed-flow’ model that fixes all EE-flows to 1. For this model the estimation gap such as defined by (2.4) is zero. At the other extreme consider the ‘empty’ model that does not impose any restriction on the EE-flows. For this model the estimation gap reaches a maximum.

Due to the triangle inequality, a lower and an upper bound for (2.2) can be expressed using (2.3) and (2.4):

$$\mathrm{E}_s[\ d(T(s),T^{k}*(s))\ ] \le \mathrm{E}_s[\ d(f(s),\hat{T}_i(o))\ ] \le \mathrm{E}_s[\ d(f(s),T^{k}*(s))\ ] + \mathrm{E}_s[\ d(\hat{T}^{k}(o),T^{k}*(s))\ ] \tag{2.5}$$

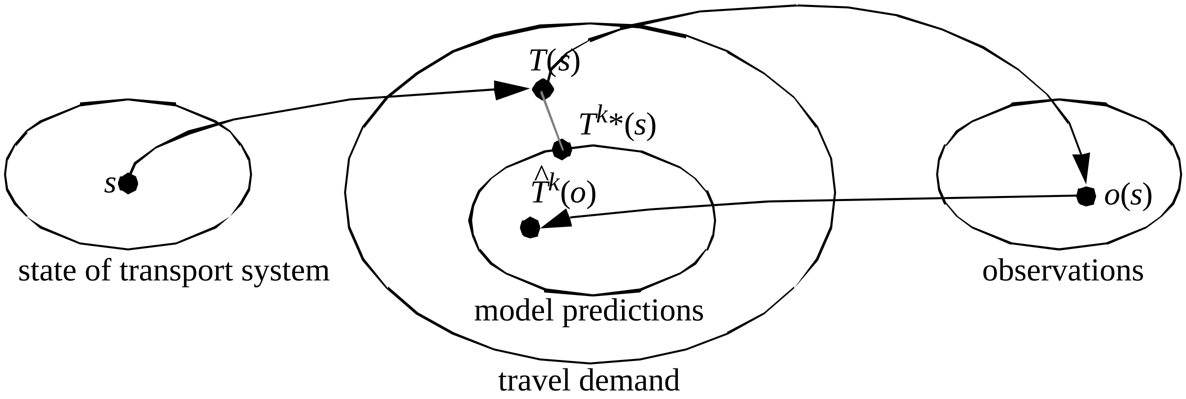


Figure 2.2: Travel demand, T(s), is a part of the system state, s. Observations, o(s), are derived from travel demand.

-end of example-

Above example illustrates that the best model choice depends on a trade-off between the specification gap and the estimation gap, and that the outcome of this trade-off largely depends on the amount and quality of data that are available. As a guideline for handling this trade-off the following principle will be used:

> “*of the models, that are identifiable in view of a given set of data, use the one that represents the weakest set of assumptions*”

Applying this principle to the EE-estimation problem at hand results in using the assumption of slowly varying split probabilities as a point of departure.

2.4 The motorway model

2.4.1 Model assumptions

In this section the model is specified that will allow us to estimate the EE-flows from time series of observations. At the heart of the model are the split probabilities $b_{ij}(t)$, defined by the probability that a trip that enters at i in period t is destined for exit j, i.e.:

$$b_{ij}(t) \equiv \text{Prob}[\text{ exit}=j \mid \text{entry}=i \wedge \text{period}=t\ ] \tag{2.6}$$

The split probabilities are summarized in a vector $b(t)$. Although the split probabilities are assumed to be fixed within each period, they are allowed to vary from period to period. This variation is expressed by the equation:

$$b(t+1)=b(t)+u(t)+w(t) \tag{2.7}$$

where $u(t)$ denotes the systematic component, and $w(t)$ denotes a zero mean random component. The twenty-four-hour average of $u(t)+w(t)$ is zero. However, historic information may give rise to specifying nonzero values of $u(t)$, see chapter 6. The random variation in the split probabilities is only small, i.e.:

$$\text{E}[w_{ij}(t)^2] << 1 \tag{2.8}$$

The covariance matrix of $w(t)$ will be denoted with Q_t, the properties of this matrix are considered in more detail in chapter 6. The distribution of the initial state $b(1)$ is assumed to be uniform over the range of permitted values.

A difference exists between on the one hand the idealized entry flows $\tilde{q}(t)$ and observed-link flows $\tilde{y}(t)$ defined by (1.5) and (1.6), and on the other hand their corresponding traffic counts $q(t)$ and $y(t)$. These differences are due to physical observation errors, i.e. miscounting, and in case of $y(t)$, also to assignment errors. Assignment errors occur as a result of an inaccurate specification of the MTCS (see section 1.3), or as a result of travel time dispersion among vehicles. The relationship between $[\tilde{q}'(t)\ \tilde{y}'(t)]$ and $[q'(t)\ y'(t)]$ is defined by:

$$q(t) = \tilde{q}(t)+r(t) \tag{2.9}$$

and:

$$y(t) = \tilde{y}(t)+s(t) \tag{2.10}$$

where $r(t)$ and $s(t)$ are zero mean noise terms.

Finally, it is assumed that the noise terms $w(t)$, $r(t)$ and $s(t)$, t=1,2,… not only are zero mean, but also mutually independent, i.e:

$$\text{E}\begin{bmatrix} w(t) \\ r(t) \\ s(t) \end{bmatrix} = \mathbf{0}, \quad \text{E}\left[\begin{bmatrix} w(t) \\ r(t) \\ s(t) \end{bmatrix} \begin{bmatrix} w'(p) & r'(p) & s'(p) \end{bmatrix}\right] = \begin{bmatrix} Q_t & & \\ & \Phi_t & \\ & & \Theta_t \end{bmatrix} \delta_{tp} \tag{2.11}$$

The definition (2.6) and assumptions (2.7-2.11) will be referred to as the *motorway model.* A graphical representation of the causal dependencies specified in this model is given in figure 2.3.

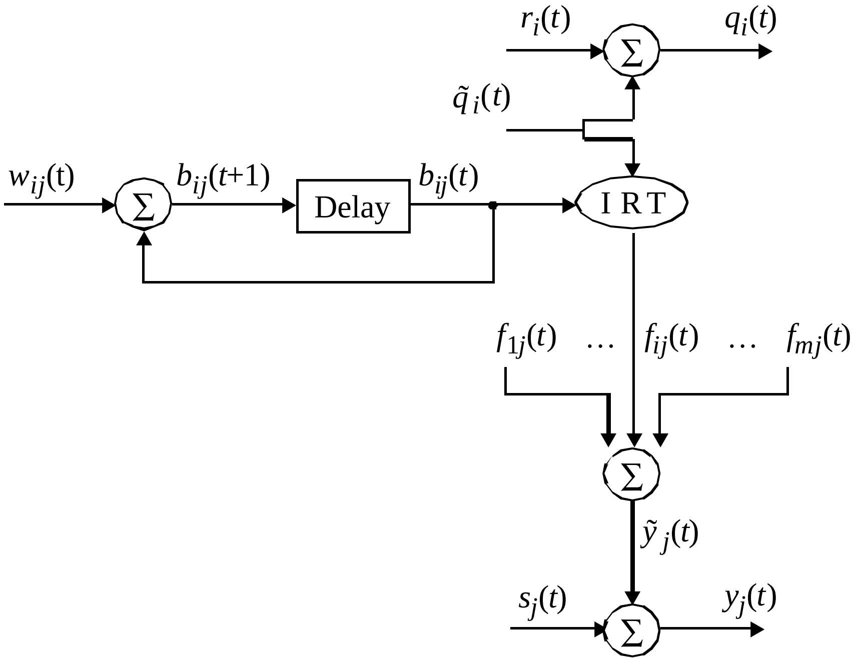


Figure 2.3: Causal dependencies in the motorway model. According to the model, the idealized entry flows $\tilde{q}_i(t)$ are distributed over the EE flows $f_{ij}(t)$ in a series of Independent Random Trials (IRT), using the split probabilities $b_{ij}(t)$.

2.4.2 Model implications

The motorway model specifies a number of causal dependencies, and has the vectors of split probabilities $b(t)$, $t \in \mathbb{N}^+$, as its unknowns. In this thesis we consider the problem of determining the most up to date estimate of $b(t)$ given the observations $\{q(1), q(2)\ldots q(t)\}$ and $\{y(1), y(2)\ldots y(t)\}$, i.e. the problem of *filtering* the process $\{b(t)\}$ from the observation processes $\{q(t)\}$and $\{y(t)\}$.

The split probabilities can only be observed via the EE flows $f_{ij}(t)$. In the light of definition (2.6), the EE flows $f_{ij}(t)$ should be considered as realizations of random variables; According to (2.6), the EE flows entering at *i* during period *t* should be considered as the cumulative outcomes of $\tilde{q}_i(t)$ independent trials, where the probability of contributing to EE flow $f_{ij}(t)$ is given by split probability $b_{ij}(t)$. The conditional probability distribution of $f(t)$, given the idealized entry flows $\tilde{q}(t)$ and the split probabilities $b(t)$ is hence given by the following multinomial distribution, see *Lehmann (1983)*, pg.28:

$$\mathrm{P}[f(t)|\tilde{q}(t),b(t)]=\prod_{i=1}^{m}\left(\frac{\tilde{q}_i(t)!}{\prod_{j=1}^{n} f_{ij}(t)!}\prod_{j=1}^{n} b_{ij}(t)^{f_{ij}(t)}\right) \tag{2.12}$$

The flows $f(t)$ are not observed directly but give rise to observations $q(t)$ and $y(t)$. All knowledge about the dependency of $q(t)$ and $y(t)$ on $b(t)$ is captured in the likelihood function

L[$b(t)$;$q(t)$,$y(t)$] defined by:

$$L[b(t);q(t),y(t)] \equiv p[q(t),y(t)|b(t)] \tag{2.13}$$

The assumptions of the motorway model imply a characteristic distinction between the two categories of observations, $q(t)$ and $y(t)$: For $q(t)$, there is no causal dependency on $b(t)$ while for $y(t)$ there is (see also figure 2.3). The absence of causal dependencies of $q(t)$ on $b(t)$ implies that the probability distribution of the entry volume counts is invariant for the split-probabilities, i.e. $P[q(t)|b(t)]=P[q(t)]$. Therefore the likelihood (2.13) simplifies to:

$$L[b(t);q(t),y(t)]=p[y(t)|q(t),b(t)].p[q(t)] \tag{2.14}$$

In the context of estimating the split probabilities $b(t)$, the entry volume counts may hence be dealt with as if they were given constants, and instead of considering the likelihood function (2.14) we may as well consider:

$$L[b(t);y(t)] \equiv p[y(t)|b(t),q(t)] \tag{2.15}$$

It is not possible to evaluate L[$b(t)$;$y(t)$] in an analytical way. This would require extra assumptions about the distribution of the random variables $r(t)$ and $s(t)$, but even then would require an explicit analytical expression for the conditional probability distribution of $\tilde{y}(t)$, given $\tilde{q}(t)$ and $b(t)$. As the elements of $\tilde{y}(t)$ are defined by sums of the multinomially distributed flows $f(t)$, see (1.6), a tractable expression for $p[\tilde{y}(t)|b(t),\tilde{q}(t)]$, let alone for $p[y(t)|b(t),q(t)]$, is not available. The absence of an analytical expression for the likelihood (2.15) needs not be an obstacle for the maximum likelihood estimation of the split probabilities; *Nihan and Davis (1989)* describe such an estimator, details of which will be given in section 3.4.1. However, this estimator is only valid under the rather restrictive assumptions $r(t)=s(t)=w(t)=\mathbf{0}$, $t\in \mathbb{N}^+$, and hence does not apply to the general case described by the motorway model.

More widely applicable results are obtained if $p[y(t)|q(t),b(t)]$ is described in terms of the first and second moment. The first moment corresponding to $p[y(t)|q(t),b(t)]$ is given by $E[y_k(t)|q(t),b(t)]$, for $k=1,2,\dots l$. This expression will be evaluated below. In view of equation (2.10) $E[y_k(t)|q(t),b(t)]$ satisfies:

$$E[y_k(t)|q(t),b(t)]=E[\tilde{y}_k(t)|q(t),b(t)]+E[s_k(t)|q(t),b(t)]=E[\tilde{y}_k(t)|q(t),b(t)] \tag{2.16}$$

As a result of lemma (C.1) it follows that:

$$E[y_k(t)|q(t),b(t)]=E_{\tilde{q}}[\ E[\tilde{y}_k(t)|\tilde{q}(t),b(t)]\ |q(t)\ ] \tag{2.17}$$

When expanding $\tilde{y}_k(t)$ applying (1.6), and using:

$$E[f_{ij}(t)|\tilde{q}_i(t),b_{ij}(t)]=\tilde{q}_i(t)b_{ij}(t), \tag{2.18}$$

and:

$$E_{\tilde{q}}[\tilde{q}_i(t)|q(t)]=q_i(t), \tag{2.19}$$

the following result is obtained:

$$E[y_k(t)|q(t),b(t)]=E_{\tilde{q}}[\ \sum_{i=1}^{m}\sum_{j=1}^{n}\tilde{q}_i(t)\,b_{ij}(t)\,\tau_{ijk}\ |q(t)\ ]=\sum_{i=1}^{m}\sum_{j=1}^{n}q_i(t)\,b_{ij}(t)\,\tau_{ijk} \tag{2.20}$$

Result (2.20) may be expressed in a more compact way using matrix notation:

$$\mathrm{E}[y(t)|q(t),b(t)]=H'(t)b(t) \tag{2.21}$$

where $H(t)$ is a matrix of which the height equals the number of elements in $b(t)$ and the width equals the number of elements in $y(t)$, and of which the nonzero elements are given by:

$$H_{x(i,j),k}(t) = \tau_{ijk}\, q_i(t) \tag{2.22}$$

for i=1,…m, j=1,…n, k=1,…l, and $x(i,j)$ representing the location of $b_{ij}(t)$ in the vector $b(t)$.

The fact that the expectation of $y(t)$ is linear in $b(t)$ gives rise to the use of the following *measurement equation*:

$$y(t)=H'(t)b(t)+v(t) \tag{2.23}$$

where $v(t)$ is a zero mean 'measurement error' accounting for:

1. specification errors in $H'(t)$ that are caused by observation errors in the entry volumes (only applicable if $r(t)\neq\mathbf{0}$),
2. random variation in the EE-patterns as a result of the uncoordinated choices of motorists, and
3. observation errors in the link volumes (only applicable if $s(t)\neq\mathbf{0}$)

From (2.11) it follows that $v(t)$ and $w(t)$ are independent and $v(t)$ and $v(p)$ are independent if $t\neq p$. Moreover $v(t)$ and $w(t)$ are zero mean, therefore:

$$\mathrm{E}\begin{bmatrix} w(t) \\ v(t) \end{bmatrix}=\mathbf{0},\quad \mathrm{E}\left[\begin{bmatrix} w(t) \\ v(t) \end{bmatrix}\begin{bmatrix} w'(p) & v'(p) \end{bmatrix}\right]=\begin{bmatrix} Q_t & \\ & R_t \end{bmatrix}\delta_{tp} \tag{2.24}$$

for some matrix R_t.

In chapter 5 it will be shown that the assumptions of the motorway model allow for the derivation of the covariance matrix R_t. In combination with appropriate statistical methods this should, at least in theory, lead to better estimates of $b(t)$, relative to those obtained when using an arbitrary matrix.

Besides the traffic counts $q(t)$ and $y(t)$ another source of information about $b(t)$ stems from the definition (2.6). From this definition it follows that the split probabilities should be nonnegative and smaller than unity;

$$0\leq b_{ij}(t)\leq 1 \tag{2.25}$$

for i=1,2,…m and j=1,2,…n. Moreover, for each entry, its associated split probabilities add up to one:

$$\sum_{j=1}^{n} b_{ij}(t)=1 \tag{2.26}$$

for i=1,2,…m. Constraints (2.25) and (2.26) are referred to as the *natural inequality and equality constraints*. Again, (2.25) and (2.26) may be written in a more compact format in matrix notation with:

$$\mathbf{0}\leq b(t), \tag{2.27}$$

and:

$$\mathrm{p}'b(t)=\mathbf{1} \qquad (2.28)$$

where p is a matrix with the height *mn* and the width of *m* of which the nonzero elements are given by:

$$\mathrm{p}_{x(i,j),i} = 1 \qquad (2.29)$$

for i=1,…m, j=1,…n, and $x(i,j)$ representing the earlier introduced arrangement of the elements $b_{ij}(t)$ in the vector $b(t)$. This matrix will be referred to as a *repeating column* matrix

2.5 Conclusions and further research

Jointly, equations (2.7), (2.23), (2.24), (2.27) and (2.28) describe a system, with the traffic counts $q(t)$ and $y(t)$ as its observations, and the split probabilities $b(t)$ as the parameters that are to be estimated, $t\in \mathbb{N}^+$.

In view of this estimation problem the following research approach was chosen:

- A summary is given of the methods known from literature that have in common the assumption of slowly varying split probabilities, and it is described how the underlying assumptions of these methods relate to the assumptions of the motorway model (see chapter 3).
- A new estimator is derived for the split probabilities $b(t)$ (see chapter 4). Derivation of such an estimator is needed as none of the known estimators deals with the inequality constraints (2.27) in a satisfactory manner.
- The assumptions of the motorway model are utilized to derive an approximation R_t of the covariance matrix for the noise term $v(t)$ (see chapter 5).
- Completely observed EE-matrices derived from toll tickets are analysed to get some quantitative insight into the rate of change represented by the covariance matrix Q_t of $w(t)$, and a mechanism is proposed to derive the vector representing the systematic component $u(t)$ in that change from historic data (see chapter 6).

The results obtained in chapters 4-5 can be extended to allow for the usage of a new category of AVI-based traffic observations as described in chapter 7. A comparison between the error of estimation of the methods known from literature and the new method proposed in chapter 4, using synthetic data is made in chapter 8, and in chapter 9 using empirical data.

3. Review of Split Ratio Methods

3.1 Introduction

This chapter reviews existing dynamic EE-estimation methods. The emphasis in the discussion will be on methods that are based on the assumption of slowly varying split fractions, referred to as split ratio methods. Many of such procedures have been proposed after the first publications on this subject, see e.g. *Cremer and Keller (1981).* A wide range of estimation techniques is used, varying from parameter optimization techniques such as least squares and constrained optimization to statistically based techniques like maximum likelihood estimation and Kalman filtering.

A few remarks apply to the contents of this chapter:

- Initially, split ratio methods were applied to intersections only. Later this was extended to linear networks such as motorway corridors. Both network structures have in common that each EE-pair is connected via at most one path. However in the case of motorway corridors the issue of determining travel delays and synchronising the observation periods arises. This issue was discussed in chapter 1 and will be ignored in this chapter except for section 3.4.5 where a number of approaches are discussed that deal with variation in travel time.
- Many of the earlier introduced methods assume that the availability of observations is limited to entry and exit counts. However, on motorway corridors, traffic counts of internal link volumes may also be available. The method that will be developed in later chapters will therefore be able to use internal link counts as a part of its input. To make a fair comparison possible, the presentation of the existing methods in this chapter also involves their extension to the use of internal link counts.
- Some of the methods described in this chapter, require that a quadratic function is minimized. As these minimizations tend to be computationally demanding and eventually will have to be performed in real time, a considerable amount of attention has been paid to the implementation of efficient algorithms. The findings on this subject have been reported in appendix A.

3.2 The existing methods in the context of the motorway model

In several ways, the assumptions in the motorway model deviate from those that are usually adopted when applying a split ratio method. This section discusses these differences in assumptions and their consequences for the resulting estimation problem.

One of these differences is the interpretation of the elements of $b(t)$. These elements are usually referred to as turning proportions, split ratio's, split parameters or split fractions, and are defined as ‘the proportion of traffic from entrance i destined for exit j’ (*Bell, 1991b, Bell et al., 1991, Nihan and Davis, 1987*), or simply with the formula $b_{ij}(t)=f_{ij}(t)/q_i(t)$ (*Cremer, 1983, Cremer and Keller, 1981, 1984, 1987, Gang-Len Chang and Jifeng Wu, 1994, Keller and Ploss, 1987*), as opposed to the definition of split probabilities given in equation (2.6).

Another difference is that usually no distinction between idealized flows and traffic counts is made, i.e. it is assumed that the traffic counts are error free observations of the idealized flows, and hence satisfy $q(t)=\tilde{q}(t)$ and $y(t)=\tilde{y}(t)$. In some instances the assumption $y(t)=\tilde{y}(t)$ is relaxed to $y(t)=\tilde{y}(t)+e(t)$, where $e(t)$ accounts for ‘travel time lags, counter errors, etc.’ (e.g. *Nihan and Davis, 1987*).

These assumptions are usually combined with the assumption that the state $b(t)$ is constant or slowly varying, i.e. $b(t)=b(t\text{-}1)$ or $b(t)\approx b(t\text{-}1)$.

Applying these assumptions leads to a measurement equation identical to (2.23). Also the state equation (2.7), inequality constraints (2.27) and equality constraints (2.28) are still applicable. Therefore the estimators that have been proposed in literature and are described in this chapter can also be applied to the estimation problem (2.7), (2.23), (2.24), (2.27) and (2.28) that is central to this thesis.

However, as a consequence of the traditional assumptions, a large portion of the variation in the quotient $f_{ij}(t)/q_i(t)$ will be attributed to variation in $b_{ij}(t)$ rather than to random effects inherent to the uncoordinated choices of motorists. Also observation errors contained in $q(t)$ and $y(t)$ are fully attributed to variations in $b(t)$. The traditional assumptions do not imply particular recommendations about the covariance matrix R_t of the measurement error $v(t)$, and leave the choice of this matrix open. Not all publications are clear about the covariance matrix that is assumed for $v(t)$. For example *Cremer and Keller (1987)* only mention that a time-independent covariance matrix was assumed. *Nihan and Davis (1987)* are more explicit, and put into words the method that is presumably used by most researchers. They report that a diagonal matrix is used which is defined by the identity matrix multiplied by a factor. This factor is treated as a design parameter and is fixed after some experimenting.

If the motorway model is adopted, the assumption $f_{ij}(t)=q_i(t)b_{ij}(t)$ is replaced with the assumption of a conditional distribution of $f(t)$ given by (2.12), and observation errors in $q(t)$ and $y(t)$ are modelled with the zero mean random variables $r(t)$ and $s(t)$. These assumptions allow for the derivation of a covariance matrix of the measurement error $v(t)$ in (2.23), as will be shown in chapter 5.

3.3 Implemented methods

Only a subset of all methods ever proposed in literature has been implemented as a reference for the method that will be developed in later chapters. These methods have been selected on the basis of the type of input data that they use, the system behaviour that is taken into account, and the results that were claimed in literature for these methods. Another criterion is that they can be applied in practical context i.e. no computational or data requirements should prevent the method from being used. In an attempt to make the review complete, methods that have not been implemented are discussed in the section ‘other methods’.

3.3.1 Least Squares Method

The least squares estimate is defined as follows:

$$\bar{b}^{\mathrm{LS}}(t) \equiv \underset{b}{\operatorname{argmin}}\ J(b, t) \tag{3.1}$$

where:

$$J(b, t) = \sum_{k=1}^{t} \|y(k) - H'(k)b\|^2 \tag{3.2}$$

where $H(k)$ is the measurement matrix defined in (2.22). Note that the split vector b is considered to be a constant here, but that the estimate of this vector, $\bar{b}^{\mathrm{LS}}(t)$, has a time index because with every new set of observations, the estimate is adjusted. When this method was first applied to EE-estimation, see *Cremer and Keller (1981)* and *Cremer and Keller (1984)*, the vector of observations consisted of exit volume counts only. Rather than simultaneously solving all split parameters, the split parameters associated with each exit were solved separately through minimization of:

$$\sum_{k=1}^{t} \left\| y_j(k) - \sum_{i=1}^{m} q_i(k)\bar{b}_{ij}(t) \right\|^2 \tag{3.3}$$

It can be shown that if $\{y(k)\}$ are vectors of exit flow counts, minimization of (3.1) and (3.3) will yield identical estimates. The matrix notation (3.1) however allows the method to deal with internal link counts too. The target J(b,t) can be rearranged to:

$$\mathrm{J}(t) = \left(\sum_{k=1}^{t} y(k)'y(k) \right) - 2\left(\sum_{k=1}^{t} y(k)'H'(k) \right) b + b' \left(\sum_{k=1}^{t} H(k)H'(k) \right) b \tag{3.4}$$

Therefore (3.1) may be written as:

$$\bar{b}^{\mathrm{LS}}(t) \equiv \underset{b}{\operatorname{argmin}}\ - 2\Psi(t)'b + b'\Omega(t)b$$

with:

$$\begin{aligned} \Psi(t) &= \Psi(t\text{-}1) + H(t)y(t) \\ \Omega(t) &= \Omega(t\text{-}1) + H(t)H(t)' \end{aligned} \tag{3.5}$$

The matrices $\Psi(t)$ and $\Omega(t)$ have sizes mn by 1 and mn by mn respectively. In order to compute the vector $\bar{b}^{\mathrm{LS}}(t)$ for which J(t) reaches its minimum, the gradient vector is set to zero:

$$-2\Psi(t) + 2\Omega(t)\bar{b}^{\mathrm{LS}}(t) = 0 \tag{3.6}$$

From which it follows that the solution to the minimization problem (3.1) is:

$$\bar{b}^{\mathrm{LS}}(t) = \Omega(t)^{-1}\ \Psi(t) \tag{3.7}$$

An implicit assumption in (3.7) is that the matrix of second derivatives of J(t), $\Omega(t)$, is not only invertible but also positive definite. The matrix $\Omega(t)$ is invertible if the rank of this matrix equals the number of columns in it. This is the case if *mn* independent columns can be found in the matrices $H(1)$, $H(2)$,…$H(t)$. If the matrix is invertible then the positive definite property follows from the structure of the matrix.

A non-recursive least squares approach like described above is usually referred to as *Least Squares* (LS).

From equations (3.5) and (3.7) it can be seen that it is possible to employ the least squares method using a constant amount of storage space. Moreover, it is generally known that the LS estimate can also be computed with a recursive algorithm (a derivation of such a recursive algorithm for the scalar measurement case of this problem was given in *Nihan and Davis,1987, 1989*). The recursive equivalent of (3.5) and (3.7) is given by (see *Ljung and Söderström, 1983*):

$$\begin{gathered}\bar{b}^{\mathrm{RLS}}(t) = \bar{b}^{\mathrm{RLS}}(t\text{-}1) + K_t\,[y(t)\text{-}\,H(t)'\bar{b}^{\mathrm{RLS}}(t\text{-}1)]\\ K_t = P(t\text{-}1)H(t)\,[H(t)'P(t)H(t) + \mathrm{I}]^{-1}\\ P(t) = P(t\text{-}1) - P(t\text{-}1)H(t)\,[H(t)'P(t)H(t) + \mathrm{I}]^{-1}H(t)'P(t\text{-}1)\end{gathered} \tag{3.8}$$

In this formula $P(t)$ can be shown to equal $\Omega(t)^{-1}$. The calculation time is reduced in comparison with the non-recursive algorithm (3.7), as now only a matrix with a height equal to the length of the measurement vector must be inverted rather than a matrix with a height that equals the length of the state vector. Another use for (3.8) is to point out the relationship between the least squares method and the Kalman filtering method. As is generally known, the least squares method can be considered as a special case of Kalman filtering.

Equation 3.8 defines the *Recursive Least Squares* (RLS) method.

A natural extension of the least squares method to take into account time variation in the split vector, is the introduction of a discounting factor. In this case the older observations are considered to be less relevant to the current state of the split vector and are discounted accordingly. The problem now changes to minimizing:

$$\mathrm{J}(\lambda, t) = \sum_{k=1}^{t} \lambda^{t-k} \| y(k) - H'(k)b \|^2 \tag{3.9}$$

The first practical application of this method to EE-estimation goes back to *Nihan and Davis (1987).* The solution to problem (3.9) can be derived along the lines of the derivation of the least squares method, and is given by:

$$\begin{gathered}\bar{b}(\lambda,t) = \Omega(\lambda,t)^{-1}\,\Psi(\lambda,t)\\ \Psi(\lambda,t) = \lambda\,\Psi(\lambda,t\text{-}1) + H(t)y(t)\\ \Omega(\lambda,t) = \lambda\,\Omega(\lambda,t\text{-}1) + H(t)H(t)'\end{gathered} \tag{3.10}$$

Also this algorithm has a recursive equivalent, see *Nihan and Davis (1987)* and *Ljung and Söderström (1983).* In this case however no parallel with the Kalman filter exists. Algorithm (3.10) will be further referred to as *Discounted Least Squares* (DLS). The parameter λ in this algorithm is one of the design parameters and should satisfy $0<\lambda\leq 1$. In practice some experimenting will be necessary to find the parameter value for which algorithm (3.10) gives the ‘best’ results. The DLS objective function is a generalization of the LS objective function. Therefore in the rest of this chapter with no lack of generality, the DLS objective function will be considered.

3.3.2 Inequality Constrained Least Squares Method

Formulas (3.7) and (3.10) do not guarantee that the natural inequality constraints (2.25) are being met. Imposing these conditions would therefore improve the estimate. On the other hand, this changes the minimization problem from an unconstrained minimization into an inequality constrained minimization problem:

$$\bar{b}^{\mathrm{ICLS}}(t) = \underset{b}{\mathrm{argmin}}\ - 2\Psi(t)'b + b'\Omega(t)b$$

subject to:

$$\mathbf{0} \leq b \leq \mathbf{1} \tag{3.11}$$

In fact this problem falls in the category of quadratic programming problems. It consumes much more computation time than the unconstrained problem. Note that the number of constraints is $2mn$. The constraints can also be written in the form:

$$\begin{gathered} \mathrm{e}^{ij}.b \geq 0 \wedge \mathrm{e}^{ij}.b \leq 1 \\ i=1,\dots m,\ j=1,\dots n \end{gathered} \tag{3.12}$$

where e^{ij} the unit vector with a 1 on location $x(i,j)$ and 0 elsewhere. A constraint is said to be *binding* at point b if the equality holds for that point. From (3.12) it follows that at most mn constraints can be binding at a time and that the binding constraints are *orthogonal*, i.e. $\mathrm{e}^{ij}.\mathrm{e}^{pq}=0$ if $i \neq p \vee j \neq q$. In appendix A a number of minimization procedures are described that employ this special property of the minimization problem.

3.3.3 Fully constrained least squares

Beside the inequality constraints, the split parameters must also satisfy certain equality constraints. As the split parameters denote the expected shares of entry flows that are destined for certain exits, the total of these shares should equal one by definition. Imposing this knowledge on the split estimates should therefore improve the estimate. This results in the following constraint minimization problem:

$$\bar{b}^{\mathrm{FCLS}}(t) = \underset{b}{\mathrm{argmin}}\ - 2\Psi(t)'b + b'\Omega(t)b \tag{3.13}$$

subject to:

$$\mathbf{0} \leq b \text{ and } \mathrm{p}'b = \mathbf{1} \tag{3.14}$$

where p is the repeating column matrix defined in (2.29). The method that corresponds to this minimization problem will be referred to as *Fully Constrained Least Squares* (FCLS), and was proposed in *Cremer and Keller (1987)*. Problem (3.13) may be converted to an inequality constraint problem. For this purpose, define a vector b^1 with:

$$b^1 \equiv [b_{11} \dots b_{1,n-1} \dots\dots b_{m1} \dots b_{m,n-1}]' \tag{3.15}$$

and a vector b^0 and matrix G in such a way that:

$$b = b^0 + Gb^1 \Leftrightarrow b_{x(i,n)} = 1 - \sum_{j=1}^{n-1} b_{x(i,j)} \quad , i=1,\ldots m \tag{3.16}$$

Now problem (3.13) can be restated as:

$$\bar{b}^{\text{FCLS}} = b^0 + G\bar{b}^1 \tag{3.17}$$

with:

$$\begin{gathered} \bar{b}^1 = \underset{b^1}{\text{argmin}} - 2[\Psi(t)'Gb^1 - b^{0\prime}\Omega(t)G]b^1 + b^{1\prime}G'\Omega(t)Gb^{1\prime} \\ \mathbf{0} \leq b^1 \\ \mathrm{f}^i . b^1 \leq \mathbf{1},\ i=1,\ldots m \end{gathered} \tag{3.18}$$

where the nonzero elements of the vector f^i are given by:

$$\begin{gathered} \mathrm{f}^i{}_{z(i,a)} = 1\ , \\ i=1,\ldots m\ ,\ a=1,\ldots n\text{-}1 \\ z(i,a) = (i-1)(n-1)+a \end{gathered} \tag{3.19}$$

Solving this inequality constrained problem and substituting the resulting $b^1(t)$ in (3.16) gives an estimate $\bar{b}(t)$ that satisfies all conditions. For solving this problem the algorithms of the inequality constrained problem can be used. Of these algorithms the projected conjugate gradients is the best alternative to obtain the exact solution.

Note that, in contradiction to (3.11), the constraints in problem (3.13) are no longer mutually orthogonal. This makes the projection of the search direction on space of feasible directions more complex, see appendix A for details.

3.3.4 Kalman Filtering

The previous methods can all be viewed upon as parameter optimization methods. They are designed to minimise the distance between measured and predicted values. Apart from these methods, a number of statistically based methods are identified. These methods are defined in terms of the probability distributions related to the unknown parameters $b(t)$. One of these methods is the Kalman filter. The Kalman filter is a widely applied method for parameter estimation in dynamic systems. It has been applied to the problem of EE-estimation by various researchers, starting with *Cremer and Keller (1987)* and *Nihan and Davis (1987).* Prior to using a Kalman filter, two equations should be supplied: the state equation and the measurement equation. The state equation describes how the unknown parameters evolve through time. The measurement equation describes the relation between the unknown parameters and the measurements. In both equations it is possible to specify uncertainty by way of noise terms. The first and second moments of the noise terms have to be specified.

In the present context the state parameters represent the split probabilities, and the state equation and measurement equation are given by equations (2.7) and (2.23) respectively. The properties of the noise terms are given in (2.24). For the methods described in literature the systematic component in the change of $b(t)$, represented by $u(t)$, is chosen to be zero.

Given a state equation and measurement equation, a recursive estimator for $b(t)$ is defined by the *Kalman filter* (*Kalman, 1960*). The Kalman filter equations for the problem (2.7), (2.23), (2.24) are:

$$\overline{b}(t)=\overline{b}(t\text{-}1)+K_t[y(t)\text{-}H(t)'\overline{b}(t\text{-}1)]+u(t)$$
$$K_t=\Sigma_{t\text{-}1}H(t)[H(t)'\Sigma_{t\text{-}1}H(t)+R_t]^{-1}$$
$$\Sigma_t=\Sigma_{t\text{-}1}-\Sigma_{t\text{-}1}H(t)[H(t)'\Sigma_{t\text{-}1}H(t)+R_t]^{-1}H(t)'\Sigma_{t\text{-}1}+Q_t \qquad (3.20)$$

These equations define a recursion that should be started with an initial estimate $\overline{b}(0)$ and an initial covariance matrix Σ_0. Given the assumptions (2.7), (2.23) and (2.24), the Kalman filter defines a minimum variance linear estimator, i.e. the estimate is a linear function of the measurements $y(1)\ldots y(t)$, and the filter implicitly finds the matrix A and vector c that solve the following problem:

$$\text{minimize }(A,c)\text{: E}[\|b(t)\text{-A.}[y(1), y(2)\ldots y(t)]\text{-}c\|^2] \qquad (3.21)$$

Moreover, this estimate can be shown to be unbiased. If, besides earlier assumptions, the noise terms and the initial state have Gaussian distributions, the Kalman filter can be shown to produce unbiased estimates that have minimum variance over all estimators, see *Anderson and Moore (1979)* or *Ljung and Söderström (1983).*

Reliability of estimates

Kalman filtering has numerous advantages such as the computational efficiency, the possibility to process interdependent measurements and its recursive formulation. An additional advantage is that together with the estimate for the split matrix, a covariance matrix is calculated. This matrix gives an indication of the reliability of the estimate. In theory, this matrix can tell a traffic engineer whether the accuracy of an EE-matrix estimate is sufficient or that extra observations need to be made available, for example by installing extra induction loops.

The reliability of the estimates generated with the Kalman filter however rely on the accuracy of the specifications (2.7), (2.23) and (2.24). Comparisons that have been made between Kalman filtering and other methods, by for example *Cremer and Keller (1987)*, can therefore not be seen apart from the choice of covariance matrices R_t and Q_t, and the initial state, defined by $\overline{b}(0)$ and Σ_0. Until now a satisfactory way to supply these values has not been presented. The issue of specifying proper values for R_t and Q_t is addressed in chapters 5 and 6 respectively.

Dealing with inequality constraints

Another fundamental problem with the application of the Kalman filter equations to the estimation of the split proportions is that there is no way to insure that the natural inequality and equality constraints are met. Under circumstances where one or more of the inequality constraints are binding, the existence of these constraints is in contradiction with the random walk assumption (2.7). Therefore from a theoretical viewpoint the Kalman filter can not be applied directly to the problem of EE-estimation. *Nihan and Davis (1989)* propose a scheme of ‘normalisation’ and ‘truncation’ but these operations undermine the theoretical justification of the statistical method. In *Van Der Zijpp and Hamerslag (1994a)* a number of modifications have been proposed to overcome this difficulty. These will be the basis for the development of a new estimation procedure in chapter 4.

In the present context we discuss how to prevent the traditional Kalman filter from producing estimates that do not satisfy the inequality constraints. If in (3.20), $\overline{b}(t)$ does not satisfy the inequality constraints (2.27) then the following constrained estimate represents a better choice:

$$\bar{b}^{\mathrm{CONSTR}}(t) \equiv \max(\mathbf{0}, \min(\mathbf{1}, \bar{b}(t))) \tag{3.22}$$

This can be seen as follows: Jointly (3.22) and the requirement $\mathbf{0} \le b(t) \le \mathbf{1}$ imply that for any $\bar{b}(t)$:

$$[b(t) - \bar{b}^{\mathrm{CONSTR}}(t)]'.[\bar{b}^{\mathrm{CONSTR}}(t) - \bar{b}(t)] \ge 0 \tag{3.23}$$

Hence:

$$\begin{aligned} \|b(t)-\bar{b}(t)\|^2 &= \|b(t)-\bar{b}^{\mathrm{CONSTR}}(t)+\bar{b}^{\mathrm{CONSTR}}(t)-\bar{b}(t)\|^2 \\ &= \|b(t)-\bar{b}^{\mathrm{CONSTR}}(t)\|^2 + 2[b(t)-\bar{b}^{\mathrm{CONSTR}}(t)]'.[\bar{b}^{\mathrm{CONSTR}}(t)-\bar{b}(t)] + \|\bar{b}^{\mathrm{CONSTR}}(t)-\bar{b}(t)\|^2 \\ &\ge \|b(t)-\bar{b}^{\mathrm{CONSTR}}(t)\|^2 \end{aligned} \tag{3.24}$$

and consequently:

$$\mathrm{E}[\ \|b(t)-\bar{b}^{\mathrm{CONSTR}}(t)\|^2\ ] \le \mathrm{E}[\ \|b(t)-\bar{b}(t)\|^2\ ] \tag{3.25}$$

Equation (3.25) shows that with respect to the expected error, $\bar{b}^{\mathrm{CONSTR}}(t)$ is at least as good an estimator as $\bar{b}(t)$. This fact is not contradicting the statement that $\bar{b}(t)$ is a solution to (3.21), as $\bar{b}^{\mathrm{CONSTR}}(t)$ is not linear.

As an alternative to applying statement (3.22) only to the output of the recursion (3.20), this statement may also be included *in* the recursion (3.20). In this case the constraining will affect the evolution of $\bar{b}(t)$ (but not of $\Sigma_b(t)$) in time.

We will refer to the latter strategy as *recursive constraining*. Since the state satisfies the inequality constraints (2.27) at all times, this seems like a useful thing to do. In fact, previous publications on the subject report similar strategies, see e.g. *Nihan and Davis (1987)*. At present it will not be possible to prove or disprove on theoretical grounds that recursive constraining leads to estimators with a lower expected error of estimation. Therefore the option of recursive constraining is tested separately, see chapters 8 and 9.

3.4 Other methods

Implementing and testing all split-estimation methods that have ever been proposed is not possible due to time constraints, and also not necessary to evaluate the theory that is described in this thesis. In this section a number of methods is discussed that for different reasons have not been implemented.

3.4.1 Maximum Likelihood

When applied to the problem of estimating the split probabilities in the motorway model the maximum likelihood (ML) estimate would be defined by:

$$\text{maximize: } \mathrm{P}[\mathrm{y}(1), \mathrm{y}(2), \ldots \mathrm{y}(t) | \bar{b}(t)] \tag{3.26}$$

Calculation of the ML-estimate normally requires the derivation of above probability distribution. The elements of $y(k)$ are sums of flows, and the conditional distribution of the flows given the split-probabilities and the entry flows is (see chapter 2 for more details):

$$\mathrm{P}\ [f_{i1}(t), f_{i2}(t) \ldots f_{in}(t) \,|\, q_i(t), b_{i1}(t) \ldots b_{in}(t)] \ = \ \frac{q_i(t)!}{\prod_j f_{ij}(t)!} \prod_j b_{ij}(t)^{f_{ij}(t)} \tag{3.27}$$

A tractable expression for the probability distribution of {y(1), y(2),…y(t)} is not available, as this would involve deriving a probability distribution for the sum of multiple multinomial random variables. *Nihan and Davis (1989)* presented an ML-approach that did not require this derivation, by using the ‘EM-algorithm’ as proposed in *Dempster et al. (1977).* This was done for the simplified system in which *b(t)* is constant rather than slowly varying, and in which no noise on the entrance volume observations is present. The resulting algorithm was non-recursive.

Another ML approach has been presented by *Bell et al. (1991)* (see also section 3.4.5). This approach is fully disaggregate and is particularly useful to describe the phenomenon of platoon dispersion. The proposed method needs individual vehicle data and is computationally too demanding to be useful in practice.

3.4.2 Recursive formula

The first use of a split ratio method appears to be reported by *Cremer and Keller (1981).* In this instance a recursive formula was proposed for tracking the split-parameters, and convergence was shown for this method. In later work of these authors the recursive formula was replaced by more ‘standard’ algorithms, like least squares and constrained least squares.

3.4.3 Correlation method

In order to apply the methods that have been described earlier in this chapter certain requirements with respect to the locations on which traffic is counted must be met. The minimum requirement is that traffic is counted at all entries and at at least one exit. In this case split probabilities can be estimated through minimization of (3.3). A method that does not have this limitation was proposed in *Keller and Ploss (1987).* It uses the cross correlation between entry flows and exit flows as an estimate for the split parameter:

$$\bar{b}_{ij} = \frac{\left[\sum_{k=1}^{t} (q_i(k) - \bar{q}_i) \cdot (y_j(k) - \bar{y}_j)\right]^2}{\left[\sum_{k=1}^{t} (q_i(k) - \bar{q}_i)^2\right] \cdot \left[\sum_{k=1}^{t} (y_j(k) - \bar{y}_j)^2\right]}$$

with:

$\bar{q}_i$: average value of $q_i(k)$, k=1,2,…t

$y_j(k)$: exit flow at exit j in period k

$\bar{y}_j$: average value of $y_j(k)$, k=1,2,…t (3.28)

The method has been used in a project that involved traffic prediction and network optimization, see *Ploss et al. (1990)*. The method can not be extended to deal with internal link counts. The method has not been involved in comparative tests since it is expected beforehand that its performance will be poor relative to methods such as RLS and FCLS.

3.4.4 Neural network approach

Neural networks are increasingly popular in traffic engineering and recently the first publications on EE-estimation using neural nets have appeared (*Yang et al. (1992), Vythoulkas (1993), Kikuchi et al. (1993), Shih-Miao Chin et al. (1994), Kwon and Stephanides (1994)*).

As this work does not reference the existing body of literature, comparative data between neural network methods and prediction minimization methods are not yet available.

Neural networks are expected to be successful when exit flows must be predicted from entry flows. A neural net can capture non-linear relationships between input and output data, and in practice an abundance of data are available to ‘train’ the network.

The prediction of exiting volumes implies that EE-estimation and traffic assignment are combined in one method. Practical problems arise however if the aim is to estimate EE-flows rather than link-flows. In this case there is no correspondence between the data that are available for training, which are link flows, and the data that are needed as output, which are EE-flows. The papers mentioned above, each solve this problem in their own way. Essentially, the above literature can be divided into three classes:

- Dynamic link volume predictors using upstream link volume counts
- Dynamic link volume predictors not using upstream link volume counts
- Dynamic EE-flow estimators

Dynamic link volume predictors using upstream link volume counts

In *Yang et al. (1992)* a two layered, feed-forward network was used with a ‘sigmoidal’ transfer function. Each input node corresponds to an entry, and each output node corresponds to an output. The network was trained using the squared-error as a performance criterion. The network was hence set up as a predictor of link flows. After the training was completed the weights of the connections from the input to the output layers are interpreted as the split-ratio’s. Also *Vythoulkas (1993)* sets up the neural net as a link volume predictor. However Vythoulkas also experiments with alternative training rules for the neural network.

Dynamic link volume predictors not using upstream link volume counts

Kwon and Stephanides (1994) make a comparison between a neural network based exit volume predictor and a ‘new’ model based prediction that was developed to this end. Neither of the methods use upstream link volume counts. This is an essential difference with the split ratio methods that were described in this chapter. This will very probably result in poor prediction results relative to methods that use upstream volume counts. On the other hand the prediction horizon of these methods is no longer limited to the travel time of the vehicles on the network.

Dynamic EE-flow estimators

Kikuchi et al. (1993) propose a method that from all proposed neural network predictors is the most similar to the split ratio methods. Using the example of the OD-estimation problem for a rapid transit line where the entering and exiting volumes at each station are observed and the OD-matrix is to be estimated, they use a neural network to predict split proportions (which in their paper are called ‘weights’). The penalty function that they use is equivalent to equation (3.1). The data needed for training the neural net are available from the ticket administration.- This seems a very sensible method and it would be interesting to compare the result of such a method with that of other split ratio methods.

Shih-Miao Chin et al. (1994) choose a slightly different approach: they train a neural network using an observed EE-matrix that was obtained via a license plate survey. The fact that a completely observed EE-matrix is needed for training is a major disadvantage. It makes implementation of the method expensive and sensitive to changes in the traffic patterns.

Conclusion

The neural network approaches are appealing since they offer an easy way to implement nonlinear regression, while statisticians have very much trouble doing so. On the other hand in at least a few publications, researchers seem to have been more concerned with the method itself than with the choice of data that was fed into the method.

For example some neural net based approaches predict EE-matrices or link flows on the basis of observed link flows from *one period*. In previous chapters it has already been shown that this is impossible, unless a model of travel demand is used, since the assignment of traffic is an irreversible process. So the best that can be said about these approaches is that the performance might be equal to that of other static methods; given the input data that are fed into the network it can not be expected to compete with methods that use time-series of observations.

Researchers who apply neural networks can take advantage of the models that were proposed within the framework of split ratio methods, and by doing so, offer a realistic alternative class of dynamic EE-estimation methods. A precondition to the successful application of neural networks to dynamic EE-estimation is however that time-series of observations are made available to the neural net. Many of the methods described in this chapter are defined in a recursive manner; these methods store only the last estimate and adapt this using the latest measurement. This suggests that a similar approach might be successful for neural networks too. In such a case the most recent estimate is fed back into the neural network as an input, creating the possibility for a recursion.

The development and implementation of such a method is however left as a future research topic. In this thesis we will concentrate primarily on improving EE-estimates by making use of improved traffic models and mathematical analysis.

3.4.5 Combined estimators

The methods that have been described until now assume knowledge of an unambiguous relation between dynamic EE-matrix and observed link flows. This implies that knowledge about travel times is present, see chapter 1. For the practical experiments in this thesis the travel times will be approximated from observations of vehicle speeds, and inaccuracies are compensated for by increasing the duration of the sampling periods. In literature however a few examples exist of methods that are aimed at simultaneously estimating travel times and EE-matrices. In this thesis such methods are referred to as *combined estimators*.

Some researchers have pointed out that travel times can also be determined from cumulative link flows only. The way this can be done is described in chapter 8, and is illustrated in figure 8.2. If link flows are used to determine travel time then the simple linear relation between split parameters and observations, (2.23), changes in a non-linear and highly complex relationship. *Gang-Len Chang and Jifeng Wu (1994)* describe this relationship and actually present estimation methods (based on the extended Kalman filter) to estimate the unknowns. Primarily this work is of theoretical value. In practice the estimation of travel times from link flows does not work due to accumulating errors in the observation of the cumulative link flows. Nevertheless elements of the proposed dynamic model formulation certainly have potential to improve dynamic EE-estimation methods especially in circumstances where travel time is an unknown factor.

Two much more simple approaches are proposed in *Bell (1991b)*. The first is based on the assumption of a geometrical distribution of the travel times. In this approach a *platoon dispersion factor* α_j is associated with each network exit, resulting in:

$$y_j(t) = (1-\alpha_j)y_j(t-1) + \alpha_j \Sigma_{i=1,2,\ldots m}\, b_{ij} q_i(t) \tag{3.29}$$

In other words if b_{ijk} is defined as the proportion of the flow $f_{ij}(t)$ that contributes to $y_j(t+k)$, then b_{ijk} is defined by:

$$b_{ijk}=b_{ij}\alpha_j(1-\alpha_j)^k \tag{3.30}$$

The advantage of this approach is that only one extra parameter per exit is introduced, while at the same time the model is extended to transport networks with travel time dispersion. A disadvantage is that random effects and dependencies are not modelled.

A second method proposed in *Bell (1991b)* is a method in which instead of one set of split parameters, three sets of parameters are estimated, where the first, second and third set corresponds with the fastest, middle, and slowest platoon respectively. This conceptually simple method has the disadvantage that the number of unknowns increase by a factor of three This means that in practice the number of independent equations, required to solve the unknown parameters, is multiplied by a factor of three. Moreover a serial correlation between the observations will arise with which it is hard to deal in a statistical correct manner.

Since with time passing by, the unknown parameters themselves are subject to change it might very well be impossible to estimate the unknown parameters with a satisfactory accuracy. Comparing the method proposed by *Bell (1991b)* with earlier methods such as proposed by *Cremer and Keller (1987)* and *Nihan and Davis (1987)* the main difference is that the earlier methods *assume* the travel time, for example on the basis of distances and observed speeds, while the method of *Bell (1991b)* implicitly *estimates* the travel time. It is an open question which of the methods works best in practice. This depends on the quality of the input data and the variability in the EE-demand and travel times. Presumably the method that works best in practice would be some intermediate form of the two variants.

The line of reasoning in *Bell (1991b)* can be taken one step further. In *Bell et al. (1991)* a fully disaggregate method is proposed that comes down to matching every entering vehicle with every exiting vehicle.

The method is described using the following symbols:

N	number of observed vehicles.
i_k	entry number of the *k*th entering vehicle
$t^1{}_k$	entry time of *k*th the entering vehicle
j_q	exit number of the *q*th exiting vehicle
$t^2{}_q$	exit time of the *q*th exiting vehicle
$t(i,j)$	average travel time from entrance *i* to exit *j*
σ^2	variance in travel time
Δ	matching map. $\Delta_{kq}=1$ if the *k*th entering vehicle corresponds with the *q*th exiting vehicle

Assuming that the travel times have a normal distribution, the following likelihood function follows:

$$\mathrm{L}[\Delta]=\prod\nolimits_{k=1,2,\ldots N}\prod\nolimits_{q=1,2,\ldots N}\left(\frac{b_{i_k,j_q}}{\sqrt{2\pi}\sigma}\exp-\frac{1}{2}\frac{(t_q^2-t_k^1-t(i_k,j_q))^2}{\sigma^2}\right)^{\Delta_{kq}} \tag{3.31}$$

The matching map should satisfy feasibility conditions in order to let every entering vehicle match with *exactly one* exiting vehicle:

$$\sum_{k=1,2,\ldots N} \Delta_{kq}=1, \text{ for } q=1,2,\ldots N$$
$$\sum_{q=1,2,\ldots N} \Delta_{kq}=1, \text{ for } k=1,2,\ldots N \tag{3.32}$$

Maximizing (3.31) under condition (3.32) gives the theoretical maximum likelihood estimate for the totally disaggregate EE-estimation problem. Not surprisingly computational constraints keep this method from being applied to problems of realistic size. The number of feasible matching maps is N!, and there does not seem to be a numerical method to minimise (3.31) within acceptable computation time.

