

---

# ORIGIN-DESTINATION MATRIX ESTIMATION IN OMNITRANS

ERIK-SANDER SMITS

---

June 1, 2010

Utrecht University



Omnitrans International BV





## Acknowledgements

This thesis finishes my master's degree in Mathematical Sciences at the Utrecht University. It was a very interesting experience to do this research at Omnitrans International and I hope that the supervisors of the company, as well as my supervisor at the university, are satisfied with the result.

A seven month internship at Omnitrans International in Deventer is the cradle of this piece of work. As a mathematician with special interest in optimization problems I was introduced to the field of traffic engineering and the match was perfect. Nowadays I can keep up in the field, although the flow of information and papers was sometimes a bit overwhelming. The highlight of the internship was my participation in the European Transport Conference (ETC) in October 2009. As a steward I had the opportunity to attend sessions and deliberate with an authority on my subject. The monthly technical meetings with the Transport Innovation and Modelling (TIM) group were a pleasant variation on the work, it brought insight in other transportation models and was an platform for discussion.

First of all I would like to thank my daily supervisor Maarten Schilpzand for his coaching and control. We are both mathematicians that are active in the field of transportation; this equal point of view improved the communication. I would like to thank Klaas Friso, my second supervisor, for his useful comments on my work; they were always sharp and raised the study to a higher level. Research to such a specialized topic can be quite lonesome, my colleagues at Omnitrans International offered some distraction. Thanks for all the coffee breaks, answers and questions.

Prof. dr. ir. E.J. Balder was my tutor during my master's study at Utrecht University. I enjoyed the conversations we had in the past three years, they were very guiding to me.

Mr. John Cliff contributed on the quality of the English text, I am very grateful to him.

Finally I would like to thank my family and especially Emma, my tower of strength. Thanks for supporting me, especially in the harder periods. I know you were all looking forward to the moment I finished this thesis.

— Erik-Sander Smits

## Abstract

The estimation of origin-destination (O-D) matrices is an important component of the field of transportation modelling. It encompasses the improvement of the modelled traffic demand between zones through the help of additional information (e.g. traffic counts). This study develops two methods of matrix estimation that are applicable to the software package **OmniTRANS** and these are compared with currently implemented method. Compatibility with **OmniTRANS** implies a multi dimensional (e.g. multi-modal and multi-class) model.

The first proposed method is a revision of the current method and its major advantage is the independency of the input order. This is accomplished by averaging the multiplication factors that are used in the current method. The second method is build up from scratch and has its roots at the gradient descent method. This method has better mathematical properties than the other two.

Furthermore a literature study is presented by means of a framework which encloses all the modelling aspects suggested in the last three decades. The information used to perform matrix estimation is usually of the type traffic count, this study introduces the notion of restrictions. Restrictions are the generalization of all types of input available in **OmniTRANS** and it offers an easy extension with other types.

## Samenvatting

Het schatten van herkomst-bestemming (H-B) matrices is een belangrijk component van het vakgebied verkeersmodelleren. Het omvat het verbeteren van de gemodelleerde vervoersvraag tussen zones met behulp van aanvullende informatie (zoals verkeerstellingen). Deze studie ontwikkelt twee matrix-schatmethodes<sup>1</sup> die van toepassing zijn op het software pakket **OmniTRANS** welke zullen worden vergeleken met de momenteel geïmplementeerde methode. Compatibiliteit met **OmniTRANS** impliceert een multi-dimensionaal model.<sup>2</sup>

De eerste voorgestelde methode is een aanpassing van de huidige methode en heeft als grootste voordeel de onafhankelijkheid van de invoer volgorde. Dit wordt bereikt door de vermenigvuldigingsfactoren van de huidige methode te middelen. De tweede methode is vanaf de grond opgebouwd en is afgeleid van de gradiënt daalmethode. Deze voorgestelde methode heeft betere wiskundige eigenschappen dan de andere twee.

Verder wordt er een literatuurstudie gepresenteerd door middel van een raamwerk dat alle modelleer aspecten die in de afgelopen drie decennia zijn gesuggereerd kan omvatten. Normaliter worden verkeerstellingen gebruikt als input voor matrix-schatten, deze studie introduceert het begrip restricties. Restricties zijn een generalisatie van alle types input beschikbaar in **OmniTRANS** en biedt de mogelijkheid deze eenvoudig met andere types uit te breiden.

---

<sup>1</sup>Vaak wordt in het Nederlands de term matrix calibratie gebruikt in plaats van deze directe vertaling.

<sup>2</sup>Hierin kunnen meerdere vervoerswijzen, vervoersklassen en tijdsdimensies worden meegenomen.

# Contents

<b>1</b>	<b>Introduction</b>	<b>7</b>
1.1	Four step model	7
1.2	Matrix estimation	8
<b>2</b>	<b>Literature review</b>	<b>9</b>
2.1	The framework	9
2.2	The objective function	11
<b>3</b>	<b>OmniTRANS</b>	<b>16</b>
3.1	Omnitrans International BV	16
3.2	The OmniTRANS software package	16
3.2.1	Dimensions	17
3.2.2	The network	17
3.2.3	Data	18
3.2.4	Jobs	19
3.2.5	Variants	20
<b>4</b>	<b>OmniTRANS Matrix Estimation method</b>	<b>22</b>
4.1	Model setting	22
4.2	The method	23
4.3	Revision of the algorithm	25
<b>5</b>	<b>The gradient descent method</b>	<b>27</b>
5.1	The gradient descent method in a nutshell	27
5.2	General form of the gradient descent method	27
5.3	Specification of the bilevel problem for OmniTRANS	29
5.4	The gradient method for the convex OmniTRANS subproblem (COSP)	32
5.5	Summary	34
<b>6</b>	<b>Test procedure</b>	<b>36</b>
6.1	The environment	36
6.2	The transport models	37
6.3	Quality measures	39
6.3.1	Objective function	39
6.3.2	Coefficient of determination, $R^2$	42
6.3.3	T-values	42
6.3.4	GEH statistic	43
6.3.5	Summary	44
<b>7</b>	<b>Results</b>	<b>45</b>
7.1	Tests	45
7.1.1	Test on the influence of the input order	45
7.1.2	Test on the performance of the methods	45
7.1.3	Test on the $\alpha$ values for GRADIENT	49
7.1.4	Dynamic Tactical Traffic Model Amsterdam	53
7.2	Memory reduction	57
7.2.1	8-bit unsigned integer	57
7.2.2	Sparse matrix structure	57
<b>8</b>	<b>Conclusions &amp; Recommendations</b>	<b>59</b>
8.1	Conclusions	59
8.2	Recommendations	59
8.2.1	Integration of methods	59

8.2.2	Other restrictions . . . . .	60
-------	------------------------------	----

## List of Algorithms

1	The OmniTRANS algorithm . . . . .	24
2	The revised OmniTRANS algorithm . . . . .	25
3	The gradient descent method . . . . .	29
4	A direct integration of (COSP) and the gradient method . . . . .	33
5	The OmniTRANS gradient method algorithm . . . . .	34

## Notation list

This section introduces the most important variables used in this study.

General variables:

$N$	set of nodes of the network
$A$	set of directed arcs of the network
$I$	set of all O-D pairs

Notations in the literature review:

$g_i$	traffic demand for O-D pair $i$
$\mathbf{g} = \{g_i   i \in I\}$	the O-D matrix
$\hat{g}_i$	a priori traffic demand for O-D pair $i$
$\hat{\mathbf{g}} = \{\hat{g}_i   i \in I\}$	the a priori O-D matrix
$v_a$	traffic flow on link $a$
$\mathbf{v} = \{v_a   a \in A\}$	vector of link flows
$\hat{A} \subseteq A$	set of links with count information
$\hat{v}_a$	counted flow on link $a$
$\hat{\mathbf{v}} = \{\hat{v}_a   a \in \hat{A}\}$	vector of counted link flows
$F_1(\mathbf{g}, \hat{\mathbf{g}})$	distance measure between $\mathbf{g}$ and $\hat{\mathbf{g}}$
$F_2(\mathbf{v}, \hat{\mathbf{v}})$	distance measure between $\mathbf{v}$ and $\hat{\mathbf{v}}$
$\alpha$	relative weight parameter
$K_i$	set of paths for O-D pair $i$
$p_{ik}$	proportion of demand $g_i$ that uses path $k$
$h_{ik}$	path flow of path $k$ for O-D pair $i$
$\delta_{ak}$	link $a$ - path $k$ incidence variable

Notations in the OmniTRANS model setting (chapter 4 and onwards):

$D$	set of dimensions
$g_i^d$	traffic demand for O-D pair $i$ in dimension $d$
$\mathbf{g} = \{g_i^d   i \in I, d \in D\}$	the O-D matrix
$\hat{g}_i^d$	a priori traffic demand for O-D pair $i$ in dimension $d$
$\hat{\mathbf{g}} = \{\hat{g}_i^d   i \in I, d \in D\}$	the a priori O-D matrix
$R$	set of restrictions
$C_r$	counted value for restriction $r$
$\mathbf{C} = \{C_r   r \in R\}$	vector of counted values
$D_r$	the set of dimensions belonging to restriction $r$
$P_{ir}^d$	the fraction of the demand $g_i^d$ that applies to restriction $r$
$\mathbf{P} = \{P_{ir}^d   i \in I, r \in R, d \in D_r\}$	a three dimensional variable containing all fractions
$\varepsilon_r$	weight of restriction $r$
$\mathcal{P}_i^d$	the number of restrictions that apply to O-D pair $i$ in dimension $d$
$l_r$	traffic load over restriction $r$
$\mathbf{l} = \{l_r   r \in R\}$	vector of restriction loads
$F_1(\mathbf{g}, \hat{\mathbf{g}})$	distance measure between $\mathbf{g}$ and $\hat{\mathbf{g}}$
$F_2(\mathbf{l}, \mathbf{C})$	distance measure between the counted values and the restriction loads
$\alpha$	relative weight parameter in GRADIENT
$\mathbf{s}$	search direction
$\lambda$	step length
$\lambda^*$	optimal step length

# 1 Introduction

Traffic and transportation modelling encompasses a broad range of topics, is applied in innumerable projects and its development and innovation is still intense. All models in this field have in common that they deal with displacements, from the voyage of fruit to the grocery store to the evacuation plan in case of a disaster. The models that are used daily by consultancy companies, governments and transport concerns to forecast, improve and schedule traffic and transportation are continuously subject of research at universities, research institutes and software developers. This study is about a tiny section of this field and aims to improve a model with additional information.

The focus is on modelling (road) networks and traffic, on models that reproduce and forecast the traffic flows on the network. These models are built up with a representation of the real road network and an assessment of the trip demand. It is also possible to consider transit options within the model. The demand can be assigned to the network which leads to traffic flows on the roads, which are the required final results. This was a very brief description of the main aspects of the four step model, which is the most commonly used model by traffic engineers for the investigation of an area.

## 1.1 Four step model

In this section the four step model is explained, the main objective of this study is to examine a part of this model. To start with, the study area is divided in zones (e.g. neighbourhoods, commercial districts, etc). Each zone has its own characteristics, the socio-economic data, these contain populations (and other demographic information), employment and other information on available services. Furthermore a representation of the real network in the form of a directed graph is available. The four step model consists of the following steps:

1. **One — Trip generation:** The socio-economic data for each zone is transformed to a trip production and a trip attraction, these are respectively the number of people finishing in and starting from the zone. The methods to do this are not comprehensive.
2. **Two — Trip distribution:** Currently it is clear where trips originate and terminate, but not the origin and destination of a specific trip. In this step the productions and attractions are connected, so the number of trips for each origin and destination pair is determined. This is the trip demand for an Origin-Destination (O-D) pair. The gravity model is a frequently used method in this step, it connects zones like gravitational forces connect celestial bodies. The results of this step are represented in a matrix since there is a trip demand for each O-D pair, this matrix is called the O-D matrix.
3. **Three — Modal split:** Step three splits the trips for each O-D into different modes. Examples of modes are car, public transport and bicycle. In practice this step is already done in step two.
4. **Four — Assignment:** This is probably the most interesting step of the model, it deals with how each trip charges the network. The most important decision for that is the route for each trip, with that information the traffic load on each link in the network and occupancy level in the transit lines can be determined. The number of methods that are developed to do this is huge, a book can be written on their classification. The most important methods are introduced later on in this study. An important note to make is that the route choice depends on the level of congestion but also influences the level of congestion. This is why most models search for an equilibrium, this makes the problem harder to solve.

The hard part of this model is that ideally the results of step four are already known when performing step two. Step two needs the distance between each O-D pair, this distance should be measured in travel time, and the travel times are not known before step four. So usually a very simple assignment is performed before step two and a feedback loop from step four to step two is performed.



## 1.2 Matrix estimation

One might wonder where matrix estimation fits in this problem. It comes into play between step three and four, the traffic demand is determined for each O-D pair, but the used models in previous steps do not guarantee good results. That is why additional information, from traffic counts for example, is used to improve the O-D matrix. This process is called matrix estimation. The additional information can be of many types and many sources. The most frequently used type is traffic counts. Traffic counts register the traffic flow on a certain location on the network. The hard part is that there is no one-to-one correspondence between the demand of an O-D pair and the traffic count. The comparison between the observed traffic flows and modeled traffic flows can only be made after step four.

In this study two methods of matrix estimation are developed for OmniTRANS and compared with the current method in OmniTRANS. The challenges are to make the methods compatible with all OmniTRANS functionalities and to develop it in such a way that different types of information can be introduced.

The thesis is built up as follows. In chapter 2 a literature review in the form of a framework is given. Chapter 3 describes the working of the OmniTRANS software. The current matrix estimation algorithm and a revision of it is given in chapter 4. The third method is introduced in chapter 5, this describes a gradient descent method and is build up from scratch. The constructed test environment and quality measures can be found in chapter 6. Chapter 7 describes the results of several performed tests. Finally in chapter 8 the conclusions and some recommendations are stated.

## 2 Literature review

The problem of estimating origin-destination (O-D) matrices from additional information like traffic counts and home surveys has received attention in literature since the 1970's. The topic is still intensively studied resulting in numerous techniques to deal with the problem. The purpose of this section is to give a state of the art of the matrix estimation techniques with a wide scope. This will be presented in one modelling framework (section 2.1), so the reader is not bothered with the variety of notations used in the literature. Most articles focus on matrix estimation from traffic counts, they are cheap to obtain in contrast to the costs of home surveys. Unfortunately it is far from straightforward to use the counts to improve an O-D matrix, especially when traffic congestion is taken into account. Other sources of information can be converted to fit in the presented framework, so they can be used for most practical usages.

In the late 1990's three extended literature reviews are written by [Chen & Florian (1996)], [Barceló (1997)] and [Abrahamsson (1998)]. They give a comparison based on the several choices and simplifications of a general problem model. The largest distinction in modelling approaches is on congested and uncongested networks. The problem becomes significantly more complex when congestion of the network is taken into account, the traffic assignment is then dependent on the estimated matrix. The cost function of a link will not be constant any more which results in an assignment that is not proportional. That implicates a subproblem and will actually lead to a bilevel optimization problem. Such problems have a leaders' and a followers' problem and use each others' output as their own input. There is no general solution for bilevel optimization problems available, present algorithms provide heuristics or local searches. There are several approaches to the followers' problem (i.e. the traffic assignment problem), that makes it more complicated.

Another divarication is on the statistical approach, this will lead to different objective functions of the matrix estimation problem. In general the objective function consists of two parts, a measure on the difference of the resulting and observed traffic flows and a measure on the distance from the a priori O-D matrix in the result. Particularly the early publications on the topic consider how the traffic counts and the a priori matrix should be modelled, or in other words, from which distribution do they come. Different assumptions herein lead to different objective functions, the most conventional are maximum likelihood (ML), generalized least squares (GLS) and Bayesian inference models. In [Cascetta (2001)] (sections 8.5.1 and 8.5.2) these estimators are derived and discussed. The advantage of the GLS method is that it does not depend on assumptions of the distributions where the traffic counts and a priori matrix are from. The advantage of ML and Bayesian methods on the other hand is that they have better statistical properties. The different objective functions incorporating these models will be discussed in section 2.2.

Inherent to diversity in modelling and complexity of the problem is that there are dozens and dozens of solution algorithms proposed. For this study the most important method is that of [Spiess (1990)]. It is probably the most widespread and used algorithm since it is cited frequently and a free implementation for the Emme software package is available. Spiess uses a gradient method to find a local solution for the problem. The gradient method proposed in this study is inspired by this article.

### 2.1 The framework

The purpose of this section is to provide a framework wherein all the results from the literature fit. J.T. Lundgren also worked on a framework in 1991, unfortunately it was not possible to obtain his working paper for the purpose of this study. Although it can be supposed that the framework in this study will be similar to his framework since both make a distinction on traffic congestion. Three different mathematical optimization problems will be proposed, a general, a convex and a bilevel approach. The latter two are specializations of the first one.