However it is not inconceivable that by making some proper approximations (3.31) and (3.32) can be a basis for methods that use more aggregated data and at the same time incorporate travel time dispersion.

Finally, *Ping Yu and Davis (1994)* propose a nonlinear least squares method that replaces the linear traffic assignment with a nonlinear traffic flow model. They suggest that EE-estimation 'may require the joint estimation of EE patterns and traffic flow model parameters'.Their simulation results seem to support this hypothesis, although the differences with traditional methods are only small.

3.5 Conclusions

In the past a variety of dynamic EE-estimation methods have been proposed. These methods all work with less detailed modelling assumptions than those specified in the motorway model. In the chapters 4, 5 and 6 a method will be developed that takes all elements of the motorway model into account and estimates the unknown parameters according to statistical principles. The challenge is to show that this new method performs better then the existing ones. It is envisaged that the new method can be developed by adapting the Kalman filter.

In order to be able to compare the existing methods with the new method, the existing methods have been modified in such a way they can take a set of internal link counts as their input. The following methods have been prepared for comparison with the new method:

- Least Squares (equation 3.10)
- Inequality Constrained Least Squares (equation 3.11)
- Fully Constrained Least Squares (equation 3.13)
- Kalman filter (equation 3.20)

The results of comparisons based on synthetic and empirical data will be described in chapters 8 and 9.

4. A Bayesian Estimator of Turning Proportions

4.1 Introduction

The aim of this chapter is to formulate an estimator for a class of linear discrete time systems that are characterized by the fact that their states slowly vary and at all times satisfy a number of inequality constraints.

The method was designed for the purpose of estimating the split probabilities in the motorway model, thereby making use of the covariance matrix of the measurement error $v(t)$ that will be derived in chapter 5 and the covariance matrix of the random change $w(t)$ that will be derived in chapter 6. However, the findings of this chapter can easily be applied to other models for which the presence of inequality constraints is dominating.

As a starting point the following equations are used (see also section 2.4):

$$\begin{array}{ll}
\text{(a)} & b(t+1)=b(t)+u(t)+w(t) \\
\text{(b)} & y(t)=H'(t)b(t)+v(t) \\
\text{(c)} & \mathbf{0}\le b(t)\le \mathbf{1} \\
\text{(d)} & \mathrm{p}'b(t)=\mathbf{1} \\
\text{(e)} & \mathrm{E}\begin{bmatrix} w(t) \\ v(t) \end{bmatrix}=\mathbf{0},\quad \mathrm{E}\left[\begin{bmatrix} w(t) \\ v(t) \end{bmatrix}\begin{bmatrix} w'(p) & v'(p) \end{bmatrix}\right] = \begin{bmatrix} Q_t & 0 \\ 0 & R_t \end{bmatrix}\delta_{tp}
\end{array} \tag{4.1}$$

where:

$b(t)$ is the state vector with height $m.n$, $t\in I\!N^+$
$u(t)$ is the systematic change in the state vector
$w(t)$ is the random change in the state vector
$y(t)$ is the observation vector
$H(t)$ is the measurement matrix
$v(t)$ is observation noise
p is a repeating column matrix
Q_t is the covariance matrix corresponding to $w(t)$
R_t is the covariance matrix corresponding to $v(t)$

Throughout the chapter, these equations are referred to as the motorway model. The objective is to estimate the state vector $b(t)$, given knowledge of model (4.1) and the observations $y(t)$. Except for the inequality constraints (4.1c) this model is well researched, and standard solutions exist for the state estimation for the unconstrained version of this model.

Bayesian inference will be used to estimate $b(t)$. The idea behind Bayesian inference is that rather than some point estimate, a probability distribution of $b(t)$ is computed. If $Y(t)$ represents all observations available up to and including period t then this distribution is represented by:

$$\mathrm{p}[b(t)|Y(t)] \tag{4.2}$$

In reality $b(t)$ is a fixed value. The uncertainty expressed by (4.2) only exists on the observers part. The density (4.2) is therefore referred to as the *subjective probability distribution*.

This chapter starts with explaining the ideas of Bayesian inference (section 4.2), followed by the application of Bayesian theory to derive a recursive estimator for the split probabilities in the motorway model (section 4.3). A standard method of dealing with the equality constraints (4.1d) is discussed in section 4.4. In the rest of the chapter, these constraints are ignored. A separate section (section 4.5) treats the issue of converting a solution of the form (4.2) into a point estimate for the state.

4.2 Basic steps in the Bayesian approach

In the present context, the derivation of point estimates from (4.2) is referred to as *postprocessing*. Postprocessing can be performed separately from a primary task: keeping track of (4.2) for each new time period. This task can be performed by the use of a recursion consisting of measurement updates and time extrapolations (see figure 4.1). This recursion is initialized with an initial distribution. The elements of figure 4.1, which are measurement update, time extrapolation, initial distribution and postprocessing, are discussed below.

Measurement update

Let $Y(t\text{-}1)$ represent the collection of all observations available up to and including period t-1 and $y(t)$ represent the observation of period t. Hence $Y(t)$ is defined recursively by:

$$\begin{gathered} Y(0)=\varnothing \\ Y(p)=Y(p\text{-}1) \cup y(p),\ \mathrm{p}=1,2,\ldots t \end{gathered} \tag{4.3}$$

The computation of a posterior, or *filtered*, distribution $\mathrm{p}[b(t)|Y(t)]$ from a prior distribution $\mathrm{p}[b(t)|\ Y(t\text{-}1)]$ and the likelihood function $\mathrm{p}[y(t)|\ Y(t\text{-}1),b(t)]$ is referred to as the measurement update. Information contained in a new observation may be incorporated in the subjective probability distribution using *Bayes*' rule:

$$\mathrm{p}\,[b\,(t)\,|\,Y\,(t)\,] = \frac{\mathrm{p}\,[y\,(t)\,|\,b\,(t)\,,\,Y\,(t-1)\,]\,\cdot\,\mathrm{p}\,[b\,(t)\,|\,Y\,(t-1)\,]}{\mathrm{p}\,[y\,(t)\,|\,Y\,(t-1)\,]} \tag{4.4}$$

The denominator of expression (4.4) is referred to as the *normalisation constant* as it is invariant for $b(t)$. The numerator consists of the product of the Likelihood function (left) and the prior distribution (right). The likelihood function defines the relation between the observed quantity $y(t)$ and the state parameter vector $b(t)$ and follows from the observation model that is adopted. Many model specifications are a special case of the form:

$$y(t)=\psi(\ b(t),\varepsilon(t)\ ) \tag{4.5}$$

where $\varepsilon(t)$ is a random component independent of $Y(t\text{-}1)$. For models that can be written in this form the measurement update equation simplifies to:

$$\mathrm{p}[b(t)|Y(t)]=\mathrm{p}[y(t)|b(t)].\mathrm{p}[b(t)|Y(t\text{-}1)]/\mathrm{c}[Y(t)] \tag{4.6}$$

where the value of $\mathrm{c}[Y(t)]$ follows from the requirement that integral of $\mathrm{p}[b(t)|Y(t)]$ with

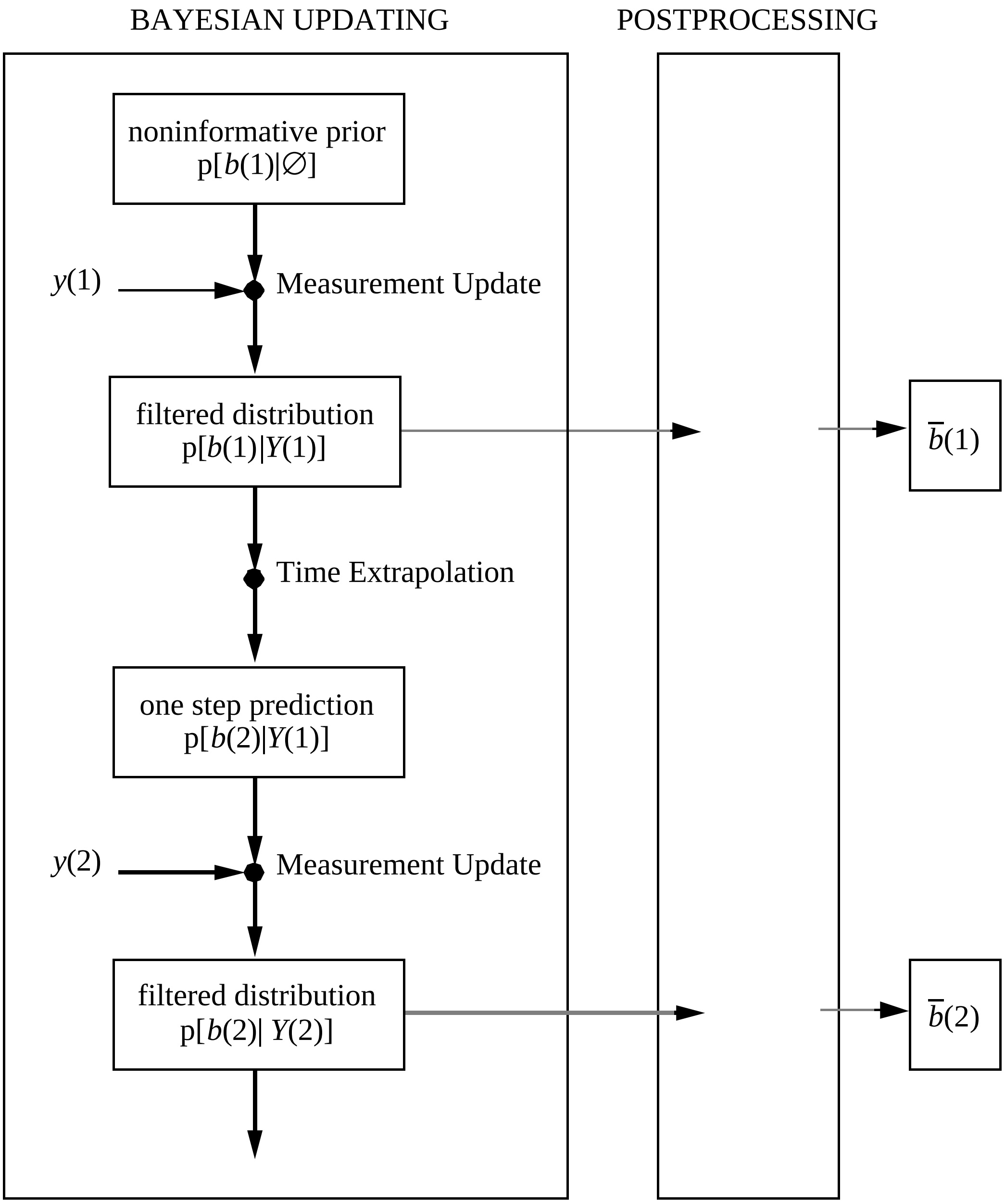


Figure 4.1: *Bayesian updating scheme. The procedure is initiated with a noninformative prior. Via a sequence of measurement updates and time extrapolations the probability distribution of the state is traced. Postprocessing is needed to make practical use of this distribution.*

respect to $b(t)$ should be unity. This implies that the storage requirements as far as the measurement update is concerned, are constant in time and equal the amount of data needed to characterize p[$b(t)|Y(t\text{-}1)$], which may be an advantage if a large amount of data is observed. For the problem at hand the simplification is justified, as (4.1b) is a special case of (4.5).

Time extrapolation

The process of deriving the, what will be called, *one step prediction* p[$b(t)|Y(t\text{-}1)$] from p[$b(t\text{-}1)|Y(t\text{-}1)$] is referred to as the *time extrapolation*. The mathematical equation that defines this update is given by:

$$\mathrm{p}[b(t)|Y(t\text{-}1)] = \int_{b(t)\in[0,1]} \mathrm{p}[b(t)|b(t\text{-}1),Y(t\text{-}1)].\mathrm{p}[b(t\text{-}1)|Y(t\text{-}1)]db(t\text{-}1) \tag{4.7}$$

In this update the relation between the previous and the current state is represented by p[$b(t)|b(t\text{-}1),Y(t\text{-}1)$]. Again in many cases the storage of $Y(t\text{-}1)$ is not necessary to determine this density. This is particularly true if the time propagation can be modelled by:

$$b(t)=\beta(\,b(t\text{-}1),\gamma(t\text{-}1)\,) \tag{4.8}$$

where $\gamma(t\text{-}1)$ is the outcome of a random process and is independent of $Y(t\text{-}1)$. In this case (4.7) simplifies to:

$$\mathrm{p}[b(t)|Y(t\text{-}1)] = \int_{b(t)\in[0,1]} \mathrm{p}[b(t)|b(t\text{-}1)].\mathrm{p}[b(t\text{-}1)|Y(t\text{-}1)]db(t\text{-}1) \tag{4.9}$$

For the problem at hand (4.9) may be used as (4.1a) is a special case of (4.8).

Initial distribution

Equations (4.6) and (4.9) constitute a recursion, that at one point should be started with an initial distribution. This is what critics point out to be a weak spot of the Bayesian approach. If (4.2) is considered as the probability distribution of $b(t)$, conditioned on the available information then the initial distribution should describe the 'ignorant' state of mind. This does however not lead to a well defined prior distribution. A general approach to this issue is the use what is called a *noninformative prior*. For the split parameters that are bound to the interval [0,1] a good candidate for such a distribution would be the uniform distribution. The discussion about the initial distribution is partly of a theoretic nature as it can be shown that for a dynamic system, such as the one shown in figure 2.3, the influence of the initial distribution on the subjective probability distribution gradually reduces in time.

Postprocessing

Equation (4.2) does not represent a *point estimate* such as, for example, the ML-estimator. However, given (4.2), point estimates can easily be derived. The two possibilities that are considered here are:

- the *subjective expectation*, defined by:

$$\overline{b}^{\mathrm{EXP}}(t) \equiv \mathrm{E}[b(t)|Y(t)] = \int_{b(t)\in[0,1]} b(t)\ \mathrm{p}\ [b(t)|\ Y(t)]\ db(t) \tag{4.10}$$

- and the *Maximum APosteriori* (MAP) estimator, defined by the argument that maximizes (4.2):

$$\overline{b}^{\mathrm{MAP}}(t) \equiv \underset{b(t)}{\mathrm{argmax}} \left(\, \mathrm{p}[b(t)|Y(t)] \, \right) \tag{4.11}$$

see e.g. *Ljung and Söderström (1983).*

Given distribution (4.2), $\overline{b}^{\mathrm{EXP}}(t)$ would be the *minimum variance estimator*, i.e. $\overline{b}^{\mathrm{EXP}}(t)$ would be equal to:

$$\underset{b}{\mathrm{argmin}} \, \mathrm{E} \left[\| \, b(t) - b \|^2 \right] \tag{4.12}$$

It should be noted however that (4.2) represents the subjective distribution of $b(t)$ which is influenced by the specification of the initial distribution. Therefore one cannot claim that $\overline{b}^{\mathrm{EXP}}(t)$ is a minimum variance estimator unless one is certain that the initial distribution is correctly specified, which only is true if the initial state is exactly known or if the initial state is the result of a well described random experiment such as the tossing of coin.

4.3 The Bayesian approach applied to the motorway model

To derive a recursion such as discussed in the previous section, that produces numerical results for the motorway model (4.1), numerical expressions for the following elements need to be provided:

- initial distribution
- measurement update equations (see equation 4.6)
- time extrapolation equations (see equation 4.9)
- postprocessing equations

Up to this point, no assumptions have been made in this chapter about the form of the distributions of the random variables that constitute the motorway model (4.1). Only results about the first and second moments of these random variables have been assumed, and are summarized in the matrices R_t and Q_t. Attempts to derive a recursion for the first two moments of the subjective distribution of $b(t)$ on this non-parametric basis are known not to be successful (*Anderson and Moore, 1979, Ljung and Söderström, 1983*). However, if the inequality constraints (4.1c) are disregarded, a recursion known as the Kalman filter can be derived with which a 'best linear minimum variance estimate' (BMVLE) can be computed (see section 3.3.4). The Kalman filter recursion relates to the mean and the covariance matrix of the subjective distribution (4.2) if, in addition to the earlier assumptions, it is assumed that $b(1)$, $w(t)$ and $v(t)$ are multivariate normal (MVN) distributed (*Anderson and Moore*, *1979*).

Although the Kalman filter has a number of advantages that were already discussed in section 3.3.4, this estimation method also exhibits a number of shortcomings when applied to the problem (4.1). All of these are related in one way or the other with the presence of the inequality constraints(4.1c).

The two main problems are:

- Which values to specify for the parameters in the initial distribution, $\overline{b}(0)$ and $\Sigma_b(0)$. It seems beneficial to specify large diagonal values of $\Sigma_b(0)$, since this expresses a lack of information about $b(0)$ and results in discarding the initial value $\overline{b}(0)$ as quickly as possible. On the other hand the initial variance is bounded above since the split parameters are bounded to the interval [0,1].
- How to perform the postprocessing task, especially if the estimate generated by the Kalman

filter does not satisfy the inequality constraints.

In this section a few modifications to the Kalman filter, aimed at overcoming these problems, are proposed that will result in a method producing estimates that deviate from those obtained with the traditional Kalman filter.

Measurement update

It is well known that if both the prior distribution and the likelihood function in (4.4) are Multivariate Normal (MVN) then the posterior distribution is also MVN. Moreover if $v(t)$ satisfies (4.1e) and is MVN then the expectation and covariance matrix characterising the posterior distribution are given by:

$$\begin{gathered} \overline{b}(t)^{+}=\overline{b}(t)^{-}+K_t[y(t)-H(t)'\overline{b}(t-1)^{+}]+u(t) \\ K_t=\Sigma_{t-1}{}^{+}H(t)[H(t)'\Sigma_{t-1}{}^{+}H(t)+R_t]^{-1} \\ \Sigma_t{}^{+}=\Sigma_t{}^{-}-\Sigma_t{}^{-}H(t)[H(t)'\Sigma_t{}^{-}H(t)+R_t]^{-1}H(t)'\Sigma_t{}^{-} \end{gathered} \tag{4.13}$$

where $\overline{b}(t)^{-}$ and $\Sigma_t{}^{-}$ are the *apriori* mean and covariance, and $\overline{b}(t)^{+}$ and $\Sigma_t{}^{+}$ are the *aposteriori* mean and covariance. This measurement update, supplemented with time extrapolation equations is known as the Kalman filter (*Kalman, 1960*).

The MVN distribution is however not a very suitable way to represent the knowledge about split probabilities, as it does not reflect the inequality constraints (4.1c). Also, usage of (4.13) could easily lead to negative estimates. This problem was already recognised in *Nihan and Davis* (*1987*), where several heuristic algorithms were proposed to constrain the estimates to the feasible region. A similar observation was made in the context of static OD-estimation by Bell (*Bell, 1991*). A remedy for this problem presented in *Van Der Zijpp and Hamerslag* (*1994*) is the usage of a *Truncated* Multivariate Normal (TMVN) distribution.

A truncated probability distribution is derived from its original by applying the truncation operator, see *Mood et al. (1963)*. The probability mass that was originally assigned to points outside the truncation interval is distributed proportionally over the points inside this interval, i.e. if $f(x)$ is the distribution corresponding to a random variable X then the truncated distribution of X to the interval T is defined by:

$$f^{\text{trunc}}(x)\equiv f(x).\mathrm{I}_T(x)/\int_{\xi\in T} f(\xi)\,d\xi \tag{4.14}$$

where $\mathrm{I}_T(x)=1$ if $x\in T$, and $\mathrm{I}_T(x)=0$ elsewhere.

The use of a truncated distribution is indicated if one beliefs that a distribution provides a reasonable model for a phenomenon inside the truncation interval while at the same time one knows that the phenomenon can never take values outside this interval. From (4.14) it can be seen that a truncated distribution is characterized by the same parameters as its original, or by a subset of these parameters. For example, the truncated MVN distribution is characterized by a vector and a matrix. Unlike the non-truncated MVN distribution these parameters do not correspond directly to the mean and variance of the truncated distribution. To indicate this fact in the notation we will mark the parameters, when used to characterise a TMVN distribution, with a symbol '*'.

If in (4.6) the prior distribution is TMVN and the likelihood function is MVN, it can be shown that the posterior distribution remains in the class of TMVN distributions. Moreover, equation (4.13) still defines the parameters that characterise the posterior distribution (*Van Der*

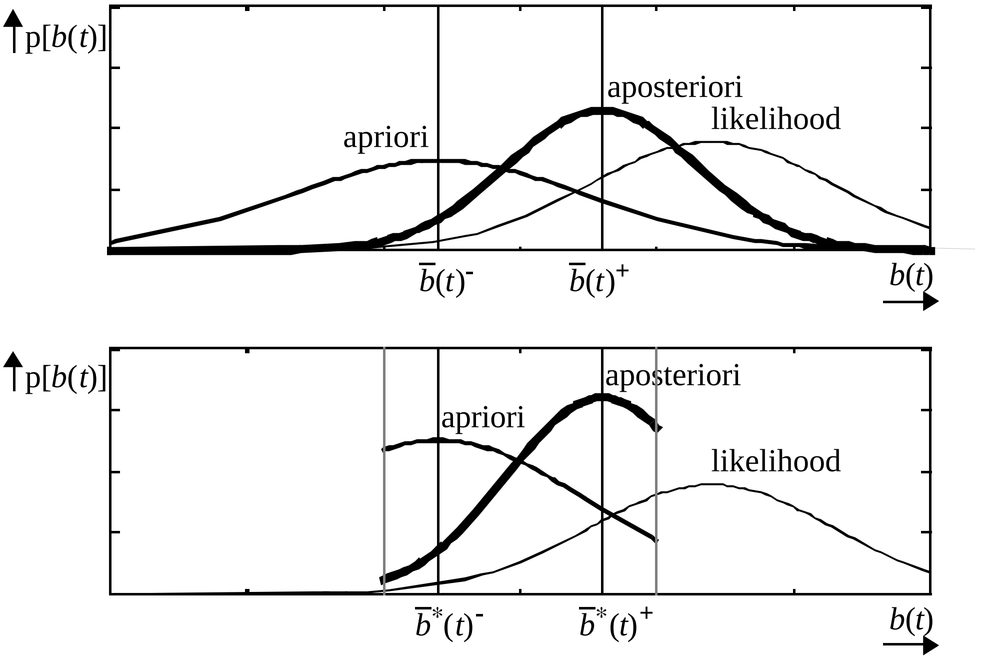


Figure 4.2: Bayesian update: aposteriori=likelihood.apriori/normalising constant. Top graph: prior distribution is MVN. Bottom graph: prior distribution is TMVN.

Zijpp and Hamerslag, 1994).

This is illustrated graphically in figure 4.2. The posterior distribution that is obtained by multiplying a normally distributed prior distribution and likelihood function and normalizing the result is a normal distribution, see top graph. If the prior distribution is replaced with a TMVN distribution then the shape of the resulting posterior distribution remains unchanged inside the truncation interval, and hence is characterised by parameters identical to those obtained when using an MVN distributed prior see bottom graph.

One remark concerns the usage of the equality constraints (4.1d). These constraints are dealt with by considering them as perfect measurements, see *Anderson and Moore* (*1979*). See section 4.4 for details of this approach in the present context.

Time extrapolation

In the above, only the measurement update has been discussed. In addition to this a time extrapolation is needed to account for changes of the split probabilities over time such as described by (4.1a). The distribution of the *one-step prediction* satisfies the general condition given by (4.10). If both factors in the integral at the right hand side of (4.10) are MVN distributions then the one-step prediction remains in the class of MVN distributions. The Kalman filter equations can be used to compute the mean and the covariance matrix of the one-step prediction. However, if the factor $p[b(t)|Y(t)]$ corresponds to a TMVN distribution then no useful analytical expression exists for the outcome.

Regardless of this fact, it was decided to use the Kalman time extrapolation equations unaltered, i.e.:

$$p[b(t+1)|Y(t)]=\mathrm{TMVN}[\,\bar{b}(t+1)^{*-}, \Sigma_{t+1}^{*-}\,]\Big|_{b(t+1)}$$

with:

$$\bar{b}(t+1)^{*-}=\bar{b}(t)^{*+}+u(t)$$
$$\Sigma_{t+1}{}^{*-}=\Sigma_t{}^{*+}+Q_t \tag{4.15}$$

Usage of this result introduces an error relative to the correct result that belongs to specification (4.1a). However, the error will only be small as jointly $u(t)$ and $w(t)$ represents only a small change in $b(t)$.

Combined measurement update/time extrapolation

For future reference the recursion that combines the measurement update and time extrapolation in one step is given below:

$$\mathrm{p}[b(t)|Y(t)]=\mathrm{TMVN}[\ \bar{b}(t)^{*-}\ ,\ \Sigma_t{}^{*-}\ ]\Big|_{b(t)}$$
$$\bar{b}(t)^{*+}=\bar{b}(t\text{-}1)^{*+}+K_t[y(t)\text{-}H(t)'\bar{b}(t\text{-}1)^{*+}]+u(t)$$
$$K_t=\Sigma_{t\text{-}1}{}^{*+}H(t)[H(t)'\Sigma_{t\text{-}1}{}^{*+}H(t)+R_t]^{-1}$$
$$\Sigma_t{}^{*+}=\Sigma_{t\text{-}1}{}^{*+}-\Sigma_{t\text{-}1}{}^{*+}H(t)[H(t)'\Sigma_{t\text{-}1}{}^{*+}H(t)+R_t]^{-1}H(t)'\Sigma_{t\text{-}1}{}^{*+}+Q_t \tag{4.16}$$

Initial distribution

As was mentioned earlier, the ideal noninformative prior for the problem of estimating split proportions is the uniform distribution. This distribution expresses that every solution is equally likely. Usage of the TMVN distribution makes it possible to define an initial distribution that is arbitrary close to the uniform distribution simply by defining $\Sigma^*{}_b(0)$ as a diagonal matrix with very large diagonal elements. Therefore the following initial distribution is chosen, with η a sufficiently large scalar:

$$\mathrm{p}[b(0)|\varnothing]\sim\mathrm{TMVN}[\ \tfrac{1}{2}\ ,\ \eta I\ ] \tag{4.17}$$

Figure 4.3 illustrates how a truncated scalar normal distribution approaches a uniform distribution if the variance increases.

Postprocessing

A practical difficulty is the computation of the mean associated with a TMVN distribution. No analytical solution exists for the multidimensional integral that needs to be solved, nor is numerical integration an option due to CPU time constraints. However the mean may be approximated by taking the average value of a large number of random numbers drawn from a TMVN distribution. TMVN random numbers may be generated simply by generating MVN random numbers and rejecting all outcomes that do not satisfy (4.1c). Details of this technique are described in section 4.5.

4.4 Equality Constraints

A second type of constraints that apply to the state parameters are the equality constraints, see equation (4.1d). In parameter optimization methods, such as constrained optimization, these constraints are used to restrict the space of possible solutions.

For the purpose of imposing these constraints to a general class of estimation procedures *Nihan and Davis (1987)* proposed a normalisation procedure. Since this procedure was meant to act separately from the active parameter estimation method, it does not take advantage of the knowledge of second moments that are available within the Bayesian procedure proposed in this chapter.

As the equality constraints apply to linear combinations of the unknown split parameters,

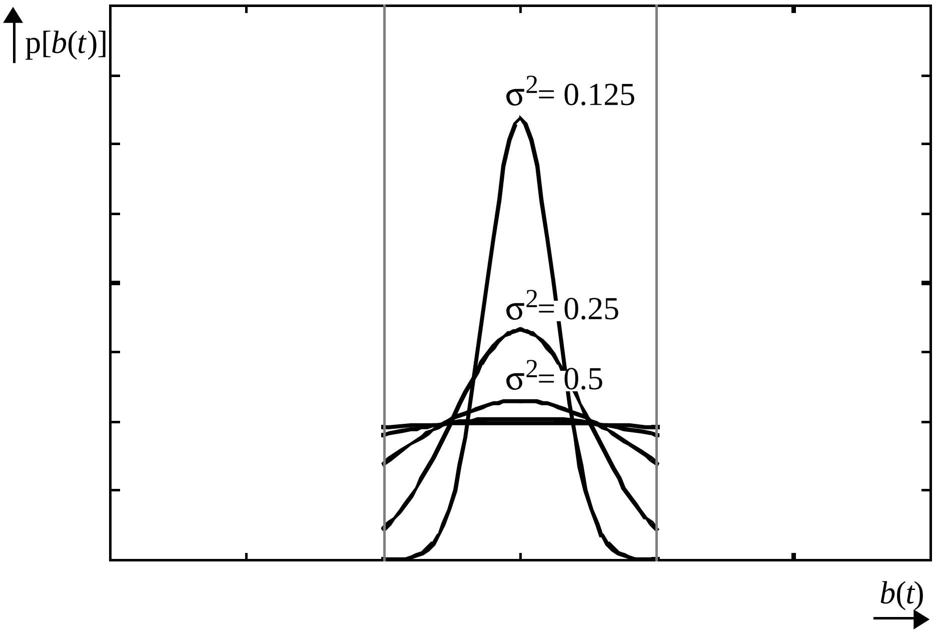


Figure 4.3: Truncated normal distributions approach the uniform distribution if variance increases

these constraints can be used as measurements in the Kalman filter. In literature, such measurements are referred to as *perfect observations* because no noise is present on these observations. In matrix notation such a measurement looks like:

$$\mathbf{1}=p'.b(t),\ \mathbf{1}=\begin{bmatrix}1\\1\\\dots\\1\end{bmatrix},\ p'=\begin{bmatrix}1, 1\dots1 & & & \\ & 1, 1\dots1 & & \\ & & \dots & \\ & & & 1, 1\dots1\end{bmatrix} \tag{4.18}$$

Anderson and Moore (1979) describe two ways to deal with this kind of observations. The first way is to reduce the order of the filter by an order *m* (*m* denotes the number of equality constraints). This can be done by a *change of coordinate basis*, similar to the one used while calculating the solution to the constrained optimization problem. The second way is to proceed as with any measurement, using a zero matrix for the measurement noise matrix. In this case the recursion (4.13) remains valid. For ease of implementation the latter method was used in this thesis. We will refer to this procedure as the *normalising measurement update*. The resulting distribution is called the *normalized* distribution.

The normalising measurement update can be combined with the regular measurement update or can be performed separately, that is, before or after the regular measurement update. In the following description we will assume that the normalising measurement update is performed after the regular measurement update.

Therefore it is assumed that the filtered distribution at period t, characterized by $\bar{b}(t)^{*+}$ and Σ_t^{*+}, is available. To indicate the difference between the normalized and non-normalized distribution, the parameters that characterize the normalized distribution are denoted with $\bar{b}(t)^{*}$ and Σ_t^{*}. Applying the standard Bayesian update parallel to equation (4.13) results in:

$$\overline{b}(t)^{*} = \overline{b}(t)^{*+} + K_t.[\mathbf{1}\text{-p}'\,\overline{b}(t)^{*+}]$$
$$K_t = \Sigma_t^{*+}\ \text{p}\ [\text{p}'\,\Sigma_t^{*+}\ \text{p}]^{-1}$$
$$\Sigma_t^{*} = \Sigma_t^{*+} - \Sigma_t^{*+}\ \text{p}\ [\text{p}'\,\Sigma_t^{*+}\ \text{p}]^{-1}\ \text{p}'\,\Sigma_t^{*+} \qquad (4.19)$$

The covariance matrix Σ_t^* defined by these update equations is singular. However $\overline{b}(t)^*$ and Σ_t^* still define the density function of $b(t)$ on the domain where $b(t)$ satisfies the equality constraints. Outside this domain the density function is zero. The probability distribution of $b(t)$ is given by:

$$\text{p}[b(t)=b]=\frac{1}{\text{C}(\mu^*,\Sigma^*)|2\pi\Sigma^*|^{mn/2}}\exp\left(-\frac{1}{2}(b-\mu^*)'\text{pinv}(\Sigma^*)(b-\mu^*)\right)\text{I}_{[0,1]}(b)\text{I}_{\text{J}_1}(b)\quad ,$$

where:

$$\text{J}_1 \equiv \{b|F'\,b=\mathbf{1}\} \qquad (4.20)$$

The operator pinv(.) denotes the *pseudo-inverse* operator. The pseudo inverse replaces the inverse operator as Σ_t^* is not invertible. See *Anderson and Moore (1979)* for details on this issue. The quantity *mn* is height of the vector *b*, and J_1 is the set of values for *x* that satisfy the equality constraints (4.1d).

The resulting sequence of initializing, time extrapolations and measurement updates is shown in figure 4.4.

4.5 Postprocessing

The recursive scheme shown in figure 4.4 defines subjective probability distributions from which point estimates should be derived. In section (4.2), two possible ways are mentioned to derive point estimates from subjective probability distributions: the subjective expectation and the maximum aposteriori estimate.

Applied to the subjective probability distribution defined by (4.16) and (4.19) these estimates are given by:

(a)
$$\overline{b}^{\text{EXP}}(t) \equiv \int_{b\,\in\,[0,1],\,F'b\,=\,\mathbf{1}} b\ \text{TMVN}[\ \overline{b}(t)^*\ ,\ \Sigma_t^*\ ]\big|_b\ db$$

and:

(b)
$$\overline{b}^{\ \text{MAP}}(t) \equiv \underset{b,\,F'b\,=\,\mathbf{1}}{\text{argmax}}\left(\ \text{TMVN}[\ \overline{b}(t)^*\ ,\ \Sigma_t^*\ ]\big|_b\ \right) \qquad (4.21)$$

Although (4.21a) defines the preferred solution, evaluation of (4.21a) might be cumbersome, as *b* is a high dimensional vector. The right hand side of (4.21b) can be shown to correspond with a constrained quadratic minimization problem which is generally more easily to solve. The possibilities to evaluate (4.21b) and (4.21a) are separately discussed in subsections 4.5.1 and 4.5.2 respectively.

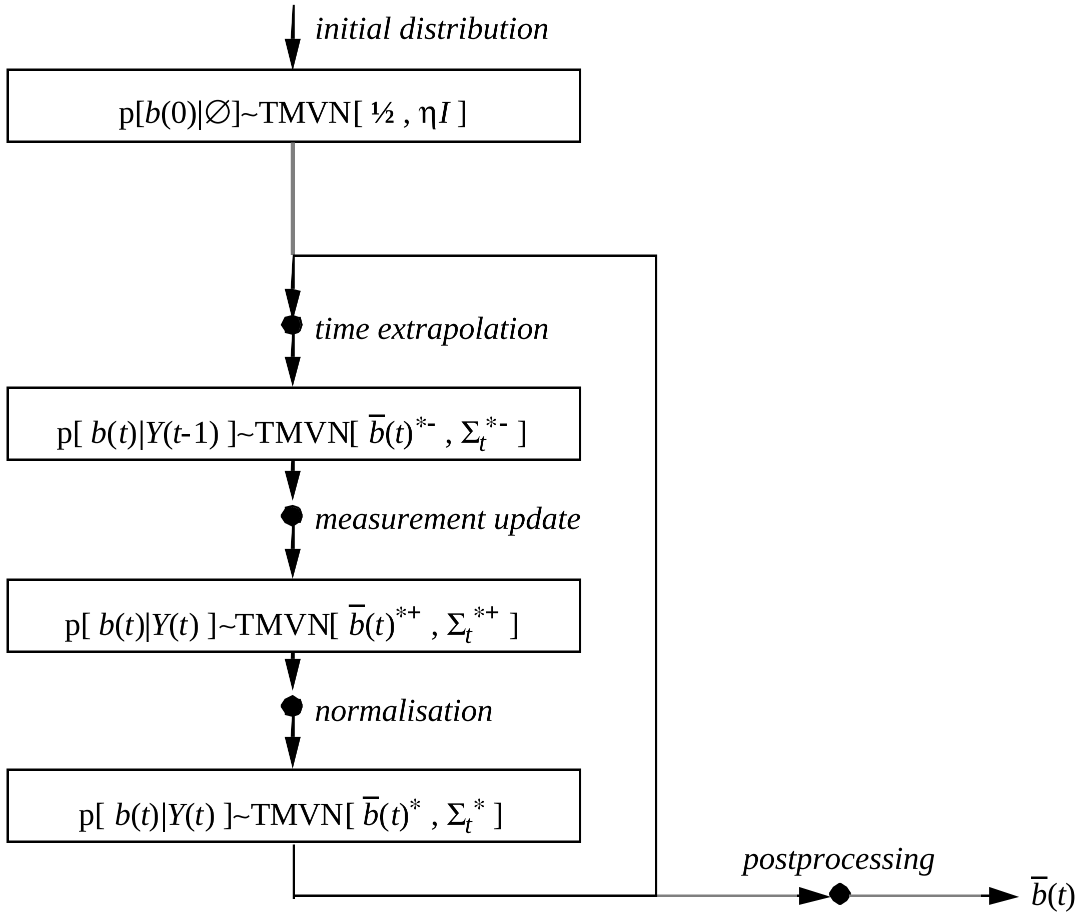

initial distribution
p[$b(0)|\varnothing$]~TMVN[½ , ηI]
time extrapolation
p[$b(t)|Y(t\text{-}1)$]~TMVN[$\bar{b}(t)^{*-}$, Σ_t^{*-}]
measurement update
p[$b(t)|Y(t)$]~TMVN[$\bar{b}(t)^{*+}$, Σ_t^{*+}]
normalisation
p[$b(t)|Y(t)$]~TMVN[$\bar{b}(t)^{*}$, Σ_t^{*}]
postprocessing
$\bar{b}(t)$

Figure 4.4: Structure of bayesian estimator for split probabilities, including initialization, time extrapolation, measurement update, normalization and postprocessing.

4.5.1 MAP estimate

Evaluating $\bar{b}^{\,\mathrm{MAP}}(t)$ in (4.21b) comes down to evaluating:

$$\bar{b}^{\,\mathrm{MAP}}(t)=\underset{b}{\mathrm{argmax}}\ (\ b\text{-}\bar{b}(t)^{*}\ )'.\mathrm{pinv}(\Sigma_t^{*}).(\ b\text{-}\bar{b}(t)^{*}\ )$$

subject to:

$$\begin{aligned}&\mathbf{0}\le b\le\mathbf{1}\\&\mathrm{p}'\,b=\mathbf{1}\end{aligned}\qquad(4.22)$$

From (4.22) it can be concluded that the MAP estimate and the solution computed with the fully constrained least squares (FCLS) procedure, see equation (3.13), are very similar. Both procedures require the solution of a constrained quadratic minimization problem.

In fact if in system (4.1) the vector u_t and matrix Q_t are specified to be zero and the matrix R_t is replaced with an identity matrix, while the discounting factor is set to one in the FCLS method, both methods should in theory produce identical results. This is confirmed by the experiments (see chapter 8). Therefore in applications to systems with a constant state, the FCLS method is a special case of (4.22).

Due to the parallels between (4.22) and the FCLS problem, the numerical procedures that are needed to compute the MAP estimator are identical to the procedures that were used for the FCLS method. This implies that two alternative solution algorithms are available, both of which have been described in appendix A:

- The computationally efficient, but not always exact method referred to as *iterative solving*, see section A.5
- The exact, but more time consuming method known as *projected conjugate gradients*, see section A.4.

In the chapters containing the results of experiments involving simulated and empirical data, both method-variants are discussed.

4.5.2 Subjective expectation

Preferably the postprocessing consists of the calculation of the subjective expectation E[$b(t)|Y(t)$], which implies that (4.21a) must be evaluated. However, this would require the evaluation of an integral of (a part of) the error function associated with the MVN distribution, for which no analytical solution exists. Numerical integration is no option either because *b(t)* is a high dimensional vector.

In this section two possibilities are discussed to obtain approximations of (4.21a) without having to evaluate the integral. The first possibility is to approximate (4.21a) with an expression that is easier to evaluate, this option is referred to as *Approximated Mean (AM)*. The second option is to sample a large number of random numbers from the distribution for which we want to determine the expected value and use the average as an approximation for the mean. This option is referred to as *Randomized Mean (RM)*. Both options are discussed below.

Approximated mean

Since no analytical or numerical ways are known to evaluate the integral defined by (4.21a), we replace (4.21a) with a simpler problem

$$\bar{b}^{\mathrm{APEXP}}(t) \equiv \int_{b \in [0,1]} b \,\mathrm{TMVN}[\,\bar{b}(t)^{*}, \mathrm{diag}(\Sigma_t^{*})\,]\big|_b \, db \qquad (4.23)$$

where the diag(.) operator replaces a matrix with another matrix only containing the diagonal elements of the original. Solving (4.23) instead of (4.21a) implies that the correlations and the equality constraints are ignored. Expression (4.23) can be evaluated element by element, i.e.:

$$\bar{b}^{\mathrm{APEXP}}{}_k(t) \equiv \int_{b_k \in [0,1]} b_k \,\mathrm{TMVN}[\,\bar{b}_k(t)^{*}, \Sigma_t{}^{*}{}_{kk}\,]\big|_{b_k} \, db_k, \quad k=1,2,\ldots mn \qquad (4.24)$$

This integral is solved in two steps. In the first step the value normalization constant associated with the TMVN distribution is determined. In the second step the mean value is determined. Lemmas B.1 and B.2 in appendix B show that for a TMVN distribution with parameters μ and σ^2 the expectation is given by:

$$\mathrm{E}[x] = \frac{\sigma}{C(\mu,\sigma)\sqrt{2\pi}}\left(\exp(-\frac{\mu^2}{2\sigma^2}) - \exp(-\frac{(-\mu+1)^2}{2\sigma^2})\right) + \mu,$$

with: $$C[\mu,\sigma] = \frac{1}{2}\mathrm{erf}(\frac{\mu}{\sigma\sqrt{2}}) + \frac{1}{2}\mathrm{erf}(\frac{1-\mu}{\sigma\sqrt{2}})$$

and: $$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)dt \tag{4.25}$$

The erf(.) operator is generally known as the error function. This integral can be evaluated in a numerical way or by making use of approximations that are known in literature. One of these approximations with an accuracy better than 2.5×10^{-5} is given by equation B.2.

The quality of (4.24) as an approximation depends on the correlations in the original distribution, and the amount of probability mass contained in the truncated tails of the distribution, see also example B.4 in appendix B.

The estimate $\bar{b}^{\mathrm{APEXP}}(t)$ does not yet satisfy the equality constraints (4.1d). In order to impose these constraints, a simple linear scaling operation was implemented.

A remarkable property of the method described above is that it produces approximations of the subjective expectation within less CPU time then is needed to evaluate MAP estimates. As the subjective expectation is preferred over the MAP, the approximated mean might very well produce better results then MAP at a lower CPU cost.

Randomized mean

A technique that can be applied to obtain the mean of any distribution is simply to sample a large set of random numbers from that distribution and using the average of this set as an approximation for the mean. In order to apply this technique, two problems need to be solved:

-a- How can we generate random numbers for a TMVN distribution with parameters μ^* and Σ^*?

-b- How many of those random numbers need to be generated in order to get a reliable estimate for the mean?

In order to address issue -a- we consider the TMVN probability distribution with parameters μ^* and Σ^* which is given by equation (4.20). In (4.20), the value of $C(\mu^*,\Sigma^*)$ follows from the requirement that p[x] integrates to one, but can not be evaluated within acceptable CPU time. The following lemma helps to generate random numbers from distribution (4.20) without having to determine $C(\mu^*,\Sigma^*)$.

Lemma 4.1: Let $X_1,X_2,X_3,\ldots$ be a series of independent and identically distributed (*iid*) random variables with distribution $f(.)$, with:

$$\int_{x\in[0,1]} f(x)dx > 0 \tag{4.26}$$

Let $x_1,x_2,x_3,\ldots$ denote the outcomes of $X_1,X_2,X_3,\ldots$, and let the random variable Y be defined by:

$$Y \equiv X_i, \text{ where i=min}\{i|x_i \in [0,1]\} \tag{4.27}$$

Then the distribution of Y is the truncated distribution of $f(.)$.

Proof we need to proof the following relationship:

$$\mathrm{p}[Y=y]=f(y)/\int_{x\in[0,1]} f(x)dx\ \mathrm{I}_{[0,1]}(y) \tag{4.28}$$

As Y is a function of a number of random variables, Y is a random variable itself. Moreover, Y cannot take on values outside the hypercube [0,1]. As definition (4.27) does not favour one value of Y over the other, for any $y\in[0,1]$ and any $y^*\in[0,1]$ with $f(y^*)>0$, the following holds:

$$\mathrm{p}[Y=y]/\mathrm{p}[Y=y^*]=f(y)/f(y^*) \tag{4.29}$$

and the probability distribution of Y should then satisfy:

$$\mathrm{p}[Y=y]=(\mathrm{p}[Y=y^*]/f(y^*)).f(y)\ \mathrm{I}_{[0,1]}(y)=c.f(y)\ \mathrm{I}_{[0,1]}(y) \tag{4.30}$$

The value of c follows from normalizing (4.30), hence equations (4.28) and (4.30) are equivalent. **End of proof**

In words, the lemma states that the truncated version of an arbitrary distribution can be obtained by defining a random variable as the first outcome of the original distribution that is an element of the truncation hypercube. Therefore it suffices to be able to generate random numbers by sampling from the original distribution.

Applied to the problem of generating TMVN random numbers, this means that these numbers can be obtained by generating a large number of MVN random numbers and rejecting all invalid samples. The latter are the numbers that are not an element of the truncation hypercube. This leaves the issue of generating random numbers for an MVN distribution with parameters μ^* and Σ^* to be solved.

The generally known approach to this is to find a matrix Q that satisfies:

$$Q.Q' =\Sigma^* \tag{4.31}$$

and to define a random variable X by:

$$X\equiv\mu^*+Q.Y \tag{4.32}$$

where Y is a vector of independent standard normal random variables. Now X is a random variable with an MVN distribution, expected value μ^* and covariance matrix $Q.Q'=\Sigma^*$. The decomposition (4.31) is known as a *Choleski* decomposition.

A precondition for a Choleski decomposition is that Σ^* is positive definite. The above approach can therefore not be applied directly since the matrix Σ^* is not invertible, let alone positive definite. The reason for the singularity of Σ^* is that corresponding to each equality constraint one element of X is defined as one minus a linear combination of the other elements of X. If these ‘dependent’ elements are dropped from the random variable X, then the distribution of the resulting r.v., say X” has an MVN distribution with expected value μ” and covariance matrix Σ”. The vector μ” and matrix Σ” can be obtained from μ^* and Σ^* by dropping the elements, rows and columns corresponding to the dependent elements. Now Σ” is a positive definite covariance matrix, and (4.32) can be applied to generate random values for the independent elements of X. The dependent elements follow directly from the requirement to satisfy the equality constraints.

In the above, a description was given, explaining how to obtain random numbers, sampled from a TMVN distribution with parameters μ* and Σ*. The second question, issue -b-, is how many of those random numbers are needed in order to get a reliable estimate of the mean. To answer this question we use the property that sample means are statistics of which

- the expected value matches the expectation from the distribution from which the samples were taken,
- the variance is given by the variance of the sampling distribution divided by the number of samples.

The sample mean is an unbiased estimator for the true mean. By taking one hundred samples it is ensured that the sample variance for the individual elements of this estimator is bound by the value 1/1200, where the value of $1 / 12$ was used as an upper limit for the variance of a truncated normal distribution and equals the variance of a uniform distribution defined on the interval [0,1]. The value 1/1200 is judged to be sufficiently low for the approximation error not to contribute significantly to the total error of estimation.

Practical notes

Concerning the practical use of randomized mean a few notes are in order:

Firstly it is recommended to use the same sequence of standard normal random variables each time the mean of a TMVN distribution is evaluated. This is to insure that the randomized mean is a continuous function of μ* and Σ*.

Secondly, for some values of μ* and Σ* the fraction of non-rejected outcomes might be very small, resulting in an unacceptably high number of trials needed to obtain a sufficient number of valid samples. If the vector of split probabilities is large or the individual elements of the vector have a low probability of being sampled in the interval [0,1] then the likelihood of a successfully computed randomized mean decreases. In such cases a recursive strategy has been applied. The problem of computing the randomized mean for the pair (μ*,Σ*) is decomposed in computing the randomized mean for the pairs (μ^*_1,Σ^*_1) and (μ^*_2,Σ^*_2), where:

$$[\mu^*_1{}' \ \mu^*_2{}'] = \mu^*{}' \tag{4.33}$$

and the approximation used for Σ* is:

$$\begin{bmatrix} \Sigma_1^* & 0 \\ 0 & \Sigma_2^* \end{bmatrix} \approx \Sigma^* \tag{4.34}$$

This implies that the correlations between the two pairs are ignored. Because of the equality constraints, it is necessary to keep the elements corresponding to identical origins together in this decomposition. This recursion continues until a solution is successfully found, or until the vector can no longer be divided. In the latter case, where the randomized mean can not even be properly computed for the splits associated with a single origin, the approximated mean is used. To recognize these cases before having wasted the CPU time on trying to generate a sufficient number of valid samples, it is advised to test on the ratio of the valid samples and the invalid samples during the process, and stop the process as soon as this ratio is below a certain threshold.

4.6 Conclusions

The practical implications of the theory put forward in this chapter for discrete time models of the form (4.1) with slowly varying, inequality constrained states, can be summarized as follows: As far as the recursive updating process is concerned, the presence of the inequality constraints (4.1c) should be ignored and the standard Kalman equations (4.16) and (4.19) should be applied. The presence of the inequality constraints should be used in a special postprocessing step.

The above statement gives a simple recipe that easily can be applied in many practical situations. Due to the postprocessing step, the actual estimates will deviate from those obtained with the traditional Kalman filter. The parameters that are recursively being updated characterize a truncated multivariate normal distribution. Therefore these parameters do not coincide with the true mean and maximum aposteriori estimates, as is the case in the traditional Kalman filter.

The maximum aposteriori estimate can either be evaluated exactly, using the earlier described Projected Conjugated Gradient (PCG) method, or can be approximated via the method of Iterative Solving (IS), see appendix A. The subjective expectation can not be evaluated exactly, but can be approximated in a CPU efficient manner with a method known as *approximated mean*, or can be approximated even more accurate, but at the cost of considerable more CPU time with a method known as *randomized mean*.

5. Spatial Correlations Between Link Volume Observations

5.1 Introduction

This chapter is concerned with the statistical analysis of the random variable defined by:

$$v(t)=y(t)-H'(t)b(t) \tag{5.1}$$

where $y(t)$ is the vector of traffic counts observed on internal or exit links, $H(t)$ is the measurement matrix defined in (2.22), and $b(t)$ is the vector of split probabilities.

The random variable $v(t)$ was introduced as the measurement error in the measurement equation (2.23). Given $q(t)$ and $b(t)$, its properties follow from the assumptions in the motorway model. Section 5.2 contains a discussion on the terms contributing to $v(t)$. We will assume that the vector $q(t)$ is observed. The vector $y(t)$ might be observed also, but should not be used to derive properties of $v(t)$. The vector of the split probabilities $b(t)$ is not known. Nevertheless, we will need this vector to determine the properties of $v(t)$. Therefore, we will subsequently assume that:

1. only time aggregated information is available (section 5.3),
2. $b(t)$ equals its most up to date estimate (section 5.4),
   or:
3. $b(t)$ is sampled from a distribution of which the mean equals the most up to date estimate, and the covariance matrix equals the covariance matrix associated with the subjective probability distribution p[$b(t)|y(1)\ldots y(t)$] (section 5.5).

Ways to derive a subjective probability distribution for $b(t)$ were discussed in chapter 4. In subsection 2.4.2 it was already argued that it will not be possible to derive any useful analytical expression for the probability density of $v(t)$. Instead we confine ourselves to deriving expressions for its expectation and covariance matrix. The covariance matrix, or its approximation will be denoted with R_t. The expectation of $v(t)$ will be zero for both case 1, 2 and 3, as according to (2.21) it holds for any $b(t)$ that:

$$\mathrm{E}[y(t)|b(t),q(t)]=H'(t)b(t) \tag{5.2}$$

Mathematical preliminaries to the derivations in sections 5.4, and 5.5 can be found in appendix C.

5.2 Interpreting the observations as linear measurements

Equation (2.23) was used in chapters 3 and 4 as a point of departure to estimate the split probabilities $b(t)$ from the observations $y(t)$. This equation suggests that a vector of linear combinations of split parameters is observed, see figure 5.1. In reality however, the observations consist of the outcomes of random processes driven by the split probabilities, see figure 5.2. This clearly is a change of interpretation, and the question is if for the sake of the estimation of $b(t)$, the model in figure 5.2 can be replaced with the model from figure 5.1. The answer is affirmative. Both systems show equal behaviour when supplied with equal split proportions. The reverse also applies; if the split proportions are estimated from the observations under assumption of the linear model (2.23) then these estimates also apply to the real system. However, a requirement is that a realistic probability distribution is specified for the random variable $v(t)$.

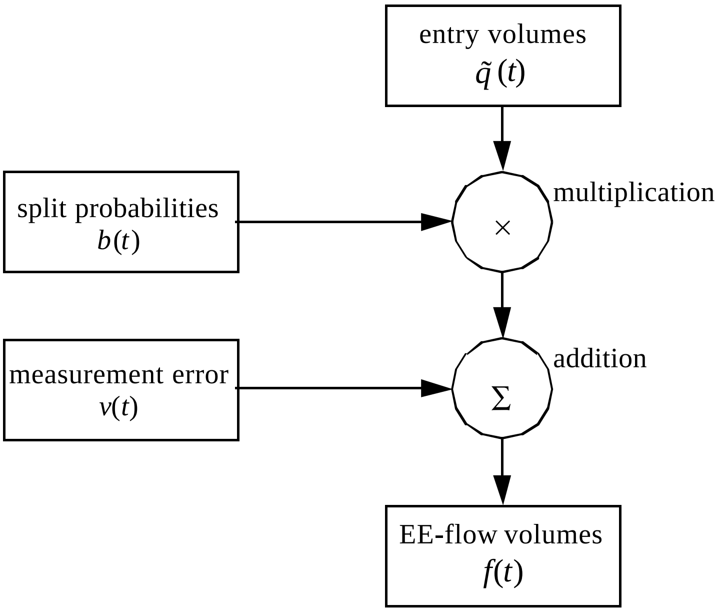


Figure 5.1: Destination choice modelled by deterministic selection, followed by addition of a measurement error.

The use of the linear model (2.23) has the advantage that standard estimation techniques are available to estimate the model parameters. One of these techniques is the Kalman filter, see section 3.3.4. A precondition to apply this technique is knowledge of the first two moments of the measurement error $v(t)$, i.e. its expectation and its covariance matrix. This measurement error can be decomposed as follows:

$$v(t) = (y(t) - \tilde{y}(t)) + (\tilde{y}(t)\text{-}\tilde{H}'(t)b(t)) + (\tilde{H}'(t)\text{-}H'(t))b(t) \qquad (5.3)$$

Interpreting the above equation reveals that the random variable $v(t)$ accounts for three independent sources of error:

- *The counting error.* This is the difference:

$$v_1(t) = y(t) - \tilde{y}(t) \qquad (5.4)$$

In the motorway model this error was denoted as $s(t)$, see figure 2.3.

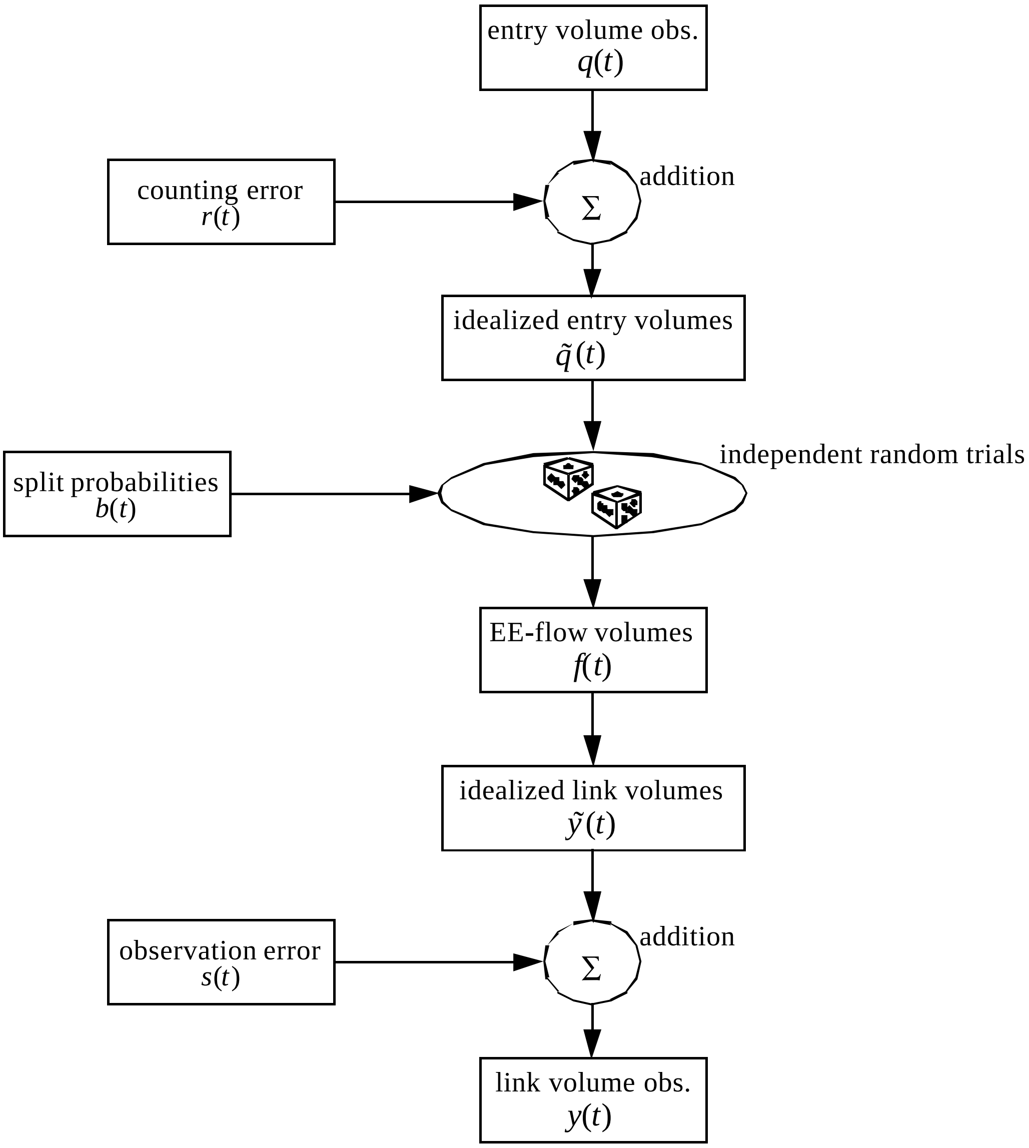


Figure 5.2: Destination choice modelled by random selection

- *The random choice error*. This is the difference between the expected and the idealized link volumes, i.e.:

$$v_2(t)=\tilde{y}(t)-\tilde{H}'(t)b(t) \tag{5.5}$$

Such a discrepancy exists since the actual shares of chosen destinations within one period are not exactly equal to the split probabilities. This component contributes slightly to the (negative) spatial correlation of the elements of $v(t)$. If by a process of random selection a collection of objects is distributed over a number of disjunct classes then the random variables defined by the number of each class are jointly distributed with a multinomial distribution, see *Lehmann (1983).*

- *The mis-specification error*. This error is due to the fact that the matrix $H'(t)$ contains observed rather than idealized entry values, and equals:

$$v_3(t)=(\tilde{H}'(t)-H'(t))b(t). \tag{5.6}$$

This error causes a strong (positive) spatial correlation since errors in entry volume observations are attributed to errors on link observation level. An overestimation of an entry volume is attributed to all involved link level errors.

The covariance matrix R_t of $v(t)=v_1(t)+v_2(t)+v_3(t)$ clearly depends on $b(t)$. As this vector is not exactly known, the matrix R_t is approximated. The advantage of a good approximation is twofold. Firstly, it is hoped that a better approximation of R_t will result in a better point estimate of $b(t)$. Secondly, with a better approximation for R_t, combined with the use of certain statistical techniques, it will be possible to quantify the reliability of the estimate more accurately.

5.3 The constant covariance matrix

In this section the time-independent approximation of the covariance of $v(t)$ is considered. A straight forward way to obtain a covariance matrix for the measurement error, is simply to observe one. If for a number of periods the split proportions denoted with $\{b^{ob}(1), b^{ob}(2),\ldots b^{ob}(t)\}$ are observed then an approximation of R_t would be:

$$R_t \cong \sum_{p=1}^{t} \frac{1}{t}\,[y(p)-H'(p)b^{ob}(p)][y(p)-H'(p)b^{ob}(p)]' \tag{5.7}$$

This can be done if the true matrix is observed from toll tickets, see e.g. *Ashok and Ben-Akiva (1993).*

Most publications are not clear about the covariance matrix that is assumed for the measurement error. For example *Cremer and Keller (1987)* only mention that a constant matrix was assumed. *Nihan and Davis (1987)* are more explicit, and put into words the method that is presumably used by most researchers. They report that a diagonal matrix is used which is defined by the identity matrix multiplied by a factor. This factor is treated as a design parameter and is fixed by some experimenting.

In this thesis two ways of defining a constant covariance are tested, which are:

1. use the identity matrix as a covariance matrix, i.e.

$$R_t = I \tag{5.8}$$

This choice of matrix is referred to as ‘Unity’ (U). Testing the method that was proposed in chapter 4 in combination with this matrix, and using the ‘maximum aposteriori’ (MAP) postprocessing option results in a method with similar properties as the FCLS method (3.13). See section 4.5.1 for a further discussion on this subject.

2. use a diagonal matrix of which the elements are chosen proportional to the average observed values, i.e.:

$$R_t = \mathrm{diag}(\bar{y}) \tag{5.9}$$

where $\bar{y}$ is the average of $\{y(1)\ldots y(t)\}$, and diag(.) is the diag operator defined in appendix C, equation (C.9). The choice of this matrix is referred to as ‘Average Link Flow’ (ALF).

5.4 The point estimate based covariance matrix

A time dependent approximation of the covariance of $v(t)$ is obtained by computing:

$$\begin{aligned} R_t &= \mathrm{E}[\, v(t)\, v'(t)\ |b(t)=\bar{b}(t)] \\ &= \mathrm{E}[\, [y(t)-H'(t)b(t)][y(t)-H'(t)b(t)]'\ |b(t)=\bar{b}(t)] \end{aligned} \tag{5.10}$$

In this approximation $b(t)$ is replaced with a fixed value for which its most up to date estimate $\bar{b}(t)$ is used. The more accurate the approximation is, the more reliable is the resulting covariance matrix. In the rest of this section the notation is compressed by dropping the time index. All variables refer to time period *t*.

The r.h.s. of (5.10) can be evaluated using of lemma C.1. According to this lemma the following holds:

$$\mathrm{E}[\, v\, v'\ |b=\bar{b}] = \mathrm{E}_{\tilde{q}}\ [\ \mathrm{E}[\ v\, v'\ |b=\bar{b}, q=\tilde{q}]\ ] \tag{5.11}$$

The covariance matrix defined in (5.10) can then be derived in the following two steps:

1. compute $\mathrm{E}[\ v\, v'\ |b=\bar{b}, q=\tilde{q}]$
2. compute $\mathrm{E}_{\tilde{q}}\ [\ \mathrm{E}[\ v\, v'\ |b=\bar{b}, q=\tilde{q}]\ ]$

In other words: first the idealized entry volumes, $\tilde{q}$, are assumed known (step 1), and subsequently the expectation with respect to the value $\tilde{q}$ is taken (step 2).

Step 1

If $\tilde{q}$ is known then the following result can be derived.

$$\begin{aligned} &\mathrm{E}[\, (y-H'b)\, (y-H'b)'\ |b=\bar{b}, q=\tilde{q}\ ] \\ &= \mathrm{E}[\, y\, y'\ |\bar{b},\tilde{q}\ ] - \mathrm{E}[\, y|\bar{b},\tilde{q}](H'\bar{b})' - (H'\bar{b})\mathrm{E}[\, y'\ |\bar{b},\tilde{q}\ ] + (H'\bar{b})(H'\bar{b})' \\ &= \mathrm{cov}[\, y,y\ |\bar{b},\tilde{q}\ ] + \mathrm{E}[\, y|\bar{b},\tilde{q}]\mathrm{E}[\, y'\ |\bar{b},\tilde{q}\ ] - \mathrm{E}[\, y|\bar{b},\tilde{q}](H'\bar{b})' - (H'\bar{b})\mathrm{E}[\, y'\ |\bar{b},\tilde{q}\ ] + (H'\bar{b})(H'\bar{b})' \\ &= \mathrm{cov}[\, y,y\ |\bar{b},\tilde{q}\ ] + (\ \tilde{H}'\bar{b}-H'\bar{b}\ )(\ \tilde{H}'\bar{b}-H'\bar{b}\ )' \end{aligned} \tag{5.12}$$

where $\tilde{H}$ denotes a matrix that equals *H*, but with all elements of *q* replaced with their corresponding elements in $\tilde{q}$, i.e.:

$$\begin{gathered} \tilde{H}_{x(i,j),k}(t) = \tau_{ijk}\, \tilde{q}_i(t) \\ i=1,\ldots m,\ \ j=1,\ldots n,\ \ k=1,\ldots l \end{gathered}$$

Consider the first r.h.s. term of (5.12). The computation of $\mathrm{cov}[\, y,y\ |\bar{b},\tilde{q}\ ]$ again takes a number of substeps:

1.1 Compute cov[$f,f\,|\bar{b},\tilde{q}$]. This matrix follows from the assumption of the multinomial distribution of the flows, given in (2.12).

1.2 Compute cov[$\tilde{y},\tilde{y}|\bar{b},\tilde{q}$]. In this step the fact that all elements of $\tilde{y}$ are linear combinations of f, i.e. $\tilde{y}=U'f$, can be used.

1.3 Compute cov[$y,y|\bar{b},\tilde{q}$]. This step is straightforward since y equals $\tilde{y}$ increased with an independent random component.

Substep 1.1

Equation (2.12) defines a multinomial distribution that can be considered as the multivariate extension of the binomial distribution. On a scalar level the following first and second moments follow from (2.12):

$$\begin{gathered}
\mathrm{E}[f_{ij}|\bar{b},\tilde{q}] = \tilde{q}_i\,\bar{b}_{ij} \\
\mathrm{E}[(f_{ij}\text{-}\mathrm{E}(f_{ij}))^2|\bar{b},\tilde{q}] = \tilde{q}_i\,\bar{b}_{ij}(1\text{-}\bar{b}_{ij}) \\
\mathrm{E}[(f_{ij}\text{-}\mathrm{E}(f_{ij}))(f_{ik}\text{-}\mathrm{E}(f_{ik}))|\bar{b},\tilde{q}] = -\,\tilde{q}_i\,\bar{b}_{ij}\,\bar{b}_{ik}\,,\ \mathrm{j}\neq\mathrm{k} \\
\mathrm{E}[(f_{ij}\text{-}\mathrm{E}(f_{ij}))(f_{hk}\text{-}\mathrm{E}(f_{hk}))|\bar{b},\tilde{q}] = 0,\ \mathrm{i}\neq\mathrm{h}
\end{gathered} \tag{5.13}$$

Using the kronecker operator δ, defined by {δ:δ_{jk}=1 if j=k and zero otherwise}, equation (5.13) can be written as:

$$\begin{gathered}
\mathrm{E}[(f_{ij}\text{-}\mathrm{E}(f_{ij}))(f_{hk}\text{-}\mathrm{E}(f_{hk}))|\bar{b},\tilde{q}] = \delta_{ih}(\ \delta_{jk}\,\tilde{q}_i\,\bar{b}_{ij}\text{-}\ \tilde{q}_i\,\bar{b}_{ij}\,\bar{b}_{ik}\,) \\
i=1,2,\ldots m,\ j=1,2,\ldots n,\ h=1,2,\ldots m,\ k=1,2,\ldots n
\end{gathered} \tag{5.14}$$

Equation (5.14) defines a block diagonal matrix. This is in accordance with the fact that no interdependencies exist between flows that originate from different entrances. The structure of the matrix is illustrated in figure 5.3. The blocks B_i in this figure only depend on the split proportions associated with entry i.

$$\begin{bmatrix}
\tilde{q}_1 B_1 & & & \\
 & \tilde{q}_2 B_2 & & \\
 & & \cdots & \\
 & & & \tilde{q}_m B_m
\end{bmatrix}$$

Figure 5.3: Structure of flow covariance matrix

For a more efficient notation it is required that (5.14) is summarized in matrix notation. Although (5.14) is linear in $\tilde{q}$, it is not possible to write (5.14) as a simple matrix multiplication. However with the aid of some notational conventions involving the use of the diag(.) operator and a repeating column matrix, p (see sections C.2 and C.3 in appendix C), we may replace (5.14) with:

$$\mathrm{cov}[f,f|\bar{b},\tilde{q}]=\mathrm{diag}(\mathrm{p}\tilde{q})(\mathrm{diag}(\bar{b})\text{-}\mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})) \tag{5.15}$$

Substep 1.2

Definition (1.6) of the idealized link flows can be expressed in matrix notation with:

$$\tilde{y}=U'f \tag{5.16}$$

where U is a matrix of height mn and with l, of which the nonzero elements are given by:

$$U_{x(i,j),k}=\tau_{ijk} \tag{5.17}$$

for i=1,…m, j=1,…n, k=1,…l, and $x(i,j)$ representing the position of f_{ij} in the vector f. From a property of the covariance operator given in lemma C.4, it follows that:

$$\mathrm{cov}[\tilde{y},\tilde{y}|\bar{b},\tilde{q}]=U'\,\mathrm{cov}[f,f|\bar{b},\tilde{q}]U \tag{5.18}$$

The moments of $\{\tilde{y}|\bar{b},\tilde{q}\}$ are obtained by combining (5.18) and (5.15):

$$\mathrm{cov}[\tilde{y},\tilde{y}|\bar{b},\tilde{q}]=U'\mathrm{diag}(\mathrm{p}\tilde{q})\ (\mathrm{diag}(\bar{b})\text{-}\mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})')\ U \tag{5.19}$$

This concludes Step 1.

<u>Substep 1.3</u>

If also the physical measurement errors, of which the properties are given by (2.11), are taken into account, the following moments can be derived for $\{y|\bar{b},\tilde{q}\}$:

$$\mathrm{cov}[y,y|\bar{b},\tilde{q}]=U'\,\mathrm{diag}(\mathrm{p}\tilde{q})\ (\mathrm{diag}(\bar{b})\text{-}\mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})')\ U+\Theta \tag{5.20}$$

-end of step substep 1.3-

Substituting result (5.20) in (5.12) gives:

$$\begin{gathered}\mathrm{E}[\ v\,v'\ |b=\bar{b},q=\tilde{q}\ ]=\\ U'\,\mathrm{diag}(\mathrm{p}\tilde{q})\ (\mathrm{diag}(\bar{b})\text{-}\mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})')\ U+\Theta\\ +(\ \tilde{H}'\bar{b}\text{-}H'\bar{b}\ )(\ \tilde{H}'\bar{b}\text{-}H'\bar{b}\ )'\end{gathered} \tag{5.21}$$

<u>Step 2</u>

The second step is to eliminate the idealized entry volume $\tilde{q}$ from equation (5.21) by evaluating (5.11). Information about $\tilde{q}$ is obtained via entry volume observations q. As a result of equations (2.9) and (2.11) in the motorway model we may use the following expected value and covariance for $\tilde{q}$:

$$\begin{gathered}\mathrm{E}[\tilde{q}|q]=q\\ \mathrm{E}[(\tilde{q}\text{-}q)'(\tilde{q}\text{-}q)]=\Phi\end{gathered} \tag{5.22}$$

When the expectation of (5.21) with respect to $\tilde{q}$ is evaluated, it is needed to expand all terms that depend on $\tilde{q}$. The dependency of $\tilde{H}$ on $\tilde{q}$ can be captured in the following matrix notation:

$$\tilde{H}'=U'\,\mathrm{diag}(\mathrm{p}\tilde{q})$$

likewise:

$$H'=U'\,\mathrm{diag}(\mathrm{p}q) \tag{5.23}$$

Therefore the following rewrite of (5.21) is permitted:

$$\begin{gathered}\mathrm{E}[\ v\,v'\ |b=\bar{b},q=\tilde{q}\ ]=U'\mathrm{diag}(\mathrm{p}\tilde{q})\ (\mathrm{diag}(\bar{b})\text{-}\mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})')\ U+\Theta\\ +U'\,\mathrm{diag}(\ \mathrm{p}(\tilde{q}\text{-}q)\ )\bar{b}\,\bar{b}'\mathrm{diag}(\ \mathrm{p}(\tilde{q}\text{-}q)\ )U\end{gathered} \tag{5.24}$$

Since above equation is a polynomial in $\tilde{q}$, with no terms of a degree exceeding 2, it should be possible to express (5.10) in terms of E[$\tilde{q}$] and cov[$\tilde{q}$,$\tilde{q}$]. We have however some difficulty doing this since $\tilde{q}$ appears as an operand of the diag(.) operator. Therefore a result that follows from lemma C.5 is substituted in (5.24):

$$\mathrm{diag}(\mathrm{p}\tilde{q})\bar{b}=\mathrm{diag}(\bar{b})\mathrm{p}\tilde{q} \tag{5.25}$$

This results in the following expression for the matrix R_t as defined in (5.10):

$$R_t = \mathrm{E}[\ v(t)\ v(t)'\ |b(t)=\bar{b}]= U'[\ \boldsymbol{Q_t}(\boldsymbol{B_t}\text{-}\boldsymbol{B_t}\mathrm{pp}'\boldsymbol{B_t}')+\boldsymbol{B_t}\mathrm{p}\Phi\mathrm{p}'\boldsymbol{B_t}\ ]U + \Theta \tag{5.26}$$

where $\boldsymbol{Q_t}$=diag(p$q(t)$) and $\boldsymbol{B_t}$=diag($\bar{b}(t)$). The expression [$\boldsymbol{Q_t}(\boldsymbol{B_t}\text{-}\boldsymbol{B_t}\mathrm{pp}'\boldsymbol{B_t}')+\boldsymbol{B_t}\mathrm{p}\Phi\mathrm{p}'\boldsymbol{B_t}$] can be shown be equal to cov[$f(t)$,$f(t)|b(t)=\bar{b}(t)$].

<u>*Interpretation of result* </u>*(5.26)*

Equation (5.26) is a generalisation of the approximation of the covariance matrix that was proposed in *Van Der Zijpp and Hamerslag (1994b)*. In fact it simplifies to the latter result if Φ is assumed to be a diagonal matrix with the values σ_q^2 on the diagonal, and the matrix Θ is assumed to be zero. In this case (5.26) may be written as:

$$\begin{gathered}\mathrm{cov}[v,v|b]=U'\,\mathrm{cov}[f,f|b]U \\ \mathrm{cov}[f_{ij},f_{hk}|b] = \delta_{ih}[\ \delta_{jk}\, q_i\, \bar{b}_{ij} +(\sigma_q^2\text{-}\ q_i\ )\bar{b}_{ij}\, \bar{b}_{ik}\ ] \\ i=1,2,\ldots m,\ j=1,2,\ldots n,\ h=1,2,\ldots m,\ k=1,2,\ldots n\end{gathered} \tag{5.27}$$

The practical value of approximation (5.26) depends on the accuracy of the estimate $\bar{b}$. A potential pitfall is that a circularity exist, consisting of split parameter estimates influencing covariance matrices which in turn influence split parameter estimates for the next period (see chapter 4). In such a way a self amplifying process might lead to diverging estimates. Ideally, some sort of stability analysis should clarify this issue. This analysis is however expected to be highly complex, and has not been pursued. Instead (5.26) has been compared with other choices of covariance matrices in a number of experiments with simulated data, see chapter 8.

In the experiments the covariance matrix that follows from (5.26) will be referred to as '*Point Estimate Based Approximation*' (PEBA) of the covariance matrix. In the experiments we will not only test the usage of the PEBA matrix, but also test the usage of a '*Diagonalized Point Estimate Based Approximation*' (DPEBA). The diagonal elements of the DPEBA matrix are equal to those of PEBA matrix but the non-diagonal elements are zero. The idea behind the usage of the DPEBA matrix is that assumptions of the motorway model are only used to gain insight into the size of the measurement errors, not into their spatial correlations.

5.5 The distribution based covariance matrix

The result derived in the previous section is based on a fixed value of $b(t)$ for which the most recent estimate was substituted. This result can be refined by using a probability distribution of $b(t)$. Instead of evaluating (5.10) the objective now is to evaluate:

$$R_t = \mathrm{E}[\ v(t)\ v'(t)\ ] \tag{5.28}$$

As a point of departure for the evaluation of (5.10), the expected value and covariance matrix of $b(t)$ are used, denoted by $\bar{b}(t)$ and $\Sigma_b(t)$ respectively. It is assumed that $\bar{b}(t)$ and $\Sigma_b(t)$ satisfy the constraints that follow from the fact that $b(t)$ is bounded to the hypercube [0,1]. It should be noted that if these constraints are not adhered to, this might lead to unrealistic results, like a matrix of R_t that is not positive definite. The matrix R_t as defined in (5.10) will be referred to as the '*Distribution Based Approximation*' (DBA).

Applying lemma C.1 again implies:

$$\mathrm{E}[\ v\, v'\ ]=\mathrm{E}_b[\ \ \mathrm{E}[\ v\, v'\ |b\ ]\ \ ] \tag{5.29}$$

To the steps 1 and 2 presented in the previous section, a third step is now added:

3. compute $E_b[\ E[\ v\,v' \,|b\ ]\ ]$

A description of this step is given in this section.

Step 3

Substituting the earlier derived result (5.26) in (5.29) implies:

$$E[\ v\,v'\ ]=E_b[\ U'\,\text{diag}(\text{p}q)(\text{diag}(b)\text{-diag}(b)\text{pp}'\text{diag}(b))\ U+\Theta+\ U'\ \text{diag}(b)\text{p}\Phi\ \text{p}'\text{diag}(b)U\ ] \qquad (5.30)$$

The E_b operator, applied to the terms that are linear in b, can be evaluated directly. This gives (maintaining the order of the terms in (5.30)):

$$\begin{gathered} E[\ v\,v'\ ]=U'\text{diag}(\text{p}q)\text{diag}(\bar{b})U+\Theta \\ -U'E_b[\ \ \text{diag}(b)\text{diag}(\text{p}q)\text{pp}'\text{diag}(b)\ \ ]U \\ +U'E_b[\ \ \text{diag}(b)\text{p}\Phi\ \text{p}'\text{diag}(b)\ \ ]U \end{gathered} \qquad (5.31)$$

In the second and third line of expression (5.31) the expected value is taken of a matrix of which the elements are polynomials in b of degree 2. Therefore it should be possible to express (5.31) in terms of $\bar{b}$ and Σ_b. Just like in section 5.4 the vector, with respect to which a covariance matrix is to be computed, appears as an argument in the diag(.) operator. This time lemma (C.7) solves the problem. According to this lemma it holds that:

$$\text{diag}(b)\text{diag}(\text{p}q)\text{pp}'\text{diag}(b)=(bb')\otimes(\text{diag}(\text{p}q)\text{pp}')$$

and:

$$\text{diag}(b)\text{p}\Phi\ \text{p}'\text{diag}(b)=(bb')\otimes(\text{p}\Phi\ \text{p}') \qquad (5.32)$$

where '⊗' represents the *array multiplication* (see appendix C for the definition). Substituting these inequalities in (5.31) leads to:

$$\begin{gathered} E[\ v\,v'\ ]=U'\text{diag}(\text{p}q)\text{diag}(\bar{b})U+\Theta \\ -U'\ E_b[\ \ (bb')\otimes(\text{diag}(\text{p}q)\text{pp}')\ \ ]U \\ +U'\ E_b[\ \ (bb')\otimes(\text{p}\Phi\text{p}')\ \ ]U \end{gathered} \qquad (5.33)$$

The final step is to apply lemma C.3, replacing $E_b[bb']$ with $\Sigma_b+\bar{b}\,\bar{b}'$. This leads to:

$$\begin{gathered} E[\ v\,v'\ ]=U'\text{diag}(\text{p}q)\text{diag}(\bar{b})U+\Theta \\ -U'\ ((\Sigma_b+\bar{b}\,\bar{b}')\otimes(\text{diag}(\text{p}q)\text{pp}'))U \\ +U'((\Sigma_b+\bar{b}\,\bar{b}')\otimes(\text{p}\Phi\text{p}'))U \end{gathered} \qquad (5.34)$$

To make this result more comparable with earlier results, the transformation (5.32) is reversed, by applying lemma (C.7) a second time. This leads to:

$$(\bar{b}\,\bar{b}')\otimes(\text{diag}(\text{p}q)\text{pp}'))=\text{diag}(\bar{b})\text{diag}(\text{p}q)\text{pp}'\text{diag}(\bar{b})$$

and:

$$(\bar{b}\,\bar{b}')\otimes(\text{p}\Phi\text{p}')=\text{diag}(\bar{b})(\text{p}\Phi\text{p}')\text{diag}(\bar{b}) \qquad (5.35)$$

Substitution of this result in (5.31), results after some manipulation in the following approximation of R_t:

$$\begin{aligned} R_t \cong \mathrm{E}[\, v\, v' \,] &= U'\mathrm{diag}(\mathrm{p}q)(\ \mathrm{diag}(\bar{b}) - \mathrm{diag}(\bar{b})\mathrm{pp}'\mathrm{diag}(\bar{b})\ )U + \Theta \\ &+ U'\mathrm{diag}(\bar{b})(\mathrm{p}\Phi\mathrm{p}')\mathrm{diag}(\bar{b})U \\ &- U'\mathrm{diag}(\mathrm{p}q)\,(\Sigma_b \otimes (\mathrm{pp}'))U + U'\,(\Sigma_b \otimes (\mathrm{p}\Phi\mathrm{p}'))U \end{aligned} \tag{5.36}$$

The following equation expresses this result in a format comparable with (5.26):

$$R_t = \mathrm{E}[\, v(t)\, v(t)'] = U'[\ \boldsymbol{Q_t}(\boldsymbol{B_t}\text{-}\boldsymbol{B_t}\mathrm{pp}'\boldsymbol{B_t}') + \boldsymbol{B_t}\mathrm{p}\Phi\mathrm{p}'\boldsymbol{B_t}\ \text{-}\boldsymbol{Q_t}\,(\Sigma_b \otimes (\mathrm{pp}')) + \Sigma_b \otimes (\mathrm{p}\Phi\mathrm{p}')\ ]U + \Theta \tag{5.37}$$

where $\boldsymbol{Q_t}$=diag(p$q(t)$) and $\boldsymbol{B_t}$=diag($\bar{b}(t)$). With this result the required approximation of R_t is completely defined in terms of $\bar{b}$ and Σ_b.

<u>Interpretation of result</u> (5.36)

In (5.36), the terms not involving Σ_b are identical to (5.26). Therefore if the distribution of b contracts to a single point resulting in the covariance matrix Σ_b becoming zero, both results will be identical. Again the results can be presented in a more legible format if Φ is assumed to be a diagonal matrix with the values $\sigma_q{}^2$ on the diagonal. In such case it holds that:

$$\begin{aligned} &\mathrm{cov}[v,v] = U'\mathrm{P}\ U + \Theta, \\ &\mathrm{P}_{x(i,j),x(h,\mathrm{k})} = \delta_{ih}(\delta_{jk}\, q_i\, \bar{b}_{ij} - q_i\, \bar{b}_{ij}\, \bar{b}_{ik} + \sigma_q{}^2 \bar{b}_{ij}\, \bar{b}_{ik}) + \mathrm{cov}[b_{ij}, b_{hk}](-q_i\delta_{ih} + \sigma_q{}^2\delta_{ih}) \\ &\quad = \delta_{ih}[\delta_{jk}\, q_i\, \bar{b}_{ij} + (\sigma_q{}^2 - q_i\,)(\,\bar{b}_{ij}\, \bar{b}_{ik} + \mathrm{cov}[b_{ij}, b_{hk}])] \\ &\quad i=1,2,\ldots m,\ j=1,2,\ldots n,\ h=1,2,\ldots m,\ k=1,2,\ldots n \end{aligned} \tag{5.38}$$

If $\bar{b}$ and Σ_b satisfy the requirements that follow from $b(t)$ being constrained to the hypercube [0,1] the r.h.s. of (5.36) represents the covariance matrix of a random variable and hence represents a positive definite matrix. The possibilities to apply result (5.36) in a practical context are more limited than for result (5.26) as the matrix Σ_b, or its approximation, needs to be available. When an approximation of Σ_b is used, it should be checked that the r.h.s. of (5.36) remains positive definite.

5.6 Conclusions

This chapter considers the covariance matrix for the measurement error in (2.23). It was shown to be possible to derive such a matrix using the assumptions in the motorway model. However the matrix depends on the unknown split probabilities and therefore can not be directly used in an estimation procedure. Therefore three categories of approximations have been considered. These are time invariant approximations, point estimate based approximations (PEBA) and distribution based approximations (DBA).

6. Serial Correlations of Split Proportions

6.1 Introduction

Taking the state equation (2.7) as a point of departure, this chapter is concerned with the properties of the drift variables $w(t)$ and the systematic changes in the split probabilities $u(t)$. These properties are an important part of the system specification, as they determine the weight that is put on older observations, and the direction in which the estimates of the split probabilities tend to move. Little is known about the best specification of these properties, as direct observations of EE-tables are not available in general. This chapter contains two results on specifying the properties of $u(t)$ and $w(t)$.

The first result concerns the variance of $w(t)$. Assuming that $u_{ij}(t)$ is chosen in such a way that $\mathrm{E}[w_{ij}(t)]=0$, the *rate of change* in the split proportions is expressed with $\mathrm{E}[w_{ij}(t)^2]$. By analysing a database of vehicle trips, which was made available by the authorities responsible for a toll road in the south of France, an appropriate value for $\mathrm{E}[w_{ij}(t)^2]$ was determined for one particular network. The findings are reported in section 6.2.

The second result concerns the variable that denotes the default change in the split probabilities, $u(t)$. This variable is only non-zero if external information indicates that the split probabilities are likely to move in a certain direction. Usually such information is derived from historic data. It is envisaged that a number of *day profiles* may be distinguished, e.g. weekdays, weekends, holidays, etc., and that for each profile the historical averages of the split proportions, represented by $b^{\mathrm{p}}(t)$, are stored. The value of $u(t)$ can then be derived from this historical average. The profiles, $b^{\mathrm{p}}(t)$, may be adapted on a daily basis. An adaptation mechanism and the way this historical average influences the estimate of $b(t)$ are discussed in section 6.3.

6.2 Empirical data on the rate of change in split proportions

The common state propagation assumptions result in using (2.7) in combination with:

$$\begin{gathered} u(t)=\mathbf{0} \\ \mathrm{E}[w(t)]=\mathbf{0} \\ \mathrm{cov}[w(t),w(p)]=\sigma^2.I.\delta_{tp} \end{gathered} \tag{6.1}$$

where σ^2 is chosen as a design parameter, see e.g. *Nihan and Davis (1987), Cremer and Keller (1987), Van Der Zijpp and Hamerslag (1994b).* Figure 6.1 shows the one hour averaged split proportions plotted against the corresponding values for the previous hour. The data have

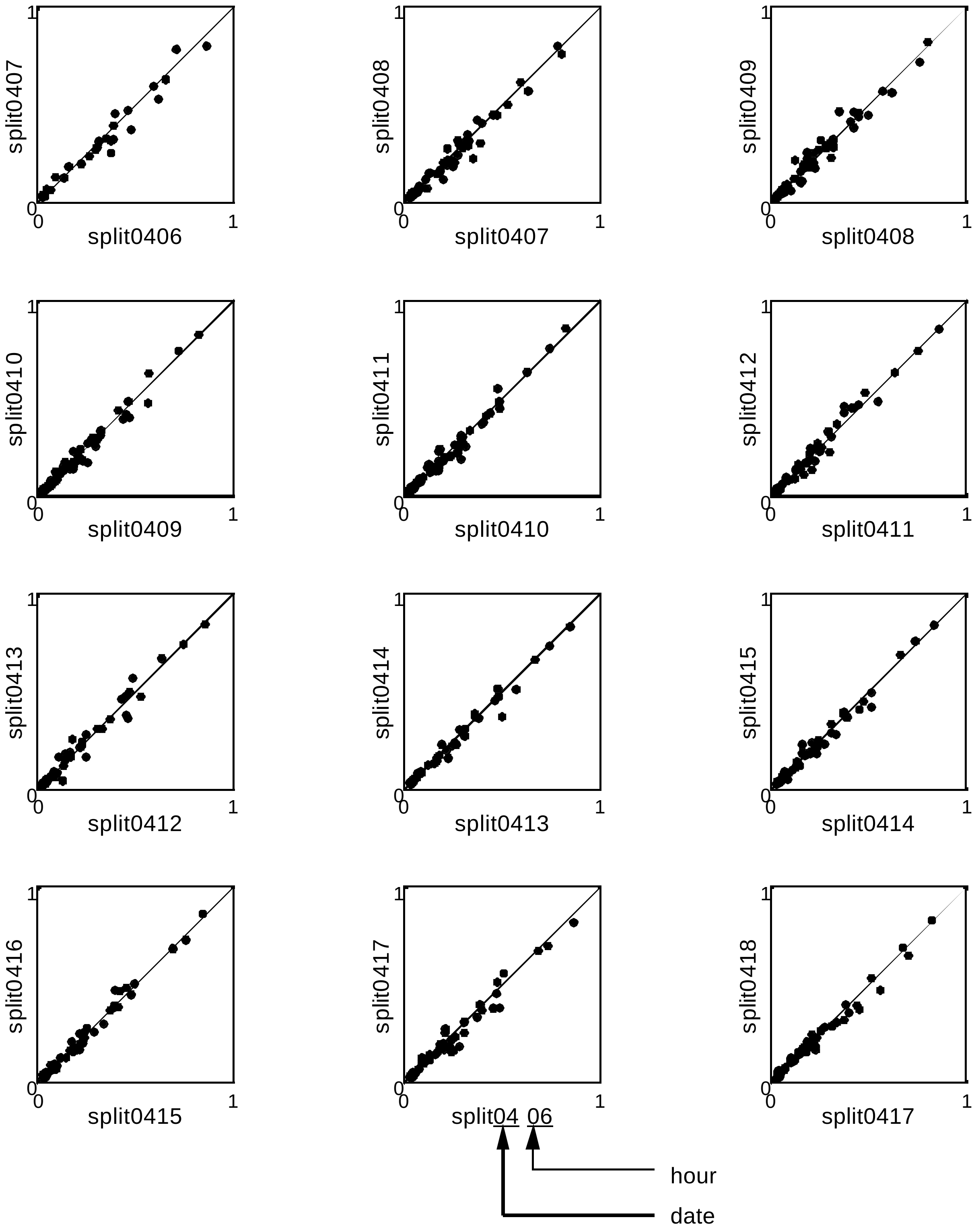


Figure 6.1: *One hour-averaged split-ratios on the tollroad 'route du soleil' in France, plotted against the split-ratios during the previous hour of departure, measured from 6.00-18.00 on tuesday, August 4, 1992*

been derived from toll tickets that were issued to motorists on the ‘route du soleil’ in France. Analysing these trips during a month on this route resulted in figure 6.2 that shows the rate of change as a function of time of day (only trips with a departure time between 7:00 and 18:00 were analysed). The rate of change is expressed in terms of the mean squared error, MSV(*t*), i.e.:

$$\mathrm{MSV}(t)=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(b^{\circ}_{ij}(t)-b^{\circ}_{ij}(t-1))^{2}}{mn} \qquad (6.2)$$

where $b^{\circ}_{ij}(t)$ represents the observed split proportions. These data provide some guidance in choosing an appropriate value for σ^2 in (6.1). The averaged value over all OD-pairs for MSV(*t*) is *0.0013* per hour. This value is based on the actual split *proportions* and hence includes random variation due to the uncoordinated choices of motorists, as well as the default change *u*(t) in (2.7). This value should therefore be viewed as an upper bound for the variation in the split *probabilities*.

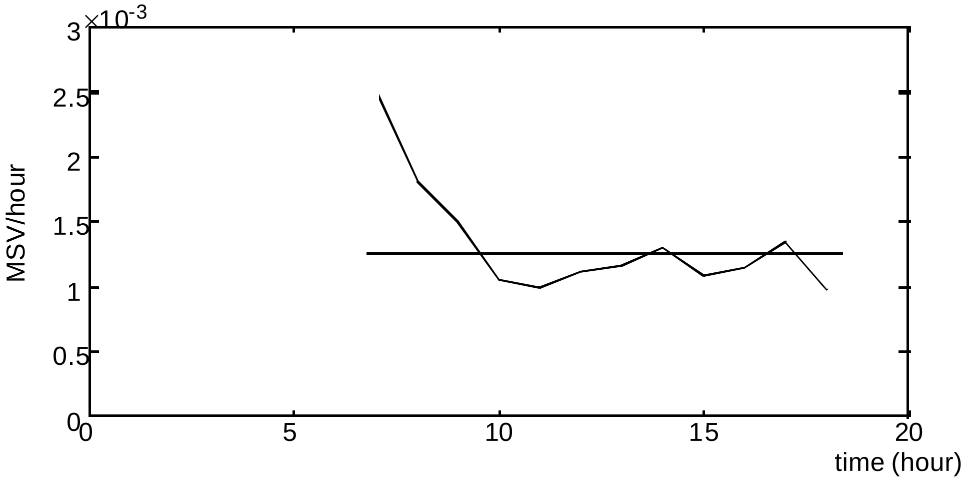


Figure 6.2: Average squared variation in actual split proportion over the month August 1992, Average squared variation in split proportion per hour: 0.001256

6.3 Default model

In the previous section the size of the random *variation* in the split probabilities was considered, while this section considers the *systematic* or *default* component, *u*(*t*), in this variation. It is envisaged that this component can be derived from historic data. For this purpose we introduce a notation where, in addition to an index for the time period, an index for the day is also included. Let $b^k(t)$ denote the vector of split probabilities in period *t* on day *k*.

Up to this point the assumption of slowly varying split probabilities was only applied to the evolution of the split probabilities within a single day, and was summarized in the following equation:

$$b^k(t+1)=b^k(t)+w(t) \qquad (6.3)$$

This evolutionary behaviour is illustrated in figure 6.3, and is referred to as the *vertical random walk* model. However, an assumption of a *horizontal random walk* is also plausible:

$$b^{k+1}(t)=b^{k}(t)+h^{k}(t) \tag{6.4}$$

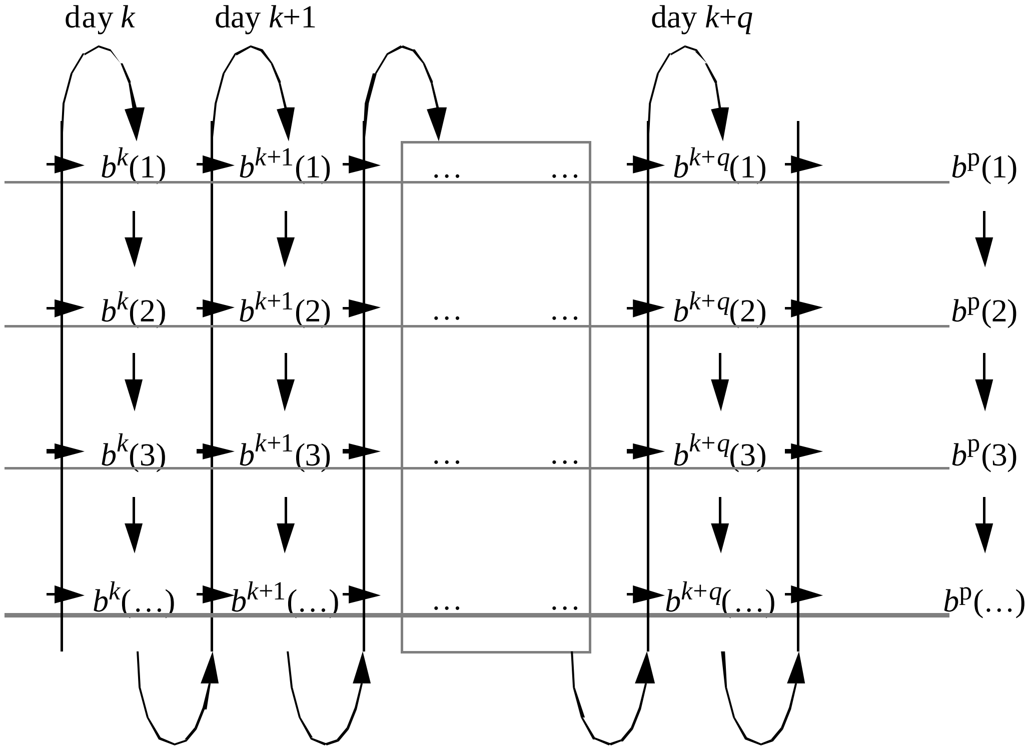


Figure 6.3: Vertical and horizontal random walk model

Equation (6.4) expresses the fact that traffic has a daily repeating pattern. Support for this hypothesis can again be derived from the tollticket data of the French tollroad. Like figure 6.1, figure 6.4 shows the one hour averaged split proportions, averaged over one hour, this time plotted against the corresponding values for the previous days. The plots show correlations between split proportions on consecutive days that are near to one.

In (6.4), the variation of $h^k(t)$ decreases if the period length increases, since an increased period length makes the model less susceptible for variations in departure time choice. Therefore (6.4) is usually associated with static methods, see for example *Maher (1983)*. The following model combines assumptions (6.3) and (6.4) and will be called the *default model*:

$$\begin{gathered} b^{k+1}(t)=b^{p,k}(t)+x(t) \\ x(t+1)=\alpha.x(t)+\varepsilon(t) \\ 0\leq\alpha\leq 1 \end{gathered} \tag{6.5}$$

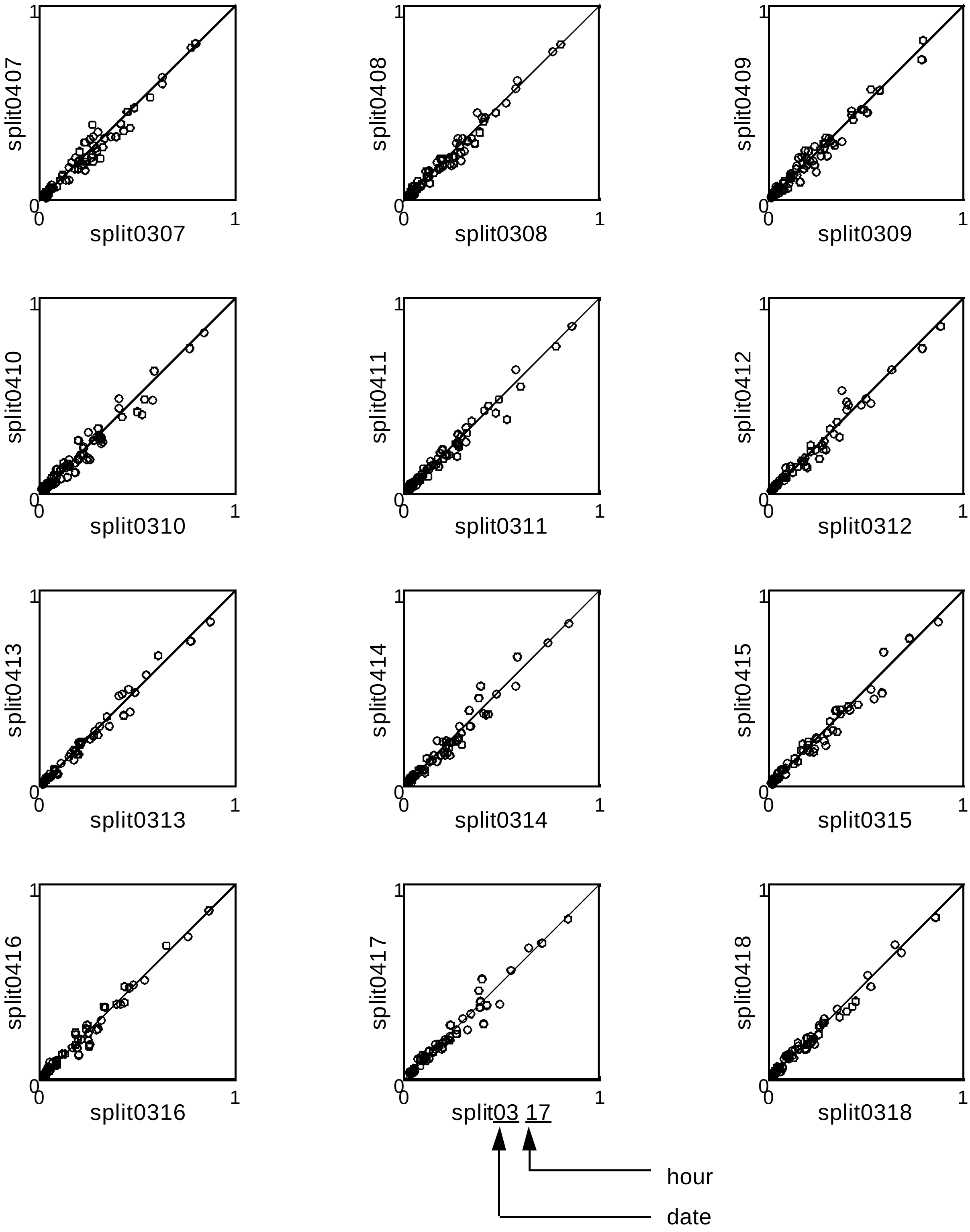


Figure 6.4: One hour-averaged split-ratios on the tollroad 'route du soleil', plotted against split-ratios during the corresponding hour of departure on the previous day, measured from 6.00-18.00 on monday and tuesday, August 3 and 4, 1992

with:

$b^k(t)$:	vector of split proportions in period t on day k
$b^{p,k}(t)$:	profile for day k, based on historic data
$x(t)$:	displacement
α:	discounting factor for displacement
$\varepsilon(t)$:	zero mean random noise

Equation (6.5) expresses that $b^k(t)$ is a process that randomly varies around the historic averages $b^{p,k}(t)$. A similar model has been proposed by *Ashok and Ben-Akiva (1993)* in a slightly different context. The vector $b^{p,k-1}$ represent a *profile* for the split probabilities updated until and including day k-1. This profile is compiled from all available historic data and may be adapted on a daily basis, for example by storing the moving average of the estimates $\overline{b}^k(t)$:

$$b^{p,k}(t)=(1-\beta)b^{p,k-1}(t)+\beta\overline{b}^k(t) \\ 0\leq\beta\leq 1 \tag{6.6}$$

An appropriate value for β should be chosen on the basis of experiments with real data.