In transportation modelling the research area is divided in several zones, each zone has its own characteristics, like population and employment. The road network is represented as a directed graph

$(N, A)$ , where  $N$  are the nodes (i.e. junctions) and  $A$  the links. Each zone is connected to the network via centroids, the centroid corresponds to one or more nodes in the centre of the zone. Trips from and to zones always start and end at those nodes. The traffic demand between zone  $i$  and zone  $j$  is stored in the origin-destination (O-D) matrix  $\mathbf{g}$ .  $I$  is the set of all O-D pairs and the vector  $\mathbf{g} = \{g_i | i \in I\}$  is the O-D matrix. The target or a priori matrix  $\hat{\mathbf{g}} = \{\hat{g}_i | i \in I\}$  is a matrix from a distribution model or an older matrix that needs to be updated.  $\mathbf{v} = \{v_a | a \in A\}$  are the link flows. Traffic counts are available on links  $\hat{A} \subseteq A$  and the counted flows on those links are  $\hat{\mathbf{v}} = \{\hat{v}_a | a \in \hat{A}\}$ . The general problem is

$$\begin{aligned} \min_{\mathbf{g}, \mathbf{v}} F(\mathbf{g}, \mathbf{v}) &= \alpha F_1(\mathbf{g}, \hat{\mathbf{g}}) + (1 - \alpha) F_2(\mathbf{v}, \hat{\mathbf{v}}) \\ \text{s.t. } \mathbf{v} &= \text{assign}(\mathbf{g}), \\ \mathbf{g} &\geq 0, \\ \mathbf{v} &\geq 0 \end{aligned} \tag{GP}$$

where  $F_1$  is a distance measure between  $\mathbf{g}$  and  $\hat{\mathbf{g}}$ ,  $F_2$  is a distance measure of between  $\mathbf{v}$  and  $\hat{\mathbf{v}}$  and parameter  $\alpha \in [0, 1]$ . The  $\alpha$  represents the relative weight for each objective. If the traffic counts are very accurate  $\alpha$  should be closer to 0, while it should be closer to 1 if there is a good a priori matrix. If there is a high confidence in both measures a good equilibrium should be found in the objective functions. (GP) is the general problem of matrix estimation and has a lot of freedom. The first specialization will now be derived.

The assignment can be done with different methods, some will take congestion into account. All-or-nothing (AON) and proportional assignment are two models that do not take congestion into account and will lead to another, specialized problem. AON is the assignment where every route is the shortest path (or minimal cost paths), so you can just apply a shortest path algorithm to get the link flows. An alternative for AON is the proportional assignment where there are several paths, each is used by some proportion of the demand. Let  $K_i$ ,  $i \in I$  be the set of paths for O-D pair  $i$ , and  $p_{ik}$ ,  $i \in I, k \in K_i$  is the proportion of demand  $g_i$  that uses path  $k$ . Path flows  $h_{ik}$  can now be calculated by

$$h_{ik} = p_{ik} g_i, \quad i \in I, k \in K_i. \tag{1}$$

The link flow on link  $a$  can be expressed as

$$v_a = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} h_{ik} = \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} p_{ik} g_i \quad \text{where } \delta_{ak} = \begin{cases} 1 & \text{if link } a \text{ is in path } k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

So the link flows are proportional with the demand, the drawback of this assumption is that it cannot deal with congestion. That is because in the congested case the assignment is dependent of the O-D matrix. Proportional assignment assumes that the cost of a link is constant, if it depends on the link flow the assignment can deal with congestion. The advantage of the uncongested approach is that under the assumption that the objective function  $F(\mathbf{g}, \mathbf{v})$  is convex we have a convex optimization problem:

$$\begin{aligned} \min_{\mathbf{g}, \mathbf{v}} F(\mathbf{g}, \mathbf{v}) &= \alpha F_1(\mathbf{g}, \hat{\mathbf{g}}) + (1 - \alpha) F_2(\mathbf{v}, \hat{\mathbf{v}}) \\ \text{s.t. } v_a &= \sum_{i \in I} \sum_{k \in K_i} \delta_{ak} p_{ik} g_i, \quad \forall a \in A \\ \mathbf{g} &\geq 0, \\ \mathbf{v} &\geq 0 \end{aligned} \tag{CP}$$

It is very unlikely that there is no congestion in a traffic study, the congestion will be the reason to start a study in a lot of cases. There are several assignment methods that can deal with congestion, examples are deterministic user equilibrium, stochastic user equilibrium, volume averaging and incremental assignment. Their mutual property is that they are iterative. User equilibrium assignment

tends to comply with Wardrop's first principle: *The journey times in all routes actually used are equal and less than those which would be experienced by a single vehicle on any unused route.* The obstacle of assignment with congestion is the inconstant link cost, it is dependent of the link flow. Generally the costs per trip on a link will increase significant if the traffic is at a level where congestion starts to occur.

In **(GP)** the assignment is therefore complicated if congestion is taken into account, actually  $assign(\mathbf{g})$  is a subproblem to be solved. An assignment method as described in the previous paragraph will be denoted as  $Assign(\mathbf{g})$ , with capital A.<sup>3</sup> A bilevel problem formulation can now be presented:

$$\begin{aligned} \min_{\mathbf{g}, \mathbf{v}} F(\mathbf{g}, \mathbf{v}) &= \alpha F_1(\mathbf{g}, \hat{\mathbf{g}}) + (1 - \alpha) F_2(\mathbf{v}, \hat{\mathbf{v}}) \\ \text{s.t. } \mathbf{v} &= Assign(\mathbf{g}), \\ \mathbf{g} &\geq 0, \\ \mathbf{v} &\geq 0 \end{aligned} \tag{BLP}$$

The remainder of this section will consist of an overview of the several prospects for the objective function, followed by a review of the solution techniques for the specific problems.

## 2.2 The objective function

The available information for matrix estimation is subject to inaccuracy. Traffic counts are a measure of the traffic over a certain period, it can be a single count or the mean over several counts on more days. Each count can be modelled as an instance of some probability distribution. The a priori matrix can have several sources, e.g. some distribution model or home surveys. A home survey can just like a count be seen as an instance of some probability distribution. Distribution models (e.g. the gravity model) will lead to errors in the a priori matrix since they are a (huge) simplification of reality. Different assumptions on the source of the information will lead to different objective functions. The first modelling approaches in the literature are very specific, [Cascetta & Nguyen (1988)] provide a framework for these approaches and discuss the various statistical approaches. The focus of the paper is on the objective function, they discuss classical (ML and GLS) estimators and Bayesian inference. The earlier mentioned literature reviews copy this approach. Recently [Chen, et al. (2009, in press)] discuss the  $L_p$ ,  $p \in \{1, 2, \infty\}$  norms to measure the error. They do this because the statistical information needed for ML and GLS estimation and Bayesian inference is often not available.

Maximum likelihood (ML) estimators are obtained by maximizing the probability that the observed results correspond to a certain distribution. There is a family of distributions to choose from which is parametrized and the likelihood function  $\mathcal{L}(\cdot|\cdot)$  is given. Maximizing  $\mathcal{L}(\cdot|\cdot)$  is equivalent to maximizing  $\log \mathcal{L}(\cdot|\cdot)$  since log is an increasing operator and we assume  $\mathcal{L}(\cdot|\cdot) > 0$ . It is supposed that the traffic counts and the a priori matrix are independent variables. So the likelihood of observing  $\hat{\mathbf{g}}$  and  $\hat{\mathbf{v}}$  with real O-D matrix  $\mathbf{g}$  is

$$\mathcal{L}(\hat{\mathbf{g}}, \hat{\mathbf{v}}|\mathbf{g}) = \mathcal{L}(\hat{\mathbf{g}}|\mathbf{g})\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g}) \tag{3}$$

and

$$\log(\mathcal{L}(\hat{\mathbf{g}}|\mathbf{g})\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g})) = \log \mathcal{L}(\hat{\mathbf{g}}|\mathbf{g}) + \log \mathcal{L}(\hat{\mathbf{v}}|\mathbf{g}). \tag{4}$$

As can be seen in problem **(GP)** there is a relation between  $\mathbf{g}$  and  $\mathbf{v}$ , via the assignment, therefore we state  $\mathcal{L}(\hat{\mathbf{v}}|\mathbf{v})$  instead of  $\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g})$  from now on. One of the supposed distributions for the a priori matrix is a multinomial distribution. This is the case if a fraction of the trips originating from each origin is sampled. Let  $\gamma_i$ ,  $i \in I$  be the fraction of trips that is sampled at the origin of  $i$ <sup>4</sup>. This study will still use  $\hat{\mathbf{g}}$  as the a priori matrix and not as the sampling results, so if  $n_i$  is the sampling result for O-D pair  $i$ ,  $\hat{g}_i = n_i/\gamma_i$ . [Abrahamsson (1998)] is unclear about the changed interpretation of  $\hat{\mathbf{g}}$ ,

<sup>3</sup>Because of the different nature of the assignment methods it is not possible to give a standard mathematical formulation. Some will have a variational model, others an optimization model.

<sup>4</sup>so  $\gamma_i = \gamma_j$  if  $i$  and  $j$  have the same origin

this is unmentioned or otherwise incorrect.<sup>5</sup> [Barceló (1997)] just ignores the multinomial case. The parameters of the multinomial distribution, the probabilities for each O-D pair are  $\hat{g}_i / \sum_{i \in I} \hat{g}_i$ ,  $i \in I$  and the log-likelihood function is

$$\log \mathcal{L}(\hat{\mathbf{g}}|\mathbf{g}) = \sum_{i \in I} \gamma_i \hat{g}_i \log \gamma_i g_i + c, \quad (5)$$

where  $c$ <sup>6</sup> can be omitted from now on because we want to maximize this function. Furthermore we need the additional constraints  $\sum_{i \in I_o} \gamma_i g_i = \sum_{i \in I_o} n_i$  for all origins  $o$  where  $I_o$  is the set of O-D pairs originating in  $o$ .

If the number of sampled trips is high, the multinomial random variable can be approximated by the Poisson random variable. The trips on each O-D pair is considered as a Poisson process, the log-likelihood function is

$$\log \mathcal{L}(\hat{\mathbf{g}}|\mathbf{g}) = \sum_{i \in I} (\gamma_i \hat{g}_i \log(\gamma_i g_i) - \gamma_i g_i) + c, \quad (6)$$

where  $c$  is omitted again. If the sampling rate is the same for each origin, then all the  $\gamma$ 's are the same, so not relevant for the optimization and can be removed from equation (5) and (6). If we suppose that the cars over a road section behave like a Poisson process then a count over that road section is a Poisson random variable. The mean for each random variable (for each counted link) is  $v_a$ . If we also suppose the random variables are independently distributed we get the following log-likelihood function:

$$\log \mathcal{L}(\hat{\mathbf{v}}|\mathbf{v}) = \sum_{a \in \hat{A}} \hat{v}_a \log v_a - v_a + c, \quad (7)$$

where  $c$ <sup>7</sup> can be omitted.

Another assumption is that the traffic counts come from a multivariate normal distribution, unlike the Poisson assumption the counts can depend on each other. The mean value for each link is  $\hat{v}_a$  and given variance-covariance matrix  $W$ <sup>8</sup> the log-likelihood function becomes

$$\log \mathcal{L}(\hat{\mathbf{v}}|\mathbf{v}) = -\frac{1}{2}(\hat{\mathbf{v}} - \mathbf{v})^T W^{-1}(\hat{\mathbf{v}} - \mathbf{v}) + c, \quad (8)$$

where  $c$  is omitted again and  $\mathbf{v}$  and  $\hat{\mathbf{v}}$  are vectors of length  $|\hat{A}|$ <sup>9</sup>. To get a proper objective function for (GP) take

$$F(\cdot, \cdot) = -\log \mathcal{L}(\cdot|\cdot). \quad (9)$$

The second classic estimator is generalized least squares (GLS). The advantage of this approach is that distribution assumptions are not needed, so it is applicable to all a priori data. The traffic counts and a priori matrix are modeled as

$$\begin{aligned} \hat{\mathbf{g}} &= \mathbf{g} + \epsilon_{\mathbf{g}} \\ \hat{\mathbf{v}} &= \mathbf{v} + \epsilon_{\mathbf{v}}, \end{aligned} \quad (10)$$

where the errors  $\epsilon_{\mathbf{g}}$  and  $\epsilon_{\mathbf{v}}$  have some distributions and  $E(\epsilon_{\mathbf{g}}) = E(\epsilon_{\mathbf{v}}) = 0$  is supposed. Unfortunately the dispersion matrices of the  $\epsilon$ 's have to be known, formally they are defined as  $\text{Var } \epsilon_{\mathbf{g}} = Z$  and  $\text{Var } \epsilon_{\mathbf{v}} = W$ . In the GLS approach the objective functions of (GP) become

$$F_1(\mathbf{g}, \hat{\mathbf{g}}) = \frac{1}{2}(\hat{\mathbf{g}} - \mathbf{g})^T Z^{-1}(\hat{\mathbf{g}} - \mathbf{g}) \quad (11)$$

$$F_2(\mathbf{v}, \hat{\mathbf{v}}) = \frac{1}{2}(\hat{\mathbf{v}} - \mathbf{v})^T W^{-1}(\hat{\mathbf{v}} - \mathbf{v}). \quad (12)$$

<sup>5</sup>This is mentioned because the derived objective functions have to fit in the framework, so it is not allowed to change the interpretation of  $\hat{\mathbf{g}}$ .

<sup>6</sup> $c$  is some constant that represents the likelihood that is independent of  $\mathbf{g}$ .

<sup>7</sup> $c$  is some constant that represents the likelihood that is independent of  $\mathbf{v}$ .

<sup>8</sup>also called a dispersion matrix

<sup>9</sup>so  $\mathbf{v}$  is truncated by removing all  $v_a$ ,  $a \in A \setminus \hat{A}$

There are some remarks in the literature that discuss how to derive  $Z$  and  $W$ , I will not go deeper into that. Although it is often assumed that the matrices are diagonal, that implies that there is no covariance between error components. The objective function then has some interesting properties, since the diagonal elements of the dispersion matrix are the variances  $\text{Var } \epsilon_{g_i}$ ,  $i \in I$  and  $\text{Var } \epsilon_{v_a}$ ,  $a \in \hat{A}$  we get

$$F(\mathbf{g}, \mathbf{v}) = \frac{1}{2} \sum_{i \in I} \frac{(\hat{g}_i - g_i)^2}{\text{Var } \epsilon_{g_i}} + \frac{1}{2} \sum_{a \in \hat{A}} \frac{(\hat{v}_a - v_a)^2}{\text{Var } \epsilon_{v_a}}. \quad (13)$$

The interesting property is that the variances are a sort of weight for each term. If the variance on a link count for example is very high its weight will become smaller.

Bayesian statistics are used to determine the feasibility or likeliness of a subjective degree of believe there is on a hypothesis. A Bayesian estimator consists of an a priori probability function  $\pi$  on some parameter (of the hypothesis) and a likelihood function  $\mathcal{L}$  that expresses the likelihood of additional information. Bayes' rule applied to these functions give a conditional (on all information) a posteriori probability function  $\theta$  on the same parameter. Maximizing over the parameter will give the Bayesian estimator. Considering O-D demand estimation the a priori matrix is used to give the a priori probability  $\pi(\mathbf{g}|\hat{\mathbf{g}})$  and the traffic counts are the additional information leading to likelihood  $\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g})$ <sup>10</sup> The a posteriori probability is

$$\theta(\mathbf{g}|\hat{\mathbf{g}}, \hat{\mathbf{v}}) \propto \mathcal{L}(\hat{\mathbf{v}}|\mathbf{g})\pi(\mathbf{g}|\hat{\mathbf{g}})^{11}. \quad (14)$$

The Bayes estimate is  $\text{argmax}_{\mathbf{g}} \theta(\mathbf{g}|\hat{\mathbf{g}}, \hat{\mathbf{v}})$ <sup>12</sup>, applying the logarithm will result in new objective function candidates for **(GP)**:

$$\log(\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g})\pi(\mathbf{g}|\hat{\mathbf{g}})) = \log \mathcal{L}(\hat{\mathbf{v}}|\mathbf{g}) + \log \pi(\mathbf{g}|\hat{\mathbf{g}}). \quad (15)$$

The objective functions for the traffic counts are the same as those obtained from the ML approach. The distinction that can be made is on the interpretation of the dispersion matrices. But for the a priori probability function  $\pi(\mathbf{g}|\hat{\mathbf{g}})$  three different assumptions are mentioned in the literature. If it is supposed to be a multinomial distribution function then  $\hat{g}_i / \sum_j \hat{g}_j$  is the probability for O-D pair  $i$  and we get

$$\log \pi(\mathbf{g}|\hat{\mathbf{g}}) = - \sum_{i \in I} g_i \log \frac{g_i}{\hat{g}_i} + c, \quad (16)$$

where  $c$  can be omitted analogical to what we have seen in the ML approach. Furthermore we need the constraint  $\sum_{i \in I} g_i = \sum_{i \in I} \hat{g}_i$ . This result can also be found in the field of entropy models, the derived equation is actually the entropy function.