It seems natural to work with multiple profiles, and to base the decision to assign a separate profile to a category of days on practical grounds, e.g. is it possible to identify a certain type of day in advance?, can characteristic traffic patterns be observed on this day?, are sufficient historical data available?, etc.

The variable $x(t)$ denotes the *displacement* of the split probabilities relative to the historic average. The expected value of the one step prediction of this variable is located between its current value and zero, depending on the value of a discounting factor α that, again, should be chosen on the basis of experiments with real data. If this factor is given a value smaller than one then the predictions based on state equation (6.5) will tend to move closer to the historic average with the increase of the prediction horizon.

Equation (6.5) has two extreme cases:

- α=0. In this case there is no serial correlation between the displacements $x(t)$
- α=1. In this case there is no force that drives back the displacements $x(t)$.

In equation (6.5), $x(t)$ rather than $b(t)$ is the variable that is to be estimated. In order to fit (6.5) into the existing framework (2.7), a number of manipulations are performed. According to (2.7), the following holds:

$$b(t+1)-b(t)=u(t)+w(t) \tag{6.7}$$

On the other hand, according to (6.5) the following holds:

$$\begin{aligned} b(t+1)-b(t)&=b^{p,k}(t+1)+x(t+1)-b^{p,k}(t)-x(t) \\ &=b^{p,k}(t+1)-b^{p,k}(t)+\varepsilon(t)+(\alpha-1)x(t) \\ &=b^{p,k}(t+1)-b^{p,k}(t)+(\alpha-1)[b(t)-b^{p,k}(t)]+\varepsilon(t) \end{aligned} \tag{6.8}$$

By matching the systematic and the random components of (6.7) and (6.8) it follows that:

$$u(t)\equiv b^{p,k}(t+1)-b^{p,k}(t)-(1-\alpha)[b(t)-b^{p,k}(t)] \\ w(t)\equiv\varepsilon(t) \tag{6.9}$$

Jointly equations (2.7), and (6.9) define a straightforwardly applicable time propagation model that takes into account historic experience. To illustrate how the use of these equations works out, substitute (6.9) in (2.7). The result can then be rearranged as follows:

$$\begin{aligned} b(t+1) &= b(t)+b^{\mathrm{p},k}(t+1)-b^{\mathrm{p},k}(t)-(1-\alpha)[b(t)-b^{\mathrm{p},k}(t)]+w(t) \\ &= \alpha b(t)+b^{\mathrm{p},k}(t+1)-\alpha b^{\mathrm{p},k}(t)+w(t) \\ &= \alpha[b(t)+b^{\mathrm{p},k}(t+1)-b^{\mathrm{p},k}(t)]+(1-\alpha)b^{\mathrm{p},k}(t+1)+w(t) \end{aligned} \tag{6.10}$$

The one step prediction is hence a weighted average of an extrapolation of the current estimate and the historical mean. With parameter α one can tune how fast the prediction will be pulled back to the historical value.

6.4 Conclusions

Empirical data reveal a serial correlation between split-ratios that is near to one. This correlation is observed both between split ratios referring to consecutive time periods within a single day (see figure 6.3), and between split ratios referring to a single time period on different days (see figure 6.4).

The simultaneous usage of both properties leads to equation (6.10). This novel expression combines usage of the assumption of slowly varying split probabilities with usage of historical data.

7. Combined Data Sources

7.1 Introduction

Up to this point we have exclusively dealt with dynamic EE-flow estimation from time series of traffic counts.Traffic counts represent also the most widely available class of data that can be obtained using automated data collection techniques. However, it is expected that in the near future the information generated by automated observation techniques will no longer be limited to volume counts only but will also comprise trajectory information. Trajectory information relates to the usage of *partial paths*, i.e. sequences of two or more adjacent links.

The availability of this type of information may stem from new devices such as automated license plate readers and transponders, or as a side effect of new products such as in vehicle route guidance. These collection techniques will be discussed in section 8.4. In this chapter it will be shown how the processing of a combination of trajectory information and link volumes fits in the Bayesian EE-estimation framework presented in chapter 4. As an example of the use of trajectory information, the processing of license plate surveys will be considered. Extensions to other sources of trajectory information are discussed in section 7.4.

7.2 Historical overview of the analysis of registration plate data

It has long since been recognized that license plates can be a valuable source of (static) OD-information (*Kryger and Ottesen, 1956, Brenner et al., 1957*). Until recently, the recording of license plate registrations had to be done manually. To economize on manpower as well to minimize errors of (the manual) recording it has become usual to only record a part of the license plate, for example a combination of the last three digits or letters. Such a survey is known as a *partial license plate survey.*

A problem that occurs if only a part of the license plate is recorded is that of the *spurious matches* which occur if two different vehicles have identical partial registrations. The extent to which spurious matches influence the OD-estimates was studied by *Makowski and Sinha (1976)* and *Hauer (1979)*, who also proposed approximate statistical procedures to estimate OD-matrices from partial license plate surveys in the presence of spurious matches. These procedures were later improved by *Maher (1985)*.

Another area of work has addressed the *elimination* of spurious matches. This has resulted in procedures that are usually heuristic and require not only the partial registration, but also other data, such as time stamp, vehicle type etc. Spurious matches are then eliminated by imposing extra conditions to a ‘match’, for example by requiring corresponding vehicle types, arrival within a certain time window, etc. Methods proposed by *Shewey (1983)* and *Evans et al. (1993)* can be classified in this category. Also most commercially available number plate matching programs are of this category, e.g. *Buchanan and Partners (1986)*, *Lucas (1986)*, and *MVA Systematica (1987)*.

Statistical procedures based on partial registration data *and* time stamp information were proposed in *Watling and Maher (1988)*, *Watling (1990)*, and *Watling and Maher (1992)*.

Some effort has also been spent on the recovery from observation errors. Some types of errors are frequently made by human observers, for example mixing up the characters ‘3’ and ‘8’ or transposing characters, e.g. writing down ‘xx-34-np’ instead of ‘xx-43-np’. Heuristic procedures to compensate or correct for this type of recording errors are common practice, see e.g. *Hamerslag (1978)*, the NOPCOP (*Lucas, 1986*) and MicroMatch computer programs (*Buchanan and Partners, 1986*), and proposed improvements by *Evans et al. (1993)*.

Reviewing the literature on this subject, it catches the eye that there is an emphasis on the statistical treatment of the phenomenon of spurious matches, while very little attention is paid to the statistical treatment of recording errors. Noted exceptions are *Geva et al., (1982)* who elegantly treat the problem of recording errors, and *Watling (1990, 1994)*. Watling discusses the case where a combination of spurious matches and recording errors occurs, but omits a tests for this combination due to computational constraints.

7.3 Processing combined data

The problem that is considered in this chapter deviates from the traditional problem in a number of ways:

- The license plate surveys will be used in the context of *dynamic* EE-estimation rather than static OD-estimation. In the existing approaches the time stamp data are used only to eliminate spurious matches. A trivial extension is however to use the time stamp data also to obtain time varying EE-estimates.
- Instead of recording partial registration numbers, *complete* registrations will be read using automated license plate readers. This eliminates the problem of spurious matches, since each registration uniquely identifies a vehicle. On the other hand this introduces the problem of *recording errors*. If license plates are to be read in a cost effective manner then less than perfect recognition rates have to be taken into account. Researchers report recognition rates during field tests of 90% at daytime and 65% at night (*Kanayama et al., 1991*) from CCTV (front-view) images obtained from a single floodlight assisted camera.
- We will be interested in using a *combination* of license plate surveys and link volume counts rather than exclusively using license plate surveys.
- Contrary to the usual assumptions, where the locations of recording the license plates jointly define a cordon (*Ortúzar and Willumsen, 1990*) and individually correspond to either origins or destinations, recording stations need not correspond with origins or destinations, and an arbitrary number of recording stations may be present on each EE-path (see figure 7.1).

These conditions give a completely different perspective to the problem, and open up a new field of research. This section introduces some new notation (section 7.3.1). Subsequently the problems of updating from trajectory counts and of updating from combined data are discussed (sections 7.3.2 and 7.3.3)

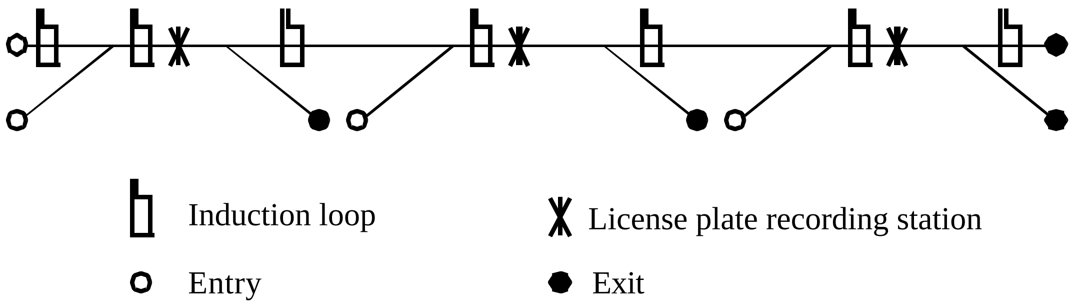


Figure 7.1: *Sample network. Traffic data are collected with induction loops and license plate readers.*

7.3.1 Notation

The following new symbols are used in this chapter:

h	Number of license plate readers
$e_{rs}(t)$	Trajectory count: Number of vehicles observed during period t at location r and at location s, but not at any site upstream of r or downstream of s. The numbering of the license plate readers is chosen in such a way that $r<s$ implies that r is not reachable from s.
$e(t)$, $\theta(r,s)$	Vector of trajectory counts, location of element $e_{rs}(t)$ in the vector $e(t)$.
α_k	Recognition rate at site k.
k	Path-license plate reader incidence map. $\kappa_{ijr}=1$ if route i-j uses license plate reader r and zero otherwise, $r=1,2,\ldots h$
$g_{rs}^{ij}(t)$	Trajectory count contribution, the number of trips that contribute to both EE flow $f_{ij}(t)$ and trajectory count $e_{rs}(t)$.
p_{rs}^{ij}	Trajectory count contribution probability; the probability that a trip in flow $f_{ij}(t)$ contributes to $e_{rs}(t)$.
$g(t)$, $\phi(r,s,i,j)$	Vector of trajectory count contributions, location of $g_{rs}^{ij}(t)$ in this vector.
p	Vector of trajectory count contribution probabilities.
$y^{\mathrm{H}}(t)$	Combined observation vector, $y^{\mathrm{H}\prime}(t)=[y'(t)\ e'(t)]'$

7.3.2 Analysis; trajectory counts

In the analysis it is assumed that each image processor produces a list of, partly erroneous, registrations. Only the *number* of matches is used in the estimation process. These numbers are summarized in a vector $e(t)$ consisting of *trajectory counts* $e_{rs}(t)$, $r<s$, which denote the number of vehicles that were recognized at site r and site s during period t, but not at any site upstream of r or downstream of s. An extra category $e_{00}(t)$ accounts for those vehicles that are detected correctly once or not at all.

Whether or not a vehicle travelling on EE-pair i-j contributes to $e_{rs}(t)$ depends on a number of nested selections. In order to contribute, it is required that no recognition (NR) occurs upstream of r or downstream of s, and recognition (R) occurs at both r and s, so the probability that a trip in flow $f_{ij}(t)$ contributes to $e_{rs}(t)$ equals:

$$\mathrm{P}[(\mathrm{NR\ upstream\ of\ }r) \wedge (\mathrm{R\ at\ }r) \wedge (\mathrm{R\ at\ }s) \wedge (\mathrm{NR\ downstream\ of\ }s)] \tag{7.1}$$

This probability is given by:

$$p^{ij}_{rs} = \left(\prod_{a=1}^{r-1} (1-\alpha_a)^{\kappa_{ija}}\right) \alpha_r^{\kappa_{ijr}} \alpha_s^{\kappa_{ijs}} \left(\prod_{b=s+1}^{h} (1-\alpha_b)^{\kappa_{ijb}}\right) \tag{7.2}$$

for $1 \le r < s \le h$, and complemented with p^{ij}_{00}.

As an aid in deriving the joint probability distribution of $\{e_{rs}(t),\ 1 \le r < s \le h\}$, *trajectory count contributions,* denoted by $g^{ij}_{rs}(t)$, are introduced which denote the contribution of flow $f_{ij}(t)$ to trajectory count $e_{rs}(t)$. The probability that an arbitrary vehicle entering at i will contribute to trajectory count $g^{ij}_{rs}(t)$ is equal to $b_{ij}(t).p^{ij}_{rs}$. Hence, the joint distribution of the trajectory count contributions is given by the following multinomial distribution, see *Lehmann (1983),* pg.28:

$$\mathrm{P}[g(t)|\tilde{q}(t),b(t)] = \prod_{i=1}^{m} \tilde{q}_i! \prod_{j=1}^{n} \prod_{r<s \vee rs=0} \frac{(p^{ij}_{rs} b_{ij}(t))^{g^{ij}_{rs}(t)}}{g^{ij}_{rs}(t)!} \tag{7.3}$$

Now let $\theta(r,s)$ denote the position of the element $e_{rs}(t)$ in the vector of trajectory counts $e(t)$, then the expectation of $e(t)$ is given by:

$$\mathrm{E}[e(t)|q(t)] = \mathrm{E}_{\tilde{q}(t)}[\ \mathrm{E}[e(t)|\tilde{q}(t)]\ \ |q(t)\ \ ] = G'(t)b(t) \tag{7.4}$$

where $G(t)$ is defined by:

$$G_{\varphi(i,j),\theta(r,s)}(t) = q_i(t).p_{rs}{}^{ij}$$
$$i=1,2\ldots m,\ \ i=1,2\ldots n,\ \ r=1,2\ldots h,\ \ s=r+1,r+2\ldots h \tag{7.5}$$

Analogous to (2.23), equation (7.4) gives rise to the specification of a measurement equation:

$$e(t) = G'(t)b(t) + z(t) \tag{7.6}$$

where $z(t)$ is a zero mean random variable that accounts for random effects and mis-specification of $G(t)$ due to observation errors contained in $q(t)$. Like described in chapter 5, one can derive the covariance matrix of $z(t)$ from the multinomial distribution of the trajectory count contributions, or one can choose a more simple approach, replacing this matrix with a diagonal matrix derived from the average observed values.

Instead of analysing this issue in detail at this stage we will first consider the more general case of estimating split probabilities from combined link volume counts - trajectory counts.

7.3.3 Analysis; combined data

The combined usage of volume counts and trajectory information is a yet unexplored possibility to collect dynamic EE-information. It is expected that the two sources of data effectively complement each other, provided that an appropriate statistical estimation procedure is used.

Equations (2.23) and (7.6) can be combined in the following measurement equation:

$$y^{\mathsf{H}}(t)=H^{\mathsf{H}\prime}(t)b(t)+v^{\mathsf{H}}(t) \tag{7.7}$$

with:

$$y^{\mathsf{H}}(t)=\begin{bmatrix} y(t) \\ e(t) \end{bmatrix}, \ H^{\mathsf{H}\prime}(t)=\begin{bmatrix} H'(t) \\ G'(t) \end{bmatrix}, \text{ and } v^{\mathsf{H}}(t)=\begin{bmatrix} v(t) \\ z(t) \end{bmatrix} \tag{7.8}$$

Dependencies between elements of the measurement error vector $v^{\mathsf{H}}(t)$ follow from the fact that both $y(t)$ and $g(t)$ can be written as linear combinations of trajectory count contributions which are multinomially distributed according to (7.3). The estimates of the split probabilities may benefit from knowledge of these dependencies, which may be summarized in a covariance matrix R_t^{H}. Parallel to (2.10) and (5.16) it holds that:

$$y^{\mathsf{H}}(t)=U^{\mathsf{H}\prime}g(t)+\begin{bmatrix} s\,(t) \\ \mathbf{0} \end{bmatrix} \tag{7.9}$$

for some matrix U^{H} of which each element is either one or zero.

The traditional problem of estimating EE-matrices from link volumes is characterized by the dependencies $b \to f \to y$, where b needs to be estimated, the conditional distribution of f given b is multinomial, and y is a linear combination of f.

The problem of estimating EE-matrices from combined data can be characterized in a similar way. In fact for each network with combined observations one can define a ‘hypernetwork’ with link volume observations only, in which the observations behave equivalently. For this purpose divide each flow $f_{ij}(t)$ into subflows $g_{rs}^{ij}(t)$ that each travel over their own imaginary path with a probability $b_{ij}(t)p_{rs}^{ij}$ and suppose that the observations $o(t)$ satisfy (7.9).

Now $o(t)$ has the statistical properties that belong to a combined observation but at the same time is a linear combination of (notional) subflows. The statistical properties can hence be captured in the equations derived in chapter 5.

7.3.4 Computing the covariance matrices

Using the above described hypernetwork approach opens up the possibility of computing a point-estimate based approximation (PEBA) or a distribution based approximation (DBA) for R_t^{H}, using the earlier derived equations (5.26) and (5.37) respectively.

However, this necessitates that the equivalent is supplied for the matrices U', p, $\boldsymbol{Q_t}$, $\boldsymbol{B_t}$ and Θ in (5.26), and in addition to that, for Σ_b in (5.37). These equivalents will be marked with the symbol ‘$^{\mathsf{H}}$’. For the hypernetwork where every trajectory count contribution $g_{rs}^{ij}(t)$ travels over its own path, it follows that these matrices are given by:

$$U^{\mathsf{H}}=\begin{bmatrix} V^1 & V^2 \end{bmatrix} \tag{7.10}$$

where the nonzero elements of V^1 and V^2 satisfy:

$$\begin{gathered} V^1{}_{\phi(r,s,i,j),k}=\tau_{ijk} \\ V^2{}_{\phi(r,s,i,j),\theta(r,s)}=1 \\ 1\leq r<s\leq h,\ i=1,2,\ldots m,\ j=1,2,\ldots n \end{gathered} \tag{7.11}$$

The equivalent of p is given by:

$$\mathrm{p}^{\mathrm{H}}{}_{\phi(r,s,i,j),i}=1 \tag{7.12}$$

for $1 \le r < s \le h$, $i=1,2,\dots m$, $j=1,2,\dots n$, and zero for all elements not defined by (7.12). The equivalent of $\boldsymbol{Q_t}$ is given by:

$$\boldsymbol{Q_t}^{\mathrm{H}}=\mathrm{diag}(\mathrm{p}^{\mathrm{H}}q(t)) \tag{7.13}$$

The equivalent of $\boldsymbol{B_t}$ is given by:

$$\boldsymbol{B_t}^{\mathrm{H}}=\mathrm{diag}(\bar{b}^{\mathrm{H}}(t)) \tag{7.14}$$

where:

$$\bar{b}^{\mathrm{H}}{}_{\phi(r,s,i,j)}=\bar{b}_{ij}(t)\ p^{ij}_{rs} \tag{7.15}$$

for $1 \le r < s \le h$, $i=1,2,\dots m$, $j=1,2,\dots n$, and zero for all elements not defined by (7.15). The equivalent of Θ is given by:

$$\Theta^{\mathrm{H}}=\begin{bmatrix}\Theta & 0\\ 0 & 0\end{bmatrix} \tag{7.16}$$

The matrix Φ can be used unaltered. Substituting these results in (5.26), gives the following point-estimate based approximation for the covariance of the observation error associated with (7.7):

$$\begin{aligned} R_t^{\mathrm{H}} &= \mathrm{E}[v^{\mathrm{H}}(t)\, v^{\mathrm{H}\prime}(t) | b(t)=\bar{b}(t)] \\ &= U^{\mathrm{H}\prime}[\ \boldsymbol{Q_t}^{\mathrm{H}}(\boldsymbol{B_t}^{\mathrm{H}}-\boldsymbol{B_t}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}\prime}\boldsymbol{B_t}^{\mathrm{H}\prime})+\boldsymbol{B_t}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}}\Phi\mathrm{p}^{\mathrm{H}\prime}\boldsymbol{B_t}^{\mathrm{H}}\ ]U^{\mathrm{H}} + \Theta^{\mathrm{H}} \end{aligned} \tag{7.17}$$

Instead of (7.17) one may want to compute a distribution-based approximation of R_t^{H}. In this case also the hypernetwork equivalent of Σ_b, denoted with Σ_b^{H}, needs to be supplied. This matrix denotes the covariance of the vector $b^{\mathrm{H}}(t)$ of which the nonzero elements are defined by:

$$b^{\mathrm{H}}{}_{\phi(r,s,i,j)}=b_{ij}(t)\ p^{ij}_{rs} \tag{7.18}$$

for $1 \le r < s \le h$, $i=1,2,\dots m$, and $j=1,2,\dots n$. Expression (7.18) may be written in a way that is computationally more convenient:

$$\begin{gathered} b^{\mathrm{H}}(t)=P^{\mathrm{H}}b(t) \\ P^{\mathrm{H}}=\mathrm{diag}(\ p\ )\mathrm{p}^{\mathrm{H}} \end{gathered} \tag{7.19}$$

where p is the vector of (fixed) trajectory count contribution probabilities p^{ij}_{rs}. As result of (7.19) the covariance of $b^{\mathrm{H}}(t)$ can now be derived from the covariance of $b(t)$ according to:

$$\Sigma_b^{\mathrm{H}}=P^{\mathrm{H}}\ \Sigma_b\ P^{\mathrm{H}\prime} \tag{7.20}$$

Substituting result (7.20), jointly with the earlier obtained results in (5.37), results in an expression for the distribution based approximation of R_t^{H}:

$$R_t^{\mathrm{H}} = \mathrm{E}[v^{\mathrm{H}}(t)\, v^{\mathrm{H}\prime}(t)]$$
$$= U^{\mathrm{H}\prime}[\ \boldsymbol{Q_t}^{\mathrm{H}}(\boldsymbol{B_t}^{\mathrm{H}}\text{-}\boldsymbol{B_t}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}\prime}\boldsymbol{B_t}^{\mathrm{H}\prime})+\boldsymbol{B_t}^{\mathrm{H}}\mathrm{p}^{\mathrm{H}}\Phi\mathrm{p}^{\mathrm{H}\prime}\boldsymbol{B_t}^{\mathrm{H}}\ \text{-}\boldsymbol{Q_t}^{\mathrm{H}}\,(\Sigma_b^{\mathrm{H}}\otimes(\mathrm{p}\mathrm{p}'))+\Sigma_b^{\mathrm{H}}\otimes(\mathrm{p}\Phi\mathrm{p}')\ \ ]U^{\mathrm{H}} + \Theta^{\mathrm{H}} \tag{7.21}$$

7.3.5 Estimating the split probabilities

The estimation of the split probabilities from combined data poses no new problems. If the measurement equation (2.23) is replaced with (7.7), and either the PEBA covariance matrix (7.17) or the DBA covariance matrix (7.21) is used, then the estimators that were described in chapter 3 and the new Bayesian updating method that was described in chapter 4 can be applied unaltered.

The above described method to estimate split probabilities from combined data has been tested in experiments using synthesized data. These experiments show that using combined data consistently leads to lower errors of estimation relative to the case where only traffic counts are used. Likewise, usage of the Bayesian updating method described in chapter 4 results in lower errors of estimation than usage of the parameter optimization methods and the traditional Kalman filter described in chapter 3 (*Van Der Zijpp, 1996*).

7.4 Extension to other sources of trajectory information

Up to this point, the chapter has dealt explicitly with trajectory information obtained from license plate readers. Other possibilities to obtain this type of information are the use of probe vehicles transmitting their trajectories, the extraction of information from route guidance equipment, the use of vehicles equipped with transponders, or the use of vehicle classifications, using the acoustic or electromagnetic properties of the vehicle. These data collection techniques can be categorized on the basis of the following properties (see table 7.1):

- *The trajectory to which the information relates*. This is either fixed or variable. If the data collection depends on infrastructure based detectors then the observed trajectory is usually fixed. If floating car data are used, e.g. obtained from vehicles actively participating in the data collection then the trajectories may refer to any part of a feasible path.
- *The time at which information becomes available*. If there is active participation of vehicles in the observation process then information about intended destinations can become available before completion of the trip. For example in order to receive in-vehicle route guidance motorists have to make their destinations known in advance.
- *The coverage of the observation*. License plates have a complete coverage since each vehicle is equipped with one. Alternatively a trajectory observation may relate to a sample of all vehicles, for example those equipped with a transponder.
- *The possibility of recording errors or mis-classification*. Although no method is totally exempt from all error, in some cases these errors can be safely neglected while in other cases they can not.
- *The possibility of spurious matches*. This possibility is excluded if vehicles are uniquely identified, but plays a dominant role in methods where vehicles are classified in a limited number of categories, for example based on the number of axles, or on the basis of parts of their registrations.

The theory that is presented in the previous sections assumes that the properties of the first row of table 7.1 apply. In the following we will examine to what extent other sources of trajectory information can be used.

Table 7.1: Typology of trajectory observation techniques

group([a])	trajectory	time of reporting	coverage	possible recording errors	possible spurious matches
License plate survey	fixed	after trip	100%	yes	no
Active probes	variable	after trip	sample	no	no
Transponders	fixed	after trip	sample	no	no
Vehicle classification	fixed	after trip	100%	yes	yes
Route guidance	fixed	before trip	sample	no	no

a. : The term active probes refers to vehicles storing and transmitting their trajectory, see *Westerman (1995)*, other vehicles marked up for the sole purpose of collecting traffic data with batches, bar-codes or transponders are referred to as passive probes.

With respect to trajectory and time of reporting, data obtained from active probes and route guidance equipment exceed the requirements. Therefore, as far as these aspects are concerned these sources of trajectory information may be used instead of license plate data, although this would come down to disregarding or delaying the processing of a part of the information.

The issue of incomplete coverage can be approached by introducing a sampling criterion. Traditionally this sampling criterion was used to lower the burden on the human observer, see *Geva et al. (1982).* An example is limiting a license plate survey to only white vehicles. The same theory can however be applied to any source of trajectory information. The trajectory information only applies to vehicles that satisfy the sampling criterion. Let β be the probability that a vehicle satisfies the sampling criterion, e.g. the fraction of vehicles equipped as probe vehicles, then multiplying all instances of p_{rs}^{ij} (except p_{00}^{ij}) in the previous section with the factor β would result in a theoretically correct method.

With respect to the possibility of recording errors, all methods other than automated license plate surveys have equal or better specifications, therefore as far as this aspect is considered all methods fit in the theory that is described in this chapter.

Finally the possibility of spurious matches, that is two different vehicles being mistaken for one single vehicle, does only occur with methods that use vehicle classification. The theory described in the previous sections does not allow for this possibility and the issue has not been further investigated. A theoretical treatment of the problem of processing license plate surveys where both recording errors and spurious matches occur was given in *Watling (1990).*

Summarizing, the theory described in this chapter not only applies to license plate surveys, but also to trajectory information that may be retrieved if active probes, transponders or route guidance equipment, or in fact any combination of these, is used. The theory does not apply to trajectory information that could be derived from vehicle classification methods.

7.5 Conclusions

Trajectory counts such as license plate surveys have been used for a long time in traffic engineering, but the situation in which this information is used in *combination* with traffic volumes has not been investigated earlier. Also the utilisation of trajectory information for *dynamic* EE-estimation is a new problem. Finally, the use of *multiple recording stations*, positioned at arbitrary locations and the possibility of *recording errors* adds another new dimension to the problem.

Nevertheless, combined volume counts/trajectory observations fit smoothly into the Bayesian framework of EE-estimation. The trajectory information and volume counts have dependencies that stem from the fact that both can be interpreted as a sum of *trajectory count contributions*; those parts of an EE-flow that are observed at at least two sites. To make optimal use of all available observations their mutual correlations should be specified. This chapter describes how this can be done.

The theory of updating from combined observations that is described in this chapter can also be applied to other sources of trajectory information such as probe vehicles, transponders and route guidance equipment.

8. Simulations and Sensitivity Analysis

8.1 Introduction

The objective of this chapter is to make a quantitative comparison between the split ratio methods known from literature of which an overview was given chapter 3, and the new Bayesian method and its variants that is discussed in chapter 4. A second objective is to gain insight into the sensitivity of the methods for characteristics of the traffic system, such as the topology of the network, the size of the EE flows and the rate of change in the split probabilities.

8.2 Methodology

In order to be able to test the methods under a range of circumstances, testdata are generated according to a number of specifications that can be given in advance. The testdata comprise link volumes as well as EE-flows. This makes it possible to evaluate and compare the various methods. The process of generating the testdata is described in detail in section 8.2.1. The multiple separate aspects of the estimation methods that have been discussed in the previous chapter give rise to a large number of possible method variants. The method variants that will be tested are described in section 8.2.2.

Jointly, the different network specifications and method variants give rise to an array of combinations, all of which are evaluated using criteria that are described in section 8.2.3. To reduce the influence of random effects, the above described process of generating datasets and evaluating method variants will be repeated ten times, while averaging the evaluation results.

8.2.1 Generation of testdata

The testdata are generated randomly according to the scheme of the motorway model, see figure 2.3. This involves the following steps:

1. generate the network topology,
2. generate the split probabilities,
3. generate the entry flow rates,
4. generate and assign the EE-flows,
5. generate the entrance volume observation errors and add these to the entry volumes,
6. generate the link volume observation errors and add these to the link volumes.

Above steps result in a set of EE-flows and a set of entry-volume and link-volume observations. The options that can be specified prior to the generation of network and testdata are described in table 8.1.

Table 8.1: Network specifications
-description of parameters-

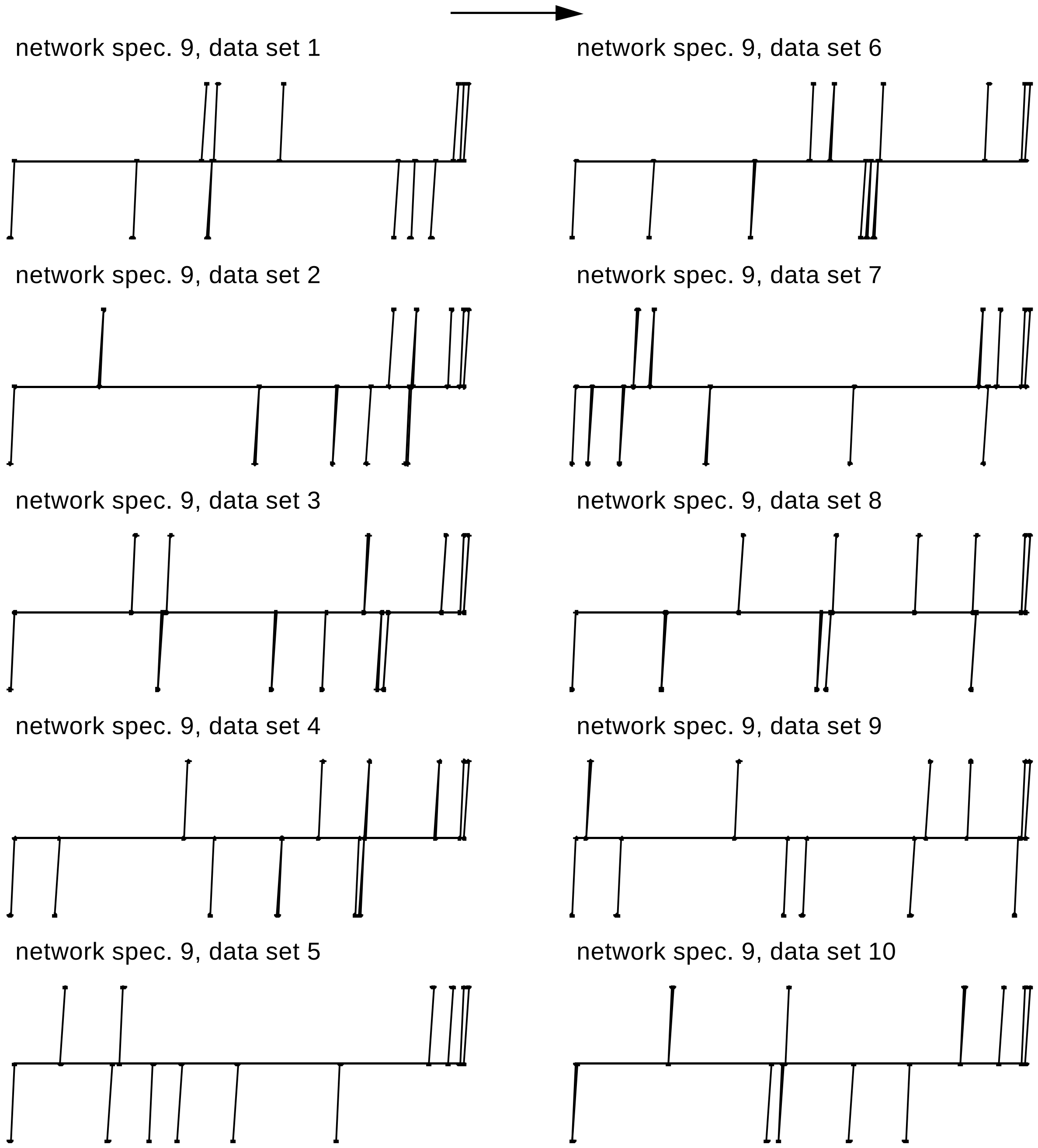
Parameter	Description
m	*The number of entrances.* All generated networks are linear networks with entrances and exits connected to a corridor (see figure 8.1). The locations of all entrances, except the first, are generated randomly. The first entrance is always positioned at the beginning of the corridor. This is to prevent the generation of networks with one or more unreachable exits.
n	*The number of exits.* The positions of the exits except for the last two, are generated randomly. The last two exits are positioned at the end of the corridor.
T	*The maximum number of periods.* This is the number of periods for which data are generated.
σ_b^2	*The rate of change in the split parameters.* The split parameters are initially generated by simulating a multi dimensional random walk according to: $b'(t+1)=b(t)+w'(t),$ $w(t)\sim\text{MVN}[0,\sigma_b^2 I],$ $t=0,1,\ldots T \qquad (8.1)$ Subsequently the values $b'(t)$ are mapped to the hypercube [0,1] using: $b''(t)=\mathbf{1}-\text{abs}(\mathbf{1}-\text{abs}(\ \text{rem}(b'(t),2)\ )) \qquad (8.2)$ Finally the result is normalized using: $b_j(t)=b''_j(t)/\Sigma_{k\in J}\, b''_k(t) \qquad (8.3)$ where J is the index set of EE-pairs that share their origin with EE-pair j. The recursion (8.1), (8.2), (8.3) is initialized with a random value $b(0)$. Examples of sequences of split probabilities generated in this way are shown in figure 8.2.
$\bar{q}$	*The mean entry flow rate.* This scalar value refers to all periods and all entrances and refers to the entry flow rates.The entry flow rates in turn are used as an average when generating the actual entry volumes. These volumes are generated randomly using a normal distribution of which mean and variance are chosen in such a way that a Poisson distribution is approximated.
$\bar{q}^{\text{range}}$	*The range within which the entry flow rates alter.* The entry flow rates may differ among entrances between $(1-\bar{q}^{\text{range}}).\bar{q}$ and $(1+\bar{q}^{\text{range}}).\bar{q}$

Table 8.1: Network specifications
-description of parameters-

Parameter	Description
$\bar{q}^{\text{mode}}$	*The way in which the entry flows alter.* The simulation program has two modes of operation. In the first mode ($\bar{q}^{\text{mode}}$=0) the entry flow rates are constant during the simulation but differ randomly among entrances within the range prescribed by $\bar{q}^{\text{range}}$. In the second mode ($\bar{q}^{\text{mode}}$=1) the flow rates change within time according to the function: $\text{flowrate}_i(t) = \bar{q} \,.\, (1+\bar{q}^{\text{range}}.\cos(\, 2\pi t/T+\text{offset}_i \,)) \qquad (8.4)$ The offsets are generated randomly in the interval [0,π/2]. Examples of sequences of entryflows that are generated using the parameters $\bar{q}$, $\bar{q}^{\text{range}}$, and $\bar{q}^{\text{mode}}$ are shown in figure 8.3.
σ_q^2	*The magnitude of the observation error in the entry volume observations*. After the entering volumes have been generated, an observation error is added to these volumes. This error is randomly generated using: $r(t)\sim\text{MVN}[0,\sigma_q^2 I] \qquad (8.5)$
σ_y^2	*The magnitude of the observation error in the link volume observations*. After generating and assigning the EE-flows, the link-flow volumes are known. To these volumes a randomly generated error is added. These errors satisfy: $s(t)\sim\text{MVN}[0,\sigma_y^2 I] \qquad (8.6)$

Table 8.2: Network specifications
-values of parameters-

	Network								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
$\boldsymbol{m}$	4	4	4	4	4	4	4	4	*6*
$\boldsymbol{n}$	4	4	4	4	4	4	4	4	*6*
$\boldsymbol{T}$	48	48	48	48	48	48	48	48	48
$\sigma_{\boldsymbol{b}}^{\mathbf{2}}$	0.0001	*0.01*	*0*	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
$\bar{\boldsymbol{q}}$	100	100	100	*200*	100	100	100	100	100
$\bar{\boldsymbol{q}}^{\textbf{range}}$	0.5	0.5	0.5	0.5	*0.05*	0.5	0.5	0.5	0.5
$\bar{\boldsymbol{q}}^{\textbf{mode}}$	0	0	0	0	0	*1*	0	0	0
$\sigma_{\boldsymbol{q}}^{\mathbf{2}}$	100	100	100	100	100	100	*10*	100	100
$\sigma_{\boldsymbol{y}}^{\mathbf{2}}$	100	100	100	100	100	100	100	*10*	100



network spec. 9, data set 1

network spec. 9, data set 6

network spec. 9, data set 2

network spec. 9, data set 7

network spec. 9, data set 3

network spec. 9, data set 8

network spec. 9, data set 4

network spec. 9, data set 9

network spec. 9, data set 5

network spec. 9, data set 10

Figure 8.1: The ten randomly generated networks for network specification 9

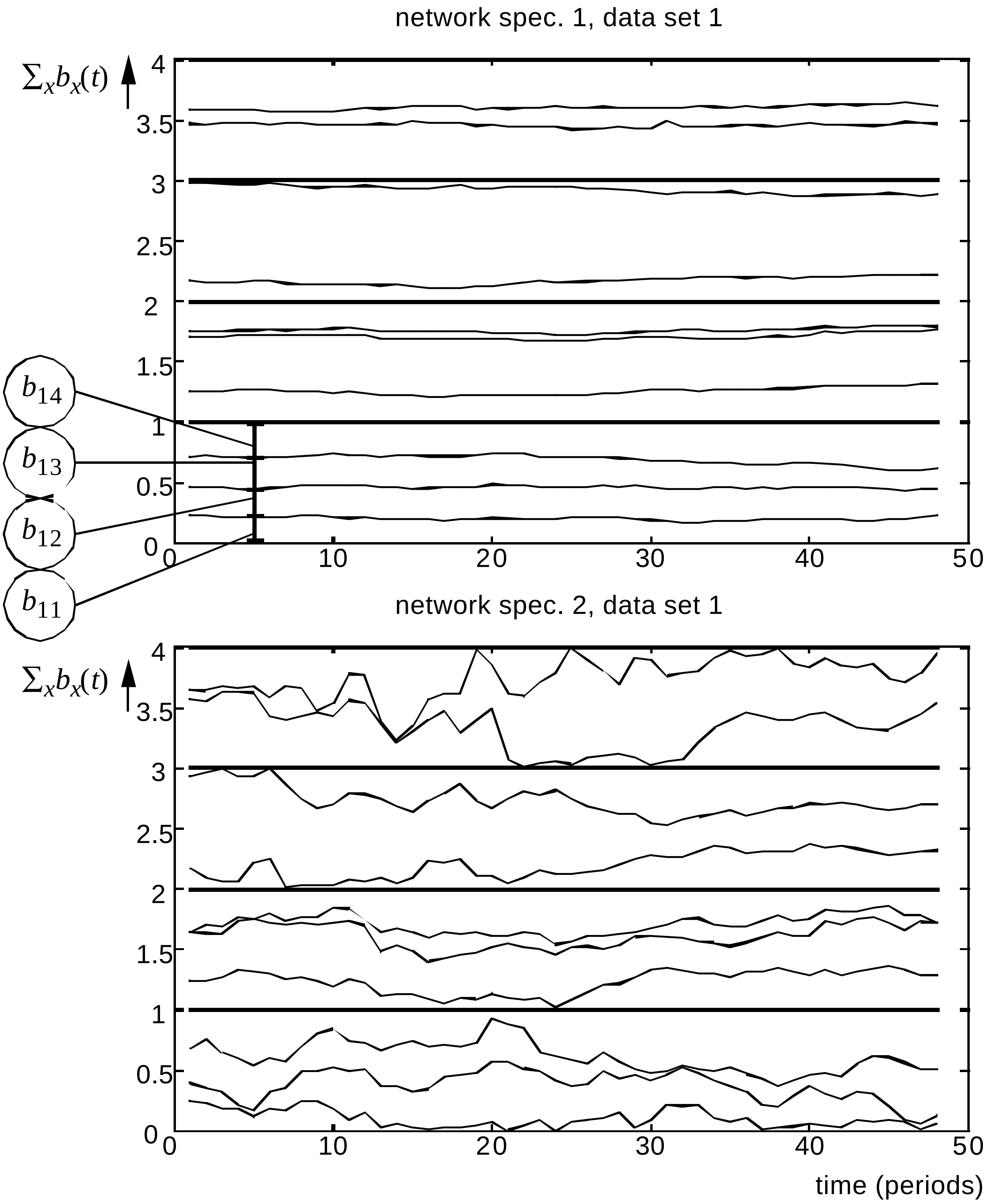


Figure 8.2: *Stacked plots of the split probabilities, generated for σ_b^2=0.0001 (above) and σ_b^2=0.01 (below). The plots contain the split probabilities in the sequence $\{b_{11},b_{12},\ldots b_{44}\}$. The specifications according to which the split probabilities were generated correspond to specification 1 and specification 2 in table 8.2.*

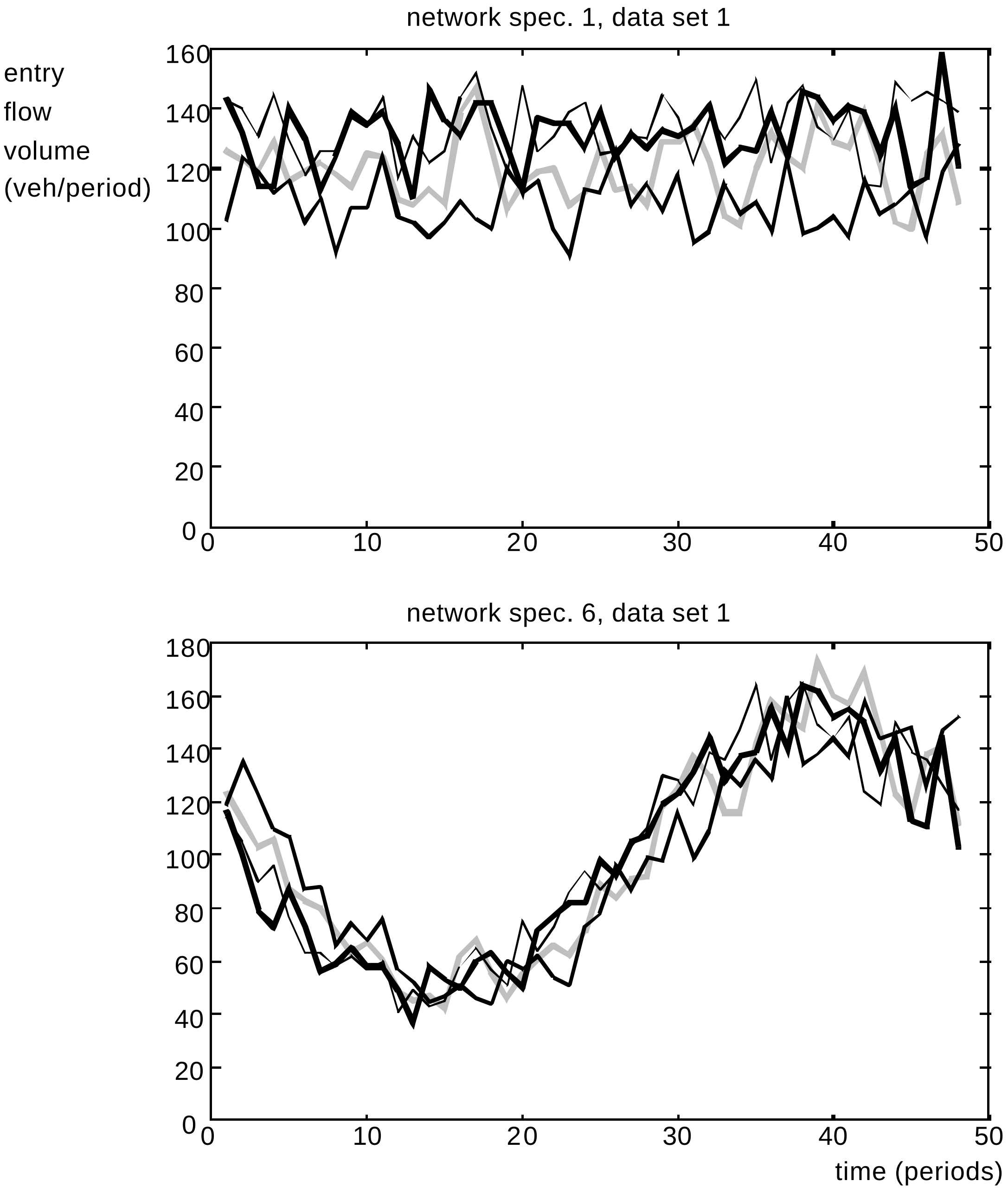


Figure 8.3: *entry volumes generated for* $\bar{\boldsymbol{q}}^{\mathbf{mode}}$*=0(above) and* $\bar{\boldsymbol{q}}^{\mathbf{mode}}$=1 *(below). The specifications according to which these entry flows were generated correspond to specification 1 and specification 6 in table 8.2.*

The network specifications that have been used to generate the testdata are listed in table 8.2. The numbers have been specified bearing in mind a period length of ten minutes. An average entry flow rate $\bar{q}$=100 hence corresponds with 600 vehicles per hour, and 48 periods correspond with a total duration of 8 hours. A rate of change corresponding to $\sigma_b{}^2$=0.0001 corresponds with an hourly rate of change of 0.0006 which is a conservative, but realistic value in view of the upper bound 0.0013 that was derived from the empirical data that were available (see section 6.2). A rate of change corresponding to $\sigma_b{}^2$=0.01 per ten minutes should be considered as a high value in view of the empirical data.

To reduce the influence of random effects, for each network specification ten data sets are generated. By means of an example, figure 8.1 shows the ten networks that were generated for specification 9. Each estimation method will be applied to all these ten sets after which the errors of estimation will be averaged.

The seeds with which the random generators are initialized only depend on the simulation number, not on the network specification. Therefore if one element in the network specification is changed, this effects only one aspect of the generated data. For example increasing the mean entry flow rate $\bar{q}$ or changing the mode of operation $\bar{q}^{\text{mode}}$ does not affect the split proportions that are generated. Moreover, changes in parameters result in changes in the simulated data according to a *continuous* mapping. This is known as the *variance reduction* technique.

Networks two and above are identical to specification one except for one property per network variant, which is printed italic in table 8.2. This is to gain insight into the relation between system properties and estimation accuracy. The findings on this subject are summarized in tables 8.5-8.8 in a later section.

8.2.2 Solution algorithm alternatives

Split ratio methods can be divided into four main categories:

- Least Squares (LS)
- Inequality Constrained Least Squares (ICLS)
- Fully Constrained Least Squares (FCLS)
- Bayesian updating (BU)

Within these categories different variants have been discussed in earlier chapters. Table 8.3 describes for each category which options can be specified, referring to earlier chapters if necessary.

It is not possible nor useful to test every combination of options. Therefore in each category a default method is defined and a number of variants on this method are tested. In each variant only one parameter is changed. In table 8.4 the resulting methods have been listed. Every method has been given a number, that will be referred to when the estimation results are presented. For each method, the options that deviate from the standard method in the category are printed italic. The default methods are method 10, 20, 30 and 47.

Table 8.3: Solution algorithms
-description of options-

Category	**Description**
Options	
Least Squares (LS)	The (discounted) LS solution corresponds to problem (3.10). This solution might not satisfy the inequality and equality constraints (2.27) and (2.28). In the implementation the computed solution is forced to satisfy the inequality constraints by setting: $\overline{b}(t)=\min(\mathbf{1},\max(\ \mathbf{0},\overline{b}'(t)\ ))$ (8.7) where $\overline{b}'(t)$ is the computed LS solution.
a_1	The method has one design parameter to be specified which is the discounting factor λ (see section 3.3.1). The optimal choice of λ is a monotone decreasing function of the rate of change parameter $\sigma_b{}^2$ that was used during the generation of the testdata, see (8.1). Moreover $\sigma_b{}^2=0$ should imply λ=1. Other than that, very little is known about the optimal value for λ, and λ should be determined by experimenting. In the tests described in this chapter the experiments are limited to determining the value for a_1 that gives the best results in combination with the following formula: $\lambda=1-(\sigma_b{}^2)^{a_1}$ (8.8)
Inequality Constrained Least Squares (ICLS)	This solution follows from solving problem (3.11).
a_1	A discounting factor λ is specified by choosing parameter a_1 and using (8.8).
solution algorithm: PCG, IS	For the solution of the constrained minimization problem an algorithm must be chosen from two alternatives. The first algorithm is the *projected conjugate gradient* (PCG) algorithm that was described in section A.4 in appendix A. This algorithm is designed to compute the exact minimum. The second algorithm is the *iterative solving* (IS) algorithm that was described in section A.5 in appendix A. This algorithm is CPU time efficient, but may produce an inexact solution if the unconstrained solution violates multiple inequality constraints.
Fully Constrained Least Squares (FCLS)	This solution corresponds to problem (3.13)
a_1	This parameter defines the discounting factor according to (8.8)

Table 8.3: Solution algorithms
-description of options-

Category	**Description**
Options	
solution algorithm: PCG, IS	Either one of these minimization procedures can be used.
Bayesian updating (BU)	This method is defined recursively with equations (4.16) and (4.19).
a_2	The matrix Q_t in (4.16) is specified with $Q_t=a_2\sigma_b^{2}I$ (8.9) The choice a_2=1 would be consistent with (8.1). However, by specifying alternative values for a_2 the sensitivity of the method for Q_t can be tested.
covariance matrix: PEBA DBA U ALF DPEBA	In chapter 5, expressions for the spatial correlation between link-flow observations have been derived, resulting in various approximations for the covariance matrix R_t that is used in (4.16). For the purpose of testing and sensitivity analysis, the following methods have been implemented: -a- the Point Estimate Based Approximation (PEBA) of R_t, see (5.26) -b- the Distribution Based Approximation (DBA) of R_t, see (5.36) -c- the Diagonalized Point Estimate Based Approximation (DPEBA). This option is equal to -a-, only this time the covariances are ignored. Usage of the DPEBA matrix comes down to ignoring the spatial correlations between measurements. The theory of chapter 5 is used, but only to determine the variance of the observation error. -d- the Average Link-Flow (ALF), see (5.9): $R_t=\mathrm{diag}(\bar{y})$, (8.10) where $\bar{y}$ is the average observed link-flow. This method does not use any of the new results of chapter 5, and can be seen as a first step in an experimental process to find a satisfactory covariance matrix. -e- Unity (U), see (5.8): $R_t=I$ (8.11) This choice can be seen as an alternative to -d-. Instead of specifying the error variance proportional to the average observed value, the error variance is assumed to be equal for all observations.

Table 8.3: Solution algorithms
-description of options-

Category	Description
Options	
a_3, a_4	If R_t is specified using either strategy -a-, -b-, or -c-, covariance matrices need to be specified for the observation errors of the entrance volumes and link volumes. These matrices are specified via the parameters a_4 and a_5, using the following equation: $\Phi = a_3 \sigma_q{}^2 I$ (8.12) $\Theta = a_4 \sigma_y{}^2 I$ (8.13) Specifying a_3=a_4=1 would be consistent with the way the testdata are generated.
recursive constr. (RC)	The theory developed in chapter 4 suggests that Recursive Constraining (RC), see equation (3.22), should not be applied. The RC option has been implemented to check this theory.
true covariance (TC)	If R_t is specified using strategy -a- the possibility exist to use the true split probabilities rather than a point estimate. The behaviour of the method on this point is determined by the option *True Covariances* (TC). If this option is selected then the true split probabilities that were generated during the simulation, are used while determining the covariance matrix. This option can be set to find the best possible specification of R_t. Running the filter with this option therefore produces a lower limit for the estimation error that can be reached with the BU method.
$u(t)$	The steering parameter $u(t)$ is set to zero for all variants. This is consistent with the way the testdata were generated. In the chapter that describes the experiments with empirical data, this parameter is used to express historical experience.
post processing: MAP-IS MAP-PCG SE-AM SE-RM	In chapter 4 a number of ways are described to perform the post-processing step. These are: • Computation of the Maximum APosteriori solution using the method of Iterative Solving (MAP-IS), see sections 4.5.1 and A.5. • Computation of the Maximum APosteriori solution using the Projected Conjugate Gradient Method (MAP-PCG), see sections 4.5.1 and A.4. • Computation of the Subjective Expectation using the Approximated Mean (SE-AM), see section 4.5.2, equation (4.24). • Computation of the Subjective Expectation using the fRandomized Mean (SE-RM), see section 4.5.2, equation (4.32) and lemma 4.1.

Table 8.4: Solution algorithms
- parameter values-

category	method	a_1	solution algorithm	a_2	variance-covariance	a_3	a_4	recursive constraining	true covariance	post processing
LS	10	1	IS							
ICLS	20	1	PCG							
	21	1	*IS*							
FCLS	30	1	PCG							
	31	1	*IS*							
	32	*0.5*	PCG							
	33	*2*	PCG							
BU	40			1	*PEBA*	1	1			*MAP-PCG*
	41			1	*PEBA*	1	1			*MAP-IS*
	42			1	*PEBA*	1	1			*SE-AM*
	43			1	*PEBA*	1	1			SE-RM
	44			1	*DBA*	1	1			SE-RM
	45			1	*U*					SE-RM
	46			1	*ALF*					SE-RM
	47			1	DPEBA	1	1			SE-RM
	48			1	DPEBA	*0*	1			SE-RM
	49			1	DPEBA	1	*0*			SE-RM
	50			*0*	DPEBA	1	1			SE-RM
	51			*0.1*	DPEBA	1	1			SE-RM
	52			*10*	DPEBA	1	1			SE-RM
	53			1	DPEBA	1	1	*RC*		SE-RM
	54			1	PEBA	1	1		*TC*	SE-RM
	55			1	*ALF*			*RC*		*MAP-PCG*
	56			1	*ALF*					*MAP-PCG*
	57			1	*U*					*MAP-PCG*

8.2.3 Evaluation criteria

The testdata that have been generated represent a *completely observed EE-table*: not only are the observations available from which the split parameters can be estimated, but also can the estimation result be compared with the true splitprobabilities and EE-flows. This enables the evaluation of the following two criteria:

$$\mathrm{RMSE}^{\mathrm{EEflow}}(t) = \sqrt{\frac{\sum_{i,j} \left(q_i(t)\bar{b}_{ij}(t) - f_{ij}(t)\right)^2}{mn}} \tag{8.14}$$

$$\mathrm{RMSE}^{\mathrm{split}}(t) = \sqrt{\frac{\sum_{i,j} \left(\bar{b}_{ij}(t) - b_{ij}(t)\right)^2}{mn}} \tag{8.15}$$

Criterion (8.15) has a better *transferability* since this criterion does not depend on the mean entry flow rate. On the other hand criterion (8.14) relates more directly to the error in EE-flow prediction, which is the quantity we are interested in. The experiments however show consistency between the values of both criteria. The results are therefore discussed using criterion (8.15). The outcomes of criterions (8.14) and (8.15) have been summarized in tables 8.6 and 8.5 respectively.

In view of testing the methods with empirical data, for which no completely observed EE-table is available, a third evaluation criterion is proposed:

$$\mathrm{RMSE}^{\mathrm{linkflow}}(t) = \sqrt{\frac{\sum_{k \in K} \left(\sum_{i,j} q_i(t)\bar{b}_{ij}(t-1)\tau_{ijk} - y_k(t)\right)^2}{K\#}} \tag{8.16}$$

with: K set of reference locations, possibly a subset of all observations,

$K\#$ number of reference locations.

τ_{ijk} assignment map, $\tau_{ijk}=1$ if EE-pair ij contributes to observation k, and zero otherwise.

Note that the data that are used to evaluate $\bar{b}(t\text{-}1)$, $q(t)$ and $y(t)$, are not used to compute $\bar{b}(t\text{-}1)$. Criterion (8.16) isreferred to as the *link-flow error criterion*. In order to test the usefulness of this criterion, the outcomes are compared with those of the other criteria. Table 8.7 contains the outcomes of the link-flow error criterion.

8.3 Results

The results that are presented in this section are based on the datasets, estimation methods and evaluation criteria that were described in sections 8.2.1-8.2.3. In the text, network specifications and methods are briefly characterized, and referenced with a number that can be found in tables 8.2 and 8.4 respectively. How the various datasets and methods relate to the theory that was described in this thesis can subsequently be looked up in tables 8.1 and 8.3.

Ten networks are generated with every network specification. Therefore every combination of network specification and method specification is evaluated ten times. The results of these evaluations are averaged, and can be displayed in a graph. Comparison of these graphs gives insight in the relative performance of the various methods. A computer program has been implemented that makes inspection of arbitrary combinations of graphs possible. Some illustrative combinations are plotted in figures 8.4-8.6, and are discussed below.

The performance of other combinations can be inspected in less detail through table 8.5. This table contains the average value of the last forty periods of the graph, i.e.:

$$\sum_{\text{network}=1}^{10} \sum_{t=9}^{T} \text{RMSE}_{\text{network}}^{\text{split}}(t) / (10 \cdot (T-8)) \tag{8.17}$$

The first eight periods have been ignored in order to get a better view of errors that would occur in a round the clock operation of the methods.

In a similar way criteria (8.15) and (8.16) have been summarized in tables 8.6 and 8.7. The average computation time in CPU seconds per time period is shown for each combination of method and network specification in table 8.8.

Parameter optimization methods (LS, ICLS, FCLS)

Method 10 (least squares), 20 (inequality least squares), and 30 (fully constrained least squares) are parameter optimization methods with different levels of constraining the parameter space. Comparison of these methods, see figure 8.4a, was expected to show an improved performance when extra constraints are imposed. This improvement occurs if the inequality constraints are imposed, but is not followed by another improvement when the equality constraints are imposed. Presumably the explanation for this is that the parameter optimization methods reach the best value for criterion (8.15) by systematically underestimating the split parameters. Imposing the equality constraints (method 30) prevents the method from reaching this biased solution. This issue will also be discussed in the next chapter. Figure 9.6 illustrates this phenomenon.

The difference between methods 21 and 31 on the one hand and methods 20 and 30 on the other hand is that the latter two methods use an exact solution algorithm (projected conjugate gradients) while the first two methods use an inexact algorithm (iterative solving). Comparing these methods will tell us whether or not the use of a suboptimal optimization algorithm has significant negative influence on the performance of these parameter optimization methods. Figure 8.4b shows that this is the case, although the effect fades away as a sufficient number of observations is available. Table 8.8 shows that the use of the IS method is a factor 10 more efficient in terms of CPU time. However the table also shows that CPU requirements will hardly be a constraint for any of the methods that are proposed, given the sizes of the networks that are considered.

Method 32 and 33 are equal to method 30, except for the choice of the discounting parameter a_1. Comparison of method 30, 32, and 33, see figure 8.4c, shows that the parameter optimization methods are barely sensitive to this value.

Method 57 is a BU method where the MAP postprocessing is applied and the covariance matrix is set to the identity matrix. Figure 8.4d confirms the claim made in section 4.5.1 that the FCLS method is a special case of a BU method.

Given the results presented in figure 8.4, methods 20 and 30 were selected as representatives of the parameter optimization methods that will be used for further comparisons.

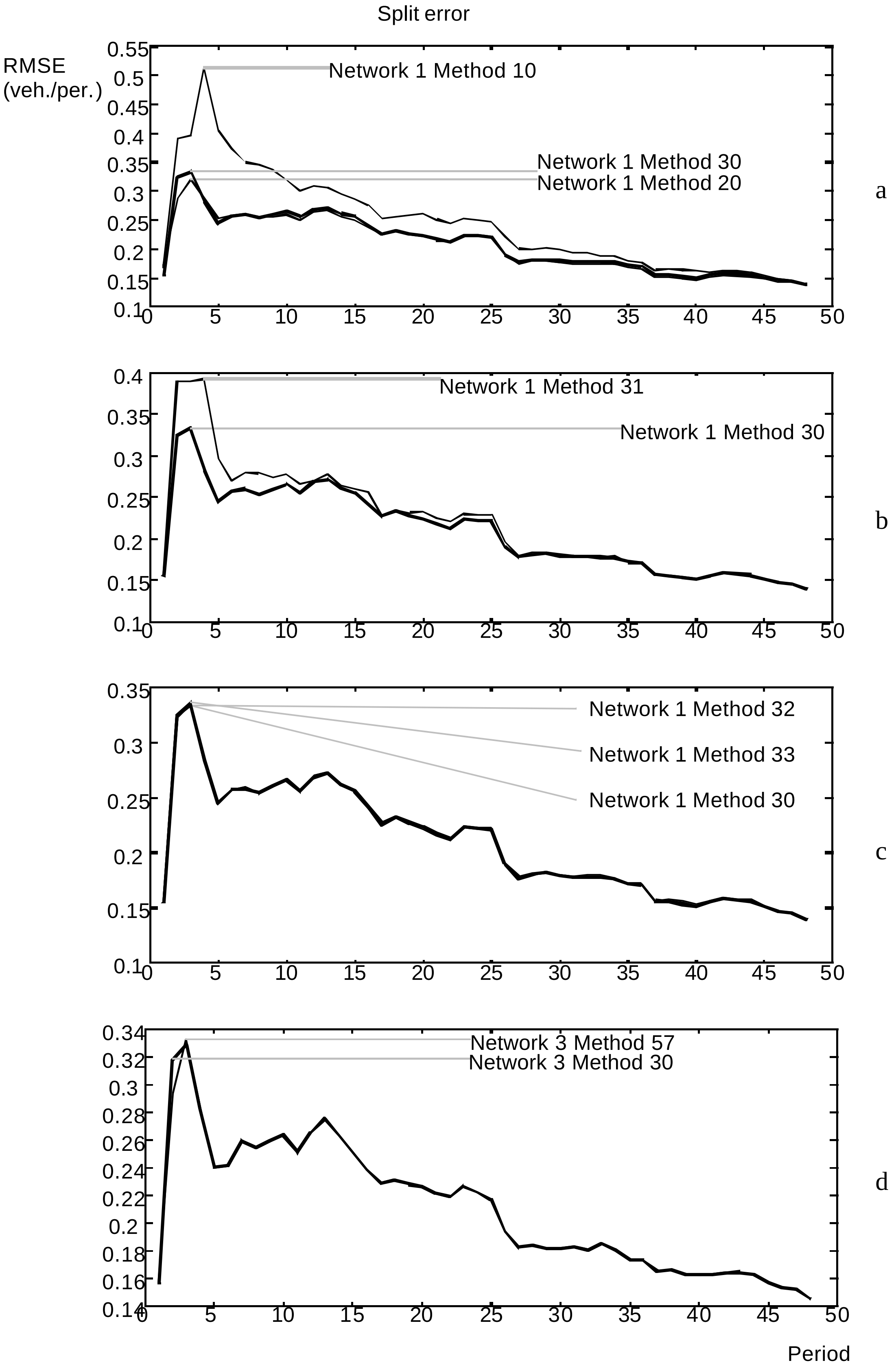


Figure 8.4: Comparison of parameter optimization methods. On the y-axis: average RMSE corresponding to the estimates of the split probability, see (8.15)

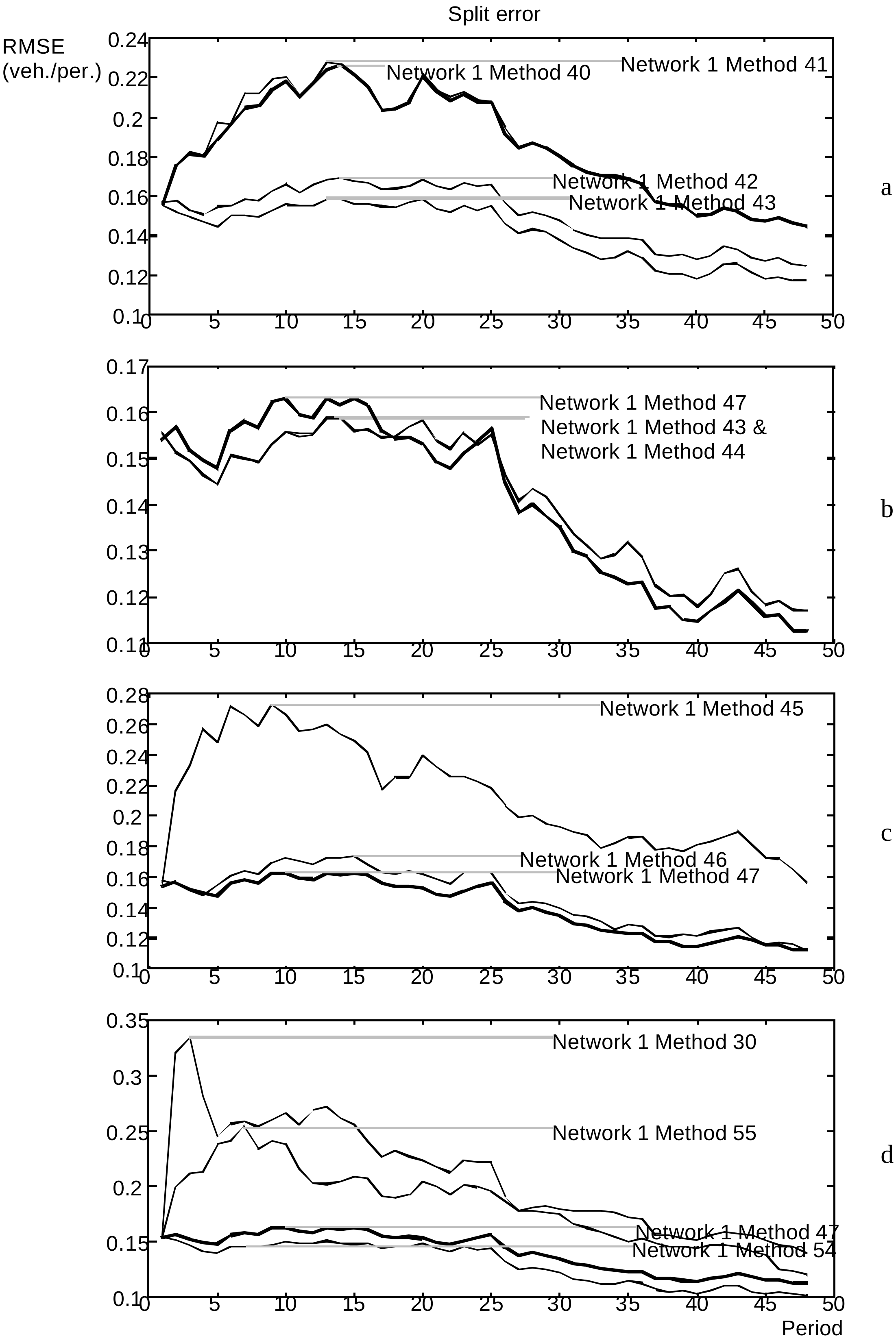


Figure 8.5: Comparison of bayesian updating methods, -I-

Bayesian updating (BU)

Method 40, 41, 42 and 43 are identical BU methods, except for postprocessing algorithm that is used. Clearly the methods that evaluate approximations of the subjective expectation (SE), methods 42 and 43, give the lowest error of estimation, see figure 8.5a. This is consistent with theory. Method 42 uses the approximated mean while method 43 uses the randomized mean. The randomized mean gives slightly better results than the approximated mean, at the cost of an increase of the computation time with a factor of twelve, as can be seen in table 8.8. The first part of the curves of methods 42 and 43 coincide. This should be subscribed to the fact that the randomized mean has not been computed if the probability of sampling a feasible outcome is too low, a situation that apparently occurred during the first few periods that the filter was running. For these cases the approximated mean has been used instead. Given the better performance of the postprocessing method based on SE combined with randomized mean (SE-RM) (method 43), which also shows the method is applied to other networks (see table 8.5), this postprocessing method was selected for further comparisons and sensitivity analysis.

As far as the MAP postprocessing methods are concerned, figure 8.5a shows that the use of a suboptimal optimization method (method 41) is at the cost of the performance, but that the effects are minor and stay limited to the first few periods.

Figure 8.5b shows the performance of three method variants that are identical, except for the covariance matrix that is used. Method 43 uses the point estimated based covariance matrix (PEBA), while method 44 uses the more complex distribution based method (DBA). A difference in performance between both methods can hardly be observed. Method 47 uses only the diagonal elements of the point estimate based approximation (DPEBA) of the covariance matrix. This method seems to have a better convergence and a slightly worse initial response. The overall averaged error however is equal to that of method 43 or 44, also when the method is applied to other networks, see table 8.5. Although it was not expected in advance that using the DPEBA covariance matrix would lead to equally good results as those obtained with the PEBA or DBA matrix, this needs not necessarily be conflicting with theory: the result can be subscribed to errors in the point estimates of the split probabilities from which both matrices are derived. Another possible explanation is that much of the effect that the use of the covariances would have had, is already reached by explicitly imposing the equality constraints.

Given the comparable results, the DPEBA option was chosen to be a part of the standard method, as it is expected that this method is less sensitive to specification errors then the others.

In figure 8.5c the DPEBA method of choosing a covariance matrix for the observation error is compared with two very simple strategies, the first of which is choosing the identity matrix (method 45) and the second of which is choosing a diagonal matrix containing the average link-flow (ALF) volumes (method 46). The latter strategy seems to work nearly as well as method 47. Since method 46 is far more easy to implement and does not need the specification of covariance matrices for the physical observation errors in the entry volumes and link volumes, Φ and Θ respectively, method 46 might be preferable in practice.

Figure 8.5d relates the performance of method 47 to that of alternative methods. Method 30 represents the FCLS method. This method was quoted by *Cremer and Keller (1987)* as the best available method. Applied to the networks generated according to the specification in table 8.2 however, this method fails to give good results, and is outperformed convincingly by the BU method.

Method 55 is a special case of a BU method that coincides with a traditional Kalman filter, it represents a number of choices that seem sensible at first sight. It uses the ALF covariance matrix, and it resets the estimates to the nearest boundary if an inequality constraint is violated, a process that was referred to as *recursive constraining* earlier, see section 3.3.4. It uses the MAP-PCG postprocessing, which implies that it just outputs the parameter that traditionally represents the mean in the Kalman filter, as the recursive constraining ensures that this parameter will always be in the hypercube [0,1] and hence represents the location of the maximum. The error of estimation of method 55 is significantly higher than that of the new BU method, although the method performs better than the FCLS method.

Method 54 uses the true spit-parameters to derive a covariance matrix, and therefore is a method that could never be applied in practice. However, it serves as a means to determine a lower boundary for the error of estimation. Consequently, this method should consistently produce the lowest error of estimation, which can be checked in table 8.5. The difference between the curves of method 47 and method 54 represents the room for further improvement. Figure 8.4d shows that this is small compared to the improvement of the existing methods that already has been reached.

Figure 8.6 contains graphs that relate to the sensitivity analysis for mis-specifying different system properties. Each time, method 47 is taken as a point of departure. This is a BU method that uses the DPEBA covariance matrix and the SE-RM postprocessing routine. The parameters $a_2=a_3=a_4=1$ imply that the specification of the method is consistent with the way the test-data are generated.

In figure 8.6a, method 48 is a variant of 47 that ignores errors in the entry volume counts ($a_3=0$), while method 49 is a variant of 47 that ignores errors in observations of other link volumes ($a_4=0$). Both mis-specifications lead to slightly higher errors but reveal no particularly high sensitivity to this kind of mis-specification.

In figure 8.6b, method 51 underestimates the change in the split probabilities with a factor 10 ($a_2=0.1$), while method 52 represents the assumption that the split probabilities change ten times faster then in reality ($a_2=10$). At the rate of change of network specification one, difference in performance between method 47, 51, and 52 can hardly be observed. For network specification 2 in which a 100 times higher rate of change was specified, mis-specification of this rate does lead to a slightly higher error, see table 8.5.

Figures 8.6c and 8.6d evaluate the influence of applying the recursive constraining option. Method 53 is a variant of method 47 that applies recursive constraining. Likewise method 56 is a variant of the traditional Kalman method mentioned earlier, where method 55 uses recursive constraining and method 56 does not. It turns out that the influence of the RC option largely depends on the network specification. Figure 8.6c shows a network with a high rate of change. For this network, specification of the RC option actually leads to better results. Figure 8.6d shows a network for which the state does not change. For this network the specification of the RC option works out very wrong. This convincingly confirms the theory of section 4.3. A practical guideline that follows from this results is that the specification of the RC option should not be combined with specifying a zero rate of change, even if the state is known to be constant.

Comparing evaluation criteria

In the next chapter the methods will be evaluated using empirical data that were collected on the Amsterdam beltway during three weeks. These data consist of observations of link-flows, but do not contain any direct observations of EE-flows. To create the possibility to evaluate the estimation methods, criterion (8.16) was proposed. It is hoped that this criterion can be used as a measure of relative performance. To check whether this is true, the outcomes

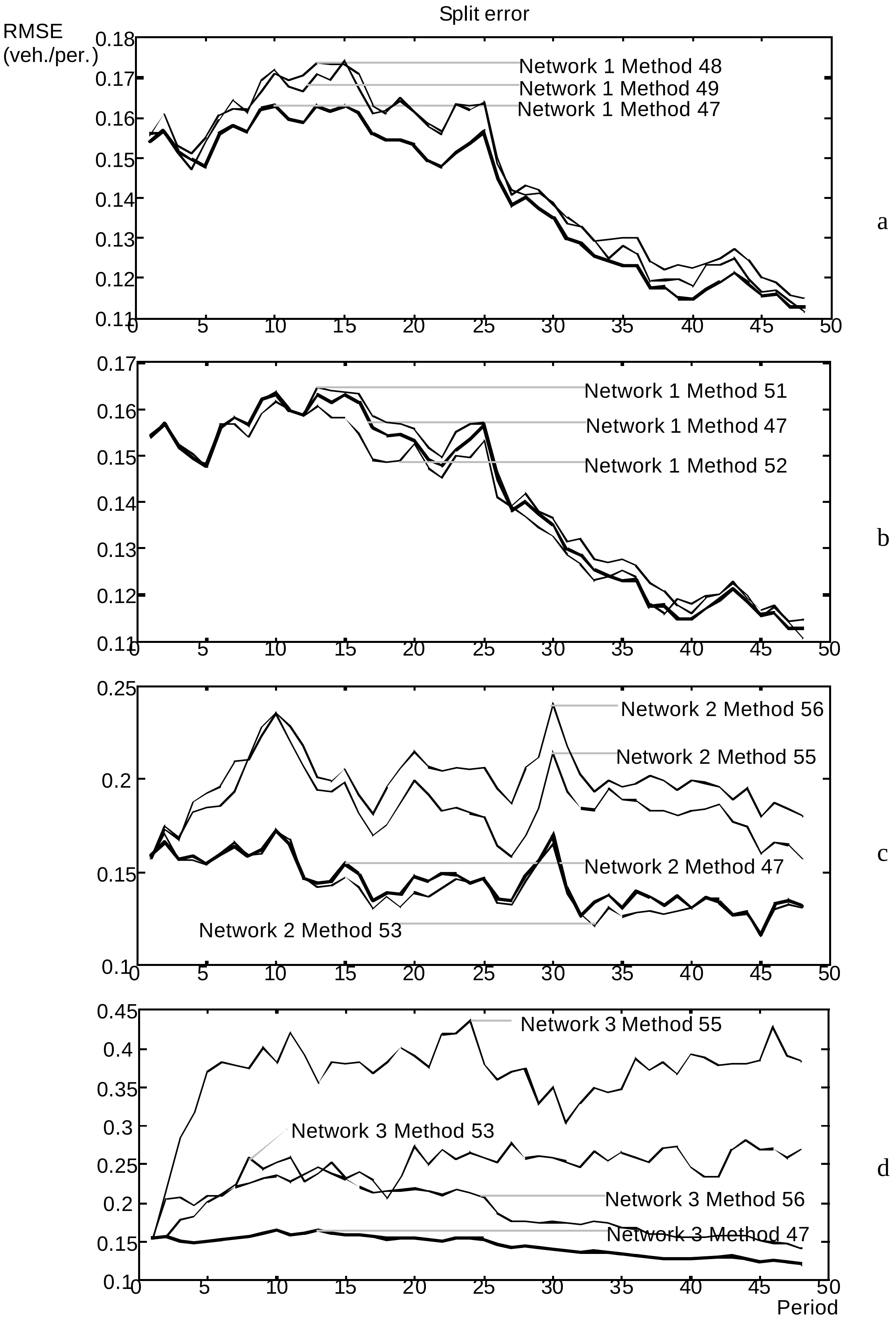


Figure 8.6: Comparison of bayesian updating methods, -II-

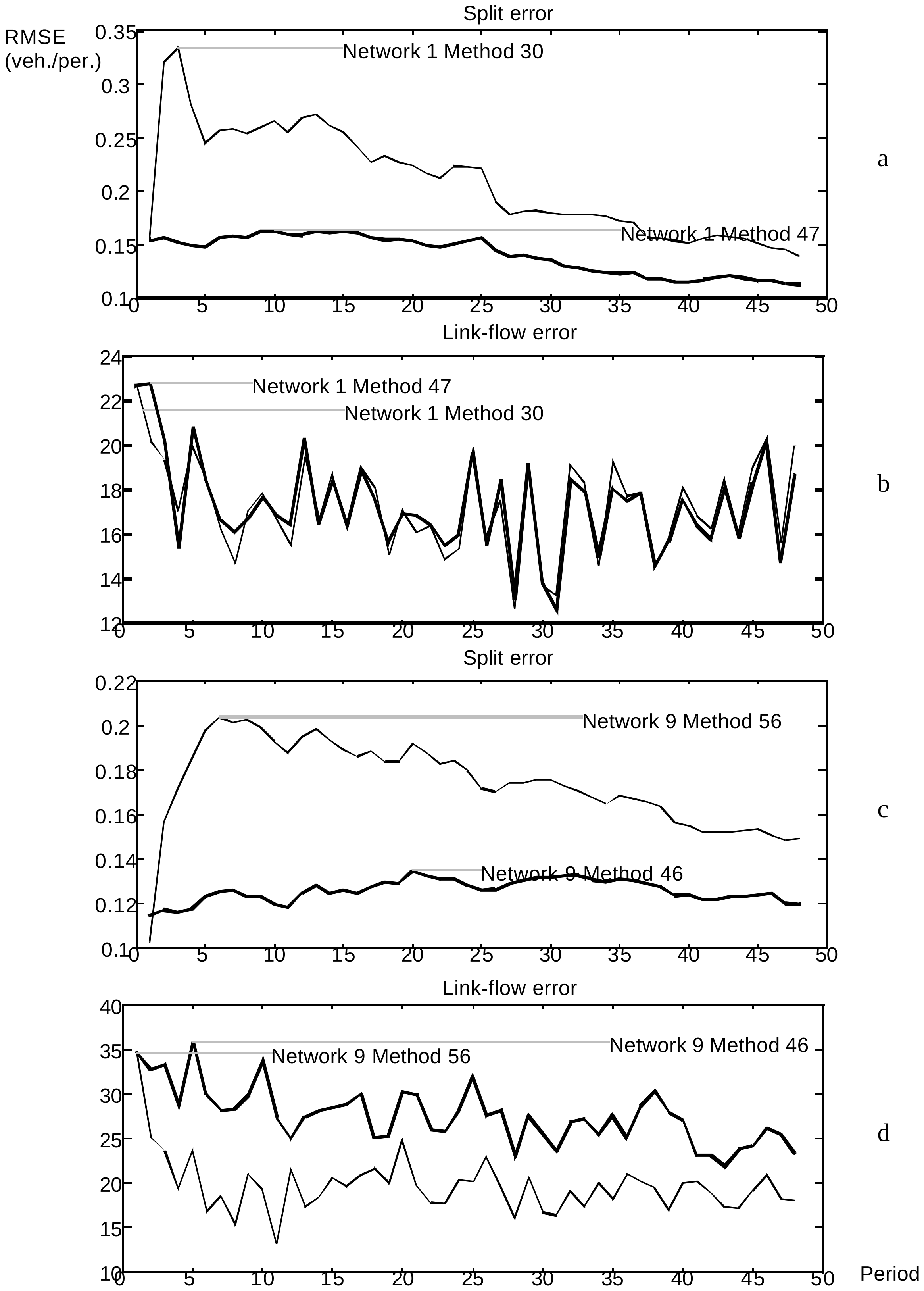


Figure 8.7: *Plot a&b: Comparison of method 30 and method 47, the first one using the split-error criterion and the second one using the link-flow error criterion. Plot c&d: Comparison of method 46 and 56.*

of all criteria used in this chapter have been ranked from 1 to 25, where 1 corresponds with the ‘best’ outcome, and 25 corresponds with the ‘worst’ outcome. This ranking is shown in table 8.9 for the split error, EE-flow error, and link-flow error, separated by comma’s. Table 8.9 shows that the rankings that originate from the split error and the EE-flow error are largely consistent, but that the rankings that originate from the link-flow errors deviate. In particular the link-flow error criterion seems to favour the parameter optimization methods (method 10-33), and the BU methods that use MAP as a part of the postprocessing (methods 40-41 and 55-57). The performance of methods 46 and 47 is underestimated in a systematic way. An illustrative example of this is given in figure 8.7a&b.

Another illustrative comparison is that between method 46 and 56. This comparison is of special interest because internally both methods work with the same subjective probability distribution. The difference in performance should therefore be entirely explained by the difference in the postprocessing algorithm that is applied. Comparing the performance of both methods according to the link-flow error criterion (see table 8.7), method 56 scores slightly better for network 1-8 and much better for network 9 (see figure 8.7d). According to the split error criterion (see table 8.5), the score of method 56 is very poor compared to that of method 46 for all networks (see figure 8.7c to compare the results for network 9).

Figure 8.7d is a problematic plot since in theory the score of method 46 according to the link-flow error criterion should be equal or better under *all* circumstances. This can be seen as follows:

It is commonly known that for any scalar or vector valued random variable y, $\mu \equiv E[y]$ defines a minimum variance estimator in the sense that μ minimises $E[(y-\mu)'.(y-\mu)]$, regardless of the distribution of y. Given the expectation of the split probabilities based on past observations, $\bar{b}_{ij}$, and the observed on ramp volumes q, the expected link volumes at the reference locations are given by (using the symbols of equation 8.16.):

$$E[y_k] = \sum_{i,j} q_i(t)\bar{b}_{ij}\tau_{ijk} \tag{8.18}$$

see section 2.4.2 for more details. Since method 46 produces the subjective expectation (SE) one would expect this estimate to minimize criterion (8.16) relative to all other methods.

Given this discrepancy between theory and experiment, the mechanism leading to figures 8.7c&d has been further investigated. A series of experiments was done comparing method 46 and 56 under different circumstances. It turns out that all cases where method 46 scores poorly on the link-flow error compared with method 56 coincide with a failure to compute the randomized mean. As was mentioned in section 4.5.2 the randomized mean can only be evaluated if the probability of sampling a valid solution is sufficiently high. In other cases a suboptimal procedure has been used, see section 4.5.2. It now appears that although the use of this procedure leads to only slightly higher split-errors, the score on the link-flow error criterion deteriorates considerably as a result of this procedure. The reason that this effect can be most clearly observed in combination with network specification 9 is that this specification includes more origins and destinations than the other specifications. Therefore the state vector that must be estimated is relatively large which decreases the likelihood of successfully applying the randomized mean procedure.

8.4 Summary of results

The most important findings are summarized once more in figure 8.8, which shows the split errors from table 8.5 averaged over all networks. Relative to the FCLS method that was formerly considered to be the best available method *(Cremer and Keller, 1987),* the error of estimation has decreased by 0.065 units from 0.199 (method 30) to 0.134 (method 43), due to the introduction of the new BU method. This is a reduction in the error of estimation with 31%. Roughly 32% (0.021 units) of this reduction should be subscribed to the use of the 'PEBA' covariance matrix as opposed to the use of the 'Unity' matrix (arrow c in figure 8.8) which was shown to be implicit to the FCLS method (see figure 8.4d). The largest part of the improvement (56%, .036 units) is obtained when the approximated mean (SE-AM) is used instead of the MAP estimate (arrow d). Moreover this adaptation results in computation times faster by a factor ten. Finally, the last 12% (.008 units) of the improvement is reached by computing the randomized mean rather than the approximated mean (arrow e). An extra improvement of 0.011 units (17% of the initial reduction) would still be possible if the true covariance matrix would be available (method 54). By making use of the ALF matrix the results only deteriorate 0.005 units (8% of the initial reduction) relative to method 43 (arrow g + i), while the latter matrix can be derived without using knowledge about the system specification.

8.5 Conclusions

The experiments with synthesized data that have been reported in this chapter have convincingly shown the advantages of the new BU method that was presented in chapter 4 in terms of the error of estimation. An average reduction in the error of estimation of 31% relative to the FCLS method was obtained. As was mentioned in section 4.6 the practical implications of using this method are that, taking the traditional Kalman filter as a point of departure, one refrains from recursive constraining (see section 3.3.4), and one uses a special post-processing routine (see section 4.5). The individual effects of these two adaptations can be observed clearly and result in a decrease of the error of estimation, see respectively arrow 'h' and arrow 'j' in figure 8.8.

This BU method was tested in combination with a variety of covariance matrices for the measurement error. As expected, usage of the matrix that is ideal on theoretical grounds (method 54) leads to the lowest error of estimation. However, this matrix will not be available in practice as exact knowledge of the split probabilities is required for its computation. Within the group of BU methods that use an approximation of the ideal matrix, the sensitivity for the (mis)specification of the covariance matrix of the measurement error appears to be small. No difference in error of estimation of any significance exists between method 43 that uses the point estimate based approximation (PEBA) and method 44 that uses the distribution based (DBA) matrix (see figure 8.5b and tables 8.5-8.7). Moreover, if the theory in chapter 5 is only used to compute the height of the variance but not the covariance (as implied by the DPEBA matrix in method 47), no significant deterioration of the error of estimation can be observed (arrow 'g' in figure 8.8). Therefore, for practical applications this approach is recommended as it is expected that the results obtained with the DPEBA matrix are less susceptible to misspecifications of system properties than those obtained with the PEBA matrix. An even simpler alternative is to use the ALF covariance matrix, as in method 46. This matrix has the average link-flows on its diagonal. Usage of this matrix results in only a small increase in the error of estimation (arrow 'i' in figure 8.8) and does not require the specification of any system properties, except for the mean of the traffic counts.

All methods have been tested in combination with network specifications that represent only a small variation in the split probabilities. At this level of variation, none of the methods appeared to be particularly sensitive to mis-specification of the rate of change in the state equation.

A final conclusion concerns the usage of the link-flow error criterion 8.16. In absence of sufficient data to evaluate the split error and the EE-flow error criteria, this criterion might be the only one available for the evaluation of EE-estimation errors in practice. The link-flow criterion systematically favours parameter optimization methods (such as LS, ICLS and FCLS), and the BU methods that use the MAP routine for postprocessing over BU methods that use one of the SE routines for postprocessing.

According to the link-flow error criterion the average error (excluding network 9) improves from 17.19 to 16.56 (.63 units) if the new BU method (method 43) is used instead of FCLS (method 30). This means an improvement of only 3.7%, as opposed to an improvement of 31% according to the split error criterion (see above).

Therefore, to make the superiority of the new BU method using SE postprocessing plausible, it is sufficient to show that this method implies a lower link-flow error. However, the experiments in this chapter show that the new method does not necessarily imply a lower link-flow error. Therefore, it might be impossible to demonstrate the superiority of the new BU method in a straightforward manner on the basis of the link-flow error criterion only.

A practical ‘work-around’ for this problem is based on the observation that without any exception, replacing the MAP postprocessing routine with an SE postprocessing routine leads to a significant decrease of the error of estimation according to the split error criterion. Therefore a strategy to demonstrate the superiority in terms of the split error criterion of a BU method using SE postprocessing over another method such as FCLS, is to demonstrate that the variant of the BU method that uses MAP postprocessing outperforms the FCLS method according to the link-flow error criterion.

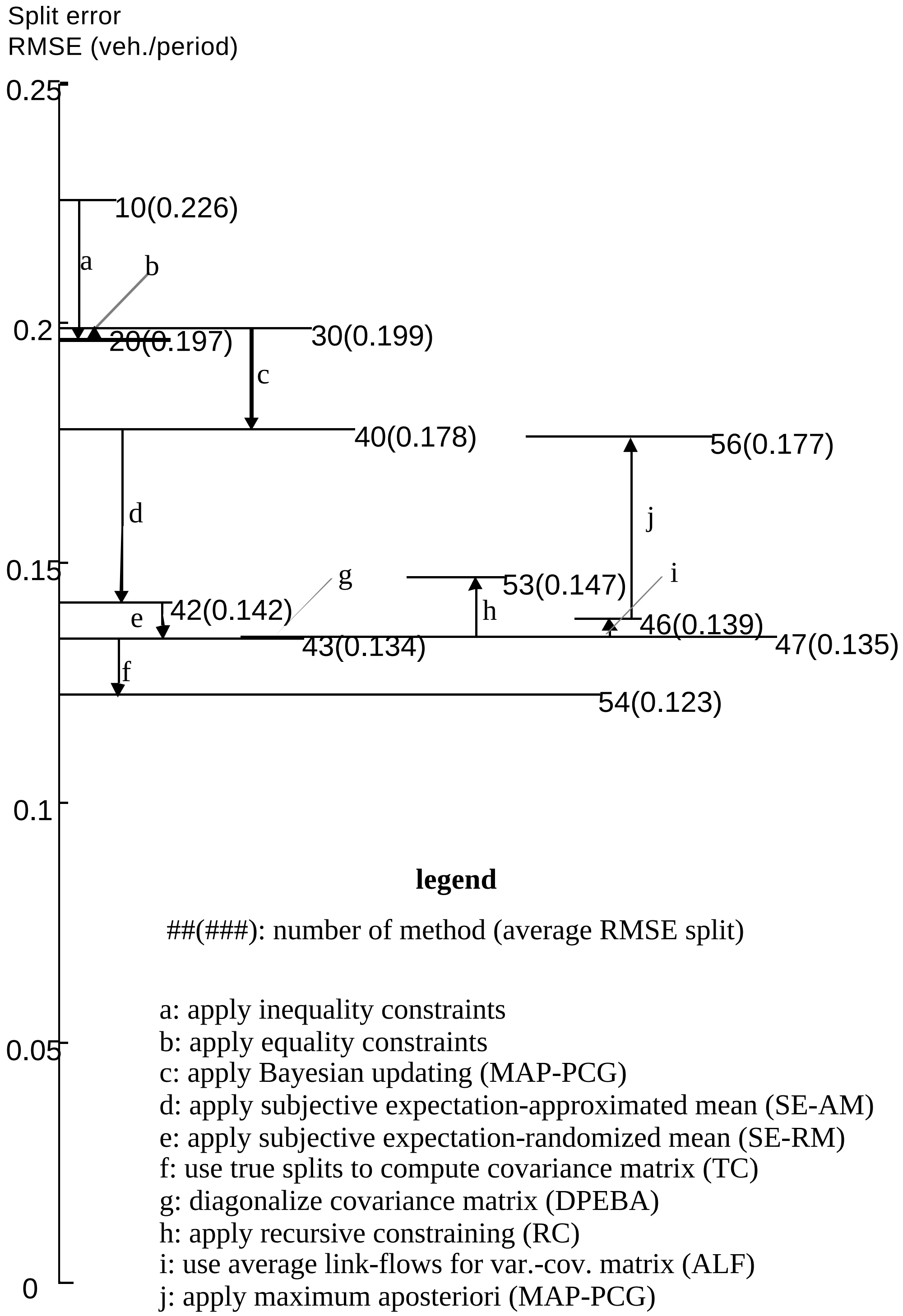


Figure 8.8: Summary of the simulation results. The split error is read from table 8.5, and averaged over all networks. Each arrow represents the change of one option.

Table 8.5: Split Errors
- average over 10 networks of criterion (8.15) -

method		network								
		1	2	3	4	5	6	7	8	9
LS	10	0.218	0.302	0.221	0.217	0.217	0.214	0.231	0.175	0.237
ICLS	20	0.194	0.250	0.197	0.192	0.196	0.188	0.196	0.163	0.193
	21	0.198	0.259	0.200	0.194	0.198	0.192	0.202	0.164	0.199
FCLS	30	0.197	0.252	0.200	0.193	0.198	0.190	0.197	0.168	0.195
	31	0.200	0.261	0.202	0.195	0.200	0.194	0.203	0.169	0.201
	32	0.197	0.253	0.200	0.193	0.199	0.190	0.197	0.169	0.195
	33	0.197	0.253	0.200	0.193	0.198	0.190	0.197	0.168	0.195
BU	40	0.185	0.208	0.191	0.173	0.172	0.173	0.169	0.165	0.166
	41	0.186	0.210	0.192	0.173	0.172	0.174	0.170	0.165	0.167
	42	0.149	0.151	0.154	0.140	0.146	0.145	0.135	0.134	0.122
	43	0.140	0.143	0.144	0.131	0.139	0.135	0.127	0.129	0.121
	44	0.140	0.143	0.144	0.131	0.139	0.135	0.139	0.129	0.121
	45	0.208	0.172	0.203	0.199	0.204	0.208	0.208	0.174	0.183
	46	0.144	0.146	0.151	0.129	0.148	0.142	0.128	0.132	0.127
	47	0.138	0.142	0.143	0.127	0.142	0.137	0.128	0.132	0.122
	48	0.145	0.146	0.151	0.132	0.149	0.144	0.129	0.138	0.128
	49	0.143	0.146	0.150	0.131	0.147	0.144	0.141	0.132	0.128
	50	0.141	0.186	0.143	0.136	0.145	0.140	0.130	0.136	0.125
	51	0.140	0.146	0.143	0.131	0.144	0.139	0.129	0.135	0.125
	52	0.137	0.154	0.143	0.127	0.140	0.137	0.128	0.129	0.117
	53	0.137	0.138	0.255	0.140	0.133	0.138	0.131	0.131	0.121
	54	0.127	0.140	0.131	0.123	0.127	0.123	0.126	0.093	0.114
	55	0.175	0.186	0.379	0.188	0.174	0.187	0.184	0.153	0.190
	56	0.183	0.202	0.189	0.165	0.182	0.173	0.175	0.153	0.173
	57	0.203	0.228	0.200	0.200	0.200	0.206	0.204	0.176	0.189

Table 8.6: EE-Flow Errors
- average over 10 networks of criterion (8.16) (trips/period)-

method		network								
		1	2	3	4	5	6	7	8	9
LS	10	21.49	30.09	21.98	41.52	21.74	19.42	22.28	16.98	25.59
ICLS	20	19.03	25.09	19.61	36.44	19.69	17.22	18.77	15.82	20.79
	21	19.42	25.92	19.86	36.86	19.90	17.56	19.34	15.94	21.57
FCLS	30	19.34	25.22	19.88	36.57	19.93	17.42	18.86	16.25	20.96
	31	19.66	26.06	20.06	37.00	20.06	17.76	19.43	16.39	21.69
	32	19.30	24.96	19.88	36.66	19.94	17.36	18.84	16.34	20.85
	33	19.34	25.30	19.88	36.56	19.93	17.43	18.87	16.25	20.96
BU	40	18.08	20.71	18.73	32.98	17.24	16.08	16.32	15.85	17.73
	41	18.14	20.85	18.89	33.08	17.27	16.14	16.41	15.87	17.83
	42	14.86	15.41	15.46	27.31	14.62	13.68	13.44	13.18	13.49
	43	13.72	14.47	14.26	25.31	13.96	12.63	12.43	12.47	13.46
	44	13.72	14.48	14.23	25.36	13.95	12.61	13.71	12.45	13.47
	45	20.71	17.46	20.39	38.33	20.52	19.19	20.81	16.96	19.87
	46	14.28	14.70	15.11	24.91	14.82	13.18	12.58	13.00	14.09
	47	13.69	14.27	14.45	24.69	14.22	12.81	12.57	13.01	13.64
	48	14.34	14.67	15.14	25.36	14.91	13.38	12.66	13.38	14.22
	49	14.16	14.76	15.09	25.21	14.77	13.34	13.80	13.00	14.17
	50	13.91	18.83	14.45	26.39	14.53	13.11	12.70	13.41	13.99
	51	13.84	14.87	14.45	25.41	14.43	12.97	12.63	13.29	13.90
	52	13.67	15.56	14.45	24.71	14.10	12.92	12.70	12.79	13.09
	53	13.63	13.86	26.68	27.73	13.39	12.99	12.80	12.85	13.52
	54	12.47	14.19	12.98	23.70	12.75	11.36	12.30	9.14	12.63
	55	17.01	18.59	39.17	36.81	17.46	17.10	17.77	14.75	20.22
	56	17.97	20.07	18.85	31.32	18.32	15.78	16.71	14.74	18.70
	57	20.09	22.73	19.88	38.11	20.05	18.84	20.01	16.99	20.06

Table 8.7: Link-flow Errors
- average over 10 networks of criterion (8.17) (vehicles/period)-

method		Network								
		1	2	3	4	5	6	7	8	9
LS	10	29.06	50.19	29.94	50.42	26.39	25.82	34.95	19.02	79.53
ICLS	20	17.14	23.82	16.70	19.72	17.15	17.23	12.67	14.17	19.71
	21	17.17	23.87	16.73	19.74	17.16	17.25	12.69	14.18	19.70
FCLS	30	17.01	23.69	16.56	19.57	16.97	17.06	12.53	14.14	19.52
	31	17.04	23.74	16.58	19.59	16.98	17.08	12.55	14.15	19.51
	32	16.98	21.74	16.56	19.48	16.93	17.02	12.49	14.10	19.48
	33	17.01	24.02	16.56	19.57	16.97	17.06	12.53	14.14	19.52
BU	40	16.66	20.14	16.44	18.34	16.42	16.60	12.24	13.71	18.93
	41	16.65	20.15	16.44	18.34	16.42	16.60	12.24	13.71	18.94
	42	17.58	21.85	17.82	22.26	17.21	17.37	14.15	14.32	25.15
	43	16.66	20.93	16.69	18.54	16.53	16.75	12.82	13.58	25.11
	44	16.66	20.92	16.69	18.54	16.50	16.75	13.15	13.57	25.09
	45	20.21	22.11	18.21	24.57	19.62	19.21	16.74	16.09	24.81
	46	17.03	21.48	16.84	19.03	16.98	16.92	12.72	13.81	26.85
	47	16.90	21.25	16.76	19.09	16.91	16.90	12.81	13.83	27.37
	48	16.98	21.28	16.77	19.11	16.96	16.98	12.82	14.06	26.52
	49	17.01	21.45	16.89	19.01	16.99	16.91	13.18	13.81	26.86
	50	17.10	24.19	16.76	19.82	17.10	17.12	12.99	14.03	27.59
	51	17.01	20.68	16.76	19.27	17.00	17.04	12.88	13.89	27.50
	52	17.18	24.57	16.76	19.76	17.04	17.08	13.21	14.15	27.73
	53	17.53	20.93	29.06	20.85	17.01	17.58	13.08	14.58	29.95
	54	16.60	21.02	16.57	18.49	16.48	16.69	12.72	13.50	24.48
	55	18.40	20.86	71.61	23.55	17.89	18.28	13.50	14.99	31.50
	56	16.82	20.55	16.52	18.61	16.71	16.85	12.25	13.86	19.21
	57	18.64	21.67	16.56	21.11	18.33	18.60	13.67	15.53	21.96

Table 8.8: Average CPU usage per time period (sec/period)

method		Network								
		1	2	3	4	5	6	7	8	9
LS	10	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.13
ICLS	20	1.30	1.30	1.36	1.88	1.45	1.40	1.49	1.30	5.60
	21	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.19
FCLS	30	0.71	0.84	0.71	1.03	0.77	0.75	0.82	0.71	3.90
	31	0.07	0.09	0.07	0.08	0.08	0.08	0.08	0.07	0.31
	32	0.70	0.89	0.71	1.05	0.79	0.77	0.81	0.71	3.89
	33	0.72	0.80	0.72	1.02	0.77	0.74	0.83	0.71	3.98
BU	40	0.47	0.47	0.59	0.54	0.50	0.49	0.52	0.47	1.61
	41	0.16	0.17	0.17	0.17	0.17	0.17	0.16	0.16	0.51
	42	0.10	0.10	0.11	0.11	0.11	0.11	0.10	0.11	0.23
	43	1.20	1.41	1.10	1.14	1.01	0.88	1.16	0.81	1.30
	44	1.14	1.42	1.13	1.15	1.05	0.90	0.89	0.80	1.16
	45	0.57	0.88	0.55	0.58	0.55	0.56	0.54	0.70	1.02
	46	0.98	1.47	0.96	1.10	0.87	0.93	1.07	0.92	1.12
	47	1.17	1.48	1.08	1.09	0.90	1.06	1.17	0.89	1.18
	48	1.16	1.36	1.05	0.99	0.84	0.92	1.14	0.97	1.17
	49	1.09	1.45	1.02	1.09	0.84	0.96	1.04	0.92	1.14
	50	1.04	1.08	1.08	1.11	1.00	0.90	1.04	0.94	1.15
	51	1.07	1.56	1.08	1.14	1.00	0.90	1.04	0.94	1.15
	52	1.12	0.61	1.08	1.26	0.95	1.04	1.05	0.92	1.05
	53	0.97	1.39	0.61	0.84	1.01	0.95	1.15	0.94	1.13
	54	1.17	1.45	1.19	1.13	1.05	0.93	1.26	0.87	1.11
	55	0.59	0.50	0.81	0.68	0.58	0.56	0.62	0.50	2.47
	56	0.55	0.50	0.72	0.61	0.53	0.48	0.53	0.49	1.97
	57	0.64	1.08	0.86	0.84	0.64	0.60	0.65	0.57	2.79

Table 8.9: Rankings of evaluation criteria average over 10 networks: - Split Error, EE-Flow Error, Link-flow Error-

method		Network								
		1	2	3	4	5	6	7	8	9
LS	10	25,25,25	25,25,25	23,23,24	25,25,25	25,25,25	25,25,25	25,25,25	24,24,25	25,25,25
ICLS	20	17,17,17	19,20,20	15,15,12	17,16,16	17,17,19	17,17,18	17,17,8	15,15,18	19,19,9
	21	21,21,18	23,23,21	16,16,13	21,21,17	18,18,20	21,21,19	21,21,9	16,18,19	23,23,8
FCLS	30	19,19,12	20,21,18	18,18,4	19,18,14	19,19,10	19,19,13	19,19,5	20,19,14	22,22,6
	31	22,22,15	24,24,19	21,21,9	22,22,15	23,23,12	22,22,16	22,22,7	22,22,17	24,24,5
	32	20,18,9	22,19,15	18,18,4	20,19,12	21,21,8	18,18,11	18,18,4	21,21,13	20,20,4
	33	18,20,13	21,22,22	18,18,4	18,17,13	20,20,11	20,20,14	20,20,6	19,20,15	21,21,7
BU	40	15,15,3	16,16,1	13,12,1	14,14,1	13,13,1	14,14,1	13,13,2	17,16,4	13,13,1
	41	16,16,2	17,17,2	14,14,2	15,15,2	14,14,2	15,15,2	14,14,1	18,17,5	14,14,2
	42	12,12,21	10,10,16	11,11,21	12,11,22	9,9,21	12,12,20	10,10,23	9,9,20	6,5,15
	43	6,6,5	4,4,7	7,3,11	5,6,5	4,4,5	3,3,4	2,2,13	3,3,3	4,3,14
	44	5,5,4	5,5,6	6,2,10	8,7,4	3,3,4	2,2,5	11,11,18	2,2,2	5,4,13
	45	24,24,24	12,12,17	22,22,22	23,24,24	24,24,24	24,24,24	24,24,24	23,23,24	16,16,12
	46	10,10,14	9,7,13	9,9,19	4,4,8	11,11,13	9,9,9	5,4,11	8,7,6	10,10,17
	47	4,4,7	3,3,10	2,4,14	3,2,9	6,6,7	5,4,7	4,3,12	6,8,8	7,7,19
	48	11,11,8	6,6,11	10,10,18	9,8,10	12,12,9	10,11,10	6,6,14	12,11,12	12,12,16
	49	9,9,10	8,8,12	8,8,20	6,5,7	10,10,14	11,10,8	12,12,19	7,6,7	11,11,18
	50	8,8,16	13,14,23	2,4,14	10,10,19	8,8,18	8,8,17	8,7,16	11,12,11	9,9,21
	51	7,7,11	7,9,4	2,4,14	7,9,11	7,7,15	7,6,12	7,5,15	10,10,10	8,8,20
	52	3,3,19	11,11,24	2,4,14	2,3,18	5,5,17	4,5,15	3,8,20	4,4,16	2,2,22
	53	2,2,20	1,1,8	24,24,23	11,12,20	2,2,16	6,7,21	9,9,17	5,5,21	3,6,23
	54	1,1,1	2,2,9	1,1,8	1,1,3	1,1,3	1,1,3	1,1,10	1,1,1	1,1,11
	55	13,13,22	14,13,5	25,25,25	16,20,23	15,15,22	16,16,22	16,16,21	14,14,22	18,18,24
	56	14,14,6	15,15,3	12,13,3	13,13,6	16,16,6	13,13,6	15,15,3	13,13,9	15,15,3
	57	23,23,23	18,18,14	17,17,7	24,23,21	22,22,23	23,23,23	23,23,22	25,25,23	17,17,10

9. Experiments with Empirical Data

9.1 Introduction

In this chapter the experiments of the previous chapter will be repeated with empirical data. The findings of the previous chapter are used to limit the experiments to the methods that were shown to give good results with simulated data. The parameter optimization methods and the traditional Kalman method are used as a reference. In addition to the variants that were tested in the previous chapter, variants are also tested that use historical data. A complete overview of the methods that are tested is given in section 9.2.