If  $\pi(\mathbf{g}|\hat{\mathbf{g}})$  is assumed to be the Poisson distribution function with parameters  $\hat{g}_i$  for each O-D pair  $i$  then we derive

$$\log \pi(\mathbf{g}|\hat{\mathbf{g}}) = - \sum_{i \in I} g_i \left( \log \frac{g_i}{\hat{g}_i} - 1 \right) + c, \quad (17)$$

where  $c$  can be omitted again.

The last possibility is that  $\pi(\mathbf{g}|\hat{\mathbf{g}})$  is a multivariate normal probability function with mean  $\hat{\mathbf{g}}$ . Let  $Z$  be the variance-covariance matrix then we get

$$\log \pi(\mathbf{g}|\hat{\mathbf{g}}) = -\frac{1}{2}(\hat{\mathbf{g}} - \mathbf{g})^T Z^{-1}(\hat{\mathbf{g}} - \mathbf{g}) + c, \quad (18)$$

where  $c$  is omitted. If it is supposed that  $W$  in equation (8) and  $Z$  in equation (18) are diagonal (i.e. the covariances are ignored) then the equations can be simplified and will be similar to equation (13). If both the likelihood and a priori function are multivariate normal distribution function the Bayesian

<sup>10</sup>As we have seen before  $\mathcal{L}(\hat{\mathbf{v}}|\mathbf{g}) = \mathcal{L}(\hat{\mathbf{v}}|\mathbf{v})$ .

<sup>11</sup>' $\propto$ ' means in proportion to. It is used here because after applying Bayes the right hand side also has a factor  $\mathbb{P}(\hat{\mathbf{v}})^{-1}$  (the pdf of the traffic counts). So it assumes that  $\mathbb{P}(\hat{\mathbf{v}})$  is independent of  $\mathbf{g}$ , this simplification is not mentioned in any literature review.

<sup>12</sup> $\text{argmax}_x f(x) = \{x | f(x) \geq f(y), \forall y \in \text{dom } f\}$

approach is very similar with the GLS method. Although the interpretations of the dispersion matrices are very different. in the GLS method the dispersion matrices are about errors in the sampling method and in the Bayesian approach they represent the confidence of the analyst in the a priori matrix.

Precedent approaches are all based on several statistical assumptions. If a practitioner of matrix estimation has statistical information (e.g. variances) about the used data, an appropriate model (i.e. objective function) can be chosen. But probably he or she will not have this information and it is not sure from which distribution the data comes. In line with this [Chen, et al. (2009, in press)] proposes three  $L_p$  norms as objective function, so it is just a measure on the error of the demand and link flows. The supposed norms are  $L_1$ ,  $L_2$  and  $L_\infty$  on  $\mathbb{R}^n$ , they are applied on the difference between the target and resulting values. The  $L_1$ -norm is the sum of the absolute errors; the  $L_2$ -norm is the square root of the sum of the squared errors; the  $L_\infty$ -norm is the maximum of the absolute errors. This leads to the following objective functions.<sup>13</sup>

$$F(\mathbf{g}, \mathbf{v}) = \|\mathbf{g} - \hat{\mathbf{g}}\|_1 + \|\mathbf{v} - \hat{\mathbf{v}}\|_1 = \sum_{i \in I} |g_i - \hat{g}_i| + \sum_{a \in \hat{A}} |v_a - \hat{v}_a| \quad (19)$$

$$F(\mathbf{g}, \mathbf{v}) = \|\mathbf{g} - \hat{\mathbf{g}}\|_2 + \|\mathbf{v} - \hat{\mathbf{v}}\|_2 = \sqrt{\sum_{i \in I} (g_i - \hat{g}_i)^2} + \sqrt{\sum_{a \in \hat{A}} (v_a - \hat{v}_a)^2} \quad (20)$$

$$F(\mathbf{g}, \mathbf{v}) = \|\mathbf{g} - \hat{\mathbf{g}}\|_\infty + \|\mathbf{v} - \hat{\mathbf{v}}\|_\infty = \max_{i \in I} |g_i - \hat{g}_i| + \max_{a \in \hat{A}} |v_a - \hat{v}_a| \quad (21)$$

The ML, GLS and Bayesian inference methods presented in this section are discussed in more detail in [Cascetta & Nguyen (1988)]. To summarize this section Table 1 is presented. A drawback for the multinomial likelihood functions is that it gives additional constraints, namely that the total counted value must be met.

---

<sup>13</sup>The  $\alpha$ 's are omitted here.

Objective function $F_1(\mathbf{g}, \hat{\mathbf{g}})$	Description	Objective function $F_2(\mathbf{v}, \hat{\mathbf{v}})$	Description
$-\sum_{i \in I} \gamma_i \hat{g}_i \log \gamma_i g_i$	ML: multinomial		
$-\sum_{i \in I} (\gamma_i \hat{g}_i \log(\gamma_i g_i) - \gamma_i g_i)$	ML: Poisson	$-\sum_{a \in \hat{A}} \hat{v}_a \log v_a - v_a$	$\begin{cases} \text{ML: Poisson} \\ \text{Bayes: Poisson} \end{cases}$
$\frac{1}{2}(\hat{\mathbf{g}} - \mathbf{g})^T Z^{-1}(\hat{\mathbf{g}} - \mathbf{g})$	$\begin{cases} \text{GLS} \\ \text{Bayes: MVN} \end{cases}$	$\frac{1}{2}(\hat{\mathbf{v}} - \mathbf{v})^T W^{-1}(\hat{\mathbf{v}} - \mathbf{v})$	$\begin{cases} \text{ML: MVN} \\ \text{GLS} \\ \text{Bayes: MVN} \end{cases}$
$\sum_{i \in I} g_i \log \frac{g_i}{\hat{g}_i}$	Bayes: multinomial		
$\sum_{i \in I} g_i \left( \log \frac{g_i}{\hat{g}_i} - 1 \right)$	Bayes: Poisson		
$\sum_{i \in I}  g_i - \hat{g}_i $	$L_1$ -norm	$\sum_{a \in \hat{A}}  v_a - \hat{v}_a $	$L_1$ -norm
$\sqrt{\sum_{i \in I} (g_i - \hat{g}_i)^2}$	$L_2$ -norm	$\sqrt{\sum_{a \in \hat{A}} (v_a - \hat{v}_a)^2}$	$L_2$ -norm
$\max_{i \in I}  g_i - \hat{g}_i $	$L_\infty$ -norm	$\max_{a \in \hat{A}}  v_a - \hat{v}_a $	$L_\infty$ -norm

Table 1: Several objective functions



## 3 OmniTRANS

### 3.1 Omnitrans International BV

Omnitrans International BV is a software company established in 2003 as a subsidiary company of Goudappel Coffeng BV. The staff consists of fifteen high qualified IT engineers, mathematicians and transport engineers. The main activity is the maintenance and development of the OmniTRANS software package. It has been the Dutch market leader in transport modelling and planning since the start of the company. Besides the software package, Omnitrans International also delivers advanced specialized software consultancy, this varies from plug-ins for OmniTRANS via real time applications for traffic control centres to configuration management systems. Within the company there are five groups:

- **Concepts:** This group is entrusted with the research for OmniTRANS and most of the consultancy. To make the software the state-of-the-art in the field new traffic concepts are delivered. Matrix estimation is one of the current research projects.
- **Development:** Approximately every year a major new version of OmniTRANS is released, the development group continuously implements new features. Some novelties in the latest version are a multi-class assignment method, split screen functionality and a parameter manager.
- **Maintenance:** This group resolves the problems with the software. It includes the customer helpdesk where assistance is extended and requirements and issues are collected.
- **Marketing:** Software still doesn't sell itself, therefore a marketing team keeps up relations with current and potential customers.
- **Training:** Several courses are organized to educate the users. There is a wide variety of courses.

As mentioned earlier, Omnitrans International is a subsidiary company of Goudappel Coffeng BV. Goudappel Coffeng, founded in 1963, is the largest traffic and transport consultancy company in the Netherlands. It has offices in Deventer, Leeuwarden, Eindhoven, The Hague and Amsterdam; Omnitrans International is established in Deventer. The Netherlands are among the countries with the highest population density, this requires an efficient arrangement of the available space where mobility is crucial. Goudappel Coffeng advises all kinds of authorities, from small city councils to the European Commission, about all kinds of traffic, transport and mobility problems. The secondment of Goudappel Coffeng is accommodated in sister company Tiem BV.

### 3.2 The OmniTRANS software package

There are two ways to approach the study of traffic flows, microscopic and macroscopic. In the microscopic model each vehicle is modelled, it has a place and velocity and anticipates on the other nearby vehicles. The behaviour of each driver is simulated as well as possible. Such models cannot be applied on large scale networks with hundreds of kilometres of roads. It will simply take too much memory and processing time. The macroscopic model focuses on three variables for road sections; the traffic load, velocity and density are considered. Not the individual road users but the traffic flows are considered, this allows much larger networks.

Another dichotomy in traffic engineering concerns the static and dynamic models. Static models have a single (equilibrium) result for the whole time scope, changes in the traffic load within the single time interval are not considered. Dynamic models do consider these changes, the time scope is divided into several smaller intervals, whose length is usually 2 to 15 minutes. In dynamic models the results are specified for each interval, which allows visualisation of the output as a movie in which the traffic builds up and resolves. There is a wide variety of dynamic traffic engineering tools available, the more

detailed they are, the smaller the workable network size is. OmniTRANS can deal with dynamic and static models, there are different assignment methods for each type. This study only deals with static models.

OmniTRANS is a macroscopic modelling package that distinguishes oneself for the management of large amounts of data and its job-engine which gives the user a very powerful customization tool. These features are discussed later, the main structure is described first. The fundamentals of an OmniTRANS project are the dimensions, the network, the data, the jobs and variants. The dimensions define what kind of traffic is considered. The network represents the physical road network, traffic enters and leaves the network through centroids that represents areas. The most important datasets are the origin-destination (O-D) matrices that include the traffic flow information between each O-D pair. Usually one O-D matrix is defined for each dimension. The software includes a job engine which enables the user to exploit a wide variety of customized algorithms. A variant is a combination of a network and a data set, they are used to consider several scenarios. That can be a change or extension in the network or different year (possibly a prediction for the future). The purpose of this section is to give an idea of the working of the software and give an explanation of most aspects, although it is not extensive and complete. The reader should refer to the OmniTRANS manual<sup>14</sup> for the details.

### 3.2.1 Dimensions

The setup of a model begins with defining the dimensions, a dimension is a 4-tuple (purpose, mode, time, user), denoted as PMTU. The dimensions represent the scope of the model, they prescribe how the traffic is classified or divided. It is important for analysts that the results are specified per dimension, so the source of the traffic is known. A PMTU consists of:

- **P, purpose:** This can be the different objectives for a trip, e.g. home  $\rightarrow$  work, work  $\rightarrow$  home, business, shopping or school.
- **M, mode:** The mode describes the kind of transport used, e.g. car, freight, public transport or cycling. When the traffic is assigned to the network it is of great importance which mode is used, different modes can place a load on different pieces of the network.
- **T, time:** This describes the time that a trip occurs, e.g. AM peak, PM peak or the remainder of a day. In dynamic models smaller intervals are used, this leads to a drastic increase of the number of dimensions and thereby the complexity of a project.
- **U, user:** The user is free to define this, a new partition of the traffic can be made. One example is car availability, this has a great influence on the mode choice. Another example is a disambiguation on internal, external and through traffic, this is especially important for local authorities that want to take counter measures against cut-through traffic.

The majority of the data is associated with a dimension, in this way it is easier to manage the data. The dimensions can be related to each other, it is common to have dimensions with totals. For mode for example one can have, car, light freight, heavy freight and in addition motorized vehicles which is the sum of the three. It is the responsibility of the model builder and user to keep the associated data sets consistent when changes are made.

### 3.2.2 The network

One of the main parts of a model is the network, it basically consists of centroids, nodes and links. The study area is divided in several smaller zones, each of these zones is represented by a centroid. A centroid is a point in the network where traffic can originate or terminate, it can be a origin or a

---

<sup>14</sup>Digital, delivered with the software package

destination for a trip. The real road network is represented with nodes and links, this can be seen as a directed graph. The centroids are connected with one or more nodes (connectors), traffic can enter and exit the network through these connectors. In general not every road is modelled, small residential streets and no-through roads are omitted. The model builder decides how precise the network becomes, in general the larger the study area, the less dense the network is. In a project for a small community almost every street will be in the network and there will be several centroids in the area, but for a national project probably only the main roads of the community are considered and the whole community is taken as a centroid.

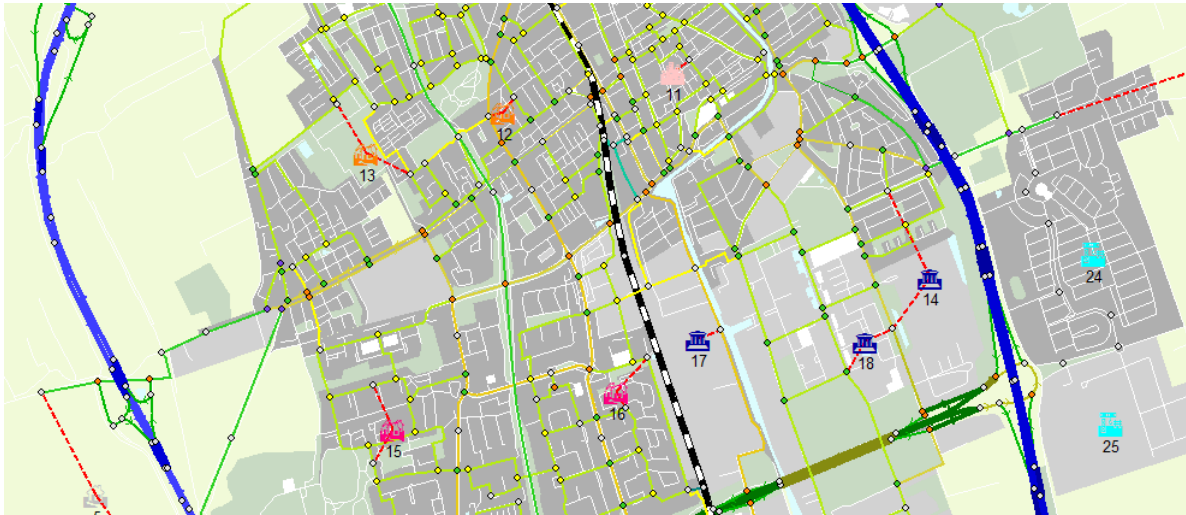


Figure 1: The Delft network in OmniTRANS with centroids, nodes and links.

Figure 1 shows a part of the Delft network. The numbered icons are the centroids, the coloured dots are the nodes, they are connected with links. The background image shows a map of the city, it is easy to see that not all roads are modelled. The discontinuous lines are the connectors between centroids and nodes.

In this section only a small part of the possibilities is described, OmniTRANS delivers a versatile set of tools to represent the real road network as closely as possible. Every element of the network has several attributes that describe its properties. Consider links for example: capacity, maximum speed and length are some of its attributes. The speed and capacity can be defined for each dimension. On junctions each approach lane can be specified and the traffic signal program can be set. A new feature of OmniTRANS is controls, they can alternate the network depending on several variables, such as current load and time. For example with controls, a controlled access to motorway with traffic lights<sup>15</sup> can be added.

### 3.2.3 Data

A project is dependent on different kind of data sets which are generally pretty large. Several kinds of information are:

- **O-D matrices:** Probably the most important information for each project are the O-D matrices. They define the number of trips between each O-D pair for each dimension. The matrices can be obtained by, for example, the gravity model, discrete choice models, large surveys or growth models.<sup>16</sup> In OmniTRANS the O-D matrices are grouped in matrixcubes, which can contain a matrix for each dimension. A project can have several matrix cubes, each with another context, which can be different years or the information before and after an operation (like matrix estimation).

<sup>15</sup>’Toeritdosering’ in Dutch

<sup>16</sup>More information on these models can be found in all good traffic engineering books (e.g. [Cascetta (2001)]).

- **Socio-economic zonal data:** The socio-economic data are for example the number of residents and the number of available jobs of a zone. It is related with a specific matrix cube, this automatically specifies the context of the socio-economic data. This data is mainly used in methods to create O-D matrices.
- **Productions and attractions of zones:** The production and attraction of a zone are respectively the number of trips that originate and terminate from that zone. It is an important information source for the gravity model. Like the socio-economic data, this dataset is also related to matrixcubes.
- **Loads and other results on the network:** Final results of a model are usually several parameters on the elements of the network. It can be the delay at a junction, or the average speed on a road section. The load, the number of cars on a road section is very important information. The results are stored according to the dimension, the context and iteration. The iteration is added because most algorithms that generate these results are iterative and users want to compare results from different iterations. The results can be visualized in the network, with pie or bar charts on junctions and zones and with bandwidth and colours on the links.
- **Screenline matrices:** Important data for matrix estimation is the route choice for O-D pairs. It is not straightforward to retrieve all route choices from an assignment. But it is possible to store the proportion of the O-D pair flow that uses a particular link. Particularly this information is stored for links with count information. For each dimension and each chosen link a screenline matrix is stored. In that matrix the proportion that passes the link is stored for each O-D pair. Generally such a matrix will contain a majority of zeroes since for most O-D pairs there will be no route over the link.
- **Traffic counts:** Traffic counts are specified in the network of a project, they can be attached to a link. For each of the (one or two) directions of the link one count can be added. The count contains count data for several designated dimensions and sets of dimensions. For example a count can contain data for the dimension car and freight and as well hold count data for the sum of these two. The dimension also defines the time period of the count. The user is responsible for the consistency of the data. Another attribute of a count is the weight that specifies the reliability of the data. This parameter between 0 and 1 is used within the matrix estimation algorithms, as will become clear in later sections.

In traffic models there is a lot of data, the management of the data is a task that needs to be done with precision. An aspiration of **OmniTRANS** is to make the management of data as easy as possible. For each variant the available data is presented in an overview, this shows the used matrix cube, the results and the screenline matrices. There is a matrixcube editor that gives access to each matrix of the cube and related zonal data. The concept of matrixcubes is also designed to make the data management easier, it addresses the context of the matrices.

### 3.2.4 Jobs

A distinguishing feature of **OmniTRANS** is the job engine, it delivers another dimension of freedom to the transport engineer. The job engine consists of a scripting language and a library of classes. The classes encompass a wide variety of **OmniTRANS** objects and algorithms. The language is called the **OmniTRANS Job Language (OJL)** and it essentially is an extension of Ruby, an object oriented scripting language. The extension provides five kinds of classes:

- **Data access classes:** To access all objects described in this section within the OJL, there is a class available for each object type. This is the bridge between an **OmniTRANS** project and the script, it enables users to retrieve, modify and store information in their project with OJL. An example is `OtMatrixCube`, a class particularly used to access O-D matrices. A request for

a specific PMTU in a matrixcube will deliver an instance of `OtMatrix`, the OmniTRANS matrix class. Another key class is `OtTable`, the best way to obtain OmniTRANS database tables in the OJL engine. A lot of information is stored in the database of OmniTRANS, it consists of almost ninety tables. It contains all the information about the network, but also the dimension definitions and traffic count values.

- **Charting classes:** It consists of all classes that can visualize data in different sorts of charts. These are not applicable or used for this study,
- **Data import/export classes:** Import and export classes are tools to import and export information to and from external sources and file formats. The compatibility with formats like ESRI shapefiles<sup>17</sup> and Saturn<sup>18</sup> networks makes the user less dependent on one particular package.
- **Modelling classes:** The core of each traffic and transport package is the set of algorithms, the arithmetic methods. The model of a project generally consists of a sequence of steps, where each step uses some algorithm. This study is aimed at one single modelling class: Matrix Estimation. OmniTRANS comes with seven other modelling classes, they are mentioned here, please see [Cascetta (2001)] for more information on the theory of transportation models.
  - `OtChoice`, four different discrete choice models of the logit type.
  - `OtGravity`, the simultaneous multi-modal gravity model.
  - `OtGrowthFactor`, the growth factor model.
  - `OtMadam`<sup>19</sup>, a dynamic assignment method
  - `OtTraffic`, several static assignment methods including equilibrium and volume averaging.
  - `OtTransit`, a public transport assignment method
  - `OtTripEnd`, the trip end model
- **Utility classes:** There are two technical classes that can manage OmniTRANS software parameters and suchlike.

With the standard Ruby classes combined with the OmniTRANS classes a user can build a project for almost every purpose in traffic and transport. The OJL has a simple syntax, it is fairly intuitive and together with the OmniTRANS user manual every user should be able to use all modelling classes of OmniTRANS. Furthermore SciTE, a source code editor, is installed with OmniTRANS, it supports syntax highlighting for the OJL.

### 3.2.5 Variants

A variant is the combination of a network and a matrixcube. A project can have several variants, each variant shows another situation, that can be a difference in the network or in the data. Examples are new estate or different years. A variant can have subvariants, a subvariant has the same network as the variant, but different data. Figure 2 shows the relationship between the three. The different scenario's that are considered for a study are translated to variants and subvariants. Since OmniTRANS stores results per variant and subvariant, conclusions are drawn from comparing variants and subvariants.

---

<sup>17</sup>For GIS applications like ArcView

<sup>18</sup>Traffic simulation and assignment software package from Great Britain

<sup>19</sup>**Macroscopic Dynamic Assignment Model.** Its successor `OtStreamline` is in the final stage of development, it is an all-embracing dynamic modelling package. It comprehends the propagation, route choice and junction modelling and it includes MaDam

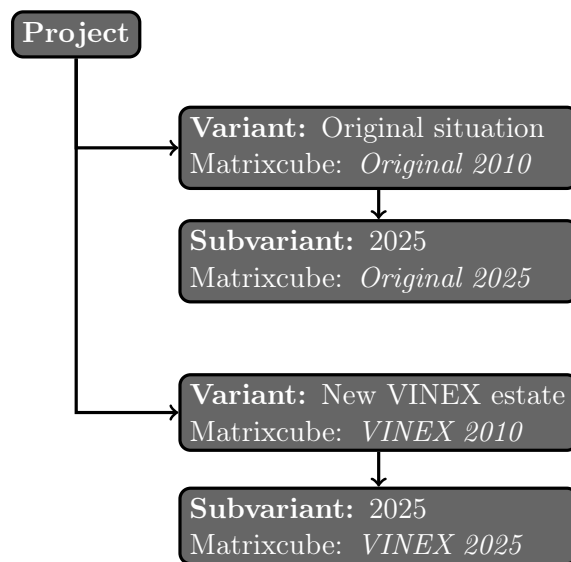


Figure 2: Project with variants and subvariants

## 4 OmniTRANS Matrix Estimation method

This chapter describes how the current matrix estimation method in OmniTRANS works and a revision of this method is proposed. This is done in a variable setting that is compatible with OmniTRANS, some OmniTRANS features are generalized, this makes it easier to describe the methods and makes it easier to extend the functionalities of the software.

### 4.1 Model setting

OmniTRANS can use four different types of information as input for matrix estimation: traffic counts, screenlines, blocks and trip ends.

- **Traffic count:** A traffic count is located on a link of the network and contains information on the number of vehicles which pass that position. The counted value is specified with one or more PMTU's which defines its context, for example car or freight. This is the most fundamental type of input.
- **Screenline:** A screenline contains information on the total number of vehicles that pass several count locations. It practically is a set of traffic counts with a single counted value. As well as the traffic counts a screenline is specified with one or more PMTU's. In practice this is useful on highways when there are parallel highway sections (e.g. for express and local traffic), then it is important that the total traffic fits the total counted value rather than the specific counts for both sections. A screenline can also be useful if a network is divided by a river, then it contains the information on the total traffic over all bridges and tunnels crossing the river. Model builders should be very careful when defining screenlines, if a route for any O-D pair passes the screenline more than once the demand for this O-D pair is counted double. This leads to a bias in the information.
- **Block:** A block contains information on the total demand for a set O-D pairs and is also specified with one or more PMTU's. The information of blocks is usually obtained from home surveys.
- **Trip end:** A trip end is a special case of a block, it contains information on the total production or attraction of a specific zone. So it is a block where the set of O-D pairs consists of O-D pairs with the same origin or destination. This information usually comes from the socio-economic data of a zone.

In this study all types mentioned above are generalized to restrictions. They all share the same properties and can all be treated the same way with the presented model setting. As shown later this also delivers possibilities to easily extend the software with additional restriction types.

To describe the matrix calibration method some notations will be introduced:

- $I$  is the set of O-D pairs.
- $D$  is the set of all dimension that are defined in an OmniTRANS project,  $D$  will be a set of PMTU's.
- $g_i^d$ ,  $i \in I, d \in D$  is the demand for O-D pair  $i$  in dimension  $d$ .
- $\hat{g}_i^d$ ,  $i \in I, d \in D$  is the a priori demand for O-D pair  $i$  in dimension  $d$ .
- $R$  is the set of restrictions.
  - $C_r$ ,  $r \in R$  is the counted value of restriction  $r$ .
  - $D_r$ ,  $r \in R$  is the set of dimensions belonging to  $r$ .

- $P_{ir}^d$ ,  $r \in R, i \in I, d \in D_r$  is the fraction of the demand of O-D pair  $i$  in dimension  $d$  that applies to restriction  $r$ .
- $\varepsilon_r \in [0, 1]$ ,  $r \in R$  is the elasticity or weight of  $r$ .

To make clear what  $P_{ir}^d$  the derivation for each restriction type is explained:

- **Traffic count:** For each traffic count and each dimension a proportion matrix<sup>20</sup> is stored that gives the fraction of the demand that pass the count, this comes from an assignment. These matrices actually store all  $P_{ir}^d$  values for traffic counts.
- **Screenline:** A screenline is a set of traffic counts, denote this set as  $T$ . The  $P_{ir}^d$  values for a screenline can now be calculated with  $P_{ir}^d = \sum_{t \in T} P_{it}^d$ .
- **Block:** For a block  $P_{ir}^d = 1$  if the O-D pair  $i$  is in the block, otherwise it is equal to zero.
- **Trip end:** For a trip end  $P_{ir}^d = 1$  if the O-D pair  $i$  is in the trip end, otherwise it is equal to zero.

Note that  $P_{ir}^d$  is only defined for  $d \in D_r$ .

**EXAMPLE 1** (Proportion values  $P_{ir}^d$ )

Figure 3 shows illustrative examples of the  $P_{ir}^d$  values for each of the available restriction types in OmniTRANS. The matrices for a specific restriction and a single dimension are drawn. The displayed traffic count is included in the screenline. Block and trip end values are either 0 or 1 and a the trip end always consists of a whole row or column of ones.

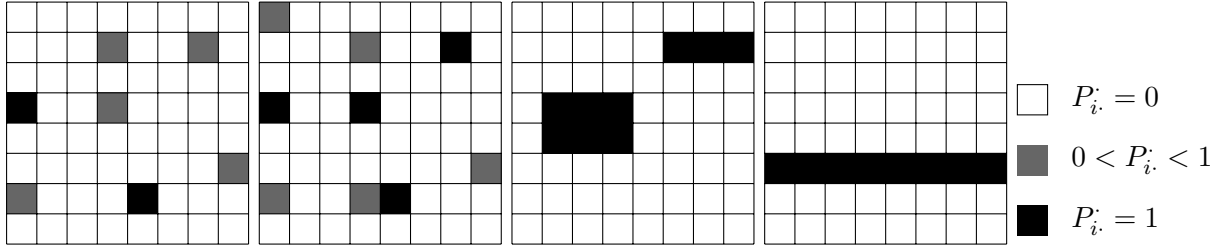


Figure 3: Illustrative examples of the  $P_i^d$  matrix for a (a) traffic count, (b) screenline, (c) block and (d) trip end.

## 4.2 The method

The matrix estimation modelling class, the arithmetic package to improve the O-D matrix with additional restriction information is analyzed in this section. Another frequently used phrase for this method is simultaneous matrix calibration (SMC), where simultaneous refers to the fact that multiple dimension are considered at once. The OmniTRANS class that is used in the OJL for this purpose is `OtMatrixEstimation`. There are several disadvantages of the current method, these will also be addressed in this section. The class is developed in the early years of the software package, it did not change in the recent period.

The current matrix calibration method consists of several multiplications on the a priori O-D matrices. First a global multiplication factor is derived and then each restriction is successively treated in an iterative process. Algorithm 1 shows how the current method works in pseudocode. Stopping criteria can be a maximum number of iterations or a convergence factor. On line 4 a  $\delta_i^d$ ,  $i \in I, d \in D$  is used to determine for which O-D pairs  $X_0$  should be applied, it is defined as follows

$$\delta_i^d = \begin{cases} 1 & \text{if } \exists r \in R \text{ such that } d \in D_r \text{ and } P_{ir}^d > 0 \\ 0 & \text{otherwise} \end{cases} \quad .21 \quad (22)$$



---

**Algorithm 1** The OmniTRANS algorithm
 

---

```

1:  $X_0 \leftarrow \frac{\sum_{r \in R} C_r}{\sum_{r \in R} \sum_{d \in D_r} \sum_{i \in I} P_{ir}^d \hat{g}_i^d}$ 
2: for all  $d \in D$  do
3:   for all  $i \in I$  do
4:      $g_i^d \leftarrow (X_0)^{\delta_i^d} \hat{g}_i^d$ 
5:   end for
6: end for
7:  $stopCrit \leftarrow \text{false}$ 
8: while not  $stopCrit$  do
9:   for all  $r \in R$  do
10:     $X_r \leftarrow \frac{C_r}{\sum_{d \in D_r} \sum_{i \in I} P_{ir}^d g_i^d}$ 
11:    for all  $d \in D_r$  do
12:      for all  $i \in I$  do
13:         $g_i^d \leftarrow g_i^d (X_r)^{\varepsilon_r P_{ir}^d}$ 
14:      end for
15:    end for
16:  end for
17:  update  $stopCrit$ 
18: end while

```

---

Figure 4 gives a simplified illustration of this method.

In OmniTRANS the order of the restrictions can be defined on two levels. First a global order in the restriction type (i.e. counts, screenlines, trip ends and blocks) can be made and secondly a more specific order can be made for each count, block and screenline.

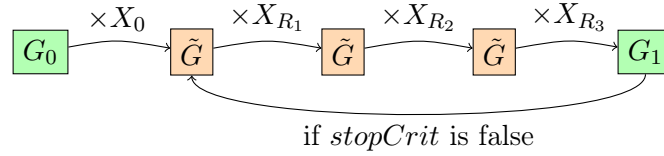


Figure 4: An illustration of the original method

One of the biggest drawbacks of the current matrix calibration method is that each restriction (e.g. count, screenline) is treated successively. This has following consequences for the results:

- Results depend on the order of the input.
- Corrections on the O-D matrix made by the first restrictions can be overridden by later restrictions.
- As a result of the latter the final O-D matrix will better represent the restrictions that are treated later. This will be explained in more detail in section 7.1.

Another drawback is that the method does not use the a priori matrix actively. The difference between the generated O-D matrix and a priori matrix can become very large. Current users sometimes apply the trip ends at the end of the algorithm to partly conquer this problem.

---

<sup>20</sup>This matrix is called a screenline matrix in OmniTRANS

<sup>21</sup>If blocks are used there used to be a fault in OmniTRANS, then  $\delta_i^d \equiv 1$ .

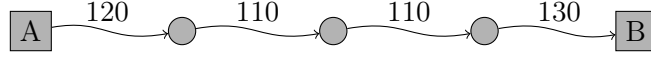


Figure 5: Example network

### 4.3 Revision of the algorithm

This section presents a revision of the algorithm where all restriction can be treated at once. If all factors  $X_r$  are based on the input matrix and we will then multiply the matrix with all these factors the results will be incorrect. A correction on the factors based on the total number of restrictions that apply to an O-D pair will result in a better estimation.

**EXAMPLE 2** (Network with undesired results)

Take for example the network in figure 5, there are two centroids and four links, each with a count, from left to right numbered 1, 2, 3 and 4. The counted values are above the links. Suppose the value for O-D pair  $AB$  in the a priori matrix is 100. The original method with the order 1-2-3-4 and  $X_0 = 1$  will result in a multiplication with factors  $C_1/g_{AB} = 120/100 = 1.2$ ,  $C_2/g_{AB} = 110/120 \approx 0.917$ ,  $C_3/g_{AB} = 110/110 = 1$  and  $C_4/g_{AB} = 130/110 \approx 1.182$  and in an a posteriori matrix with value  $g_{AB} = 130$  for O-D pair  $AB$ . So the only relevant restriction is count 4, the others are dummy.

If we base all factors on the a priori matrix then then  $X_1, X_2, X_3$  and  $X_4$  are respectively 1.2, 1.1, 1.1 and 1.3. Now  $100 \times 1.2 \times 1.1 \times 1.1 \times 1.3 = 188.76$  is a bad candidate for the a posteriori matrix. In the revised algorithm the weighted geometrical mean of the factors is used for the multiplication, this results in  $100 \times (1.2 \times 1.1 \times 1.1 \times 1.3)^{\frac{1}{4}} \approx 117.21$ .

This idea is presented in pseudocode in algorithm 2 and figure 6 gives a simplified illustration of the revised method.