Data have been collected on the Amsterdam beltway for three weeks. A number of practicalities are dealt with in section 9.3. These concern the selection of a study area, checking the data for this area, the conversion of the input data in order to match the moving time coordinate system, and the recovery (if possible) from missing data.

A completely observed EE-table is not available, e.g. the true EE-flows can not be used in the evaluation. Instead the link-flow error criterion that was discussed in the previous chapter is used. The results are presented in section 9.5.

9.2 Solution algorithm alternatives

The results presented in the previous chapter give rise to the selection of BU methods 46 and 47 (see tables 8.3, 8.4) as methods with the best properties for practical use. These methods will therefore be tested with empirical data. As a reference the parameter optimization methods 20 (ICLS) and 30 (FCLS) will be used, as well as the standard Kalman method 55 (presented as a special case of the BU method).

Method 56 is the method that is identical to method 46, except for the fact that it uses the MAP-PCG postprocessing routine rather then the SE-RM postprocessing routine. Therefore, method 56 can be used as a part of the strategy described in section 8.5, to make plausible that method 46 produces more accurate EE-flow estimates than the FCLS method, without using direct observations of EE-flows.

In addition to these comparisons, the use of historical data using the recipe described in chapter 7 is evaluated. This requires the specification of two extra method parameters. The purpose and values of these parameters are described in tables 9.1 and 9.2 that supplement tables 8.3 and 8.4.

Table 9.1: Solution algorithms
-description of options-

Category	**Description**
Options	
Bayesian updating (BU)	see table 8.3.
M	Number of the method in table 8.4 that contains the options and parameter values
α	Historical experience is imposed through a steering parameter $u(t)$. In chapter 6 it was shown that the use of the default model (6.5) implies the following value for $u(t)$ (see equation (6.9)): $u(t) \equiv b^{\mathrm{p},k}(t+1) - b^{\mathrm{p},k}(t) - (1-\alpha)(b(t) - b^{\mathrm{p},k}(t))$ where $b^{\mathrm{p},k}(t)$ represents a historical pattern of type p, updated until and including day k, and α is a design parameter that represents the strength with which the state is pulled back to its historical average. In our examples, only one historical pattern is used, which is the weekday pattern. This pattern has been computed by averaging the estimates over all available weekdays, i.e.: $b^{\mathrm{p},k}(t) \equiv \sum_{d \in \text{ weekdays}} \bar{b}^{d,\,M}(t)$ where $\bar{b}^{d,\,M}(t)$ is the estimated vector of split probabilities, computed with method M.

Little can be said in advance about the best value for α. Therefore three values (1, 0.9 and 0.8) are tested. Both method 46 and method 47 are tested with and without the use of historical patterns. The methods with which these patterns are computed are identical to the methods from which the parameters are used. All resulting method variants are listed in table 9.2.

Table 9.2: Solution algorithms
- parameter values-

category	method	parameter values in table 8.4:	method used to prepare historical data:	pull-back factor (α)
ICLS	20	20		
FCLS	30	30		
BU	43	43		
	46	46		
	47	47		
	55	55		
	56	56		
	61	46	46	1
	62			0.9
	63			0.8
	64	47	47	1
	65			0.9
	66			0.8

9.3 Network and data

For the collection of data the MARE system has been used. MARE is a research facility connected to the national Motorway Traffic Management (MTM) system. Data was collected between the 12 th and the 30 th of April 1994. Figure 9.1 gives an impression of the network that was originally taken into consideration. Each link that is supplied with one or more induction loops is indicated with a number. This number corresponds with the number that is used internally in the MARE system. For each induction loop one minute aggregated volume counts and arithmetic speed averages are stored by the MARE system.

9.3.1 Collection of motorway traffic data in the Netherlands

In the Netherlands the MTM system plays an important role in collecting traffic data. The primary objectives of MTM is the upstream warning for congestion and slow traffic. It is claimed that this leads to a significant decrease in risk of pile ups and incidents (-50%), an increase of flow (+5%) and a decrease of traveltime (-15%) (Source: *AVV, 1994*). Earlier versions of this system are known as MCSS and MCSS+, see *Westerman (1994).*

Motorway sections on which MTM is installed are equipped with dual induction loops on all lanes with intervals of approximately 500 meters. For research purposes a data collection facility exists, the MARE system. MARE enables the storage in a file of traffic data for a number of pre-specified locations. These data are aggregated to one-minute periods, and involve average speed, flow, and a number of status variables, *AVV (1992).* The MARE system is intended for off-line applications, but with some minor technical modifications data can also be obtained on-line, as was demonstrated in the DYNA project *Hague Consulting Group (1994).*

Not all collected data are available at a central computer. Data is aggregated and smoothed

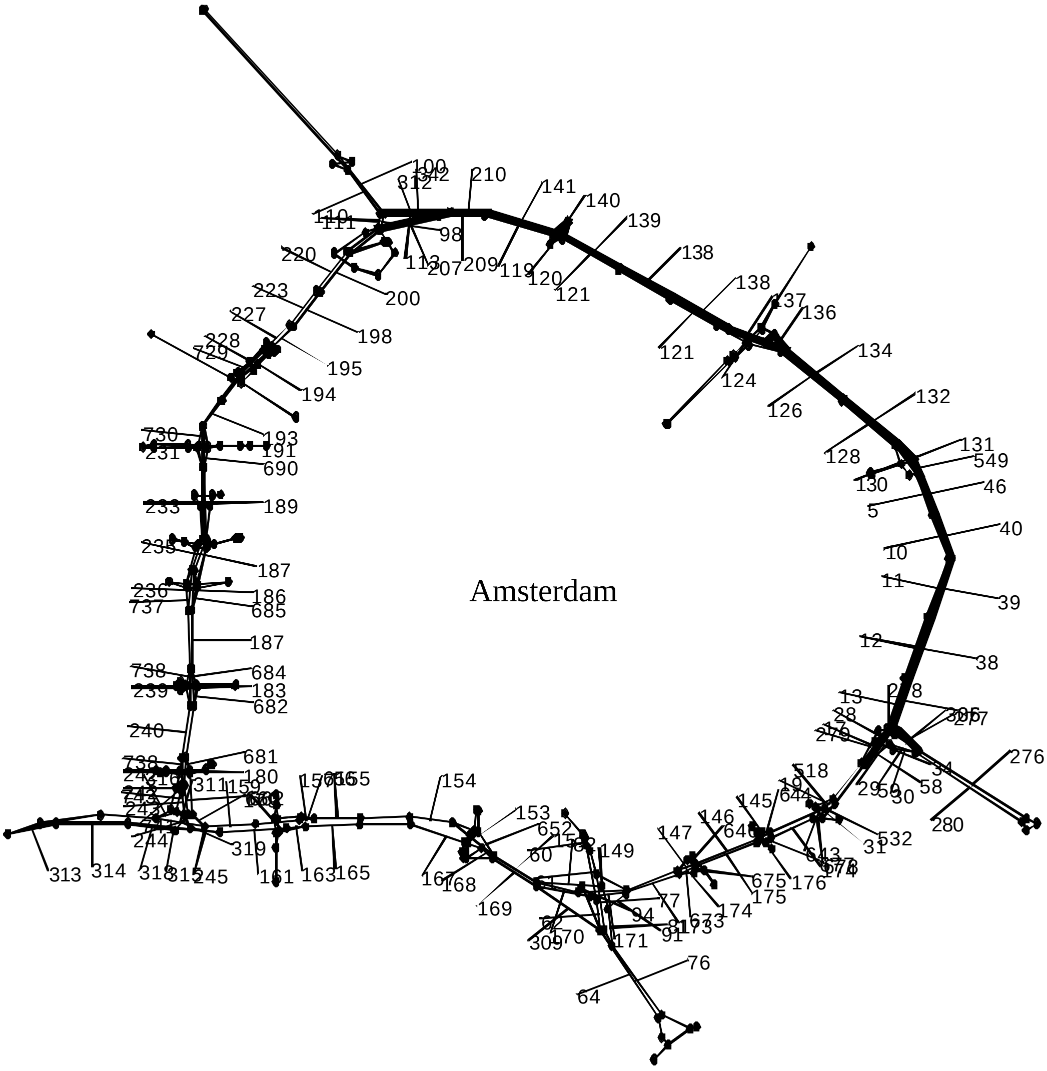


Figure 9.1: Representation of the Amsterdam network. The numbers represent a subset of the induction loop locations. The emphasized links represent the network that was eventually selected for further tests.

in road-side processors (substations) before transmitted to the central computer. Applications that use the MARE data, should take the following properties of the MARE data into account.

- Data are aggregated to one minute periods.
- The data in the road side substations are examined approximately once a minute. The substations have no internal clock. As a consequence the length of the observation periods depends on the exact time of interrogation. Time stamps are logged together with the observations.
- Observed flows are converted to hourly flows and rounded to units of 50 vehicles. This implies that vehicle counts are rounded to units of 5/6 vehicle.
- A minimum of 4 vehicles per minute is ‘observed’ in every substation. For some technical reason early versions of MCSS generated dummy vehicles every 15 seconds, if no real vehicles were observed. This property is preserved in later versions of MCSS+ and MTM and is a constant source of errors, especially at night and on less frequently used links.
- Flow levels are observed through exponential smoothing of ‘gaps’ in the traffic flow (first generation of MCSS substations only). Rather than counting the number of vehicles that traverse a screenline the average gap between vehicles is computed through the mechanism of exponential smoothing. The smoothing factors can vary per location. As this process is irreversible, the exact flows are not known at the central computer.
- Arithmetic averages of observed speeds are stored whereas harmonic averages are needed to compute average travel-times.
- One minute average speeds are approximated using exponential smoothing (first generation of MCSS substations only).
- In general, motorway on- and off-ramps are not explicitly monitored. The detectors of the monitoring system are located on the internal links of the motorway. At an aggregate level on- and off-ramp volumes can be reconstructed from the observations at the adjacent links, if present. Applying this technique to one minute data however, introduces significant errors, regularly leading to negative flows. A factor that makes the problem worse is that the accuracy of induction loop detectors decreases in areas with many lane changes, which is typical for the vicinity of on- and off ramps.

9.3.2 Selection of the data

A prerequisite to apply split ratio methods is that for each on-ramp, the entry volumes can be reconstructed. Preferably the same applies to the off-ramps. Detailed inspection of the network configuration and the locations on which link volumes have been counted has revealed that for many corridors, observations corresponding to one or more on-ramps or off ramps are unavailable, either due to the absence of induction loops or due to apparent errors in the loop-data. A relatively long corridor that is sufficiently monitored is highlighted in figure 9.1 and is shown in detail in figure 9.2. This corridor was used for further analysis.

After selecting the study area the next step is to check the data and to decide which days to use. As a group of comparable days is needed to compute a historical pattern, it was decided not to use data from days in the weekends.

The remaining days were tested for missing data or severe incidents. Data are labelled as ‘missing’ when the detection status that is stored along with the other data indicates that the detection system has been dead or not been working properly.

For inspecting the data a computer program was implemented. This program collects data from various files and then presents a contour plot of the speed as a function of time and location, overlaid with a plot of the missing data, see figure 9.3. The speeds that are presented in

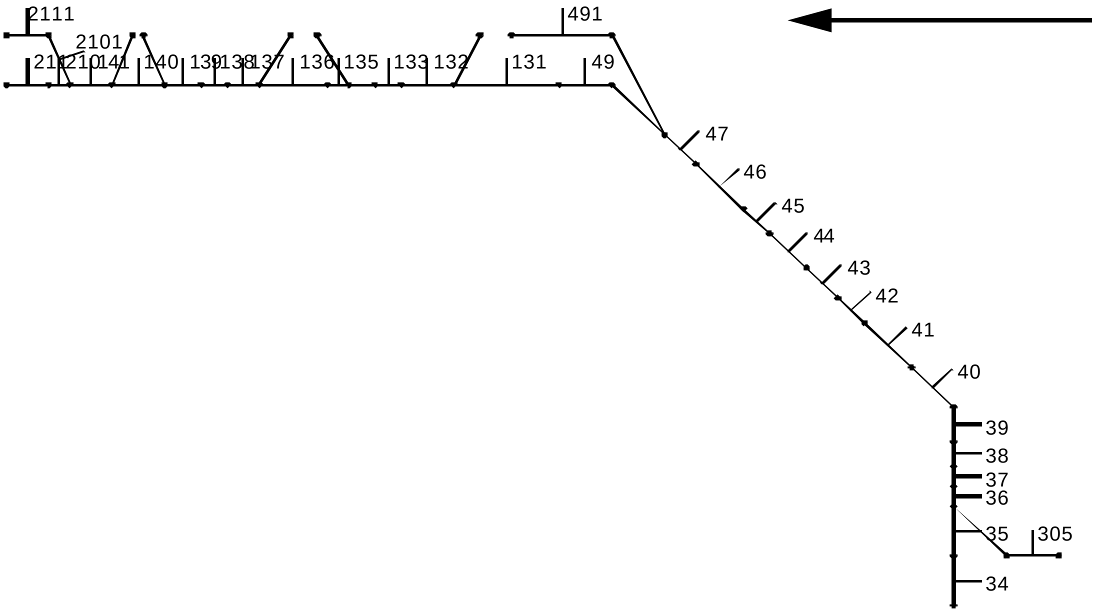


Figure 9.2: Detailed view of study area. The network contains 5 entrances and 5 exits and has a total length of 11 km.

this figure are smoothed in time (5 minutes) and space (2 locations). Any severe incident or massive occurrence of missing data can easily be spotted by viewing the graph.

Figure 9.3 and similar plots that can be produced for the other days in the dataset show that datasets without missing data do not exist. Although the Bayesian updating methods that have been implemented ignore data that have been labelled as missing, the missing data would be disruptive to the evaluation criterion, and should therefore not be used while evaluating the methods.

In many cases it is possible to recover from missing data especially as far as the link volumes is considered. For example some of the detectors in the sequence 36-47 (see figure 9.2) are not working, a link volume can still be produced by using the remaining detectors. The three graphs on the right in figure 9.3 show datasets for which the loopdata have been merged. Eventually, from the twelve available days, nine have been selected on the basis of the graphs for further use.

To judge the effects of the network size, tests have been performed with three network variants. These are shown in figure 9.4. Each variant is a subnetwork of the network shown in figure 9.2. In addition to this, experiments were carried out with aggregation levels of 10 minutes and 5 minutes. This leads to a total of six network variants which are listed in table 9.3. The quantities σ_b^2, σ_q^2 and σ_y^2 could not be directly observed. The values that have been used were determined by experimenting.

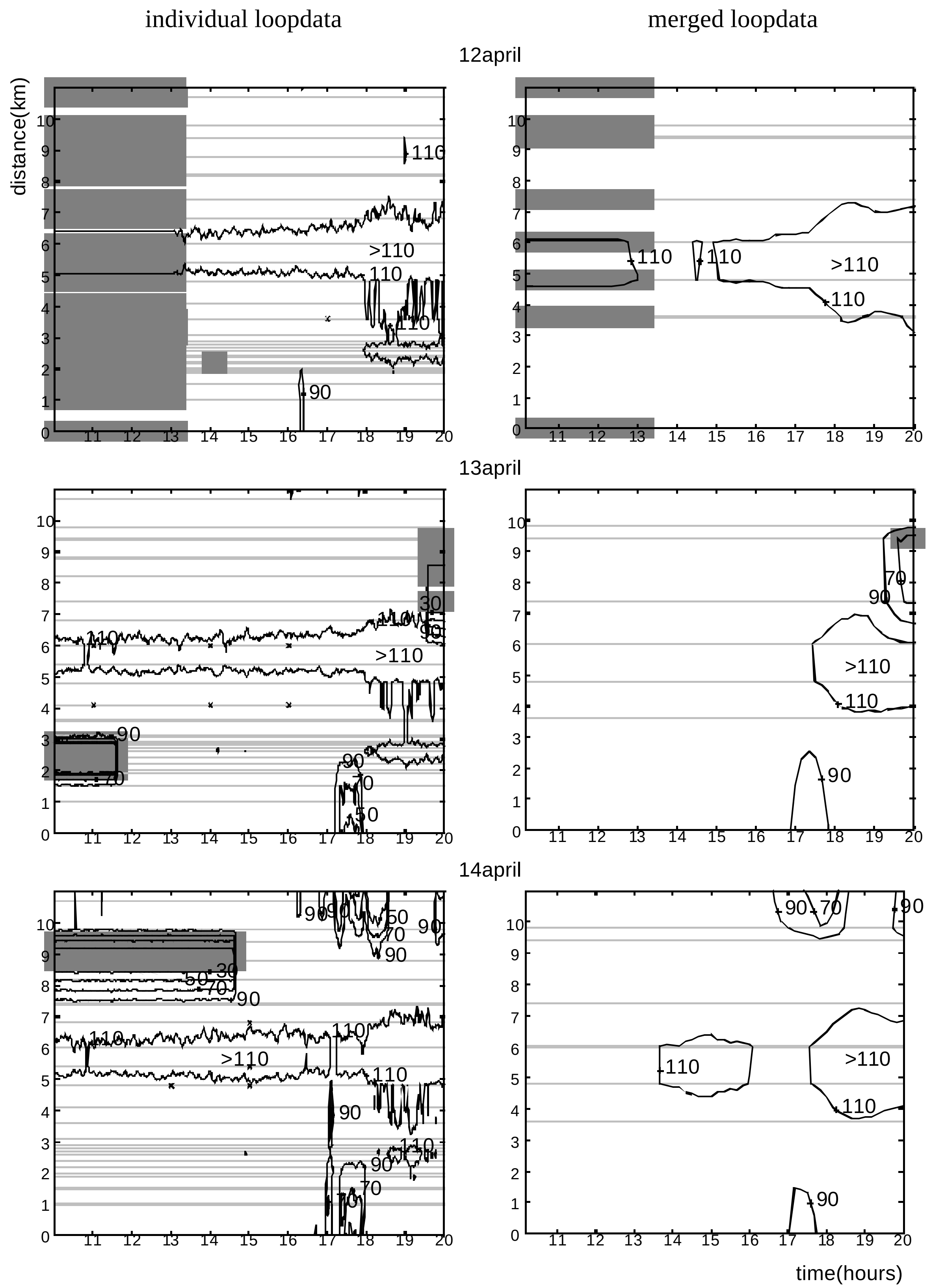


Figure 9.3: Contours of the observed velocities as a function of time and location. The dotted lines represent locations of induction loops. The gray bars represent missing data. Left: all induction loops are used. Right: induction loop data have been merged where appropriate, eliminating some of the missing data.

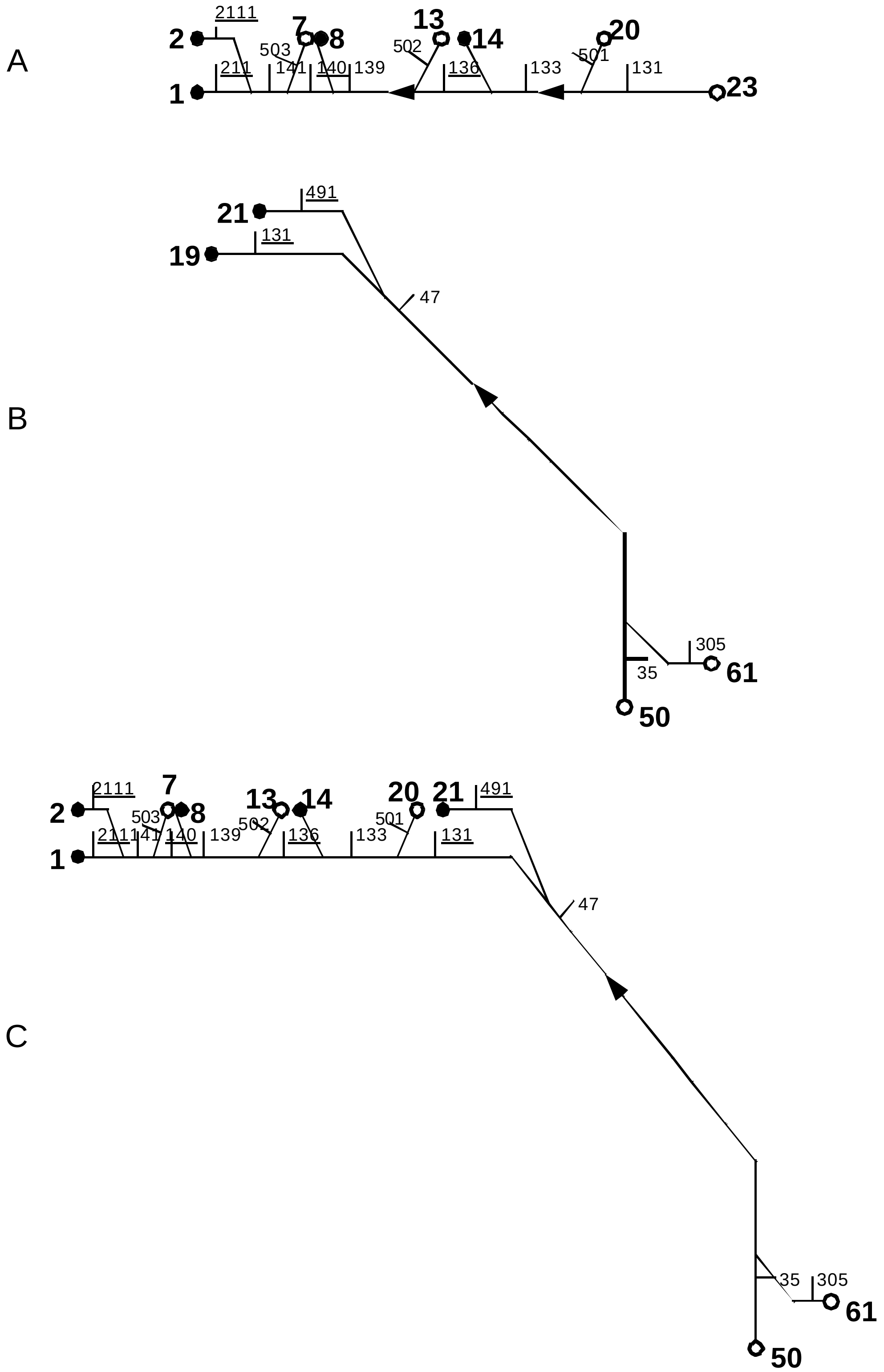


Figure 9.4: Network variants A, B and C. The bold numbers are node numbers. Closed circles are origins. Open circles are destinations. The small numbers indicate the locations of the induction loops. The underlined numbers indicate the induction loops that were used as c reference location.

Table 9.3: Testnetworks

		Network specification					
		A10	**B10**	**C10**	**A5**	**B5**	**C5**
network layout (see figure 9.4)		A	B	C	A	B	C
aggregation level (minutes)		10	10	10	5	5	5
m	*(*)see tables 8.1, 8.2*	4	2	5	4	2	5
n		4	2	5	4	2	5
T		60	60	60	120	120	120
β		0.0002	0.0002	0.0002	0.0001	0.0001	0.0001
$\bar{q}$		182.2	279.8	162.9	91.1	139.9	81.5
σ_q^2		20	20	20	10	10	10
σ_y^2		20	20	20	10	10	10
Days		1994, April 13, 14, 18, 19, 21, 22, 26, 28, 29					

9.3.3 Preparation of the data

<u>*Conversion of the raw data to a moving time coordinate system*</u>

To adjust the data to the moving time coordinate system that is used internally in the estimation methods the data have to be converted from ‘real time’ to ‘moving time coordinate time’. The idea behind this conversion is that each vehicle that travels from A to B on the network contributes to observations only in one period according to the moving time coordinate system. This conversion was performed on the one minute data. Let $y'_k(t)$ represent the observations represented in the real time coordinate system, and let d_k represent the travel delay in periods from location 1 to location k, then the conversion to the moving time coordinate system was done in the following way:

$$y_k(t)=(1-\beta).y'_k(t+\lfloor d_k \rfloor)+\beta.y'_k(t+\lfloor d_k \rfloor+1)$$

with:

$$\beta=d_k-\lfloor d_k \rfloor \qquad (9.1)$$

where $\lfloor d_k \rfloor$ represents the largest integer smaller then d_k. The travel delays d_k on which the conversion is based have been chosen time-invariant. These delays are based on the location of the induction loop and an average speed that was computed using all available data. It is recognized that this approach may be improved by using a more advanced way of computing the travel delays, but for the purpose of comparing the different methods the approach with the fixed delays was judged to be sufficient.

Determining the average travel delay

For the calculation of the average travel delay we need to know the harmonic average of the speeds, which unfortunately is not stored in the MARE system. Instead, the harmonic average is approximated in two independent ways. The first method only uses the speeds that have been observed with the induction loops. The arithmetic averages for the days that have been considered are shown in table 9.4.

Table 9.4: Two approximations of the average speed

date (april 1994)	average speed (km./h.)	
	arithmetic average	maximum inproduct
13	101.8	110.0
14	101.3	96.5
18	100.6	103.1
19	105.5	102.2
21	103.9	103.4
22	105.9	97.6
26	103.9	106.1
28	104.2	107.1
29	103.9	106.1
average	103.4	103.6

The second method only uses traffic counts. For each pair of locations $\{k,p\}$, $p>k$, the delay k-p is approximated with *maximum inproduct* (MIP) delay d_{kp}^{MIP} as follows:

$$d_{kp}^{\mathrm{MIP}} = \underset{d}{\operatorname{argmax}} \sum_t y'_k(t).[(1-\beta).y'_p(t+\lfloor d \rfloor)+\beta.y'_k(t+\lfloor d \rfloor+1)]$$

$$p>k,\ p\leq l,\ k=1,2,\ldots l$$

where:

$$\beta = d - \lfloor d \rfloor \tag{9.2}$$

i.e. d_{kp}^{MIP} is chosen to be the argument that maximizes the inproduct between the vector of observations at location k and the vector of observations at location p shifted in time by the amount d_{kp}^{MIP}. The criterion (9.2) is piecewise linear in d_{kp}^{MIP}. Therefore (9.2) will always result in an integer approximation of d_{kp}. Moreover, as (9.2) is applied to all pairs of locations the outcomes might not be consistent, i.e. for some locations $k_1 < k_2 < k_3$:

$$d_{k_1k_3}^{\mathrm{MIP}} \neq d_{k_1k_2}^{\mathrm{MIP}} + d_{k_2k_3}^{\mathrm{MIP}}$$

Therefore the consistency of the travel delays is imposed as a separate boundary condition and the delays $\{d_{k,k+1},\ k=1,2,\ldots l-1\}$ are solved in a least squares manner from the $(l^2-l)/2$ equations that are generated by (9.2). Adding up all delays gives an average travel time for the 11 km. long motorway corridor. This result can be converted in an harmonic speed average. This average is shown in the column ‘maximum inproduct’ of table 9.4. Based on the data in table 9.4 it was decided to use an average speed of 104 km./h. to compute the travel delays d_k.

Generating missing entry volumes

The observations 501, 502 and 503, present in networks A and C (see figure 9.4) do not exist in the original network shown in figure 9.2. These entry volumes needed to be reconstructed from the internal link volumes.

9.4 Evaluation criterion

The results have been expressed in the link-flow error criterion that has been proposed and tested in the previous chapter, see equation (8.16). This error criterion measures the accuracy with which the link volumes are predicted on a number of reference locations.

The reference locations that have been used are (see figure 9.4) locations 136, 140, 211 and 2111 for network A, locations 491 and 131 for network B, and locations 491, 131, 136, 140, 211 and 2111 for network C.

To some extent this measure gives an impression of the relative performance of various estimation methods. Figure 9.5 plots the link volumes that have been observed at the reference locations, averaged over all reference locations and all selected days. The values of criterion (8.16) should be related to these averages.

9.5 Results

A summary of the results is given in table 9.5. Referring to these results, this section contains discussions on a number of topics. Firstly the optimal length of the aggregation period is discussed. After that we discuss the influence of using the link-flow error as an evaluation criterion. Subsequently the results obtained with the new BU method are compared with those obtained with the traditional methods. Finally the influence of using historic data is discussed.

Five-minute aggregation versus ten-minute aggregation level

First we consider the differences between the results for the datasets with ten minute aggregation (A10, B10 and C10) and those with five minutes aggregation (A5, B5 and C5).

In theory, reducing the aggregation level increases the number of observations and the amount of information that can be retrieved per observation. The latter effect occurs as with the decrease of the aggregation, the fluctuations in the observed entry volumes increase. On the other hand if the aggregation level is too low, effects of travel time dispersion will start to dominate.

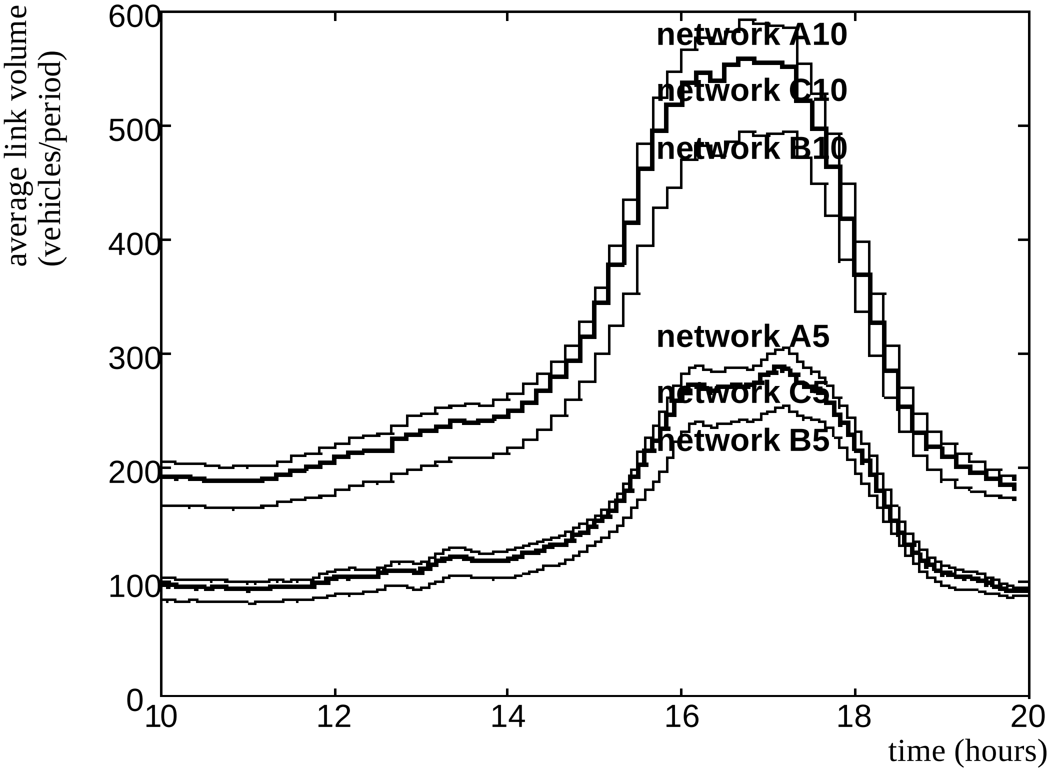
Before answering the question which aggregation level to use, the effect of the increase of aggregation level on the square of the link-flow error criterion (8.16) is analysed, while assuming that the estimate of the split proportions remains constant. The expectation of the squared link-flow error is given by:

$$\mathrm{E}[\mathrm{MSE}^{\mathrm{linkflow}}(t)]=c.\mathrm{E}[||H'(t)b(t)\text{-}y(t)||^2] \tag{9.3}$$

for some constant c. This can be decomposed into a systematic and a random component:

$$\mathrm{E}[\mathrm{MSE}^{\mathrm{linkflow}}(t)]=c.||.\mathrm{E}[H'(t)b(t)\text{-}y(t)]||^2+c.\mathrm{E}[||\ \mathrm{E}[H'(t)b(t)\text{-}y(t)]\ \text{-}\ (H'(t)b(t)\text{-}y(t))||^2\ ] \tag{9.4}$$

If the aggregation period is doubled from five to ten minutes, the systematic component quadruples, and the random component doubles in size. Hence, the factor with which the square of the link-flow error increases if the aggregation period is doubled while the estimates remain unaltered, is given by:



	Network					
	A5	**B5**	**C5**	**A10**	**B10**	**C10**
Average ref. volume	327	271.9	308.6	163.5	136	154.3

Figure 9.5: Observed link volumes at the reference locations as a function of time of day, and averaged over all reference locations (see figures 9.4A,B,C) and all days in dataset (see table 9.3).

$$(4C_1+2C_2)/(C_1+C_2) \tag{9.5}$$

with:

$$C_1=||.E[H'(t)b(t)-y(t)]||^2$$

$$C_2=E[||\ E[H'(t)b(t)-y(t)] - (H'(t)b(t)-y(t))||^2\ ] \tag{9.6}$$

This factor is bounded by:

$$2<(4C_1+2C_2)/(C_1+C_2)<4 \tag{9.7}$$

If in reality this factor would be higher then four, then this would indicate a deterioration of the estimates as a result of the increase of aggregation period. On the other hand if this factor is lower than two, then this can only be explained by an improvement of the estimates. Considering the data in table 9.5, the factor with which the *square* of the link-flow error increases when doubling the aggregation period usually is in the interval [2,4] but occasionally exceeds the value of 4, for example for the combination of method 46 and networks C5 (29.95) and C10 (13.48) this factor is 4.94. Therefore the use of an aggregation period of five minutes seems to be justified.

Table 9.5: Link-flow errors (veh./period)
- average over 9 networks of criterion (8.16)-

category	method			network					
	-	M	α	A10	B10	C10	A5	B5	C5
ICLS	20			20.80	14.03	22.00	12.18	8.81	12.92
FCLS	30			20.87	16.44	22.19	12.16	9.86	13.02
BU	43			18.05	15.39	25.79	10.51	9.28	12.68
	46			18.56	16.02	29.95	10.58	9.53	13.48
	47			18.59	15.37	27.09	10.81	9.29	12.96
	55			17.65	16.05	35.60	10.61	9.54	14.83
	56			16.92	16.05	19.92	10.39	9.54	11.82
	61	46	1	18.72	15.99	33.18	10.45	9.52	13.81
	62		0.9	15.78	15.86	20.21	9.86	9.43	11.76
	63		0.8	15.89	15.74	22.55	9.92	9.39	12.35
	64	47	1	18.96	15.37	30.10	10.82	9.27	13.32
	65		0.9	15.99	15.22	19.64	9.93	9.17	11.56
	66		0.8	15.80	15.15	20.52	9.91	9.17	11.73

Influence of the link-flow error criterion

As direct observations of EE-flows are not available with the Amsterdam dataset, the only evaluation criterion that can be evaluated is the link-flow error criterion (8.16). The properties of this error criterion have been extensively discussed in section 8.3 under the title ‘comparing evaluation criteria’ (page 93). Some of the pitfalls that have been mentioned in that discussion could easily lead to misinterpreting the results in table 9.5. Examples will be given below.

The main problem with the link-flow error criterion is that is systematically favours the ICLS, FCLS and BU methods that use MAP postprocessing. In table 9.5 method 20 (ICLS), 30 (FCLS) and 55, 56 (BU+MAP) are of this type, see table 8.4.

As an example of this phenomenon consider the fact that for all datasets, method 56 (BU+MAP) has a lower link-flow error than the FCLS method 30. As method 46 (BU+SE) is identical to method 56 except that it uses the subjective expectation (SE) postprocessing routine instead of the maximum aposteriori (MAP) postprocessing routine, the error of estimation of method 46 in terms of EE-flow error may be assumed to be lower than that of method 56 (see conclusions chapter 8). Nevertheless for dataset C10, and to a lesser extent, network C5 method 46 results in a higher link-flow error then method 30. A similar phenomenon is observed with method 43 and 47 in combination with dataset C10.

It should be noted that this phenomenon is related to inaccuracies in evaluating the expectation of a TMVN distribution, and only occurs if the probability of sampling a value from this distribution within the hypercube [0,1] is too low (see section 8.3, page 93). It might be possible to overcome this problem by developing a more advanced version of the SE postprocessing routine.

Another example where the link-flow error criterion seems to favour one method over the other stems from comparing the performance of method 20 (ICLS) and method 30 (FCLS) on datasets B10 and B5. The split probabilities estimated by these methods may be inspected using stacked plots, see figure 9.6. In these plots, the top lines represent the sum of all split estimates. It is apperent that method 20 systematically underestimates the split proportions. This problem is corrected by imposing the equality constraints (method 30), however this appears to have an adverse effect on the link-flow error.

Comparison of the new BU method with traditional methods

The main novelties of the method developed in chapters 4-6 relative to the traditional Kalman filter are the absence of recursive constraining, the use of an SE postprocessing routine, the use of an approximation for the covariance matrix of the measurement error, and the possibility for using historic information. In table 9.5 the methods that incorporate these novelties are methods 43, 46 and 47, and their variants using historic information, methods 61-66.

Methods 20 and 30 may be considered as representatives of the traditional ICLS and FCLS parameter optimization methods, method 55 coincides with the traditional Kalman filter. Although the method uses a covariance matrix identical to method 46. Finally, method 56 is some intermediate form, it is identical to method 55, except for the fact that it does not use recursive constraining.

With the exception of the results for dataset C10, which were discussed above, the results for the methods 43, 46 and 47 tend to be better then those for the traditional methods 20, 30 and 55, despite the biased link-flow error evaluation criterion. Figure 9.7 shows a comparison between method 30 and method 46. This plot is representative for the comparison of the Bayesian update methods with FCLS methods. During the first periods the FCLS method corresponds with a lower link-flow error but, after a number of periods the estimates of methods 46 result in more accurate link-flow predictions.

As far as it is possible to analyse the influences of the novelties of the new methods separately using the link-flow error criterion, these will be discussed below.

The advantages of not applying recursive constraining become clear by comparing the results of method 55 and 56. In all instance method 56 leads to equal or lower link-flow errors.

The advantages of applying SE postprocessing can not be made clear by using the link-flow error criterion only. For example, on the basis of the link-flow error criterion, one would conclude that method 56 produces more accurate estimates of split proportions than method 46. However, in all instances where this could be checked, method 46 produced the most accurate split estimates in terms of EE-flow error.

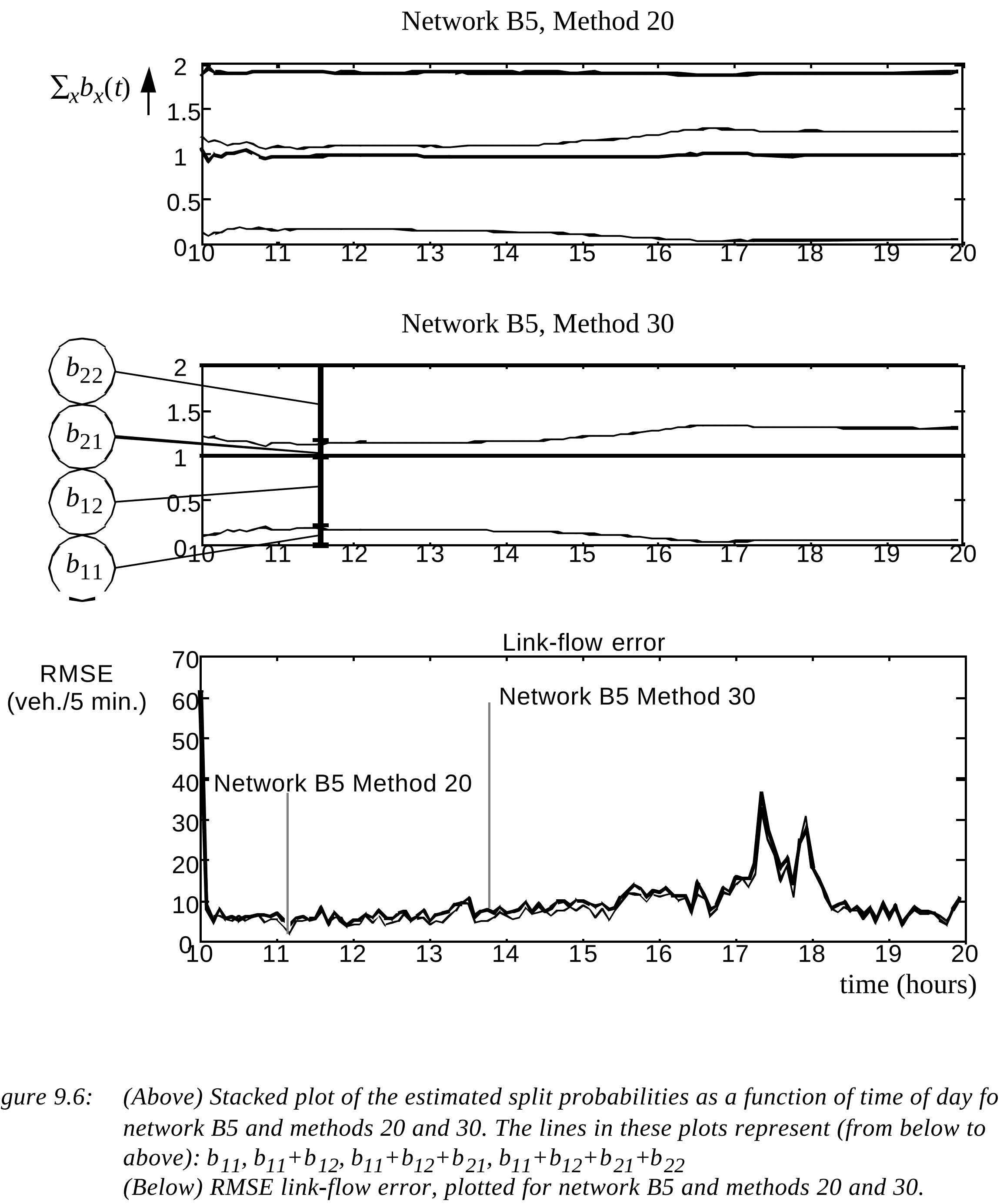


Figure 9.6: *(Above) Stacked plot of the estimated split probabilities as a function of time of day for network B5 and methods 20 and 30. The lines in these plots represent (from below to above): b_{11}, $b_{11}+b_{12}$, $b_{11}+b_{12}+b_{21}$, $b_{11}+b_{12}+b_{21}+b_{22}$*
(Below) RMSE link-flow error, plotted for network B5 and methods 20 and 30. Method 20 systematically underestimates the split probabilities, but despite these bias, results in lower link flow errors

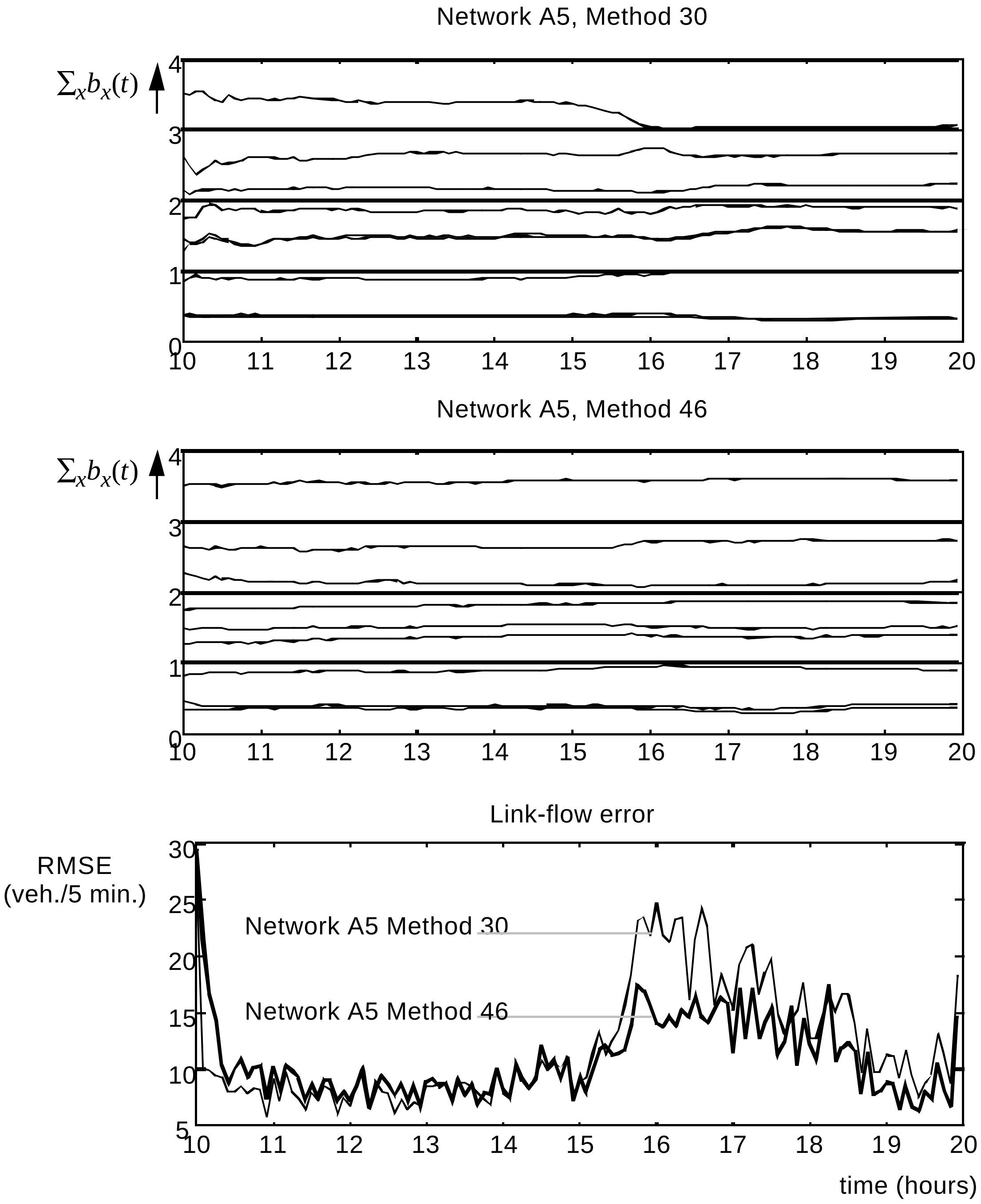


Figure 9.7: Comparison between method 30 and method 46 for network A5. The top two graphs contain stacked plots of the estimated split proportions, i.e. the lines represent from (below to above): b_{11}, $b_{11}+b_{12}$, $b_{11}+b_{12}+b_{21}$, $b_{11}+b_{12}+b_{21}+b_{22}$

The advantages of using an accurate approximation of the measurement error covariance matrix can be seen by comparing the results for, on the one hand, methods 43 and 47, and, on the other hand, method 46. These methods are identical except for the fact that methods 43 and 47 use a the PEBA and DPEBA covariance matrix respectively, while methods 46 uses the ALF covariance matrix. On average the results for methods 43 and 47 are slightly better, as was the case with synthetic data (see arrow ‘i’ in figure 8.8). Between methods 43 and 47, method 43 gives better results in terms of the link-flow error criterion. This was also found in the previous chapter, as can be seen in table 8.7, but is no reason to assume that the estimates of method 43 are ‘better’ in terms of the split error criterion.

The influence of historic information

The use of historic information clearly leads to better scores on the link-flow error criterion, see for example figure 9.8. The main improvements occur due to more accurate link-flow predictions in the first periods. The results are however highly sensitive to the parameter α, that determines with what strength the state is attracted to its historical mean. The value α=1 corresponds with the model that the state equals an historical average, increased with an independent process $x(t)$, see equation (6.5). The influence of historical information is now limited to predetermined values of $u(t)$, derived from changes in the historical pattern from period to period. For values α<1 the value of $u(t)$ also depends on a second component that is derived from the discrepancy between $b(t)$ and the historical average.The accumulated effect of $u(t)$ is now potentially larger, since the second component only fades away if $b(t)$ equals the historical average.

Of the three values that were tested, α=0.9 gave the best results according to the link-flow error criterion. The results obtained with α=1 (methods 61 and 64) hardly are an improvement over their corresponding methods (46 and 47). The probable explanation for this is that the values of $u(t)$ that correspond to α=1 are negligible.

Discussion

The large extent to which the ranking according to the link-flow error criterion and the ranking according to, for example, the EE-flow error criterion are inconsistent was unforeseen, nor have such findings been reported in earlier literature.

This phenomenon forces one to consider for which purpose EE-matrices are estimated. For example if EE-matrices are estimated for the purpose of making short term traffic predictions under average conditions the results reported in this chapter suggest that the best matrix in terms of EE-flow does not necessarily give the best link-flow predictions. On the other hand if the estimated EE-matrix is used to predict the effect of a control measure that relates to a specific group of EE-flows, e.g. a lane closure, an EE-matrix with a low EE-flow error is required.

This suggests that the general approach to EE-estimation should be not to estimate a single ‘best’ EE-matrix, but instead to define an objective function, and subsequently estimate an EE-matrix in such a way that the objective function is minimized.