---

**Algorithm 2** The revised OmniTRANS algorithm

---

```

1:  $X_0 \leftarrow \frac{\sum_{r \in R} C_r}{\sum_{r \in R} \sum_{d \in D_r} \sum_{i \in I} P_{ir}^d \hat{g}_i^d}$ 
2: for all  $d \in D$  do
3:   for all  $i \in I$  do
4:      $g_i^d \leftarrow (X_0)^{\delta_i^d} \hat{g}_i^d$ 
5:      $\mathcal{P}_i^d \leftarrow \sum_{r \in R} \varepsilon_r P_{ir}^d$ 
6:   end for
7: end for
8:  $stopCrit \leftarrow \text{false}$ 
9: while not  $stopCrit$  do
10:  for all  $r \in R$  do
11:     $X_r \leftarrow \frac{C_r}{\sum_{d \in D_r} \sum_{i \in I} P_{ir}^d g_i^d}$ 
12:  end for
13:  for all  $d \in D_r$  do
14:    for all  $i \in I$  do
15:       $g_i^d \leftarrow g_i^d \left( \prod_{r \in R} (X_r)^{\varepsilon_r P_{ir}^d} \right)^{\frac{1}{\mathcal{P}_i^d}}$ 
16:    end for
17:  end for
18:  update  $stopCrit$ 
19: end while

```

---

**EXAMPLE 3** (Network with multiple routes)

Figure 7 is another example that incorporates multiple routes and weights other than one. There are two zones ( $A$  and  $B$ ), four links, two nodes and four counts (①, ②, ③ and ④). The demand for O-D pair  $AB$  is 100 and it is supposed that 20 % of the trips take the 'upper' route and the other 80 % take the

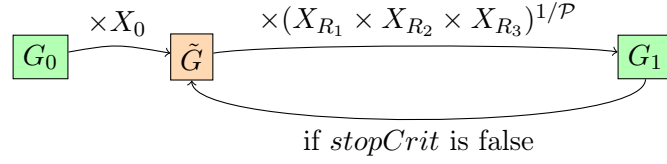


Figure 6: An illustration of the revised method

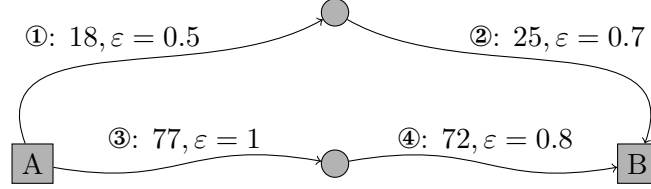


Figure 7: Another example network

'lower' route. So for the 'upper' two counts we have proportion value  $P_{AB①} = P_{AB②} = 0.2$  and for the 'lower' two counts we have proportion value  $P_{AB③} = P_{AB④} = 0.8$ . Together with the  $\varepsilon$ 's this leads to  $\mathcal{P}_{AB} = \varepsilon_① \times P_{AB①} + \varepsilon_② \times P_{AB②} + \varepsilon_③ \times P_{AB③} + \varepsilon_④ \times P_{AB④} = 0.5 \times 0.2 + 1 \times 0.2 + 0.7 \times 0.8 + 0.8 \times 0.8 = 1.5$ . Furthermore  $X_① = \frac{18}{0.2 \times 100} = 0.9$ ,  $X_② = \frac{25}{0.2 \times 100} = 1.25$ ,  $X_③ = \frac{77}{0.8 \times 100} = 0.9625$  and  $X_④ = \frac{72}{0.8 \times 100} = 0.9$ . One iteration of the revised method results in  $g_{AB} = \hat{g}_{AB} \times (X_①^{\varepsilon_① \times P_{AB①}} \times X_②^{\varepsilon_② \times P_{AB②}} \times X_③^{\varepsilon_③ \times P_{AB③}} \times X_④^{\varepsilon_④ \times P_{AB④}})^{\frac{1}{1.5}} = 100 \times (X_①^{0.5 \times 0.2} \times X_②^{0.2} \times X_③^{0.7 \times 0.8} \times X_④^{0.8 \times 0.8})^{\frac{1}{1.5}} \approx 96.4163$ .

## 5 The gradient descent method

In this chapter the gradient descent method is explained and the matrix estimation problem is subjected to it. As a basis the bilevel problem (BLP) on page 11 is taken and modified for OmniTRANS. The gradient method is applicable on the new problem since under some simplifications it has a convex subproblem. Commencing with a general introduction to gradient descent methods the development of the final algorithm is presented. It turned out that this result is a tool that delivers new possibilities for OmniTRANS users.

### 5.1 The gradient descent method in a nutshell

The remainder of this paragraph is a description of the gradient descent idea in a nutshell. A frequently used family of methods to deal with optimization problems are the descent methods. Those methods search a point in the domain of an objective function that is a local minimum in several steps. From a starting point some search direction and a step length is chosen iteratively. There are several possibilities in choosing a search direction, though the negative gradient or a negative subgradient of the objective function are high potential candidates. The step length is also very important, it is necessary that it is not too big, overshooting the minimum, or too small, slowing down the optimization process. It would be perfect if the step length minimizes the objective function over the search direction. Unfortunately this is not always possible, in that case a line search should be performed. The descent methods are of special interest in convex optimization, since then a local minimum is also a global minimum.

### 5.2 General form of the gradient descent method

A very important problem in mathematics is to find the optimum (minimum or maximum) of a function of multiple variables. In applications it often refers to the point with maximum profit in the broadest sense of the word. Because of the wide variety of properties in optimization problems they are categorized. The two most important properties of optimization problems are the form of the objective function and the existence (and type) of constraints. Several categories of functions that have special attention in the field of optimization are linear functions, quadratic functions, integer functions and convex functions. Especially the latter is interesting for this study. For some optimization problems the variables are bounded by restrictions, these can be grouped by inequality constraints, strict inequality constraints and equality constraints.

In this section the gradient descent method for convex functions of multiple variables without constraints is presented. Let  $n > 1, n \in \mathbb{N}$  and  $F : \mathbb{R}^n \rightarrow \mathbb{R} \in \mathcal{C}^1$  be a convex function<sup>22</sup>, then

$$\min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}) \quad (23)$$

is its corresponding minimization problem. The advantage of the convexity of  $F$  is that if a local minimum is found, it will be the global minimum as well. In general, problems that have multiple (local) minimizers are significantly harder, it is not straightforward to find all minimizers and there is a decision problem in the case of multiple global minimizers. In 23 it is even possible that the convex function  $F$  has a range of global minimizers. Because in this study the emphasis lies in finding the minimum, not in finding the complete range of minimizers, the goal is to find one minimizer.

**EXAMPLE 4** (A simple minimization problem)

This example considers the minimization of a quadratic function. Consider

$$\min_{x,y \in \mathbb{R}} M(x,y) = (2x + 5)^2 + (y - 6)^2 + xy,$$

---

<sup>22</sup> $\mathcal{C}^1$  is the space of functions that have continuous partial derivatives at each point.

figure 8 shows a three-dimensional plot of  $M(x, y)$ . It is not hard to see that it is a convex function from the plot. For the proof it must be shown that the Hessian

$$\begin{pmatrix} \frac{\partial^2 M(x,y)}{\partial x^2} & \frac{\partial^2 M(x,y)}{\partial x \partial y} \\ \frac{\partial^2 M(x,y)}{\partial y \partial x} & \frac{\partial^2 M(x,y)}{\partial y^2} \end{pmatrix} = \begin{pmatrix} 8 & 1 \\ 1 & 2 \end{pmatrix}$$

is positive semidefinite. This is true since the leading principal minors<sup>23</sup> are 8 and 15, both positive.

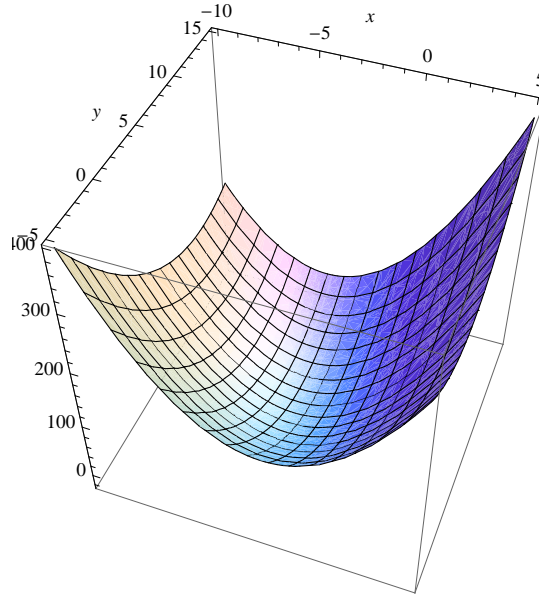


Figure 8: A three dimensional plot of  $M(x, y)$

Necessary and sufficient conditions for optimality in problem (23) are

$$\nabla F(\mathbf{x}) \equiv \mathbf{0}, \quad (24)$$

that is, all partial derivatives of  $F$  are equal to zero. The conditions are sufficient since  $F$  is convex and only one global minimum is required. This leads to a system of equations:

$$\frac{\partial}{\partial x_i} F(\mathbf{x}) = 0, \quad \forall i \in \{1, 2, \dots, n\}. \quad (25)$$

Depending on  $F$  and the number of dimensions  $n$ , this system can be solved and a solution of the problem is found. If  $F$  is a linear or quadratic then the system (25) will respectively lead to a direct answer or a system of linear equations. Systems of linear equations can be solved for considerable high dimension with current software packages.

**EXAMPLE 4 (CONTINUED)**

The partial derivatives lead to

$$\left. \begin{aligned} \frac{\partial M(x,y)}{\partial x} = 20 + 8x + y = 0 &\Rightarrow -20 - 8x = y \\ \frac{\partial M(x,y)}{\partial y} = -12 + x + 2y = 0 &\Rightarrow 6 - \frac{x}{2} = y \end{aligned} \right\} \Rightarrow \begin{cases} x = \frac{-52}{15} \approx -3.4667 \\ y = \frac{116}{15} \approx 7.7333 \end{cases},$$

which is the only global minimizer. The minimum value is  $-\frac{301}{15} \approx -20.0667$ .

If it is not possible to solve (25) with the available hardware and software, then a different approach is required. This is where the gradient descent method comes into play. It is an iterative method that

<sup>23</sup>The leading principal minors of an  $n \times n$ -matrix are the determinants of all  $i \times i$  upper left corner submatrices,  $\forall i \in \{1, \dots, n\}$ .

searches the domain step by step for a minimum. Each iteration consists of two steps, first a search direction is determined and second a step length is determined. This is the principle of each descent method and for both the direction and step length several candidates can be taken. The descent method described here takes the negative gradient as direction and calculates the optimal step length over the search direction.

---

**Algorithm 3** The gradient descent method

---

```

1: stopCrit ← false
2: choose  $\mathbf{x}$ 
3: while not stopCrit do
4:    $\mathbf{s} \leftarrow -\nabla F(\mathbf{x})$ 
5:    $\lambda^* \leftarrow \frac{\partial}{\partial \lambda} F(\mathbf{x} + \lambda \mathbf{s}) = 0$ 
6:    $\mathbf{x} \leftarrow \mathbf{x} + \lambda^* \mathbf{s}$ 
7:   update stopCrit
8: end while

```

---

Algorithm (3) describes the method. First a starting point  $\mathbf{x}$  is chosen, this can be a random point in  $\mathbb{R}^n$ . On line 4 the direction  $\mathbf{s}$  is calculated as the negative gradient of  $F$  at point  $\mathbf{x}$ . Directly afterwards the optimal step length  $\lambda^*$  is determined by finding the minimum of  $F(\mathbf{x} + \lambda \mathbf{s})$  regarding to  $\lambda$ , note that  $\mathbf{x}$  and  $\mathbf{s}$  are constants here. This can be achieved with finding the  $\lambda$  with derivative zero since  $F$  is convex. The boolean variable *stopCrit* is the stopping criterion. After each iteration this variable is updated and if it is fulfilled the algorithm ends. There are several possibilities for the criterion, those include a maximum number of iterations, absolute convergence and relative convergence.

**EXAMPLE 4 (CONTINUED)**

Albeit the global minimizer of  $M(x, y)$  is already known, it also can be found with the gradient descent method. The search direction at some fixed point  $(x, y)$  is

$$\mathbf{s} = -\nabla M(x, y) = - \begin{pmatrix} 20 + 8x + y \\ -12 + x + 2y \end{pmatrix}$$

The optimal step can now be calculated with solving

$$\frac{\partial}{\partial \lambda} M(x - (20 + 8x + y)\lambda, y - (-12 + x + 2y)\lambda) = 0.$$

Writing this out shows that the problem is simply to find the minimum of a parabola

$$\begin{aligned} & \frac{\partial}{\partial \lambda} \left( 61 + 20x + 4x^2 - 12y + xy + y^2 - \right. \\ & (544 + 296x + 65x^2 - 8y + 20xy + 5y^2) \lambda + \\ & \left. (1504 + 1180x + 265x^2 + 140y + 85xy + 10y^2) \lambda^2 \right) = 0. \end{aligned} \quad (26)$$

With this information a new point  $(x, y)$  is found, which is the input for a new iteration. Figure 9a show the contours of  $M(x, y)$  and ten gradient steps starting at the point  $(0, 0)$ , the optimal value is drawn as a point. This figure demonstrates how the gradient method finds the optimal value. Figure 9b shows the point transition for several other values. Notice that the search direction is orthogonal to the contour line at the starting point and parallel to the contour line at the end point.

### 5.3 Specification of the bilevel problem for OmniTRANS

In this section the bilevel problem (BLP) on page 11 is modified to fit the OmniTRANS model as described in section 4.1. This leads to a new, specific, optimization problem for matrix estimation in

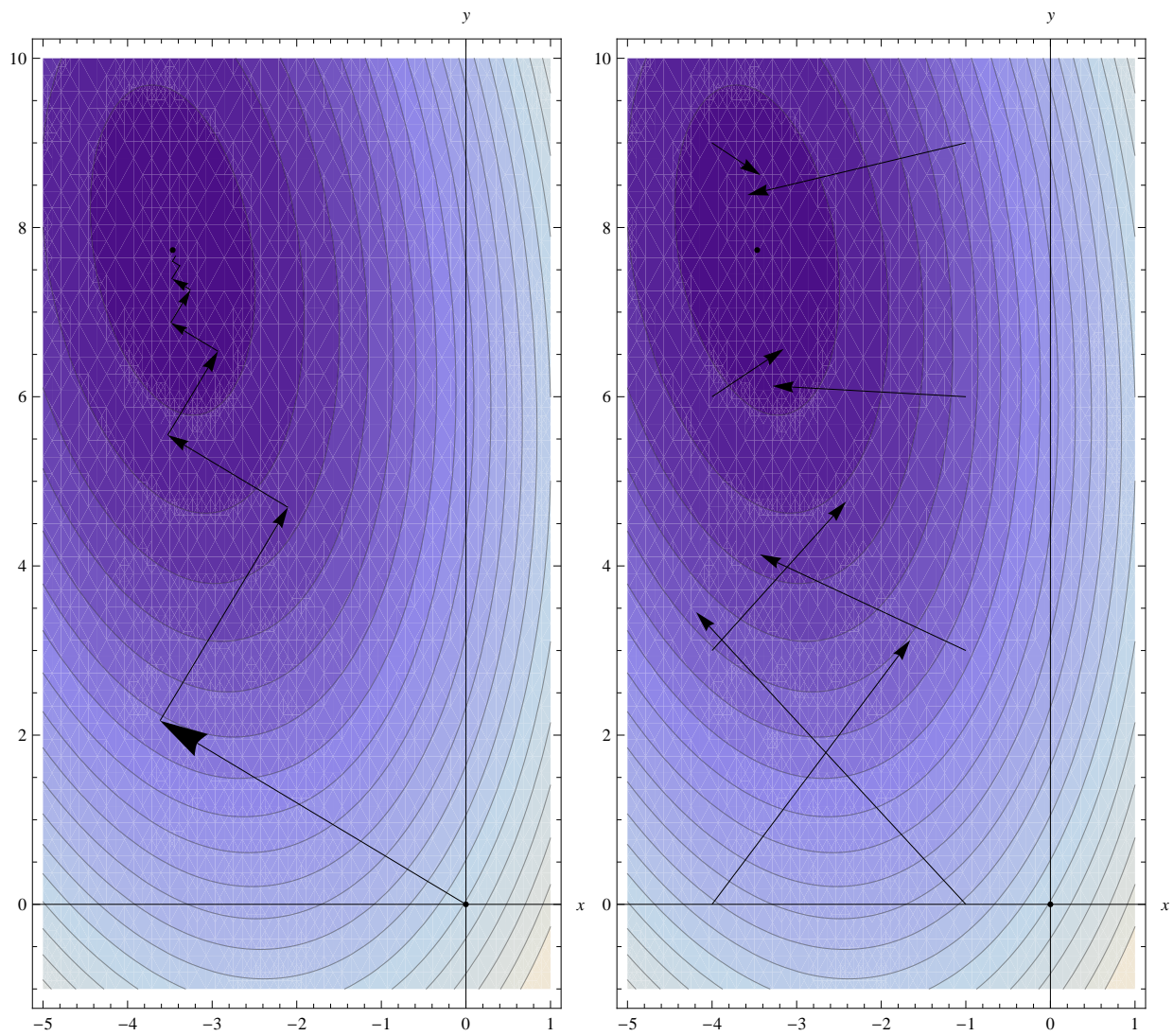


Figure 9: A contour plot of  $M(x, y)$  with a) ten gradient steps starting at  $(0, 0)$  and b) several gradient steps from different points.



OmniTRANS. The two aspects of OmniTRANS that deviate most from the framework are the use of multiple dimension and the use of other restrictions then traffic counts.

The generalization of traffic counts to restrictions makes it hard to define the problem based on traffic loads (as it is in the literature study). In the model with only traffic counts the load on a particular road section can be compared with the count value to measure the error. The loads are the direct result of a traffic assignment. The idea of traffic loads for road sections is extended to traffic loads for restrictions. This value is called the *restriction load*, the definition with a formula is stated below. The error over all restrictions can be calculated similarly as for traffic counts, since we have the restriction load and the restriction value.

The use of multiple dimensions alternate the framework in two ways. First the number demand variables is multiplied with the number of dimensions, since for each O-D pair and each dimension the number of trips is specified. Secondly the restrictions can be over several dimension, this issue is also resolved with the generalization to restrictions and the restriction load. If a traffic count for example registers the traffic of multiple dimension, the restriction load will be the sum over the traffic loads for each dimension.

The setting and notations as described in section 4.1 is used in the remainder of this chapter. For more convenience with the notations define  $\mathbf{g} = \{g_i^d | i \in I, d \in D\}$ ,  $\hat{\mathbf{g}} = \{\hat{g}_i^d | i \in I, d \in D\}$ ,  $\mathbf{C} = \{C_r | r \in R\}$  and  $\mathbf{P} = \{P_{ir}^d | i \in I, r \in R, d \in D_r\}$ . Introduce a set of restriction load variables  $\mathbf{l} = \{l_r | r \in R\}$ , that is the traffic load over restriction  $r$  (i.e. the restriction load). The variables  $\mathbf{g}, \hat{\mathbf{g}}, \mathbf{l}$  and  $\mathbf{C}$  are considered as vectors, with the usual vector operators. All variables needed to describe the problem for OmniTRANS are now available, the adjustment of the bilevel problem (**BLP**) becomes:<sup>24</sup>

$$\begin{aligned} \min_{\mathbf{g}, \mathbf{l}} F(\mathbf{g}, \mathbf{l}) &= \alpha F_1(\mathbf{g}, \hat{\mathbf{g}}) + (1 - \alpha) F_2(\mathbf{l}, \mathbf{C}) \\ \text{s.t. } \mathbf{l} &= \text{Assign}(\mathbf{g}), \\ \mathbf{g} &\geq 0, \\ \mathbf{l} &\geq 0 \end{aligned} \tag{OP}$$

$F_1, F_2$  and  $\alpha$  are inherited from (**BLP**) and are discussed below. The problem above is the major OmniTRANS matrix estimation problem<sup>25</sup>, the main goal of this study is to find a solution technique to conquer the OmniTRANS problem (**OP**).

Parameter  $\alpha \in [0, 1]$  can steer the algorithm towards either the a priori matrix or the restriction values. Smaller  $\alpha$  values put more weight on the restriction values, while larger values put more weight on the a priori matrix. It is not possible to give a meaningful default value for  $\alpha$ , since for different projects the two objective functions take different values that are not related. Properties that should be considered for the choice of  $\alpha$  are the number of demand variables, the number of restrictions, the type of restrictions, the confidence in the a priori matrix and the confidence in the restriction values. It should be calibrated through trial and error.

The functions  $F_1$  and  $F_2$  are still unspecified, to create a concrete problem they should be chosen. Table 1 on page 15 summarizes the objective functions found for the framework in the literature study. The choice for the gradient descent method became the  $L_2$ -norm. One of the arguments for this choice is that it does not rely on statistical assumptions. Additionally the  $L_2$ -norm was favoured above the  $L_1$ -norm and  $L_\infty$ -norm. For algebraic reasons this norm is squared, this will not lead to a fundamental change since it is monotonic (on  $\mathbb{R}^+$ <sup>26</sup>) and  $\alpha$  is free. The objective function is now built up with the

<sup>24</sup>Assign( $\mathbf{g}$ ) is an OmniTRANS assignment that stores the restriction loads in this case.

<sup>25</sup>Note that there is still a huge amount of freedom in this problem formulation. The weighting factor  $\alpha$ , the objective functions  $F_1$  and  $F_2$  and the assignment method should be chosen.

<sup>26</sup> $\mathbb{R}^+ = \{x | x \geq 0, x \in \mathbb{R}\}$



following parts:

$$F_1(\mathbf{g}, \hat{\mathbf{g}}) = \frac{1}{2} \|\mathbf{g} - \hat{\mathbf{g}}\|_2^2 = \frac{1}{2} \sum_{\substack{i \in I, \\ d \in D}} (g_i^d - \hat{g}_i^d)^2 \quad (27)$$

$$F_2(\mathbf{l}, \mathbf{C}) = \frac{1}{2} \|\mathbf{l} - \mathbf{C}\|_2^2 = \frac{1}{2} \sum_{r \in R} (l_r - C_r)^2 \quad (28)$$

Another adjustment on the objective function is the implementation of the weight variable  $\varepsilon_r$  for each restriction. This effects the second part of the objective function. To incorporate the relative differences of the weights they are included as a multiplicative factor. This leads to

$$F_2(\mathbf{l}, \mathbf{C}) = \frac{1}{2} \sum_{r \in R} \varepsilon_r (l_r - C_r)^2 \quad (29)$$

(OP) is a bilevel problem since the assignment is another optimization problem. As mentioned earlier there is no general solution to bilevel problems and it is hard to find a satisfactory solution. It is not possible to apply the gradient method directly. Therefore some simplifications are needed, the major assumption made is that the assignment is locally proportional. This gives rise to a subproblem that searches an O-D matrix while keeping the proportions fixed. The adjustment effectively means that for changes in the O-D matrices no new assignment is needed to calculate the restriction loads (because the assignment is proportional) and thus the lower level problem is eliminated. This new problem is convex and can be conquered with the gradient descend method.

The literature study discussed the situation without congestion (see page 10). In that case the path proportions  $p$  were independent of the O-D matrix and the flows can be expressed in terms of the proportions and demand. The  $\mathbf{P}$  values for traffic counts in the current applied variable space determine the same relation, but now between the restriction load and the demand.  $\mathbf{P}$  actually is a generalization of the path proportions to restriction level. So the fixation of  $\mathbf{P}$  makes it possible to express the restriction loads  $\mathbf{l}$  in terms of  $\mathbf{g}$ . This eliminates the  $\mathbf{l}$  variables from problem (OP), they are rewritten as

$$l_r = \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d, \quad \forall r \in R. \quad (30)$$

Now suppose that  $\mathbf{P}$  is the result of an assignment of some O-D matrix  $\tilde{\mathbf{g}}$ . This implies that the local constancy is assumed around the point  $\tilde{\mathbf{g}}$ . The result of choosing the objective function and fixing the proportions is the convex OmniTRANS subproblem:

$$\begin{aligned} \min_{\mathbf{g}} F(\mathbf{g}) &= \alpha \sum_{\substack{i \in I, \\ d \in D}} \frac{1}{2} (g_i^d - \hat{g}_i^d)^2 + (1 - \alpha) \sum_{r \in R} \frac{\varepsilon_r}{2} \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right)^2 \\ \text{s.t. } &\mathbf{g} \geq 0 \end{aligned} \quad (\text{COSP})$$

Some  $\tilde{\mathbf{g}}$  that is the basis for the subproblem should be specified, this is the point around which a local search is made. So the best known  $\tilde{\mathbf{g}}$  available should be chosen. The  $\mathbf{P}$  values in this problem either come from an assignment of  $\tilde{\mathbf{g}}$  (that is for traffic counts and screenlines) or they directly result from the definition of the restriction (that is for blocks and trip ends).

#### 5.4 The gradient method for the convex OmniTRANS subproblem (COSP)

The final step to the required algorithm is the integration of the gradient descent method and the convex OmniTRANS subproblem. In this section the final hurdle is taken and the algorithm used for

the tests that are presented in the next chapter is presented. The hurdle is the context and non-negativity of the O-D demand. The general gradient method presented in section 5.2 is not suitable for problems with constraints.

---

**Algorithm 4** A direct integration of (**COSP**) and the gradient method

---

```

1: stopCrit ← false
2:  $\mathbf{g} \leftarrow \tilde{\mathbf{g}}$ 
3: while not stopCrit do
4:    $\mathbf{s} \leftarrow -\nabla F(\mathbf{g})$ 
5:    $\lambda^* \leftarrow \frac{\partial}{\partial \lambda} F(\mathbf{g} + \lambda \mathbf{s}) = 0$ 
6:    $\mathbf{g} \leftarrow \mathbf{g} + \lambda^* \mathbf{s}$ 
7:   update stopCrit
8: end while

```

---

Directly integrating (**COSP**) and algorithm 3 without considering the constraints will lead to algorithm 4, where  $\tilde{\mathbf{g}}$  is the starting point. It is desirable that if there is no demand for an O-D pair in the start situation then there is still no demand after matrix estimation. Therefore the search direction on line 4 is adapted. The direction becomes  $-\mathbf{g} \cdot \nabla F(\mathbf{g})$ <sup>27</sup>, the entry-wise multiplication with  $\mathbf{g}$  ensures that the demands with value zero will remain zero. To ensure the non-negativity of the O-D demand a truncation step is introduced, this will set all negative values to zero after the stopping criterion is fulfilled. This is a rather harsh approach and neither a mathematically responsible one nor a sound solution. Although fortunately it has one saving grace since it never occurred in any performed test. Apparently the nature of the problem, where the values should lie in the neighbourhood of the (positive) a priori demands and try to satisfy the (also positive) restriction values, does not result in negative demand.

Let  $i \in I$  and  $r \in R$  then  $P_{ir}^d$  is not defined if  $d \notin D_r$ . To prevent notation errors, redefine

$$P_{ri}^d := \begin{cases} P_{ri}^d & \text{if } P_{ri}^d \text{ is defined} \\ 0 & \text{otherwise} \end{cases}, \quad \forall r \in R, i \in I, d \in D.$$

The same symbol is used since this does not affect the meaning of  $P_{ri}^d$ , but  $\mathbf{P}$  becomes  $\{P_{ri}^d \mid i \in I, r \in R, d \in D\}$ <sup>28</sup>.

The final method is presented in algorithm 5. To execute the actions in the algorithm a more elaborated description is needed, they are presented in this paragraph. The gradient  $\nabla F(\mathbf{g})$  consists of partial derivatives,

$$\frac{\partial F(\mathbf{g})}{\partial g_i^{\bar{d}}} = \alpha \left( g_i^{\bar{d}} - \hat{g}_i^{\bar{d}} \right) + (1 - \alpha) \sum_{r \in R} \varepsilon_r \left( \sum_{\substack{i \in I \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right) P_{r\bar{i}}^{\bar{d}}, \quad \forall \bar{i} \in I, \bar{d} \in D, \text{ } ^{29} \quad (31)$$

which lead to the search direction (for line 4 of the algorithm):

$$\mathbf{s} = -\mathbf{g} \cdot \nabla F(\mathbf{g}) \quad (32)$$

$$s_{\bar{i}}^{\bar{d}} = -\alpha \left( g_i^{\bar{d}} - \hat{g}_i^{\bar{d}} \right) g_i^{\bar{d}} - (1 - \alpha) \sum_{r \in R} \varepsilon_r \left( \sum_{\substack{i \in I \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right) P_{r\bar{i}}^{\bar{d}} g_i^{\bar{d}}, \quad \forall \bar{i} \in I, \bar{d} \in D \quad (33)$$

---

<sup>27</sup>Let  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{m \times n}$ , now  $\cdot$  is the entry-wise multiplication  $(\mathbf{a} \cdot \mathbf{b})_{ij} = \mathbf{a}_{ij} \mathbf{b}_{ij}$ ,  $\forall i \in \{1, \dots, n\}, j \in \{1, \dots, m\}$

<sup>28</sup>The only difference is the erasure of the superscripted  $r$  under  $D$

<sup>29</sup>Note that with the redefinition of  $\mathbf{P}$  the latter  $P_{ri}^d$  in this equation became well defined.

The optimal step length can then be derived with some algebra

$$\frac{\partial}{\partial \lambda} F(\mathbf{g} + \lambda \mathbf{s}) = 0, \quad (34)$$

$$\frac{\partial}{\partial \lambda} \left( \frac{\alpha}{2} \sum_{\substack{i \in I, \\ d \in D}} [g_i^d + \lambda s_i^d - \hat{g}_i^d]^2 + \frac{1-\alpha}{2} \sum_{r \in R} \varepsilon_r \left[ \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d (g_i^d + \lambda s_i^d) \right) - C_r \right]^2 \right) = 0, \quad (35)$$

$$\alpha \sum_{\substack{i \in I, \\ d \in D}} [g_i^d + \lambda s_i^d - \hat{g}_i^d] s_i^d + (1-\alpha) \sum_{r \in R} \varepsilon_r \left[ \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d + \lambda P_{ri}^d s_i^d \right) - C_r \right] \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d s_i^d = 0, \quad (36)$$

$$\alpha \left( \sum_{\substack{i \in I, \\ d \in D}} g_i^d s_i^d + \lambda (s_i^d)^2 - \hat{g}_i^d s_i^d \right) + (1-\alpha) \left( \sum_{r \in R} \left[ \varepsilon_r \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d s_i^d \right] \left[ \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d + \lambda P_{ri}^d s_i^d \right) - C_r \right] \right) = 0, \quad (37)$$

$$\frac{\alpha \left( \sum_{\substack{i \in I, \\ d \in D}} g_i^d s_i^d - \hat{g}_i^d s_i^d \right) + (1-\alpha) \left( \sum_{r \in R} \left[ \varepsilon_r \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d s_i^d \right] \left[ \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d \right) - C_r \right] \right)}{\alpha \left( \sum_{\substack{i \in I, \\ d \in D}} (s_i^d)^2 \right) + (1-\alpha) \left( \sum_{r \in R} \varepsilon_r \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d s_i^d \right)^2 \right)} = \lambda. \quad (38)$$

Equation (38) delivers the optimal step length needed in line 5. The truncation to ensure the non-negativity of the demand is done on line 9 where

$$\mathbf{g}^+ = \left\{ \max \left\{ 0, g_i^d \right\} \mid i \in I, d \in D \right\}. \quad (39)$$

If the situation occurs that there was a negative demand for some O-D pair, an error or warning should be returned. This is not a proper way to deal with this unpredictable behaviour, but as mentioned earlier, it did not occur in the test cases. The user can then decide if the returned demands are satisfactory and defensible.

---

**Algorithm 5** The OmniTRANS gradient method algorithm

---

- 1: *stopCrit*  $\leftarrow$  **false**
  - 2:  $\mathbf{g} \leftarrow \tilde{\mathbf{g}}$
  - 3: **while** not *stopCrit* **do**
  - 4:      $\mathbf{s} \leftarrow -\mathbf{g} \cdot \nabla F(\mathbf{g})$  (see (33))
  - 5:      $\lambda^* \leftarrow \frac{\partial}{\partial \lambda} F(\mathbf{g} + \lambda \mathbf{s}) = 0$  (see (38))
  - 6:      $\mathbf{g} \leftarrow \mathbf{g} + \lambda^* \mathbf{s}$
  - 7:     update *stopCrit*
  - 8: **end while**
  - 9:  $\mathbf{g} \leftarrow \mathbf{g}^+$  (see (39))
- 

## 5.5 Summary

With the final algorithm 5 a gradient technique for solving the OmniTRANS subproblem (**COSP**) is delivered. To use it a choice for  $\tilde{\mathbf{g}}$  has to be made. This should be the best available O-D matrix. With a blank problem, this will be the a priori matrix  $\hat{\mathbf{g}}$ . After the gradient method is terminated a new, better O-D matrix is generated. An assignment with the new O-D matrix leads to new proportions

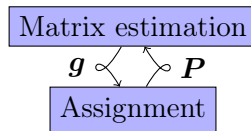


Figure 10: The outer shell iterative process

and a new subproblem, which can be tackled with the gradient method again. This outer shell iterative process is displayed in figure 10.

To deliver this algorithm several assumptions are made and several steps are taken. The  $L_2$ -norm was chosen as the objective function in **(OP)**. The proportions are locally constant to make the usage of the gradient method possible. With this assumption the bilevel origin of the matrix estimation problem is broken and with that the guarantee that an optimal solution of **(OP)** can be found is lost. This results in the convex subproblem **(COSP)** on which the gradient method is applied. The search direction in the gradient method is adjusted to ensure that zero values retain zero. Finally there is a hard check on negative demand as a last step of the method, fortunately this harsh tool was never necessary in practice.