Method 56 serves as an example of such an approach: although simulations produce strong evidence that this method results in higher EE-flow errors then method 46, of all methods that do not use historical information (including method 46), this method produces the most accurate link flow predictions.

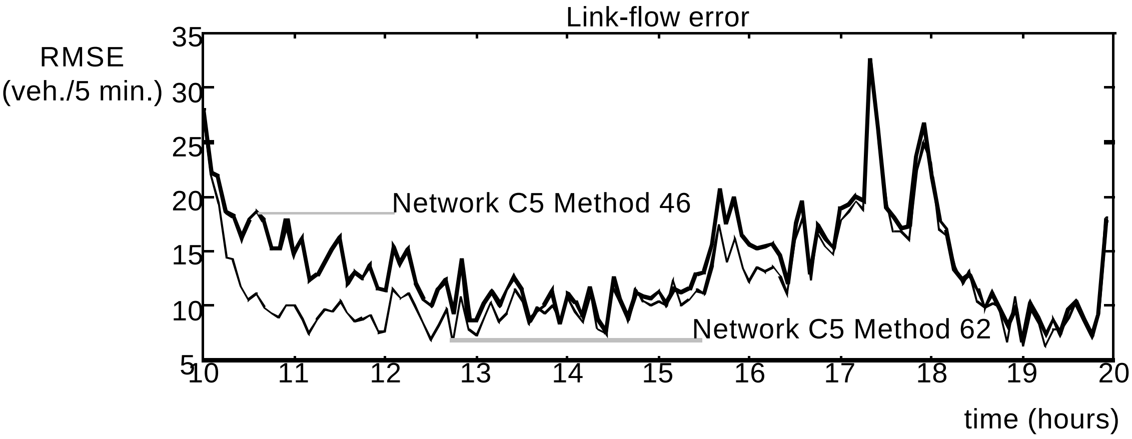


Figure 9.8: The influence of using historic information. Method 62 is identical to method 46, except for the use of historic information (α=0.9).

9.6 Conclusions

The experiments with empirical data largely confirm the outcomes of the simulations in chapter 8. Relative to the FCLS method, an average reduction of 7% in the link-flow error over the datasets with 5 minute aggregation was obtained with the BU method that was shown to give the lowest error of estimation for synthetic data.

Even though the link-flow error criterion that was used is biased in favour of parameter optimization methods that were used as a reference, such as FCLS, the new Bayesian updating methods gave rise to a lower error according to this criterion.

An aspect of the new method that has not been tested before is the use of historic information. Implementation of the default model that was described in chapter 6 has led to a significant improvement in the accuracy with which link-flows are predicted.

Within the class of methods not using historic data, the most accurate link-flow predictions are obtained with a variant of the new method of which previous experiments have shown that it performs poor in terms of EE-flow error relative to other Bayesian updating methods. This suggests that the general approach to EE-estimation should be not to estimate a single ‘best’ EE-matrix, but instead to define an objective function, and subsequently estimate an EE-matrix in such a way that the objective function is minimized.

10. Conclusions

10.1 Research findings

This thesis describes research into the problem of tracking time varying EE-travel demand from time series of observations, applied to small networks such as motorway corridors. To eliminate the underspecification that was shown to be inherent to this problem a model referred to as the motorway model was specified.

Central to this model is the assumption that entry volumes, which may vary from period to period, distribute themselves with near constant fractions over all reachable exits. From the viewpoint of an observer, EE-flows may be considered as outcomes of a number of random trials, governed by slowly varying split probabilities, $b_{ij}(t)$, that define the probability that a vehicle entering at entrance i in period t will be destined for exit j (chapter 1).

Usage of the motorway model refines a number of earlier proposed dynamic approaches. For example in the motorway model, the split *probabilities* supersede the split *proportions* as the unknown parameters. This implies that a part of the variation in EE flows that was formerly subscribed to variation in the unknown parameters, is now described statistically by a random selection process that is inherent to individual motorists making uncoordinated travel decisions. The split probabilities implicitly reflect the prevailing traffic conditions and travel demand on the full network. However, no dependencies of EE-flows on other processes in the traffic system, or implied mutual relations between EE-flows of such dependencies, are imposed on these flows (chapter 2).

The objective of this thesis was not only to improve the theoretical underpinning, but also to expand the practical applicability of dynamic EE-estimation methods. The commonly used restrictions that the entry flows are exactly known, traffic counts do not involve internal links, and traveltimes are small relative to the length of the discretisation period are relaxed by explicitly taking errors in entry volume observations into account, allowing for the use of internal link counts, and relaxing the assumption of small traveltimes into the assumption of small traveltime *dispersion* (chapters 1, 2).

Taking the motorway model as a point of departure, a second step in the research involved deriving system properties following from the model, deriving estimators for its parameters, and investigating how additional sources of information may be used to improve the estimates.

The assumptions in the motorway model imply that the expectation of traffic counts is a linear function of the unknown split probabilities, and give rise to a linear measurement equation that enables the estimation of the split probabilities using standard techniques such as least squares, constrained least squares or the Kalman filter. A novel theoretical result that was obtained is the covariance matrix for the measurement error. Knowledge of this matrix in combination with appropriate statistical techniques such as the Kalman filter helps to improve the estimates of the split proportions. As the derived matrix is a function of the unknown split proportions, the result can only be used as a basis for approximations of the true matrix, for example by substituting the most up to date point estimate for $b(t)$ (resulting in a Point Estimate Based Approximation, PEBA) or by using the most up to date subjective probability distribution of $b(t)$ (resulting in a Distribution Based Approximation, DBA) (chapter 5).

The conclusion from studying the methods that have been proposed in literature to estimate the split proportions is that none of them does so in a completely satisfactory manner. Parameter optimization methods such as Fully Constrained Least Squares (FCLS) offer no possibility to employ the derived statistical properties of the observations, while the Kalman filter does not deal with the inequality constraints that apply to $b(t)$ in a proper manner (chapter 3).

To overcome these problems a new method referred to as Bayesian updating was proposed. The new method is based on adopting a Truncated Multivariate Normal (TMVN) distribution for the split probabilities and updating this distribution using Bayes rule if new observations become available. It was shown that if the shape of the likelihood function corresponds to MVN, then the updated distribution remains in the class of TMVN distributions, and the parameters that characterise the updated distribution follow from applying the well known Kalman measurement update equations. The method is exact if $b(t)$ does not vary in time. For the time extrapolation no exact analytical expression has been found. However, it is expected that if the time variation in $b(t)$ is small, a good approximation is obtained if the standard Kalman time extrapolation equations are used (chapter 4).

To derive point estimates from the above described recursion, a postprocessing routine is needed. Two types of such routines have been identified. The first type involves computing the point estimate that maximizes the aposteriori density. The second type involves evaluating the expectation of the aposteriori density. Point estimates of the first type are referred to as Maximum APosteriori (MAP) estimates, while estimates of the second type are referred to as Subjective Expectation (SE) estimates. In theory, SE estimates minimise the ($L2$) error of estimation, but their computation is not analytically tractable. Approximations of SE estimates may be obtained by averaging a large number of vectors sampled from the TMVN aposteriori distribution. TMVN random vectors may be generated simply by generating MVN random vectors and rejecting all outcomes that do not satisfy the inequality constraints. This strategy is referred to as Randomized Mean (SE-RM), but fails if the probability of sampling valid outcomes is too low. As an alternative to the SE-RM point estimate an analytical approximation was derived, referred to as Approximated Mean (SE-AM). The computation of MAP estimates is more straightforward, and comes down to minimizing a quadratic function under inequality constraints (chapter 4).

EE-flows that are estimated from time series of traffic counts may be improved by using additional sources of information. The analysis of empirical data has revealed not only that split proportions vary little between consecutive time periods but also that there is very little variation in the splits when corresponding periods of departure on multiple days are considered. To use this property for the improvement of estimates of split probabilities, a model was proposed that expresses that the split probabilities vary randomly around historic patterns. From this model a state equation was derived that combines usage of the assumption of slowly varying split probabilities with usage of historical data (chapter 6).

In the future, new technologies will make it possible to trace individual vehicles in an automated manner either by installing Automated Vehicle Identification (AVI) equipment at multiple locations or by letting vehicles transmit their trajectories. This will generate data on traffic characteristics that could not be directly observed at low cost before, such as trajectories and traveltimes. To benefit from these data in full, new estimators are needed. As an example, the usage of automated license plate readers is considered. Although the problem of updating trip matrices from direct observations such as travel surveys and license plate surveys has been extensively studied in the past, the present context contains many new elements, such as the requirement to process the license plate data in combination with traffic counts, the presence of recording errors, the dynamic context, and the absence of any restrictions on the locations of the license plate readers. At first sight the problem of updating EE-flows from license plate data has little in common with EE-flow estimation from traffic counts. However, by constructing a fictional hypernetwork it was shown that an analogy exist, and that in fact all theory that was proposed earlier can be recycled to solve the problem of estimating EE-flows from combined data. This analogy remains valid if the trajectory information only applies to a representative *sample* of all vehicles (the probe vehicles). It may hence be concluded that the proposed Bayesian estimator may be used as a part of a framework to estimate travel demand, that is highly flexible with respect to the type of input data that are required. For example such a framework may use a mixture of historic data, traffic counts, automated license plate surveys, and data obtained from a group of vehicles that is equipped with a transponder (chapter 7).

In the last part of the thesis, many of the theoretical findings have been put to the test in two series of experiments. The first series of experiments involved synthetic EE-flows and traffic counts that were generated according to the specifications of the motorway model. The second series of experiments involved traffic counts collected on Amsterdam beltway during one month.

The tests with the synthetic data very clearly show that usage of the new BU method reduces the error of estimation considerably, relative to usage of existing methods, such as FCLS and the Kalman filter. Relative to the FCLS method a reduction of 31% of the average RMSE of the split probabilities was obtained. The largest part of this reduction (67%) should be subscribed to the improved treatment of the inequality constraints by the new estimator combined with the use of the SE-RM postprocessing routine. A smaller part of the reduction (33%) is due to the use of a more accurate system specification, and the PEBA and DBA covariance matrices that follow from this specification. If instead of the PEBA or DBA matrices, a matrix based on the Average Link Flows (ALF) is used then only a small part of the reduction is sacrificed again (-8%), while the advantage of such an approach is that it may be applied without having knowledge of the system specification (chapter 8).

An unexpected result was that a reduction of the error of estimation does not necessarily result in an increase of the accuracy with which link volumes are predicted (link-flow error). For example, otherwise identical Bayesian updating methods consistently result in a lower error of estimation but a higher link-flow error, if the SE-AM instead of the MAP postprocessing routine is used. Although in theory this phenomenon does not occur if the randomized mean (SE-RM) point estimates is computed, in practice SE-RM point estimates can not always be evaluated. In such cases the postprocessing routine reverts to computing the SE-AM value after all. If the link-flow error criterion is used as an evaluation criterion then the above mentioned reduction of 31% reduces to only 3.7% for synthetic data, and 7% for empirical data (chapter 8, 9).

In the experiments with empirical data the estimates can only be evaluated according to the link-flow error criterion, as the correct EE-matrix is not known. However, the link-flow error criterion was shown to be biased in favour of the FCLS method and BU methods that use MAP postprocessing. Nevertheless a few conclusions can be drawn. Firstly, in all cases, usage of historic information reduces the link-flow errors. Secondly: a variant of the new BU method that uses MAP postprocessing outperformed the FCLS and the traditional Kalman filter in all cases. Thirdly: except for one of the six datasets considered, the BU methods that use SE-RM postprocessing outperformed the FCLS method, despite the biased error criterion (chapter 9).

Although this does not show from the link-flow errors, from experience it is known that SE-RM postprocessing should be preferred over MAP postprocessing. Therefore the results suggest that it may be possible to develop a third postprocessing routine that estimates link flows even more accurate than can be done using MAP estimates. More in general this means that the postprocessing routine should be chosen depending on the purpose for which the EE matrix is estimated.

10.2 Practical recommendations

In this thesis new theories have been proposed that enable dynamic EE-estimation procedures to take into account properties of realistic traffic data, such as errors occurring in observations of entry volumes and spatial correlations between observations that arise when all available traffic counts are used. Although for a given set of observations these theories help to reduce the error of estimation, there is a limit to these reductions, and the best attainable accuracy eventually depends on the quality and amount of observations.

Practitioners who design and implement the data collection system hence play a key role in creating favourable circumstances for EE-estimation and applications requiring EE-flow estimates such as short term traffic prediction and control. Based on the experiences with data retrieved from the traffic surveillance system that is currently in operations on a large part of the motorways in the Netherlands, recommendations for improvements are summarized in the following list:

- Preferably induction loops should be present at all entries and exits of the motorway network. Entry and exit volumes can be reconstructed by taking the difference between the observed volumes at the two adjacent internal links, however, this would imply inclusion of the counting errors at the internal links in the observation of potentially low entry or exit volumes, resulting a very low ratio between average volume and counting errors. This problem is made worse by the fact that the accuracy of induction loop detectors decreases in areas with high rates of lane changing, which is typical for areas containing on- or off ramps. Also uncertainty about traveltime introduces extra errors in reconstructed volumes.
- If vehicle counts and observed speeds need to be aggregated in time at road side processors, for example as a result of restrictions on storage or communication capacity, this should be done with care as to minimise the information loss arising from this aggregation. In Dutch practice some older equipment aggregates traffic counts by maintaining a moving average of the gaps in the traffic flow. Observations of vehicle speeds are aggregated by computation of the arithmetic average. For EE-flow estimation, we are mainly interested in traffic counts and average traveltimes. Therefore if data are aggregated this should be done by totalling the traffic counts over the aggregation period and by storing the *harmonic* average of the observed speeds.
- Better documentation of the surveillance system is needed, and a database containing accurate information about the network layout and the exact position of the counting locations

should be maintained.

- At present, any number of observed vehicles equal to or lower than four per minute results in a *reported* traffic count of four. This is to prevent the system from labelling a set of detectors as malfunctioning. This property is a constant source of errors, especially at night and on less frequently used links. It is therefore recommended that the detection of dead detectors is done in another way.

The above recommendations apply to the existing traffic data collection system that is entirely based on induction loops. Apart from improving this existing system, the employment of techniques that enable the tracing of individual vehicles is a cost effective way to gain insight in traveltimes and routes. In this thesis it was shown that, as far as the estimation of EE-matrices is concerned, no strict requirements apply to the location and accuracy of identification. Thus, equipment may be used that does not give one hundred percent recognition, and this equipment need not be installed at all entrances and exits. This opens up the possibility to use low cost equipment such as cameras combined with image processors, possibly even using camera’s that are already in place for other purposes.

10.3 Further research

A number of limitations and assumptions were incorporated in the problem described in the first chapter of this thesis. Relaxing these would give rise to new research. Also a number of new questions have arisen during the current research project. As neither the available time, nor the size of this report allow for all of these issues to be addressed in the present context, these are left as future research topics. The following list consists of topics on which research is in progress or topics that may be expected to be taken on by researchers in the near feature.

- *Taking into account traveltime dispersion*. Traveltime dispersion occurs if vehicles travel at different speeds. As a result, a one to one correspondence between trip departure interval and interval of vehicle observation as assumed in the moving time coordinate system can no longer be maintained. Instead the vehicles that depart in one particular interval may contribute to traffic counts in multiple intervals, according to a traveltime distribution. This distribution can either be prespecified, or could be estimated.
  If the traveltime distribution is *prespecified*, the measurement equation given by (2.23) should be replaced by an equation that specifies the measurement as a linear function of a vector consisting of split probabilities in multiple periods. Such a vector is known as an *augmented* state, and a linear state equation, similar to (2.7), that describes the evolution of the augmented state in time can be given. Given the linear state- and measurement equations, all parameter estimation methods described in this thesis can again be applied, although it should be noted that these methods do not account for the serial correlation that will be displayed by the measurement error.
  A second possibility is to estimate the traveltime distribution simultaneously with the split proportions. One way this can be done is to introduce a separate split parameter for each combination of EE-pair and traveltime, see e.g. *Bell (1991b)*. However, this leads to an increase of the number of unknown parameters with a factor that equals the number of periods involved in the traveltime distribution. Typically this factor would equal two or three, making it possible to distinguish between fast and slow platoons, or between fast, average and slow platoons.
- *Taking into account route choice*. Again, route choice can be taken into account by prespecifying the route choice proportions, or by introducing a separate split parameter for

each path flow. Neither of these two approaches would necessitate fundamental changes to the methods proposed in this thesis, although the moving time coordinate system will have to be reconsidered, as different routes are unlikely to have equal traveltimes.

- *Time dependent assignment maps*. The assignment map, *U*, specifies the relation between EE-flows and link flows, and in this thesis was assumed to be given by a constant matrix. To establish the assignment map for general networks, information is needed on route choice proportions, average traveltime delays, and traveltime dispersion. The issue of specifying the assignment map is hence closely related to the two issues mentioned above. The methods presented in this thesis would not fundamentally change if the fixed assignment map were to be replaced by an estimate that depends on observations taken from the network. Such an assignment map can again be regarded as prespecified, as long as the estimation of the assignment map and the estimation of the EE-flows are done separately. Again, prespecifying the assignment map can be avoided by introducing separate split parameters for each combination of departure interval, route and traveltime at the cost of a vast increase of the number of unknown parameters.
- *Combined estimators*. The three cases mentioned above imply taking into account phenomena such as travel time dispersion, route choice and time dependency of the assignment map, and hence correspond to relaxations of assumptions used in this thesis. It was suggested that these phenomena should be incorporated either by prespecification or by the introduction of extra split parameters. Neither of these suggestions endanger the linear structure of the estimation problem, which means the estimation procedures proposed in this thesis can still be used. Moreover, in this context the notion of prespecification can be extended to cases in which a phenomenon, e.g. travel time, is estimated, e.g. from observations of speeds, as long as the estimation of this particular phenomenon is done separately from the estimation of the EE-flows.

  However, in reality EE-flows and variables representing phenomena such as travel time, route choice and travel time dispersion are strongly correlated, and it may be expected that part of the future research will address the simultaneous estimation of these variables with the EE-flows. On the other hand, this estimation problem is highly nonlinear.

  A number of techniques are likely to be applied to this estimation problem. One of these is the extended Kalman filter which is based on a linearization of the problem. Other techniques one can think of are iterative procedures based on an operator to which the solution of the estimation problem is a fixed point. Much research is still to be done into the effectiveness of these procedures for this particular class of problems.
- *Utilizing static models in dynamic EE-matrix estimation*. In the past, a wide range of OD estimation methods has been developed to estimate static OD-matrices using single, instead of time series of traffic counts. These methods rely on models that describe mutual relations between OD-flows or dependencies of OD-flows on observable data such as land use data. These models require a certain level of time and spatial aggregation to be sufficiently plausible, and this complicates the application of these models to the dynamic EE-estimation problem. Nevertheless, it is conceivable that in some form, static models can be used as an extra source of information in dynamic EE-estimation.
- *Utilizing dynamic data in static OD-matrix estimation*. When the objective is to estimate a static matrix instead of a dynamic matrix, calibration of a static model from traffic counts, survey data and land use data is a frequently used approach. The aggregation of time series of observations that occurs in such an approach results in an information loss. One possibility that has not been investigated before in this context, is to involve estimates of split pro-

portions that are derived from the time series of observations in the calibration of the static model.

- *Applying the BU method to other estimation problems*. In this thesis the Bayesian updating method was used to replace well known parameter estimation methods such as least squares and the Kalman filter, and it was shown that such a change resulted in a reduction of the error of estimation. This suggests that a similar change could also be appropriate in other cases where least squares or the Kalman filter are used under circumstances where the unknown state satisfies inequality constraints and is known to vary only slowly in time. A case in which the method can be applied straightforwardly, is that of updating a prior OD-matrix from traffic counts. For example *Bell (1991a)* reported that if a prior matrix is updated using traffic counts by the unconstrained minimization of a generalized least squares objective function, the result did not always satisfy the nonnegativity requirement. In *Bell (1991a)* it was therefore suggested to use constrained least squares instead. The results presented in this thesis suggest that the Bayesian Updating method would have been an even better alternative. Other examples of methods that could be improved in this way are found in *Maher (1983)* and *Pursula and Pastinen (1993)*.
- *Predicting link volumes*. One of the findings of the present research project was that the estimator that produces the *EE-flow estimates* with the lowest error of estimation does not necessarily produce *link flow predictions* with the lowest error of estimation. Especially least squares estimators and estimators similar to least squares produce estimates with a relatively high error of estimation but which imply link flow predictions that are more accurate then those obtained with the Bayesian updating method. It was also shown that this is caused by the postprocessing algorithm. This suggests that special postprocessing algorithms should be designed for the purpose of predicting link flows.

11. References

Abramowitz, M. and Segun, I.E. (1968) *Handbook of Mathematical Functions*, fifth edition, Dover Publications, Inc., New York

Anderson, B.D.O. and Moore, J.B. (1979) *Optimal Filtering*, Prentice Hall

Ashok, K. and Ben-Akiva, M.E. (1993) Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems, *Proc. 12th Int. Symp. on Transportation and Traffic Theory*, Berkeley, C.F. Daganzo (Ed)

AVV (1974) *Basisnetwerk Programmatuur Reistijden*, (in Dutch) Rijkswaterstaat, DVK, Centrum voor Vervoersplannen

AVV, Transport Research Centre (1992) *Mare Onderzoeker Handleiding*, (in Dutch) Rijkswaterstaat, DVK, Hfdafd. Dynamische Verkeersbeheersing

AVV, Transport Research Centre (1994) Motorway Signalling, *AVV*, Brochure, Directorate-General for Public Works and Water Management

Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. (1993) *Nonlinear Programming, Theory and Algorithms*, second edition, Wiley

Bell, M.G.H. (1991a) The Estimation of Origin-Destination Matrices by Constrained Generalised Least Squares, *Transportation Research-B*, Vol.25-B, pp.13-22

Bell, M.G.H. (1991b) The Real Time Estimation of Origin-Destination Flows in the Presence of Platoon Dispersion, *Transportation Research-B*, Vol.25-B, pp.115-125

Bell, M.G.H., Inaudi, D., Lange, J. and Maher, M. (1991) Techniques for the Dynamic Estimation of O-D Matrices in Traffic Networks *Proceedings of the Drive Conference*, Feb.4-6, 1991

Buchanan, C. and Partners (1986) Micromatch Number-plate Matching Suite, *User Guide*

Cascetta, E. and Nguyen, S. (1988) A Unified Framework for Estimating or Updating Origin/Destination Matrices from Traffic Counts, *Transportation Research-B*, Vol. 22-B, No. 6, pp. 437-455

Cascetta, E., Inaudi, D. and Marquis, G. (1993) Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts, *Transportation Science*, Vol. 27, No. 4, pp. 363-373

Catlin, D.E. (1989) Estimation, Control, and the Discrete Kalman Filter, *Applied Mathematical Sciences 71*, Springer-Verlag

Cremer, M. (1983) Determining the Time-Dependent Trip Distribution in a Complex Intersection for Traffic Responsive Control, *IFAC Control in Transportation Systems*, Baden-Baden

Cremer, M. and Keller, H. (1981) Dynamic Identification of Flows from Traffic Counts at Complex Intersections, *Proc. 8th Int. Symposium on Transportation and Traffic Theory*, University of Toronto Press, Toronto Canada

Cremer, M. and Keller, H. (1984) a Systems Dynamics Approach to the Estimation of Entry and Exit O-D Flows, *Ninth International Symposium on Transportation and Traffic Theory*

Cremer, M. and Keller, H. (1987) A New Class of Dynamic Methods for the Identification of Origin-Destination Flows, *Transportation Research-B*, Vol. 21B, No. 2, pp. 117-132

Davis, G.A. (1989) *A Stochastic, Dynamic Model of Traffic Generation and its Application to the Maximum Likelihood Estimation of Origin-Destination Parameters,*, Ph.D. Thesis, University of Washington

De Romph, E. (1994) *Dynamic Traffic Assignment, Theory and Applications*, Ph.D. Thesis, Delft University of Technology

De Romph, E., Van Grol, H.J.M. and Hamerslag, R. (1994) Application of Dynamic Assignment in Washington, D.C., Metropolitan Area, *Transportation Research Records*, No. 1443, pp. 100-109

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood Estimation from Incomplete Data via the EM algorithm, *J.Roy. Statist. Soc. B*, 39, pp. 1-38

Erlander, E. and Stewart, N.F. (1990) *The Gravity Model in Transportation Analysis, Theory and Extensions*, VSP, Utrecht, the Netherlands

Evans, R., Martin, P.T. and Bell, M.C. (1993) A Method for Analysing Partial Registration-plate Data, *Traffic Engineering and Control*, Vol. 34, pp. 76-79

Fletcher, R. and Reeves, C.M. (1964) Function Minimization by Conjugate Gradients, *The Comp. J.*, 7, pp. 149-154

Furness, K.P. (1965) Time function iteration, *Traffic Engineering and Control*, 7(7), pp. 458-460

Gang-Len Chang and Jifeng Wu (1994) Recursive Estimation of Time-Varying Origin-Destination Flows from Traffic Counts in Freeway Corridors, *Transportation Research-B*, Vol. 28B, pp. 141-160

Geva, I., Hauer, E. and Landau, U. (1982) *Maximum Likelihood and Bayesian Methods for the Estimation of O-D Flows*, University of Toronto, Publication 82-11, ISBN 0-7727-7038-7

Hague Consulting Group (1993) *Annual Project Review Report 1992- PartB*, Hague Consulting Group B.V., the Hague

Hague Consulting Group (1994) *Annual Project Review Report 1994*, Hague Consulting Group B.V., the Hague

Hamerslag, R. (1978) *Verkeerskundige Modellen I and II*, lecture notes (in Dutch), code e30, Delft University of Technology

Hamerslag, R. and Immers, L.H. (1988) Estimation of Trip Matrices: Shortcomings and Possibilities for Improvement, *Transportation Research Record*, No. 1203, pp. 27-39, National Research Council

Hauer, E. (1979) Correction of License Plate Surveys for Spurious Matches, *Transportation Research A*, Vol. 13A, pp. 71-78

Kalman, R.E. (1960) A New Approach to Linear Filtering and Prediction, *Journal of Basic Engineering (ASME)*, 82D, pp. 35-45

Kanayama, K., Fujikawa, Y., Fujimoto, K. and Horino, M. (1991) Development of Vehicle-license Number Recognition System Using Real-time Image Processing and its Application to Travel-Time Measurement, *Proc. of 41st IEEE Vehicular Technology Conference*, IEEE Service Center, USA, pp. 798-804

Keller, H. and Ploss, G. (1987) Real-Time Identification of O-D Network Flows from Counts for Urban Traffic Control, *Proceedings 10th. Symposium on Traffic Theory*

Kikuchi, S., Nanda, R. and Perincherry, V. (1993) A Method to Estimate Trip O-D Patterns Using a Neural Network Approach, *Transportation Planning and Technology*, Vol. 17, pp. 51-65

Kryger, P. and Ottesen K.A. (1956) Van the License Plate Method be Used for Traffic Studies?, *Traffic Quarterly*, Vol. 10, pp. 377-386

Kwon, E. and Stephanides, Y.J. (1994) Comparative Evaluation of Adaptive and Neural-Network Exit Demand Prediction for Freeway Control, *Transportation Research Records*, No. 1446, pp. 66-76

Lehmann E.L. (1983) *Theory of Point Estimation*, Wiley&Sons

Lipschutz, S. (1968) *Theory and Problems of Linear Algebra,*, Schaum's outline series, McGraw-Hill

Ljung, L. and Söderström, T. (1983) *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA

Lootsma, F.A. (1984) *Algorithms for Unconstrained Optimization*, Lecture Notes of the dep. of Mathematics and Informatics, Delft Univ. of Technology

Lucas, C.F. (1986) The NOPCOP Range of Number Plate Comparison Programs, *User Guide*, Transport & Road Research Laboratory

Maher, M.J. (1983) Inferences on Trip Matrices from Observations on Link Volumes: a Bayesian Statistical Approach, *Transportation Research-B*, Vol. 17B, No. 6, pp. 435-447

Maher, M.J. (1985) The Analysis of Partial Registration-plate Data, *Traffic Engineering and Control*, Vol. 26, pp. 495-497

Makowski, G.G. and Sinha, K.C. (1976) A Statistical Procedure to Ananlyze Partial License Plate Numbers, *Transportation Research*, Vol. 10, pp. 131-132

Mood, A.M., Graybill, F.A. and Boes, D.C. (1963) *Introduction to the Theory of Statistics*, 3rd edition, McGraw W-Hill International Book Company

MVA Systematica (1987) MVMACH license plate matching program, *User Guide*

Nihan, N.L. and Davis, G.A. (1987) Recursive Estimation of Origin-Destination Matrices from Input/Output Counts, *Transportation Research B*, Vol. 21B, No. 2, pp. 149-163

Nihan, N.L. and Davis, G.A. (1989) Application of Prediction-Error Minimization and Maximum Likelihood to Estimate Intersection O-D Matrices from Traffic Counts, *Transportation Science*, Vol 23, No. 2

Ortúzar, J. D. and Willumsen, L.G. (1990) *Modelling Transport*, Second Edition, JohnWiley and Sons, Chichester, England, ISBN 0 471 94193 X

Papageorgiou, M. (1991) *Concise Encyclopedia of Traffic & Transportation Systems*, Pergamon Press

Ping Yu and Davis, G.A. (1994) Estimating Freeway Origin-Destination Patterns Using Automatic Traffic Counts, *Transportation Research Board, Preprint 940112*, 73rd ann. meeting, Washington D.C.

Ploss, G., Philipps, P., Inaudi, D. and Keller, H. (1990) MOTION-A New Traffic Control Concept Based on Real-Time Origin-Destination Information, *proc. of the 11th Int. Symp. on Transp. and Traffic Theory*, Elsevier

Pursula, M. and Pastinen, V. (1993) A Bayesian Approach to Update Traffic Flows From Traffic Counts, *Proc. of the twelfth Int. Symp. on Transportation and Traffic Theory*, C.F.Daganzo (ed.), Elsevier, pp. 507-522

Rijsdijk, J. (1995) *OD-estimation on motorway networks*, masters thesis (in Dutch), Civil Engineering Department, Delft University of Technology

Sage, A.P. and Melsa J.L. (1971) *Estimation Theory with Apllications to Communications and Control*, McGraw-Hill Book Company

Sheffi, Y. (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice Hall

Sherali, H.D., Sivanandan, R. and Hobeika A.G. (1994) A Linear Programming Approach for Synthesizing Origin-Destination Tables from Link Traffic Volumes, *Transportation Research B*, Vol. 28B, No. 3, pp. 213-233

Shewey, P.J.H. (1983) An Improved Algorithm for Matching Partial Registration Numbers, *Transportation Research B*, Vol. 17B, No. 5, pp. 391-397

Shih-Miao Chin, Ho-Ling Hwang and Tzusheng Pei (1994) Using Neural Networks to Synthesize Origin-Destination Flows in a Traffic Circle, *Transportation Research Board, preprint 940353*, 73rd ann. meeting, Washington D.C.

VanAerde, M., Hellinga, B. and MacKinnon, G. (1993) QUEENSOD: A Method for Estimating Time Varying Origin-Destination Demands fro Freeway Corridors/Networks *presented at TRB annual meeting*, Washington DC

Van Der Zijpp, N.J. and Hamerslag, R. (1993) The Real Time Estimation of Origin-Destination Matrices for Freeway Corridors, *Proc. of the 26th International Symposium on Automotive technology and Automation ISATA*, Aachen, Germany

Van Der Zijpp, N.J. (1993) Optimal Road Management Using a Traffic Database, *Colloquim Vervoersplanologisch Speurwerk*, (in Dutch), Rotterdam

Van Der Zijpp, N.J. and Hamerslag, R. (1994a) A Bayesian Approach to Estimate Origin-Destination Matrices for Freeway Corridors, *presented at the Universities Transport Study Group 1994 Conference*, Leeds

Van Der Zijpp, N.J. and Hamerslag, R. (1994b) An Improved Kalman Filtering Approach to Estimate Origin-Destination Matrices for Freeway Corridors, *Transportation Research Records*, No. 1443, pp. 54-64

Van Der Zijpp, N.J. and Bovy, P.H.L. (1994a) Modelling Traveller Behaviour, *DYNA, Annual Progress Report 1994*, DRIVE project V2036, workpackage G, Hague Consulting Group, the Hague

Van Der Zijpp, N.J. and Bovy, P.H.L. (1994b) Modelling Driver Information Acquisition Behaviour in Route Forecasting, *Proc. of the 1st Erasmus-network Conference on Transportation and Traffic Engineering*, in: Mortelmans, J.F. (ed), ACCO, Leuven

Van Der Zijpp, N.J. and Bovy, P.H.L. (1994c) Driver Information Acquisition Behaviour, a Modelling Approach, *Proc. of the 27th International Symposium on Automotive technology and Automation, ISATA*, Aachen, Germany

Van Der Zijpp, N.J. and de Romph E. (1994) A Dynamic Traffic Forecasting Application on the Amsterdam Beltway, *Proc. of the Second DRIVE-II Workshop on Short-Term Traffic Forecasting*, in: VanArem, B., Kirby, H.R. and Whittaker, J.C. (ed), report INRO-VVG 1994-19, Delft

Van Der Zijpp, N.J. and de Romph E. (1994) A Dynamic Traffic Forecasting Application on the Amsterdam Beltway, *Proc. of the Second DRIVE-II Workshop on Short-Term Traffic Forecasting*, in: VanArem, B., Kirby, H.R. and Whittaker, J.C. (ed), report INRO-VVG 1994-19, Delft

Van Der Zijpp, N.J. and Heydecker, B. (1996) How many Parameters should a Traffic Model have?, *UTSG, ann.conference*, University of Huddersfield

Van Der Zijpp, N.J. (1996) Dynamic OD-Matrix Estimation from Combined Data, *Transportation Modelling for Tomorrow*, P.H.L. Bovy (ed.), Delft University

Van Der Zijpp, N.J. and de Romph E. (1996) A Dynamic Traffic Forecasting Application on the Amsterdam Beltway, *Int. Journal of Forecasting*, 1996, forthcoming

Van Vuren, T. (1984) *The Estimation of Origin-Destination Triptables with Partial Matrix Techniques: an Overview*, Masters Thesis (in Dutch), Delft University of Technology, Faculty of Civil Engineering, Transportation Section

Van Zuylen, H.J. and Willumsen, L.G. (1980) The most Likely Trip Matrix Estimated from Traffic Counts *Transportation Research-B*, Vol. 14B, pp. 281-293

Vythoulkas, P.C. (1993) Alternative Approaches to Short Term Traffic Forecasting for Use in Driver Information Systems, *Proc. of the twelfth Int. Symp. on Transportation and Traffic Theory*, C.F. Daganzo (ed), Elsevier Science Publishers, pp. 485-506

Watling, D.P. and Maher, M.H. (1988) A Graphical Procedure for Analysing Partial Registration-plate Data, *Traffic Engineering and Control*, Vol. 29, pp. 515-519

Watling, D.P. (1990) *The Statistical Analysis of Partial Registration Plate Data*, Ph.D. Thesis, Dep. of Probability and Statistics, University of Sheffield

Watling, D.P. (1994) Maximum Likelihood Estimation of an Origin-Destination Matrix from a Partial Registration Plate Survey, *Transportation Research B*, Vol. 28B, No. 4, pp. 289-314

Watling, D.P. and Maher, M.J. (1992) A Statistical Procedure for Estimating A Mean Origin-Destination Matrix from a Partial Registration Plate Survey, *Transportation Research B*, Vol. 26B, No. 3, pp. 171-193

Westerman, M. (1994) *Inwinning en Verwerking van Dynamische Verkeersgegevens*, report , Delft University of Technology

Westerman, M. (1995) *Real-time Traffic Data Collection for Transportation Telematics*, PhD Thesis, Delft University of Technology

Willumsen, L.G. (1984) Estimating Time-Dependent Trip Matrices from Traffic Counts *Ninth Int. Symposium on Transportation and Traffic Theory*Volmuller and Hamerslag eds., VNU Science Press, pp. 397-411

Yang, H., Akiyama, T. and Sasaki, T. (1992) A Neural Network Approach to the Identification of Real Time Origin-Destination Flows from Traffic Counts, *Proc. of the Int. Conf. on Artificial Intelligence Applications in Transportation Engineering*, pp. 253-269

Appendix A: Constrained minimization of quadratic functions

A.1 Introduction

A number of the methods described in chapter 3, require the solving of an inequality constrained quadratic minimization problem. Since some of these problems tend to be computational demanding a considerable amount of attention has been paid to the implementation of efficient algorithms. The findings on this subject are reported in this appendix, and result in the recommendation of two algorithms for the constrained minimization of quadratic functions. The first method finds the exact solution using an iterative search strategy with projected conjugate search directions. This method is described in section A.4. The second method finds an approximated solution using considerably less computation time, and is described in section A.5. The exact algorithm will be particularly useful to evaluate different methods for split-estimation in a laboratory environment. The second algorithm facilitates large scale application of split-estimation methods.

A.2 General problem description

The general problem under consideration is the following constrained quadratic minimization problem:

minimize: $$J= - 2\Psi' b + b'\Omega b$$

subject to either:

-a- $$\mathbf{0}\leq b\leq\mathbf{1}$$

or:

-b- $$\mathbf{0}\leq b \text{ , } \mathrm{p}' b\leq\mathbf{1} \qquad \text{(A.1)}$$

where p is a repeating column matrix of appropriate size. If the constraints (A.1-a) apply then all constraints are orthogonal. If the constraints (A.1-b) apply then (m-1) non-orthogonal constraints exist, represented by p$'b\leq$**1**.

A.3 Interior steepest descent

The first method that was tested represents a well known technique known as *interior steepest descent*. This method starts at an arbitrary point in the feasible region, $\overline{b}^{(0)}$, and then iteratively approaches a solution. Each iteration a new search direction, $s^{(k)}$, is determined and a line minimization in that search direction is performed, leading to the next iteration point $\overline{b}^{(k+1)}$, i.e.:

$$\overline{b}^{(k+1)}=\overline{b}^{(k)} + \mu s^{(k)} \tag{A.2}$$

The search direction is obtained by projecting the steepest descent direction on the space of feasible directions. The space of feasible search directions is the intersection of all half-spaces defined by the binding constraints.

For problem (A.1-a) the projection of the negative gradient on this space is quite simple due to the orthogonality of the constraints. For this purpose we use lemma A.1 that states that the projection on the intersection of orthogonal halfspaces can be computed by the sequential projection on each halfspace. i.e. the projection comes down to setting those elements of the search vector to zero that correspond with violations of the binding constraints.

Lemma A.1: Projection on the intersection of orthogonal halfspaces

If P is the projection operator, i.e. $\mathrm{P}^{F}(\mathrm{x})$ is projection of x on the space F, and if $\{F^1, F^2,\ldots F^q\}$ are halfspaces defined by mutually orthogonal constraints, then:

$$\mathrm{P}^{F^1\cap F^2\cap \ldots\cap F^q}(x)=\mathrm{P}^{F^1}(\ \mathrm{P}^{F^2}\ (\ldots\ \mathrm{P}^{F^q}(x)\ldots)) \tag{A.3}$$

Proof The constraints that define $F^1\cap F^2\cap \ldots\cap F^q$ can be written in the form $Cx\geq 0$ with C a matrix of height q and width corresponding to the length of x. Assume $Cx>0$ (Otherwise, without losing generality drop the constraint that is not violated).

Since $c_p.c_q=0$ the projection of x, x^q, satisfies $Cx^q=0$. The projection of x is hence given by:

$$x^q = x\text{-}C'(CC')^{-1}Cx \tag{A.4}$$

Since the rows of C are orthogonal, equation (A.4) can be rewritten as:

$$x^q = x - (\ c_1'c_1/c_1c_1'+ c_2'c_2/c_2c_2' +\ldots+ c_q'c_q/c_qc_q'\ )\ x \tag{A.5}$$

This is exactly the result that is obtained by the sequence:

$$\begin{gathered} x^0 = x \\ x^p = x^{p-1} - (\ c_p'c_p\ /\ c_pc_p'\ )x^{p-1} \\ p=1,2,\ldots q \end{gathered} \tag{A.6}$$

This sequence defines the consecutive projections on $F^1, F^2,\ldots F^q$. **End of proof**

For problem (A.1-b) the problem of projecting the gradient is slightly more complex since it is not trivial in advance which constraints will be active after the projection. As an example consider figure A.1. In this example the non-negativity constraint e^2 may or may not be active after the projection. However there are no more than m-1 binding non-orthogonal constraints represented by the equation $\mathrm{p}'b\leq\mathbf{1}$. It can be shown that the projected gradient can be obtained by first projecting on the non-orthogonal constraints, and subsequently on a combination of the non-orthogonal and orthogonal constraints, to prevent the latter constraints from being violated. Algorithm A.1 summarizes this approach.

After the search direction has been determined, a line minimization is performed. This is done by substituting (A.2) in (A.1). This leads to a quadratic equation in μ, which is convex due to the non-negativity of the matrix Ω Consequently a unique solution exists for the stepsize μ. To ensure that the next iteration point is in the interior of the feasible region, all constraints are checked. If any constraint is violated then μ is set to the maximum value that does

Algorithm A.1: Find search direction

```
Procedure find search direction
Step 1
    set $s$ as the negative gradient
    determine the set $B$ of binding constraints
Step 2
    determine the set $V$ of violated constraint by $s$
    if $B \cap V \neq \emptyset$
        redefine $s$ as the projection of $s$ on the constraints in the set $B \cap V$
        repeat step 2
    else
        stop
```

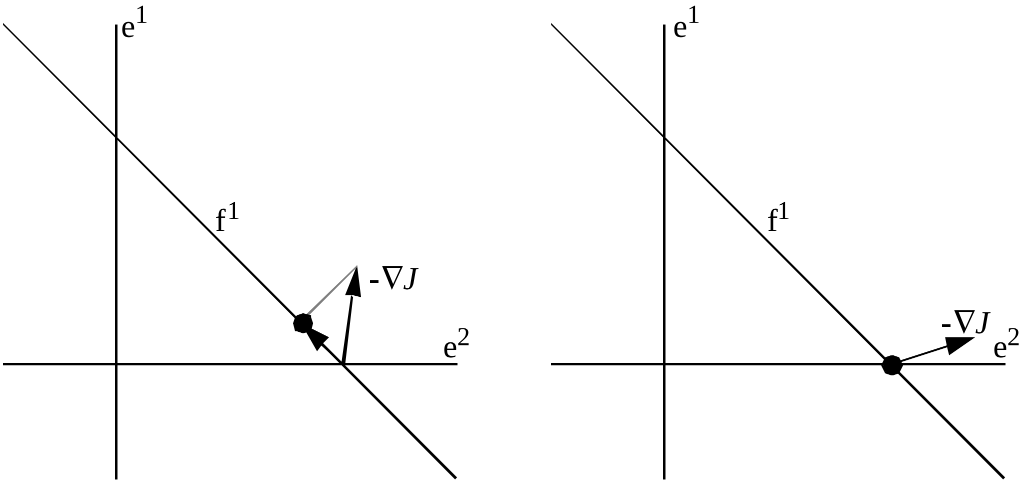


Figure A.1: Projection of the negative gradient on non-orthogonal constraints. In the first case (left) e^2 is not active, in the second case e^2 is active

not lead to violation of any constraints.

The interior steepest descent method has the advantage of ease of implementation and robust performance. However it is also known for its slow convergence. This problem deteriorates if the matrix Ω is *ill-conditioned*. Ill conditioning occurs if the rows of a matrix are nearly dependent. Translated to the practical case of estimating EE-matrices, this means that if little variations in entry volumes occur, the corresponding matrix Ω is ill-conditioned and the computation time needed to solve (A.1) increases.

Practical experiments indicate that computation times can easily exceed the time available if results need to be produced in real-time.

A.4 Projected conjugate gradients

Unconstrained conjugate gradient method

In unconstrained minimization of quadratic functions, the conjugate gradient method is a very effective method. A conjugate gradient method consists of a sequence of line searches, each search in a conjugate direction relative to Ω, i.e. if $p \neq q$,

$$s^{(p)}\Omega s^{(q)}=0 \tag{A.7}$$

It can be shown that, in case of exact line minimizations and absence of rounding errors, the exact minimum of (A.1) can be found in *N* steps, where *N* is the dimension of Ω, for details see *Lootsma (1984)*. A problem remains the determination of conjugate search directions. For this purpose a number of possibilities exist. An elegant way to generate conjugate search directions was proposed by *Fletcher and Reeves (1964)*, and is given by:

$$s^{(k)}(t) = -\nabla \mathrm{J}(\bar{b}^{(k)}(t)) + \frac{\left\| \nabla \mathrm{J}(\bar{b}^{(k)}(t)) \right\|^2}{\left\| \nabla \mathrm{J}(\bar{b}^{(k-1)}(t)) \right\|^2} s^{(k-1)}(t) \tag{A.8}$$

The advantage of this method is that only the last search direction needs to be stored, and that the method is easy to implement.

Projected Conjugate Gradient (PCG) method

To apply a method such as described above to the constrained problem (A.1) a projection strategy can be used similar to the method used with the interior steepest descent method. If the search direction is violating a binding constraint then it is truncated. However the truncation causes the new search direction to be no longer conjugate to the previous ones.

A response to this is to 'reset' the method to a steepest descent direction every time a new constraint becomes binding. As long as no alterations are made in the binding constraints set the theory described above is still valid and the solution will be found in less than *mn* steps. However, experiments have shown that in some cases constraints keep jumping in and out of the binding constraints set, causing the method to behave very similar to the steepest descent method, resulting in slow convergence.

A less well motivated but more satisfactory method in practice is to refrain from resetting to steepest descent at least during a number of searches.

Stop criterion

Another practical aspect is the stopping criterion. This aspect not only applies to the conjugate gradient method but also to iterative search methods like the interior steepest descent method. As in the constrained case no real conjugate directions are used and because rounding errors cause the line-minimizations and search directions to be inexact, the method will usually not find the exact solution of the minimization problem. Instead the method keeps finding solutions with lower target values, asymptotically touching the true minimum.

A stop criterion is needed to check if the real solution has sufficiently been approached. This involves making a trade-off between computation-time and accuracy. Also an analysis is needed of the accuracy that is attainable, given the machine accuracy and the structure of the problem. A number of possibilities exist:

- *Checking for the relative improvement of the objective function*. As long as the objective function significantly improves convergence is not reached. The relative improvement is however no sufficient condition for convergence: a number of steps with slow improvement can be alternated with steps of large improvement.
- *Checking for change in the solution vector.* Also this is a necessary but not a sufficient condition for convergence, due to the same argument as above. Above that, care should be

taken imposing this convergence criterion, as in cases with ill conditioned matrices the machine accuracy might prevent the algorithm from reaching the stop criterion resulting in an infinite loop in the program.

- *Checking for a maximum number of iterations*. This is neither a necessary nor a sufficient condition. In practice it can however be used in combination with above conditions. Especially this condition is useful as a 'safety catch', if above conditions cannot be reached this stop criterion will eventually abort the program.
- *Checking the theoretical optimality condition*s. The *Karush-Kuhn-Tucker* (KKT) theorems, see *Bazaraa et al. (1993)*, provide a necessary and sufficient criterion for optimality. This optimality check involves the inversion of a matrix with a height that corresponds with the number of binding constraints (maximum *mn*) and is therefore potentially expensive.

The optimization method that is implemented contains a mix of above stopping conditions. First a check is performed for change in the solution vector. If this change is below a certain threshold then the theoretical optimality conditions are checked. To prevent the algorithm stranding on degenerate problems, also a check for a maximum number of iterations is included.

Smart initialization

The EE-splits are re-evaluated each time slice. By initializing the iterative procedure with the solution that was obtained during the previous time slice, fast convergence is stimulated.

A.5 Iterative solving

The procedures described above are iterative search methods, aimed at generating the exact solution to the minimization problem (A.1). In this section a heuristic algorithm is presented that will either find the exact or the approximated solution, using a drastically reduced computation time relative to the exact methods.

The idea of this method is simple: first the unconstrained solution to (A.1) is computed. This can be done with one matrix inversion. Then a check is performed for any violated non-negativity constraints. If any constraints are violated, then a linear equality constraint is imposed on the corresponding variables, implicitly assuming that for these variables the non-negativity constraint will be binding in the optimal solution. This process is repeated until a solution is found that does not violate any non-negativity constraints.

There is no check performed on the violation of unitary constraints, if necessary the final solution is truncated by setting all elements exceeding one to the unitary value. The ratio behind this is that the split parameters will generally be closer to zero then to one, and that there is only a small probability that the unconstrained solution to (A.1) will violate the unitary constraint. Algorithm A.2 summarizes the approach.

If the unconstrained solution violates none or one non-negativity constraint, this algorithm finds the exact solution in at most two steps. If the unconstrained solution violates multiple non-negativity constraints then the solution that is found may be suboptimal, as is shown in figure 11.1. This figure contains a constructed example of a quadratic function of which the unconstrained minimum violates two non-negativity constraints, but for which in the constrained solution only *one* constraint is binding. Therefore the solution that is generated by the iterative solving method must be considered as an approximated solution, that may be inexact if the unconstrained solution violates multiple non-negativity constraints.

Algorithm A.2: Iterative solving

```
Iterative solving procedure
Step 1
        Solve the unconstrained problem
Step 2
        if no elements of the solution vector are negative:
                stop
        else
                introduce a constraint for each negative element,
                setting this element to zero
Step 3
        Solve the problem again, using the extra constraints from step 2.
        Go to step 2
```

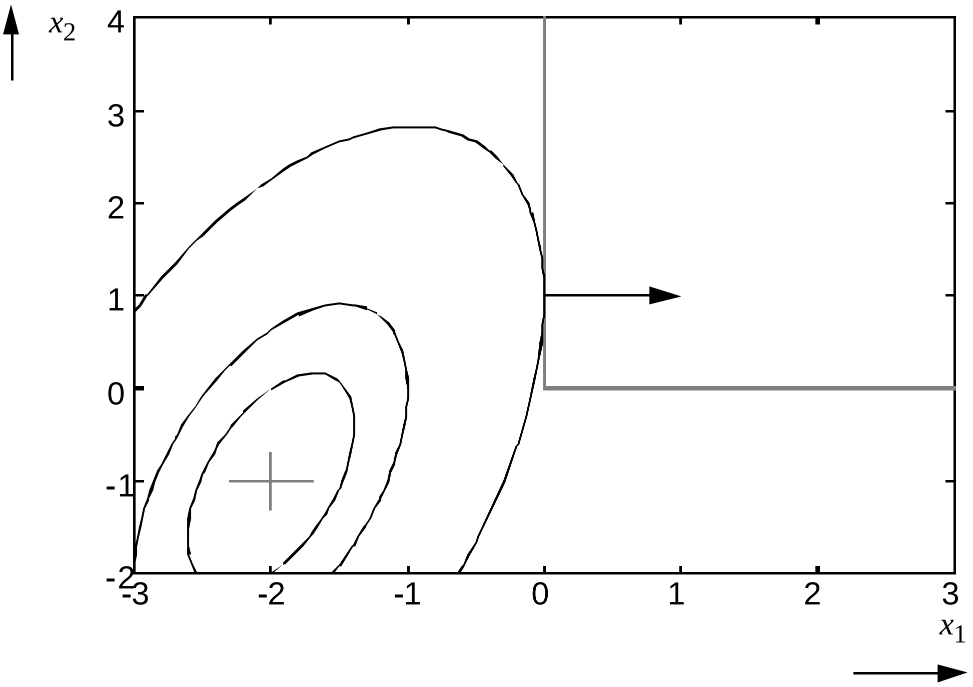


Figure 11.1: Example of a quadratic function for which the method of iterative solving does not find the exact minimum.