## 6 Test procedure

Chapters 4 and 5 provide three different solution methods for the matrix estimation part of the bilevel problem. Those are the currently implemented method in **OmniTRANS**, the provided revision of that algorithm and the gradient method. This chapter describes an environment and an approach to compare these solution algorithms. It is already clear that matrix estimation is far from standardized and so are the methods to judge the output. By defining several tests and benchmarks, this chapter aims to give a clear overview of how the algorithms are compared. In chapter 7 the results of the tests performed are presented.

### 6.1 The environment

First of all the three methods are briefly recapitulated and named to make it easier to refer to them.

- **ORIGINAL:** This is the method currently implemented in the **OmniTRANS** software. It is dependent on the order of the input and per chance outdated. Each restriction is treated successively, directly modifying the O-D matrix by multiplication of the corresponding O-D pairs with a factor. (see section 4.2)
- **REVISED:** This is the modified ORIGINAL method. The caveat of being dependent of the input order is overcome. Instead of the successive treatment an average over the factors of each restriction is taken and the O-D matrix is therefore only modified once. (see section 4.3)
- **GRADIENT:** This method is built from scratch and has an entirely different approach compared to ORIGINAL and REVISED. The foundation is a gradient descent method which is adapted for the matrix estimation problem. It is made compatible for **OmniTRANS** by modifying the problem formulation. An advantage is that the a priori matrix is considered actively in the process. (see section 5.4)

The methods ORIGINAL, REVISED and GRADIENT are all designed for the mono disciplinary matrix estimation task. The input and the output are O-D matrices and the information of restrictions and a priori matrices is used to improve the O-D matrix. This means only the matrix estimation part of the bilevel problem is solved. These methods are alternated with traffic assignments to approach the solution of the bilevel problem.

The three methods are implemented with the Matlab software package. ORIGINAL, derived from the current implementation, is re-implemented in Matlab. The **OmniTRANS** implementation of this method is not used in this study. There is no comparison made between the Matlab and **OmniTRANS** implementations, the grounds are that the result already differs with different input orders and it would be technically cumbersome. For the traffic assignments the **OmniTRANS** implementation is used, it is desirable to use this original method and it is not even possible to achieve this in Matlab since all network information is needed.

The tests presented here frequently use the assignment and matrix estimation methods in loops. Since that implies that **OmniTRANS** and Matlab routines are used successively within a test an environment is built. An overview of the environment is shown in figure 6.1. The main program is Matlab, apart from the methods also the test routines are executed in Matlab. It is possible to run **OmniTRANS** jobs in Matlab from the command line with **OmniTRANS** Real-Time (RT), these jobs are used to read/write information from/to text files and perform assignments. For an assignment a considerable amount of information from the **OmniTRANS** database is needed, especially the network information is of great importance. The information which is exchanged for the interaction between the assignment and matrix estimation are the O-D matrices and proportion values  $\mathbf{P}$ . Furthermore Matlab needs the information of the restrictions and a priori matrices, as well as some information on the dimensions of the problem. The communication between Matlab and **OmniTRANS** is performed

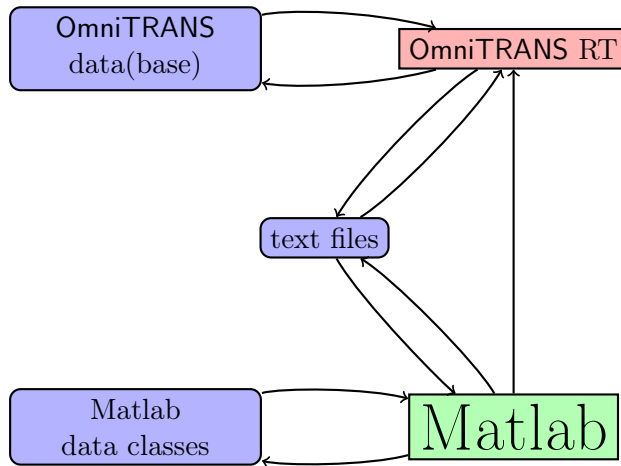


Figure 11: Test environment

using ordinary text files. To be able to write maintainable code in Matlab some of its object oriented programming functionalities are utilized. This turned out to be especially convenient to store and retrieve the information.

As described in section 3.2.4 the modelling classes of OmniTRANS are directed with jobs. For the tests several jobs are developed to assign matrices that are stored in the text files. These jobs also handle the data export and are used to restore original O-D matrices. The most important  $\mathbf{P}$  values, those of screenlines and counts, are stored in the OmniTRANS database during the assignment and are exported later.

The bottleneck in the test environment is the storage of the  $\mathbf{P}$  values in text files. Firstly it is slow to write data to the hard disk and secondly the standard matrix export to text files of OmniTRANS is exceptionally inefficient. Consider the traffic counts, OmniTRANS generates a screenlinematrix<sup>30</sup> for each count and each dimension. In this matrix the proportion value for each O-D pair is stored, but this value is only unequal to zero if there is a path passing the count location for that particular dimension. It is not hard to see that the majority of roads are not used in any path for an O-D pair. So the majority of the stored values are zeroes. On top of that each value is stored with a precision of six decimals. This bottleneck restricted this study to 730 centroids, that is still more than half a million O-D pairs. The advantage of this restriction is that it was not necessary to save on iterations inside the method.

## 6.2 The transport models

Several OmniTRANS projects<sup>31</sup> are considered during the study. Some important characteristics of the projects are the number of centroids in the network and the number of dimensions. The goal for Omnitrans is to have a matrix estimation method that can be applied to all plausible projects. One of the largest networks in OmniTRANS is the Dutch national model, it consists of about 4000 centroids. Not every reviewed model is used for the tests. The Delft model for example had a lack of count information and the model of the Maastricht area dropped out due to the bottleneck described above, this network has 2000 centroids and stretches deep into the Benelux area and Germany. Eventually it was not possible to perform tests with the desired model size of 4000 centroids, this is caused by the environment with Matlab and OmniTRANS. Three models turned out to be useful for performing tests.

- **Simple model** (figure 12): This theoretical network consists of only six centroids and is used to check the theoretical properties of the methods. It consists of an express way that is surrounded

<sup>30</sup>Given a restriction  $r$  and a dimension  $d$ , the screenlinematrix is  $\{P_{ri}^d \mid i \in I\}$

<sup>31</sup>Frequently the word *model* is used to refer to a project.

with local roads, the amount of trips on this network is sufficient to create traffic congestion on the express way. This leads to multiple routes within traffic equilibrium for some O-D pairs. Two different scenarios are examined for this model, one with a single dimension (car) and one with two dimensions (car and freight). The distinction in the two scenarios gave the opportunity to test the extension to multiple dimensions.

Consider the scenario with one dimension. The number of trips between the centroids that is representing the real case (i.e. the actual amount) is:

	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>
<i>1</i>	0	1000	1000	750	500	1000
<i>2</i>	1000	0	4000	500	750	500
<i>3</i>	1000	4000	0	500	1000	500
<i>4</i>	750	500	500	0	750	1000
<i>5</i>	500	750	1000	750	0	1000
<i>6</i>	1000	500	500	1000	1000	0

The assignment of these trips lead to traffic on each link of the network. These loads are now taken as the count values, they can be seen as the observed traffic without any error. The a priori matrix is generated by randomizing the value for each O-D pair, they are multiplied with a uniform random number from the interval  $[\frac{4}{5}, \frac{6}{5}]$ . The obtained a priori matrix and count values act as the input for the matrix estimation methods which ideally deliver a matrix that is similar with the matrix above.

- **Alphen aan den Rijn model** (figure 13): Alphen aan den Rijn is a Dutch city in the province Zuid-Holland with more than 70.000 residents. The study area is roughly determined by the triangle Schiphol, The Hague and Gouda. The network consists of 730 centroids, almost 7.000 links and there are six dimensions. The dimensions are all combinations of {car, feight} with {morning peak, evening peak, off peak}, the off peak is all traffic that does not belong to the morning peak and evening peak rush hours.

There are 208 locations with traffic counts, on most of these locations there are multiple counted values, for several dimensions and for several combinations of dimensions.<sup>32</sup> Counts do exist for the following combinations of dimensions:

1. {car-morning peak}
2. {car-evening peak}
3. {car-off peak}
4. {freight-morning peak}
5. {freight-evening peak}
6. {freight-off peak}
7. {car-morning peak, car-evening peak, car-off peak} (Car 24 hours)
8. {freight-morning peak, freight-evening peak, freight-off peak} (Freight 24 hours)
9. {car-morning peak, freight-morning peak} (Total morning peak)
10. {car-evening peak, freight-evening peak} (Total evening peak)
11. {car-off peak, freight-off peak} (Total off peak)
12. {car-morning peak, car-evening peak, car-off peak, freight-morning peak, freight-evening peak, freight-off peak} (Vehicles 24 hours)

There are 1414 count values in total, that is an average of 6.8 counts per location.

This model will be referred to as Alphen.

---

<sup>32</sup>In terms of restrictions, there is more than one restriction for each location, all with a different dimension sets  $D_r$ .

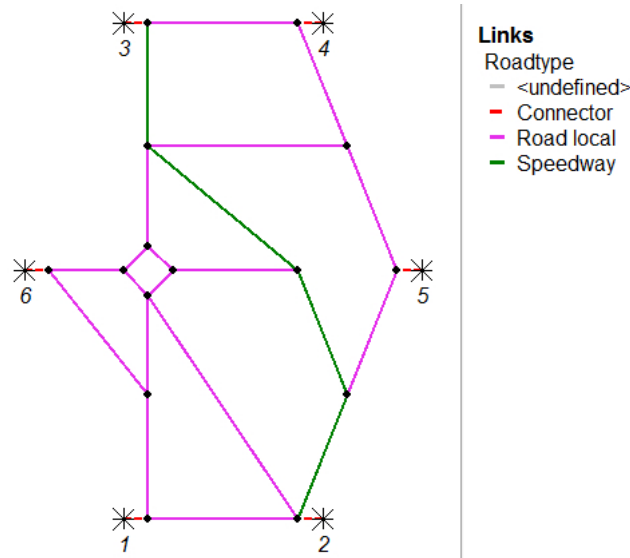


Figure 12: The simple network

- **Amsterdam model** (see figure 14): This is the Amsterdam project used for the Dynamic Tactical Traffic Model Amsterdam<sup>33</sup>. This is the model Omnitrans engineers worked on and successfully used the GRADIENT method to estimate the matrix. The study area covers the inside of ring road A10, this is a cordon excision of a larger network. There are 279 centroids, 201 counts locations and 3 screenlines in the project. There is only one dimension, so each location has one counted value resulting in a total of 204 restrictions.

The name of the model already reveals that the final result of the project is not static. The input for the dynamic assignment that brings these final results is an O-D matrix that represents the total traffic over the whole period, 15h00 - 19h00. For this study the goal is to deliver that O-D matrix for the whole period. The traffic counts for some locations were taken over the period 15h00 - 18h00. This issue is solved by extrapolating these counts using the ratio, between periods 15h00-18h00 and 18h00-19h00, of the counts that had this information available. Another point of attention is that the matrix estimation results have to be evaluated with a static instead of dynamic assignment<sup>34</sup>

## 6.3 Quality measures

Different matrix estimation methods give different results, different parameters for a method give different results as well. There is no unequivocal answer to the question how to compare these results. In practice several methods are used, the choice generally coheres with the location of the study. Each country developed its own standard that complies with the wishes of the client.<sup>35</sup> The resulting matrices of matrix estimation can be judged on two aspects, the approximation to the a priori matrix and the approximation to the counts. Model analysts should consider both factors, although there is no prescription on how to handle this dichotomy.

### 6.3.1 Objective function

The framework in the literature study delivers the most obvious measure on the results, namely the objective function. The way of modelling in optimization problems is focused on the delivery of the

<sup>33</sup>The goal of this project is to provide a tool which calculates the effects of tactical traffic measurements (e.g. traffic-light configurations) on traffic flows in the centre area of Amsterdam.

<sup>34</sup>A dynamic assignment would give time varying loads which cannot be compared with the extrapolated count values.

<sup>35</sup>The client is often a government that wants to stick to a method when it is used to it.





Figure 13: The Alphen aan den Rijn network



Figure 14: The Amsterdam network

optimal result in terms of objective functions. The problem with matrix estimation and this approach is that there is no general objective function, as shown in table 1, that is a potential measure for the results. The methods ORIGINAL and REVISED do not originate directly from an optimization problem. Method GRADIENT does originate from an optimization problem, but does not guarantee the optimal result. It is derived from problem (COSP), therefore the objective function that is used as a measure is its objective function

$$\text{OBJ} = \alpha \sum_{\substack{i \in I, \\ d \in D}} \frac{1}{2} (g_i^d - \hat{g}_i^d)^2 + (1 - \alpha) \sum_{r \in R} \frac{\varepsilon_r}{2} \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right)^2. \quad (40)$$

This measured is named OBJ, when only the left or right term is mentioned it is named respectively OBJ-MAT and OBJ-RES.

### 6.3.2 Coefficient of determination, $R^2$

The coefficient of determination is a measure derived from statistics. It is a crude measure of the strength of a relationship that has been fit by least squares. For the matrix it is defined as

$$\text{R}^2\text{-MAT} = 1 - \frac{\sum_{\substack{i \in I, \\ d \in D}} (g_i^d - \hat{g}_i^d)^2}{\sum_{\substack{i \in I, \\ d \in D}} (\hat{g}_i^d - \bar{g})^2}, \quad (41)$$

where  $\bar{g} = \frac{\sum_{\substack{i \in I, \\ d \in D}} \hat{g}_i^d}{|I||D|}$ , the mean of all elements of the a priori matrix. For the restrictions it is defined as

$$\text{R}^2\text{-RES} = 1 - \frac{\sum_{r \in R} \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right)^2}{\sum_{r \in R} (C_r - \bar{C})^2}, \quad (42)$$

where  $\bar{C} = \frac{\sum_{r \in R} C_r}{|R|}$ , the mean of the count values. These coefficients of determination can be interpreted as the proportion of the variability of the results that can be explained by the a priori information. The better  $\text{R}^2\text{-MAT}$  and  $\text{R}^2\text{-RES}$  approaches 1, the better the results are. An advantage of the coefficient of determination is that the interpretation of the result is straightforward and not dependent on the size of the model, in contradiction to the objective function. This also delivers insights in the progress of the methods over the iterations.

### 6.3.3 T-values

In the Netherlands the T-values are used frequently as the tool for judging the results. It is the criterion Goudappel Coffeng uses to determine if the matrix estimation was successful. The restriction case is treated to start with. As the plural already suggests there is no single outcome, for each restriction a T-value is calculated with

$$T_r = \log \frac{\left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right)^2}{C_r}, \quad \forall r \in R. \quad (43)$$

This gives an evaluation of each restriction, in the following tabular a general guideline is given:

T-value	
< 0.5	Excellent
0.5 - 2.5	Very good
2.5 - 3.5	Good
3.5 - 4.5	Acceptable
> 4.5	Poor

The norm is that at least 80% of the T-values should be good or better and at least 95% of the T-values should be acceptable or better. The acquired percentages corresponding to these norms are respectively named T3.5-RES and T4.5-RES, the percentage of values  $\leq 5.5$  is named T5.5-RES. This guideline holds for counts measured over a rush hour on express ways. Twenty-four hour counts for example can be treated by applying a factor  $\frac{1}{8}$  on the final matrix and count value.

The T-values are also used as a measure for the O-D matrix, but not in an element-wise way. The T-values are applied to the trip ends<sup>36</sup> of the O-D matrix. Let  $Z$  be the set of all zones and  $J = \{\text{O-D pairs with origin } z \mid z \in Z\} \cup \{\text{O-D pairs with destination } z \mid z \in Z\}$ <sup>37</sup> be the set of all rows and columns, then set of T-values is defined as

$$T_j = \log \frac{\left( \sum_{i \in j} g_i^d - \sum_{i \in j} \hat{g}_i^d \right)^2}{\sum_{i \in j} \hat{g}_i^d}, \quad \forall j \in J \text{ with } \sum_{i \in j} \hat{g}_i^d > 0. \quad (44)$$

The T-values are only used for positive trip ends, values of zero are not well defined. The norm is that at least 90% of the T-values of the trip ends should be good or better. The percentage of values that lower than 3.5, 4.5 and 5.5 are respectively named T3.5-TE, T4.5-TE and T5.5-TE.

The advantage of the T-value measures is the influence of both the relative and absolute difference. The down side is that it cannot be used in all projects, especially those with abnormal time intervals and dimensions. When for example a freight count for the AM peak and a car count for the whole day are used in one matrix estimation project, the results of the T-value norm will rely heavily on the latter one. A solution for this can be to convert all values to peak hour values.