Appendix B: Evaluation of the Mean and Variance of TMVN Distributions

This appendix contains numerical approximations for the normalization constant, true mean, and true variance associated with the truncated normal distribution that has been widely used in this thesis. These quantities have been expressed in terms of the well known error function, i.e.:

$$\mathrm{erf}(x) \;=\; \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,dt \tag{B.1}$$

For this function a number of numerical approximations exist, see for example *Abramowitz and Stegun (1968)*. An approximation that seems sufficiently accurate for our purposes is:

$$\begin{gathered}\mathrm{erf(x)}= \;1-\exp(-x^2)\,(ta_1+a_2t^2+a_3t^3)\;+\varepsilon(x)\;, \qquad t\;=\;\frac{1}{1+px}\\ |\varepsilon(x)|\le 2{,}5\ ´\ 10^{-5}\\ p\;=\;0{,}47047 \qquad a_1\;=\;0{,}3480242 \qquad a_2\;=\;-0{,}0958798 \qquad a_3\;=\;0{,}7478556\end{gathered} \tag{B.2}$$

However, in most computing environments, a library routine will be available that evaluates the error function with an even better accuracy.

Lemma B.1: Normalization constant of a truncated normal random variable. Let x be a random variable with a truncated normal distribution with parameters μ and σ^2, i.e.:

$$\mathrm{p}[x]=\frac{1}{\mathrm{C}(\mu,\,\sigma)\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad 0\le x\le 1 \tag{B.3}$$

Then the value of the normalization constant satisfies:

$$\mathrm{C}[\mu{,}\sigma] \;=\; \frac{1}{2}\mathrm{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) \;+\; \frac{1}{2}\mathrm{erf}\left(\frac{1-\mu}{\sigma\sqrt{2}}\right) \tag{B.4}$$

Proof The normalization constant is defined as:

$$C[\mu,\sigma] = \int_0^1 \frac{1}{\sigma\sqrt{2\pi}}\left(\exp(-\frac{(x-\mu)^2}{2\sigma^2})dx\right) \tag{B.5}$$

To simplify this integral we use a change of coordinates. Set:

$$t = \frac{x-\mu}{\sigma\sqrt{2}}, \text{ i.e.: } x = \sigma t\sqrt{2}+\mu \tag{B.6}$$

Consequently (B.5) changes in:

$$C[\mu,\sigma] = \frac{1}{\sqrt{\pi}}\int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} \exp(-t^2)dt \tag{B.7}$$

Result (B.4) now follows from decomposing (B.7) in two separate integrals. Note that this result also applies if μ is negative or larger than one. This follows from the property that erf(-*x*)=-erf(*x*).

End of proof

As an illustration the value of normalization constant has been plotted in figure B.1 for a range of values for μ and σ^2.

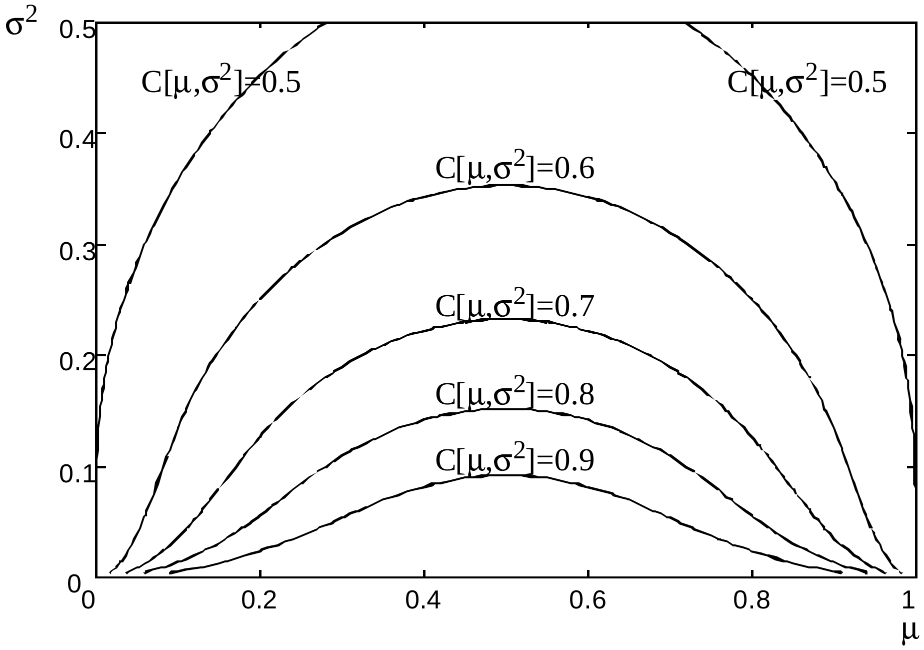


Figure B.1: Contourplot of the normalization constant, as a function of the parameters μ *and* σ^2 *of a truncated normal distribution.*

Lemma B.2: Let x be a random variable with a truncated normal distribution with parameters μ and σ^2, then the expected value of x satisfies:

$$\mathrm{E[x]}= \frac{\sigma}{C(\mu,\sigma)\sqrt{2\pi}}\left(\exp(-\frac{\mu^2}{2\sigma^2})-\exp(-\frac{(-\mu+1)^2}{2\sigma^2})\right) + \mu \tag{B.8}$$

Where C[μ,σ] is given by (B.4).

Proof The expected value of a truncated normal variable is defined as:

$$\mathrm{E[x]} = \int_0^1 \frac{x}{C(\mu,\sigma)\sigma\sqrt{2\pi}}\exp - \frac{(x-\mu)^2}{2\sigma^2}dx \tag{B.9}$$

From applying a change of coordinates, it follows that

$$\mathrm{E}[x]= \frac{1}{C(\mu,\sigma)\sigma\sqrt{\pi}}\int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} (\sigma t\sqrt{2}+\mu)\exp(-t^2)dt$$

$$= \frac{-\sigma}{C(\mu,\sigma)\sqrt{2\pi}}\int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} -2t\exp(-t^2)dt + \frac{\mu}{C(\mu,\sigma)}\frac{1}{\sqrt{\pi}}\int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} \exp(-t^2)dt \tag{B.10}$$

Observing that: $-2t\exp(-t^2) = \frac{d}{dt}\exp(-t^2)$, and combining with (B.7) leads to the required result. **End of proof**

In figure B.2 the true mean is plotted for a range of values of μ and σ^2. As an extra check on result (B.8) the true mean has also been evaluated by numerical integration for a number of points.

Lemma B.3: Let x be a random variable with a truncated normal distribution with parameters μ and σ^2, then the variance of x satisfies:

$$\mathrm{var}[x] = -\frac{\sigma^2}{C(\mu,\sigma)\sqrt{\pi}}\left(\frac{\mu}{\sigma\sqrt{2}}\exp(-\frac{\mu^2}{2\sigma^2})-\frac{1-\mu}{\sigma\sqrt{2}}\exp(-\frac{(-\mu+1)^2}{2\sigma^2})\right)+\sigma^2 \ -(\mathrm{E}[x]-\mu)^2 \tag{B.11}$$

Where C[μ,σ] follows from lemma B.1, and E[x] follows from lemma B.2.

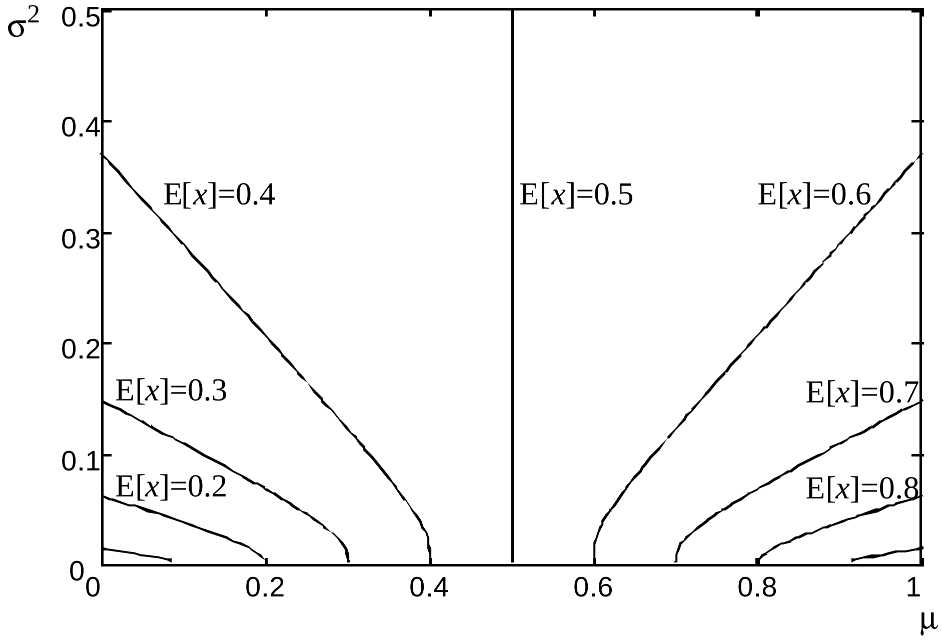


Figure B.2: Contourplot of the true mean, as a function of the parameters μ *and* σ^2 *of a truncated normal distribution.*

Proof The variance of a truncated normal variable is defined as:

$$\mathrm{var}[x] = \mathrm{E}[x^2]\text{-}\mathrm{E}^2[x] = \int_0^1 \frac{x^2}{\mathrm{C}(\mu,\sigma)\sigma\sqrt{2\pi}} \exp -\frac{(x-\mu)^2}{2\sigma^2}\,dx \;\; - \mathrm{E}^2[x] \tag{B.12}$$

From which it follows that:

$$\mathrm{var}[x] = \frac{1}{\mathrm{C}(\mu,\sigma)\sqrt{\pi}} \int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} (\sigma t\sqrt{2}+\mu)^2 \exp(-t^2)\,dt \;\; - \mathrm{E}^2[x]$$

$$=$$

$$\frac{\sigma^2}{\mathrm{C}(\mu,\sigma)\sqrt{\pi}} \int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} 2t^2\exp(-t^2)\,dt + \frac{\sigma\mu\sqrt{2}}{\mathrm{C}(\mu,\sigma)\sqrt{\pi}} \int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} 2t\exp(-t^2)\,dt + \frac{\mu^2}{\mathrm{C}(\mu,\sigma)\sqrt{\pi}} \int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} \exp(-t^2)\,dt$$

$$- \mathrm{E}^2[x] \tag{B.13}$$

Observe that: $-\frac{d}{dt}(t\exp(-t^2)) \;=\; 2t^2\exp(-t^2) - \exp(-t^2)$, and that:

$-\frac{d}{dt}\exp(-t^2) = 2t\exp(-t^2)$. Furthermore apply equation (B.7). Now it follows that:

$$\mathrm{var}[x] = \left.\frac{-t\exp(-t^2)\sigma^2}{C(\mu,\sigma)\sqrt{\pi}}\right|_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} + \frac{\sigma^2}{C(\mu,\sigma)\sqrt{\pi}}\int_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}}\exp(-t^2)dt - \left.\frac{2\sigma\mu\exp(-t^2)}{C(\mu,\sigma)\sqrt{2\pi}}\right|_{\frac{-\mu}{\sigma\sqrt{2}}}^{\frac{1-\mu}{\sigma\sqrt{2}}} + \mu^2 - E^2[x] \tag{B.14}$$

Substituting the boundaries of the integral and applying equation (B.7) once more, results in:

$$\begin{aligned}\mathrm{var}[x] = & -\frac{\sigma^2}{C(\mu,\sigma)\sqrt{\pi}}\left(\frac{\mu}{\sigma\sqrt{2}}\exp(-\frac{\mu^2}{2\sigma^2}) - \frac{1-\mu}{\sigma\sqrt{2}}\exp(-\frac{(-\mu+1)^2}{2\sigma^2})\right) + \sigma^2 \\ & +2\mu\ \frac{\sigma}{C(\mu,\sigma)\sqrt{2\pi}}\left(\exp(-\frac{\mu^2}{2\sigma^2}) - \exp(-\frac{(-\mu+1)^2}{2\sigma^2})\right) + \mu^2 - E^2[x]\end{aligned} \tag{B.15}$$

The first term of the second line in above equation exactly matches: 2μ(E[x]-μ), see equation (B.8). Now the required result follows directly. **End of proof**

In figure B.3, the true variance is plotted as a function of the parameters μ and σ^2 of a truncated normal distribution. As an extra check on result (B.12) the variance has also been evaluated numerically for a number of points.

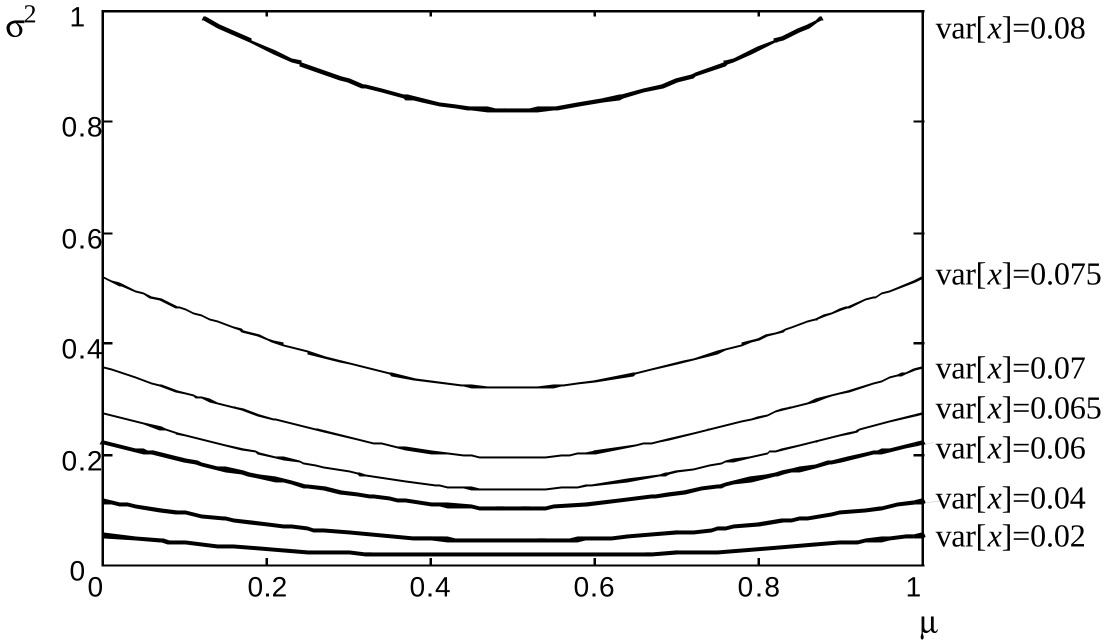


Figure B.3: Contourplot of the true variance, as a function of the parameters μ and σ^2 of a truncated normal distribution.

Above lemma's apply to univariate truncated normal distributions, whereas most truncated distributions used in this thesis are multivariate. A well known property of MVN distributions is that the individual elements of a MVN distributed vector have a normal distribution. In the following example it is checked whether a similar property applies to the marginal density of a TMVN distribution. The example will show that this is generally not the case.

Example B.1: Marginal density associated with a TMVN distribution.

Let x be a two dimensional, zero mean random variable with a TMVN distribution, i.e.:

$$x\sim\text{TMVN}[\mathbf{0},\Sigma],\ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},\ \Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix} \tag{B.16}$$

In this example we will attempt to compute the marginal density of x_1 and are particularly interested in whether or not this density is given by: $x_1\sim\text{TMVN}[0,\sigma_{11}{}^2]$.

The probability distribution of x is given by:

$$\begin{aligned} \text{p}[x] &= \frac{1}{2\pi C(\Sigma)\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix}\begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{22}^2 \end{bmatrix}^{-1}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)\text{I}_{[0,1]}(x_1)\text{I}_{[0,1]}(x_2) \\ &= \frac{1}{2\pi C(\Sigma)\sqrt{|\Sigma|}}\exp\left(-\frac{1}{2}\left(\frac{\sigma_{22}^2 x_1^2 - 2\sigma_{12}x_1x_2 + \sigma_{11}^2 x_2^2}{\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2}\right)\right)\text{I}_{[0,1]}(x_1)\text{I}_{[0,1]}(x_2) \end{aligned} \tag{B.17}$$

The marginal density of x_1 is obtained by integrating p[x] over the value of x_2, i.e.:

$$\begin{aligned} \text{p}[x_1] &= \int_0^1 \text{p}\ [x]\ dx_2 \\ &= \frac{1}{2\pi C(\Sigma)\sqrt{|\Sigma|}}\exp\left(-\frac{(\sigma_{22}^2 - \sigma_{12}^2/\sigma_{11}^2)\,x_1^2}{2\,(\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2)}\right)\int_0^1 \exp\left(-\frac{\sigma_{11}^2\,(x_2 - \sigma_{12}x_1/\sigma_{11})^2}{2\,(\sigma_{11}^2\sigma_{22}^2 - \sigma_{12}^2)}\right)dx_2\ \ \text{I}_{[0,1]}(x_1) \end{aligned} \tag{B.18}$$

After the integration has been performed, the second factor in above equation will change in a factor that still depends on x_1. In fact, according to lemma (B.1) this factor proportional to:

$$\frac{1}{2}\text{erf}\left(\frac{\mu}{\sigma\sqrt{2}}\right) + \frac{1}{2}\text{erf}\left(\frac{1-\mu}{\sigma\sqrt{2}}\right)$$

with:

$$\begin{aligned} \mu &= \sigma_{12}x/\sigma_{11} \\ \sigma^2 &= \sigma_{22}^2 - \sigma_{12}^2/\sigma_{11}^2 \end{aligned} \tag{B.19}$$

Only if σ_{12} amounts zero, i.e. x_1 and x_2 are not correlated, this factor changes in a constant value. In this case (B.18) retains the shape of a TMVN distribution with parameters 0 and $\sigma_1{}^2$. In all other cases deviating distributions arise. We must therefore conclude from this example that in general the marginal distribution of a TMVN distribution is *not* truncated normal.

Appendix C: Mathematical Preliminaries

In chapter 5 various approximations are derived of the covariance matrix that applies to the measurement error that was defined by (5.3). Each approximation is characterized by the information on which the matrix will be conditioned. This appendix contains a number of lemmas and some notational conventions that help to facilitate the derivations in chapter 5.

C.1 Properties of expectation and covariance

The derivation of the conditional expectations and covariance matrices mentioned in the introduction are straightforward, but require some lengthy derivations. The following two lemmas help to keep the length of the derivations in chapter 5 to a minimum.

Lemma C.1: Let **x** and **y** be (univariate) random variables, of which the joint density $p_{\mathbf{x},\mathbf{y}}(x,y)$ is defined for all pairs $\{x, y|x\in X, y\in Y\}$. Let the operators $E_{\mathbf{x}}[.]$ and $E_{\mathbf{y}}[.]$ be defined as the expected value of its operands with respect to distributions **x** and **y** respectively, and let $E_{\mathbf{x}|\mathbf{y}}[.|y]$ be defined as the expected value operator with respect to the conditional density **x** given $\mathbf{y} = y$. Furthermore let the random variables **x** and **y** satisfy all necessary conditions in order for $E_{\mathbf{x}}[.]$, $E_{\mathbf{y}}[.]$, and $E_{\mathbf{x}|\mathbf{y}}[.|y]$ to exist. Then:

$$E_{\mathbf{x}}[x] = E_{\mathbf{y}}[\ E_{\mathbf{x}|\mathbf{y}}[x|y]\ ]. \tag{C.1}$$

Proof (derived from *Sage and Melsa, 1971*)

$$\begin{aligned} E_{\mathbf{x}}[x] &= \int_X x\, p_{\mathbf{x}}(x)\, dx \\ &= \int_X\int_Y x\, p_{\mathbf{x},\mathbf{y}}(x,y)\, dxdy \\ &= \int_Y\int_X x\, p_{\mathbf{x}|\mathbf{y}}(x|y)\, p_{\mathbf{y}}(y)\, dxdy \\ &= \int_Y E_{\mathbf{x}|\mathbf{y}}[x|y]\, p_{\mathbf{y}}(y)\, dy \\ &= E_{\mathbf{y}}[\ E_{\mathbf{x}|\mathbf{y}}[x|y]\ ] \end{aligned}$$

A similar lemma applies to multivariate random variables. and to scalar or vector valued functions of random variables, i.e. if f(.) is a function of defined on the domain X then it holds that:

$$E_{\mathbf{x}}[f(x)]=E_{\mathbf{y}}[\ E_{\mathbf{x}|\mathbf{y}}[f(x)|y]\ ] \tag{C.2}$$

Moreover parallel to lemma C.1 it can be shown that if ***z*** is a random variable:

$$E_{\mathbf{x}|\mathbf{z}}[x|z]=E_{\mathbf{y}|\mathbf{z}}[\ E_{\mathbf{x}|\mathbf{y},\mathbf{z}}[x|y,z]\ ] \tag{C.3}$$

Lemma C.2: (Corollary): Let the covariance operator, $\text{cov}_{\mathbf{x}}[.,.]$, be defined by: $\text{cov}_{\mathbf{x}}[x,x']=\text{E}_{\mathbf{x}}[x.x']-\text{E}_{\mathbf{x}}[x]\text{E}_{\mathbf{x}}[x']$, and let the conditional covariance operator, $\text{cov}_{\mathbf{x}|\mathbf{y}}[.,.|y]$ be defined by: $\text{cov}_{\mathbf{x}|\mathbf{y}}[x,x'|y]=\text{E}_{\mathbf{x}|\mathbf{y}}[x.x'|y]-\text{E}_{\mathbf{x}|\mathbf{y}}[x|y]\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]$, then:

$$\text{cov}_{\mathbf{x}}[x,x'] = \text{E}_{\mathbf{y}}[\ \text{cov}_{\mathbf{x}|\mathbf{y}}[x,x'|y]\ ] + \text{cov}_{\mathbf{y}}[\ \text{E}_{\mathbf{x}|\mathbf{y}}[x|y],\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]\ ] \tag{C.4}$$

Proof By definition it holds that:

$$\text{cov}_{\mathbf{x}}[x,x']=\text{E}_{\mathbf{x}}[x.x']-\text{E}_{\mathbf{x}}[x]\text{E}_{\mathbf{x}}[x']$$

As a result of lemma C.1 this transforms in:

$$\text{cov}_{\mathbf{x}}[x,x']=\text{E}_{\mathbf{y}}[\ \text{E}_{\mathbf{x}|\mathbf{y}}[x.x'|y]\ ]-\ \text{E}_{\mathbf{y}}[\text{E}_{\mathbf{x}|\mathbf{y}}[x|y]].\text{E}_{\mathbf{y}}[\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]]$$

Some rearranging gives:

$$\begin{gathered}\text{cov}_{\mathbf{x}}[x,x']=\ \text{E}_{\mathbf{y}}[\ \text{E}_{\mathbf{x}|\mathbf{y}}[x.x'|y]-\ \text{E}_{\mathbf{x}|\mathbf{y}}[x|y].\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]\ ]\\ +\ \text{E}_{\mathbf{y}}[\text{E}_{\mathbf{x}|\mathbf{y}}[x|y]\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]]-\ \text{E}_{\mathbf{y}}[\text{E}_{\mathbf{x}|\mathbf{y}}[x|y]].\text{E}_{\mathbf{y}}[\text{E}_{\mathbf{x}|\mathbf{y}}[x'|y]]\end{gathered}$$

which is by definition equal to the required result. **End of proof**

Lemma C.1 states that the expected value of a random variable **x** is equal to the expected value of a random value $\text{E}_{\mathbf{x}|\mathbf{y}}[x|y]$, i.e. the r.v. that is defined by the conditional expectation of **x** based on an (arbitrary) random variable **y**. This result will turn out to be particularly useful if we need to compute $\text{E}_{\mathbf{x}|\mathbf{z}}[x|z]$ where z is an observed r.v. while we only have the conditional distributions $\text{p}_{\mathbf{x}|\mathbf{y}}$ and $\text{p}_{\mathbf{y}|\mathbf{z}}$ at our disposal. If in addition to this we may assume:

$$\text{p}_{\mathbf{x}|\mathbf{y}}(x|y)=\text{p}_{\mathbf{x}|\mathbf{y},\mathbf{z}}(x|y,z),\ \forall x,\ y,\ z \tag{C.5}$$

i.e. at all times r.v. **y** is a *sufficient statistic* (see *Mood et al., 1963*) of **z** for r.v. **x**, then the following holds:

$$\text{E}_{\mathbf{x}|\mathbf{z}}[x|z]=\text{E}_{\mathbf{y}|\mathbf{z}}[\ \text{E}_{\mathbf{x}|\mathbf{y},\mathbf{z}}[x|y,z]\ ]=\text{E}_{\mathbf{y}|\mathbf{z}}[\ \text{E}_{\mathbf{x}|\mathbf{y}}[x|y]\ ] \tag{C.6}$$

This equation is the key to obtain the conditional covariance matrix for $y(t)$ in the motorway model from figure 2.3 since the dependencies in this model define various sufficient statistics, e.g. $q\tilde{}(t)$ is a sufficient statistic of $q(t)$.

To complete this part on the properties of the expectation and covariance, two more lemma's are given. These lemma's relate to well known properties of the expectation and covariance operator. For a proof of the scalar case, see for example *Mood et al. (1963)*.

Lemma C.3: For any vector valued random variable **x** for which the first two moments exist, holds:

$$\text{E}[x\ x']=\text{cov}[x,x']+\text{E}[x]\text{E}[x'] \tag{C.7}$$

Lemma C.4: For any two vector valued random variables **x** and **y** for which the first two moments exist, and matrices A and B of appropriate size, holds:

$$\text{cov}[Ax,By]=A\ \text{cov}[x,y]\ B' \tag{C.8}$$

C.2 Matrix operators

In order to get a compact notation, some special matrices and matrix operators are needed in the later chapter 5. First introduce the linear operator diag(.). This operator takes a vector as its argument and puts the elements of this vector on the diagonal of a matrix, i.e.

Definition: diag(.) operator

$$D=\mathrm{diag}(d) \Leftrightarrow D_{ij}=d_i\,\delta_{ij},\ \forall i, j \qquad \text{(C.9)}$$

A property of the diag(.) operator that is summarized in the next lemma:

Lemma C.5: For arbitrary vectors a and b of equal length it holds that:

$$\mathrm{diag}(a)\ b = \mathrm{diag}(b)\ a \qquad \text{(C.10)}$$

Proof $c = \mathrm{diag}(a)\ b \Leftrightarrow c_i = a_i b_i,\ \forall i \Leftrightarrow c = \mathrm{diag}(b)\ a.$

End of proof

Note that diag(.) can not be replaced by a series of matrix multiplications. This can be seen as follows: for an arbitrary product of two matrices the rank of the product is lower than the minimum rank of the factors (see *Lipschutz, 1968)*, therefore a diagonal matrix (rank>1) can never be obtained as a product of a vector (rank=1) and a matrix.

The diag(.) operator is a key to building matrices composed of the elements of a vector.

A second matrix operator is the element by element multiplication of two matrices of equal size. Again, this so-called *array multiplication*, denoted as '⊗', seems to be no standard vocabulary of linear algebra.

Definition: Array multiplication

$$C=A\otimes B \Leftrightarrow C_{ij}=A_{ij}B_{ij},\forall i, j \qquad \text{(C.11)}$$

The array multiplication is related with the earlier introduced diag(.) operator via the following lemmas:

Lemma C.6: for any two vectors a and b and matrix C op appropriate size, the following holds:

$$\mathrm{diag}(a)\ C\ \mathrm{diag}(b) = (a\ b')\otimes C \qquad \text{(C.12)}$$

Proof $D = \mathrm{diag}(a)\ C\ \mathrm{diag}(b) \Leftrightarrow D_{ij} = a_i b_j C_{ij},\ \forall i, j \Leftrightarrow D = (a\ b')\otimes C$

Lemma C.7: for any vector a and two matrices B and C of appropriate size, the following holds:

$$B\otimes(\mathrm{diag}(a)\ C) = \mathrm{diag}(a)\ (B\otimes C) \qquad \text{(C.13)}$$

Proof $D = B\otimes(\mathrm{diag}(a).C) \Leftrightarrow D_{ij} = B_{ij}a_iC_{ij},\ \forall i, j \Leftrightarrow D = \mathrm{diag}(a).(B\otimes C)$

End of proof

C.3 Special matrices

To compute the flow, entry volumes are multiplied with split-proportions e.g. $f_{ij}=q_ib_{ij}$. This can be written as a matrix multiplication with the aid of the diag(.) operator and a special vector $\boldsymbol{q}$, e.g.:

$$\begin{gathered} f=\mathrm{diag}(\boldsymbol{q})b, \\ \boldsymbol{q}=(q_1,q_1,\ldots q_1,\ q_2,q_2,\ \ldots q_{2,}\ \cdots\ \cdots\ _{,}\ q_m,q_m,\ \ldots q_{m)}' \end{gathered} \qquad \text{(C.14)}$$

The vector $\boldsymbol{q}$ consists of m series of n identical elements. Converting a vector of length m to a vector of length mn consisting n series of m identical elements above each other, can be done by multiplying the vector with a so called *repeating column* matrix. The nonzero elements of such a matrix, that will be denoted as $\mathrm{p}_{m,n}$ or simply p, are defined by:

$$\begin{gathered} \mathrm{p}_{x(i,j),i} = 1 \\ x(i,j)=(i\text{-}1)n+j \\ i=1,2,\ldots m\ ,\ \ j=1,2,\ldots n \end{gathered} \tag{C.15}$$

Hence, $\boldsymbol{q} = \pi\ q$, and f=diag($\pi\ q$)b.

The structure of the matrix π is illustrated in figure C.1.

$$\pi = \underbrace{\left.\begin{bmatrix} 1 & & & & \\ 1 & & & & \\ \ldots & & & & \\ 1 & & & & \\ & 1 & & & \\ & 1 & & & \\ & \ldots & & & \\ & 1 & & & \\ & & \ldots & & \\ & & & \ldots & \\ & & & & 1 \\ & & & & 1 \\ & & & & \ldots \\ & & & & 1 \end{bmatrix}\right\}mn}_{m}$$

Figure C.1: Structure of a repeating column matrix

Nederlandse Samenvatting (summary in Dutch)

Dynamic Origin-Destination Matrix Estimation on Motorway Networks

Nanne van der Zijpp

Aanleiding voor dit onderzoek

Het verschijnsel congestie is de laatste jaren meer en meer kenmerkend voor het verkeersbeeld geworden. Ondanks het actieve beleid dat de overheid voert om de automobiliteit terug te dringen, mag op grond van factoren zoals demografische ontwikkeling, groeiende welvaart, veranderende arbeidsmarkt en een toenemende hoeveelheid vrije tijd, voorlopig nog een aanzienlijke jaarlijkse groei in de automobiliteit worden verwacht. Het traditionele antwoord op deze ontwikkeling, uitbreiding van de fysieke infrastructuur, zal gezien het ruimtebeslag dat dit met zich meebrengt, en de eisen die wij stellen aan onze leefomgeving, steeds moeilijker en vooral kostbaarder zijn in te passen. Dit heeft tot gevolg dat steeds hogere eisen aan de planning van infrastructurele uitbreidingen en aanpassingen worden gesteld, en dat steeds meer aandacht wordt besteed aan dynamische verkeersbeheersing en actuele reizigersinformatie.

Voor een modelmatige ondersteuning van zowel planning als dynamische verkeersbeheersingsegelen, zijn de laatste jaren tal van modellen ontwikkeld, zoals dynamische toedelingsmodellen, micro simulatie modellen, en modellen die in detail het route keuze gedrag van automobilisten onder invloed van pre-trip en en-route informatie beschrijven. Belangrijk kenmerk van deze modellen is dat zij expliciet rekening houden met de variatie van de diverse grootheden in de tijd, in plaats van te werken met gemiddelde waarden. Hierdoor kunnen verschijnselen zoals congestie nauwkeuriger worden gemodelleerd. Deze methoden stellen echter hoge eisen aan invoerdata, omdat ook de *vervoersvraag* nu dynamisch moet worden gespecificeerd.

Dynamische Herkomst-Bestemmings tabellen

De vervoersvraag wordt doorgaans samengevat in Herkomst-Bestemmings (HB) tabellen, ook wel HB-matrices genoemd, die voor iedere combinatie van herkomst en bestemming het geprognotiseerde aantal ritten bevat. Een dynamische HB-tabel kan worden beschouwd als een serie HB-tabellen die zijn gerangschikt op basis van de vertrekperiode. Een typische lengte van zo’n periode is vijf minuten. In dit proefschrift staat het schatten van deze dynamische HB-tabellen centraal.

Traditionele methoden om HB-tabellen te schatten zijn het uitvoeren van enquêtes, het voorspellen van verplaatsingen op basis van socio-economische gegevens, de kalibratie van een verklarend model of het aanpassen van een oude matrix aan tellingen. Voor het schatten van dynamische HB-tabellen voor real-time toepassingen gaat de belangstelling uit naar methoden die matrices schatten op basis van data die op automatische wijze kunnen worden verzameld. In de praktijk bestaan deze data met name uit tijdreeksen van telgegevens. Het schatten van HB-matrices op basis van telgegevens is echter een ondergespecificeerd probleem, wat wil zeggen dat meerdere HB-tabellen passen bij één set telgegevens.

Alhoewel alle traditionele methoden voorzien in een manier om deze onderspecificatie op te heffen, is het nadeel van deze methoden dat hun toepassing een zekere mate van aggregatie vereist. Bij het aggregeren van tijdreeksen van telgegevens gaat informatie verloren, terwijl de uitkomsten van geaggregeerde methoden niet het gewenste detail niveau bevatten.

Voor dynamische toepassingen wordt daarom sinds enkele tientallen jaren onderzoek gedaan naar een klasse van dynamische HB-schatters die ook op gedisaggregeerd niveau kan worden toegepast.

Deze schatters zijn gebaseerd op de aanname dat voor iedere toerit de toeritintensiteit zich in nagenoeg constante fracties over de afritten verdeelt. Deze fracties worden in de literatuur aangeduid als splitproporties of afslagfracties. Door de variatie van toeritintensiteiten in de tijd, kunnen de opeenvolgende telgegevens worden beschouwd als lineair onafhankelijke combinaties van splitproporties, zodat deze laatsten in theorie na verloop van tijd uit de telgegevens zijn op te lossen. Echter, als gevolg van het optreden van telfouten, alsmede random effecten in de bestemmingskeuze, en in de tijd veranderende herkomst-bestemmings patronen, moeten splitproporties in de praktijk worden *geschat*.

Probleemstelling

In dit proefschrift wordt gekeken naar toepassingen van split-ratio-methoden op eenvoudige netwerken waarbij routekeuze geen rol speelt, zoals een corridor op een autosnelweg. Bovendien wordt er vanuit gegaan dat de op het netwerk optredende reistijden voldoende nauwkeurig bekend zijn. De algemene doelstelling van het beschreven onderzoek is om de bestaande split-ratio-methoden op een aantal punten te verbeteren. Deze punten vallen uiteen in de volgende onderzoeksdoelstellingen:

- *Het meer algemeen toepasbaar maken van split-ratio-methoden*. In de literatuur worden doorgaans een aantal aannames gedaan die de toepasbaarheid van split-ratio-methoden in de praktijk sterk inperken. Zoals het uitsluiten van telfouten op toeritintensiteiten, en het uitsluiten van telpunten op interne schakels van het netwerk. Bovendien wordt doorgaans aangenomen dat de reistijden in het netwerk te verwaarlozen zijn vergeleken bij de lengte van de discretisatie periode. In dit proefschrift wordt nagegaan in hoeverre het mogelijk is om deze aannames los te laten.
- *Het beter theoretisch onderbouwen van split-ratio-methoden*. De in de literatuur beschreven split-ratio-methoden zijn gebaseerd op de aanname van constante of langzaam variërende *splitproporties*. Vanuit het standpunt van een waarnemer bezien kan er echter hooguit sprake zijn van constante of langzaam variërende *splitkansen*, omdat het kiezen van bestemmingen door automobilisten ongecoördineerd plaatsvindt. De hypothese is dat deze verbeterde beschrijving van het verkeerssysteem uiteindelijk zal leiden tot betere schattingen van de dynamische HB tabel.

  Een tweede aanpassing, die vanuit theoretisch oogpunt zou moeten leiden tot een verbeterd schattingsresultaat is het rekening houden met de afhankelijkheid van telgegevens. Vooral als telpunten op interne schakels van het netwerk liggen, vertoont hun conditionele kansdichtheid, gegeven de splitkansen, een sterke afhankelijkheid. De in de literatuur beschreven schattingsmethoden houden met deze afhankelijkheid geen rekening.

- *Het verbeteren van de schattingsmethode*. De in de literatuur beschreven schatters voor splitproporties vallen uiteen in twee categorieën; parameter optimalisatie methoden en statistische methoden. Binnen geen van de twee categorieën bestaat een schatter die in alle opzichten aan de eisen voldoet. Hieruit is de onderzoeksdoelstelling afgeleid dat een nieuwe statistische schatter dient te worden ontwikkeld die rekening kan houden met alle eigenschappen van het probleem.
- *Het benutten van additionele bronnen van informatie*. Naast tijdreeksen van telgegevens zijn er nog andere databronnen die van belang kunnen zijn voor het schatten van dynamische HB-tabellen. Twee van deze bronnen worden in dit proefschrift behandeld:
  -1- *Historische telgegevens*. HB-patronen blijken van dag tot dag sterk op elkaar te lijken. Het ligt daarom voor de hand om schattingen te baseren op telgegevens van meerdere dagen in plaats van telgegevens van slechts één dag.
  -2- *Automatische Voertuig Identificatie (AVI)*. In de toekomst zal technologie het mogelijk maken om tegen betaalbare kosten individuele voertuigen te herkennen of op automatische wijze te volgen, hetzij door de installatie van AVI apparatuur op meerdere plaatsen, hetzij door individuele voertuigen informatie over hun routes te laten verzenden. Op deze manier zal informatie beschikbaar komen over verkeerskarakteristieken, zoals routes en reistijden, die niet eerder direct kon worden waargenomen tegen acceptabele kosten.
- *Het testen van split-ratio-methoden in theorie en praktijk*. Er is tot nu toe weinig ervaring met het toepassen van split-ratio-methoden. Om meer inzicht in de eigenschappen van deze methoden te krijgen zijn experimenten nodig.

Resultaten van het onderzoek

De doelstelling om te komen tot een bredere toepasbaarheid en een betere theoretische onderbouwing van split-ratio-methoden hebben geleid tot de formulering van het *motorway model*. In dit model worden een aantal relaties binnen het verkeerssysteem expliciet vastgelegd. De belangrijkste kenmerken van het motorway model zijn:

- Er wordt rekening gehouden met telfouten, ook met telfouten in waarnemingen van toerit intensiteiten. Telfouten worden gemodelleerd als stoortermen waarvan de verwachtingswaarde nul is, en waarvoor een willekeurige variantiewaarde kan worden gespecificeerd.
- Er wordt uitgegaan van de aanname van langzaam variërende split*kansen* in plaats van de aanname van langzaam variërende split proporties. Hierdoor wordt een gedeelte van de variatie die optreedt in de bestemmings keuze die voorheen werd toegeschreven aan de variatie in de onbekende parameters, beschreven als stochastisch verschijnsel. De verandering in de splitkansen van periode tot periode is gedefinieerd als een storterm.
- Tellingen op interne schakels worden toegestaan.
- De in de literatuur gebruikelijke aanname dat *reistijden* verwaarloosbaar zijn wordt vervangen door de minder stringente aanname dat de reistijd*variatie* verwaarloosbaar is.

Met de formulering van het motorway model is het HB-schattingsprobleem gereduceerd tot het schatten van de splitkansen in het motorway model op basis van de beschikbare telgegevens.

Daarbij wordt er vanuit gegaan dat tellingen een functie zijn van HB-matrix celwaarden, en dat deze op hun beurt weer een functie zijn van *splitkansen* vermeerderd met een *stoorterm*. In deze stoorterm komt het verschil tussen split*kans* en split*proportie* tot uitdrukking. De variantie van deze stoorterm kan worden uitgedrukt in de splitkans. Als tellingen op interne schakels in het schattingsproces worden betrokken, wordt deze stoorterm op meerdere plaatsen waargenomen. Hierdoor ontstaan afhankelijkheden tussen tellingen, die kunnen worden uitgedrukt in een covariantiematrix.

Het gebruik van deze covariantiematrix zou in theorie tot betere schattingsresultaten moeten leiden. Omdat de covariantiematrix die werd afgeleid een functie is van de onbekende splitkansen, kan het resultaat alleen worden gebruikt als een basis voor een *benadering* van de echte covariantiematrix, bijvoorbeeld door substitutie van de meeste recente puntschatting van de splitkansen (voor deze benadering wordt in het proefschrift de term *Point Estimate Based Approximation*, -PEBA-, gebruikt) of door gebruik te maken van een kansverdeling van de splitkansen (deze benadering wordt een *Distribution Based Approximation*, -DBA-, genoemd).

Voor het *schatten* van de splitkansen in het motorway model zijn een aantal standaard technieken in beschouwing genomen, zoals Least Squares, Inequality Constrained Least Squares, Fully Constrained Least Squares, en het Kalman filter. De bestudering van deze schatters heeft tot de conclusie geleid dat geen van de bestaande schatters in alle opzichten voldoet. Methoden zoals Least Squares en Fully Constrained Least Squares (FCLS) bieden geen mogelijkheid om de afgeleide statistische eigenschappen te benutten, terwijl het Kalman filter geen mogelijkheid biedt om op een goede manier om te gaan met de ongelijkheidsvoorwarden die gelden voor de splitkansen.

Om deze moelijkheden te ondervangen wordt een nieuwe methode genaamd de Bayesian Updating (BU) methode voorgesteld. Het kenmerk van een Bayesian methode is dat deze niet direct resulteert in een *puntschatting*, maar dat eerst een *kansverdeling* van de onbekende grootheid wordt geschat.

De nieuwe methode is gebaseerd op het aannemen van een Truncated Multivariate Normal (TMVN) verdeling voor de splitkansen, en het aanpassen van deze verdeling volgens het beginsel van Bayesian updating wanneer nieuwe informatie beschikbaar komt. Het is aangetoond dat als de vorm van de likelihood functie overeenkomt met MVN, de aangepaste kansverdeling in de klasse van TMVN verdelingen blijft, en dat de parameters die de aangepaste verdeling karakteriseren voldoen aan de wel bekende Kalman measurement update vergelijkingen. De methode is exact als de splitkansen constant zijn in de tijd. Voor de tijd extrapolatie is echter geen exacte uitdrukking beschikbaar. Echter, wanneer de variatie in de splitkansen klein is, is de verwachting dat de standaard Kalman time extrapolation vergelijkingen een acceptabele benadering vormen.

Om puntschattingen af te leiden uit de kansverdelingen volgende uit de hierboven beschreven recursie is een z.g. *postprocessing* routine nodig. Twee klassen van dergelijke routines worden onderscheiden. De eerste klasse houdt in dat een puntschatting wordt afgeleid door de aposteriori verdeling te *maximaliseren*. De tweede klasse houdt in dat de *verwachting* van de aposteriori verdeling wordt uitgerekend. Puntschattingen van het eerste type worden aangeduid met de naam *Maximum APosteriori* (MAP) schatters, terwijl puntschattingen van het tweede type worden aangeduid met de naam *Subjective Expectation* (SE) schatters. In theorie zouden SE schatters de schattingsfout moeten minimaliseren, maar een analytische uitdrukking voor de SE schatter is niet beschikbaar. Benaderingen van SE schatters kunnen worden verkregen door het gemiddelde te nemen van een groot aantal vectoren die zijn geloot uit de TMVN aposteriori verdeling. Random vectors voor een TMVN verdeling kunnen worden gegenereerd door MVN random vectors te genereren, en alle uitkomsten die niet aan de ongelijkheids voorwaarden voldoen te negeren. Deze aanpak wordt aangeduid als *Subjective Expectation - Randomized Mean* (SE-RM), maar is alleen praktisch toe te passen als de kans op het loten van een geldige uitkomst voldoende groot is. Als alternatief voor de SE-RM punt schatter is een analytische *benadering* ontwikkeld, aangeduid als *Subjective Expectation - Approximated Mean* (SE-AM). Het berekenen van MAP punt schatters kan op meer directe wijze plaatsvinden, omdat dit neerkomt op het minimaliseren van een kwadratische functie onder een aantal ongelijkheids voorwarden.

HB-matrices die zijn geschat uit tijdseries van telgegevens kunnen worden verbeterd door *aanvullende databronnen* te gebruiken. Analyse van empirisch gegevensmateriaal toont aan dat de splitfracties niet alleen een kleine variatie vertonen van periode naar periode, maar ook van dag tot dag, wanneer corresponderende periodes worden vergeleken. Om deze eigenschap aan te wenden voor een verbetering van de schatting van de splitkansen is een model voorgesteld dat tot uitdrukking brengt dat splitkansen variëren rondom hun historische waarden. Uitgaande van dit model werd een bewegingsvergelijking voor de splitkansen afgeleid waarin de aanname van langzaam variërende splitkansen wordt gecombineerd met het gebruik van historische data.

Een andere bron van informatie is het gebruik van Automated Vehicle Identification data. Om deze data volledig te benutten, zijn nieuwe schattingsmethoden nodig. Als een voorbeeld wordt het gebruik van automatische kentekenplaat lezers beschouwd. Alhoewel het schatten van HB-matrices van directe waarnemingen zoals enquêtes en kentekenplaat onderzoeken uitgebreid onderzocht is in het verleden, bevat de huidige probleemstelling een aantal nieuwe elementen, zoals de vereiste de kentekenplaat gegevens te verwerken in *combinatie* met telgegevens, de aanwezigheid van afleesfouten in het kenteken materiaal, het dynamische aspect, en het ontbreken van voorwaarden aan de lokaties waarop kentekens worden gelezen (kentekenplaatlezers mogen op willekeurige plaatsen in het netwerk opgesteld staan).

Op het eerste gezicht heeft het schatten van HB-matrices uit kentekenplaat gegevens weinig gemeen met het schatten van HB-matrices uit telgegevens. Maar door de constructie van een denkbeeldig ‘hypernetwerk’ kan worden aangetoond dat er een analogie tussen deze twee gevallen bestaat, en dat alle theorie die eerder werd voorgesteld voor het schatten van HB-matrices uit telgegevens kan worden gemakkelijk kan worden aangepast voor het probleem van het schatten van HB-matrices uit gecombineerde tellingen-kentekengegevens. Deze analogie blijft ook bestaan als de beschikbare informatie alleen betrekking heeft op een *steekproef* uit alle voertuigen (we spreken in dit geval van probes). Er kan daarom worden geconcludeerd dat de Bayesian schatter kan worden gebruikt als een onderdeel van een raamwerk voor het dynamisch schatten van HB-matrices, dat zeer flexibel is wat betreft de vereiste waarnemingsdata. Een dergelijk raamwerk kan bijvoorbeeld gebruik maken van een combinatie van historische data, tellingen, kenteken plaat gegevens en probe vehicle gegevens.

De theoretische bevindingen uit het proefschrift zijn getest in twee series van experimenten. De eerste serie experimenten werd uitgevoerd met behulp van gesynthetiseerde HB-matrices en tellingen, welke werden gegenereerd volgens de specificaties van het motorway model. De tweede serie werd uitgevoerd met behulp van telgegevens die gedurende een maand werden verzameld op de ringweg Amsterdam.

De experimenten met de gesynthetiseerde data laten duidelijk zien dat het gebruik van de nieuwe BU methode de schattingsfout aanmerkelijk reduceert, ten opzichte van het gebruik van bestaande methodes, zoals FCLS en het Kalman filter. Ten opzichte van de FCLS methode werd een reductie van 31% in de gemiddelde schattingsfout van de splitkansen verkregen (uitgedrukt in RMSE, voert./per.). Het grootste gedeelte van deze reductie (67%) moet worden toegeschreven aan de verbeterde behandeling van de ongelijkheids voorwarden door de nieuwe schatter, gecombineerd met het gebruik van de SE-RM postprocessing routine. Een kleiner gedeelte van de reductie (33%) moet worden toegeschreven aan de betere beschrijving van de eigenschappen van de tellingen, uitgedrukt in de PEBA en DBA covariantie matrices. Als in plaats van de PEBA of DBA matrices, een matrix gebaseerd op de Average Link Flows (ALF) wordt gebruikt, hoeft overigens slechts een klein gedeelte (-7%) weer te worden opgeofferd, terwijl het voordeel van een dergelijke aanpak is dat deze kan worden toegepast zonder kennis van de verdeling van de verschillende stoortermen die bij het genereren van de testdata een rol hebben gespeeld.

Een onverwacht resultaat was dat een reductie in de schattingsfout niet noodzakelijkerwijs resulteert in een verbeterde nauwkeurigheid waarmee schakel intensiteiten kunnen worden voorspeld (de link-flow error). Bijvoorbeeld, als in twee verder geheel identieke Bayesian updating methodes de SE-AM postprocessing routine wordt verwisseld voor een MAP postprocessing routine, leidt dit consequent tot een hogere schattingsfout, maar gelijktijdig tot een lagere link-flow error. In theorie zou een dergelijk verschijnsel niet mogen gelden voor randomized mean (SE-RM) puntschattingen. Echter, de praktijk is dat SE-RM puntschattingen niet altijd kunnen worden geëvalueerd. In dergelijke gevallen grijpt deze postprocessing routine terug op de SE-AM waarde. Daarom reduceert de bovengenoemde verbetering van 31% tot slechts 3.7% voor de gesynthetiseerde data, en 7% voor de empirische data, wanneer het link-flow criterium wordt gebruikt als foutmaat.

Bij de experimenten met empirische data worden de schattingen enkel geëvalueerd op basis van het link-flow error criterium, omdat de echte HB-matrix niet bekend is. Het gebruik van het link-flow error criterium werkt systematisch in het voordeel van de FCLS methode en de BU methodes die de MAP postprocessing routine gebruiken. Desalniettemin kunnen toch een aantal conclusies worden getrokken. Ten eerste; in alle gevallen, leidt het gebruik van historische gegevens tot een reductie van de link-flow error. Ten tweede; een variant van de nieuwe BU methode die gebruik maakt van de MAP postprocessing routine, verbeterde de link-flow error ten opzichte van de traditionele methoden in alle gevallen.

Deze laatste conclusie geldt tevens als het indirecte bewijs dat de BU methode gecombineerd met de SE-RM postprocessing routine ook voor empirische data de meest nauwkeurige HB-matrix schatting oplevert, omdat de ervaring met gesimuleerde data heeft geleerd dat in alle gevallen SE-RM postprocessing moet worden verkozen boven MAP postprocessing.

About the author

Nanne van der Zijpp was born in Dacca, Pakistan, in 1963 and grew up in Haarlem and Hollandsche Rading in The Netherlands. He graduated at the Alberdingk Thijm College in Hilversum in 1981. After his secondary education he studied applied mathematics at Delft university. In 1988 he received his M.Sc. from the mathematical systems theory group for his thesis on the subject of radar tracking which was based on research carried out during a one year stay at the National Aerospace Laboratory.

After his graduation, Nanne was employed as a systems analyst at Holland Information Consultants. In 1990 he went back to Delft university, and joined the transportation planning and traffic engineering section at the faculty of Civil Engineering, where he specialized in modelling, statistics, optimization, and the computer implementation of traffic models. He has participated in projects on trip planning for road transport, logistics in the port of Rotterdam, static and dynamic OD-estimation, and modelling route choice behaviour. In 1993 he stayed at the Center for Transportation Research of the Virginia Polytechnic Institute and State University (Virginia Tech) in Blacksburg, USA for a period of five months. In 1994 he participated in the DRIVE research project DYNA, granted by the European Commission.

Nanne submitted his Ph.D. thesis on dynamic OD-estimation in 1995. His present research interests are OD-matrix estimation, dynamic assignment and traveller information systems. In 1995 Nanne was selected for a fellowship in the EC Human Capital and Mobility programme by his current employer, the Centre for Transport Studies in London.

Index