#### 6.3.4 GEH statistic

The GEH statistic is developed by Geoffrey E. Havers and is primarily used in the field of traffic and transport. Traffic engineers in the United Kingdom use this as the primary statistic. Unfortunately it not based on a real statistical test, just like the T-value. The test is only used for restrictions, it is not applied to the matrix (or trip ends). For each restriction the GEH-value is defined as

$$GEH_r = \sqrt{\frac{\left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d - C_r \right)^2}{\frac{1}{2} \left( \sum_{\substack{i \in I, \\ d \in D_r}} P_{ri}^d g_i^d + C_r \right)}}. \quad (45)$$

As a rule the GEH-value is good if it is lower than 5, it is acceptable if it is between 5 and 10 and it is unacceptable if it is 10 or higher. Guidelines to how many restrictions should be of a certain quality are not available. There are two measures defined from the GEH statistic, GEH5 and GEH10, which are the percentages of restrictions with values lower than respectively 5 and 10.

<sup>36</sup>The column and row sums of the O-D matrix, or attraction and production per zone

<sup>37</sup>So each element of  $J$  is a set of O-D pairs

### 6.3.5 Summary

---

OBJ	The objective function
OBJ-MAT	The matrix part of the objective function
OBJ-RES	The restriction part of the objective function
R <sup>2</sup> -MAT	The coefficient of determination for the matrix
R <sup>2</sup> -RES	The coefficient of determination for the restrictions
T3.5-RES	The percentage of restrictions with a T-value $\leq 3.5$
T4.5-RES	The percentage of restrictions with a T-value $\leq 4.5$
T5.5-RES	The percentage of restrictions with a T-value $\leq 5.5$
T3.5-TE	The percentage of trip ends with a T-value $\leq 3.5$
T4.5-TE	The percentage of trip ends with a T-value $\leq 4.5$
T5.5-TE	The percentage of trip ends with a T-value $\leq 5.5$
GEH5	The percentage of restrictions with a GEH-value $\leq 5$
GEH10	The percentage of restrictions with a GEH-value $\leq 10$

---

## 7 Results

This chapter describes the results of the performed tests and as well as a discussion on memory usage. The tests incorporate the verification of the REVISED method, a comparison of the three different methods and an analysis of the  $\alpha$ -value of the GRADIENT method.

### 7.1 Tests

#### 7.1.1 Test on the influence of the input order

Several tests are performed to prove the improvements of the method REVISION and how it influences the results. The first test is on the simple network, the purpose is to see what the effect of the input order is between ORIGINAL and REVISED. 500 permutations of the restrictions, leading to 500 input orders, are used for both methods. In each method 20 iterations were performed.<sup>38</sup> The algorithms result in a list of errors: the error of the first treated restriction, the error of second treated restriction, ... , the error on the 146<sup>th</sup> treated restriction. The error is the absolute difference between the restriction value and the traffic load. The  $i^{\text{th}}$  list contains 500 errors of restrictions that was on the  $i^{\text{th}}$  place in the input order. The average error for each list is calculated and represented in figure 15.

It can be seen that method ORIGINAL favours the later treated restriction since the line is declining. Furthermore the figure shows that method REVISED has the effect for which it is designed, namely giving each restriction the same priority.

#### 7.1.2 Test on the performance of the methods

A major test is performed to compare the performances of the methods ORIGINAL, REVISED, GRADIENT and as a combination of the latter two called COMBI. The tests are performed on the Alphen model. To get the best results the runs of matrix estimation and assignments are successively performed.

The assignment used in this test consists of two steps, firstly the freight traffic is assigned to the network and secondly the cars. The freight is assigned with an all-or-nothing assignment, this implies that freight traffic takes the shortest route (based on time). The assignment of the cars is done afterwards, an initial freight load from the all-or-nothing assignment is placed on the network, which reduces link capacities. The used capacity restrained assignment method is volume averaging, and ten iterations are performed. Since this has to be done for three different time periods there is a total of six assignment calls. As mentioned earlier all necessary proportions are saved during the assignments.

ORIGINAL and REVISED can be used without any additional input parameters, the only specification is the number of iterations, this is set to 10 for each matrix estimation method. GRADIENT on the other hand has a rather important parameter  $\alpha$ . ORIGINAL and REVISED do not actively use the a priori matrix within the process. Furthermore OBJ becomes incomparable if we use an  $\alpha$  different than zero. So for this test  $\alpha = 0$  and the GRADIENT method will not use the a priori matrix actively as well.

Because it sounds very interesting to do a combination of methods, COMBI is introduced. It uses both REVISED and GRADIENT, alternating per iteration. The test for each method consists of five outer loops, each test has five matrix estimation calls and six assignment calls. OBJ will be stored during the matrix estimation process, so the convergence can be reviewed later. All other measures are also stored five times, between the matrix estimation and the assignment, the fifth storage is stored after the final assignment. COMBI will use GRADIENT in odd iterations and REVISED in even iterations. Figure 16 is a schematic representation of the tests, where OBJ means that the objective

---

<sup>38</sup>For this test there was no assignment performed, the goal is to review the characteristics of the method and not the performance.

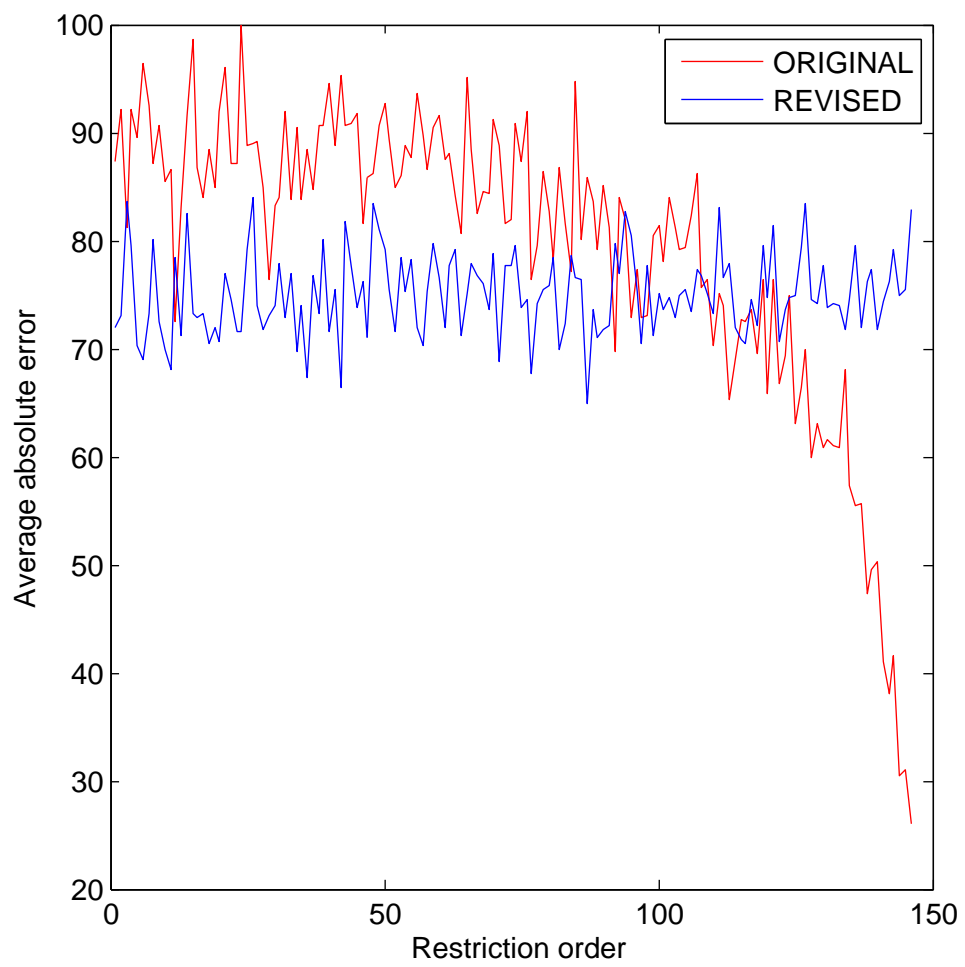


Figure 15: Average error over 500 permutations

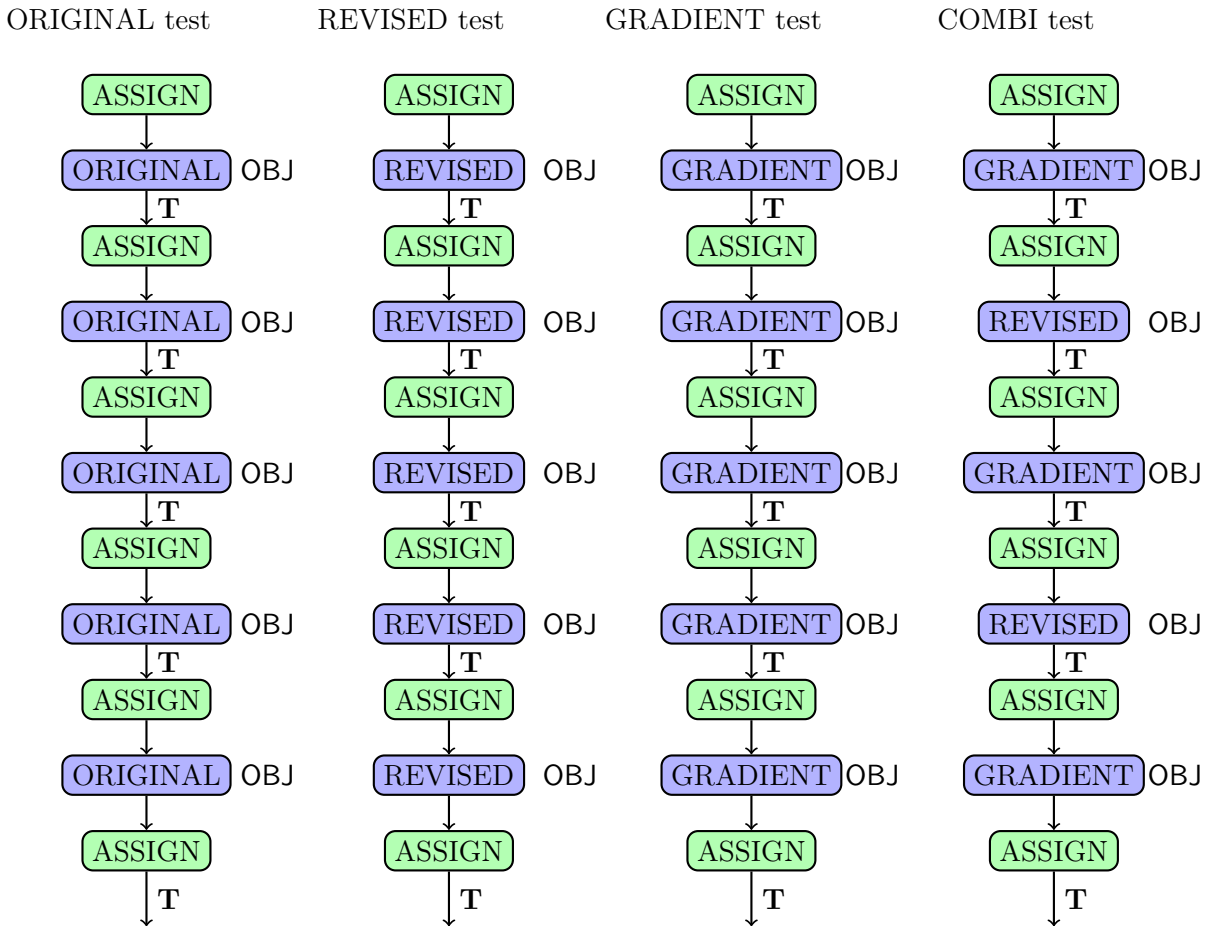


Figure 16: Diagram of the method comparison tests

function is monitored and  $\mathbf{T}$  represents the places where the measures are stored. Note the lowest  $\mathbf{T}$  which is taken irregularly.

The computation time for the different methods did not have significant differences, it took between 2400 and 2460 seconds to do one matrix estimation method with ten iterations. This is just the time that is necessary for the calculations, excluding the time to read and write the results. As mentioned earlier the bottleneck of the test environment, the reading and writing to the disk, loading the variables from the hard disk to Matlab took almost 1000 seconds (each time). It is not possible to register the time that external processes (i.e. the *OmniTRANS* assignment) take. The total time consumed by an outer iteration is around 110 minutes, so it is plausible that an assignment including writing the data took 55 minutes.<sup>39</sup>

The most intuitive result of this test is the development of *OBJ*, the progress can be plotted and shows the 'winner' directly. Figure 17 shows this plot, and at first sight *ORIGINAL* seems to be the outright winner, with *GRADIENT* trailing far behind. Since  $\alpha = 0$  there is no difference between *OBJ* and *OBJ-RES* in this case, so the only judged information (and error) is that of the restrictions. Later in this section it will be clear that the matrices tell another story. The plot shows several discontinuities at the vertical lines. Those are the effect of the assignment which renew the proportion values and that leads to an increment in the objective function. The single rightmost values are the *OBJ-RES* measures after the final assignment (i.e. the final result). *ORIGINAL*, *REVISED* and *COMBI* perform significantly better in the sense of this convergence than *GRADIENT* does. Another result is that *COMBI* shows better convergence for its *REVISED* parts, this is in line with the expectation since *REVISED* has an overall better convergence than *GRADIENT*.

An explanation for this behaviour can be found in the nature of the methods. *ORIGINAL* works

<sup>39</sup>This test was executed on a machine with 4GB memory (far above the limit of Matlab) with an Intel Q9550 processor.



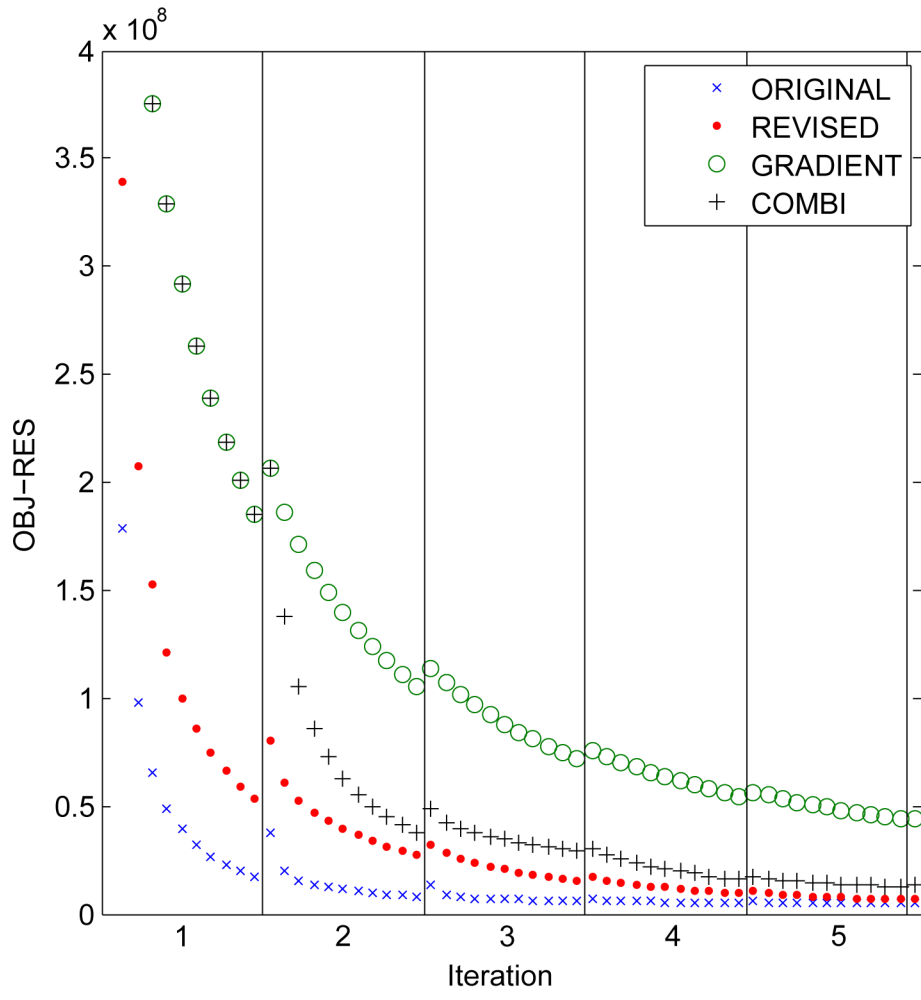


Figure 17: OBJ-RES convergence for four methods on the Alphen network

quite aggressively to achieve good restriction matchings, for each restriction, every involved O-D pair is adjusted directly to match the restriction value. This can only be disturbed by a following restriction with the same O-D pair(s) involved. So it is not surprising that ORIGINAL works so well. REVISED on the other hand is far less aggressive, the O-D pairs are adjusted with a mean factor that involved every restriction. This offers some prevention to outlier restrictions<sup>40</sup>, but also reduces the strength to match the restrictions. And finally the GRADIENT method. This method is restricted to one search direction in each iteration, this means that if one restriction goes to a better matching over that direction the others will be changed as well. By determining the optimal step length the best possible choice is made, but that does not guarantee good restriction matchings. The convergence with respect to OBJ-RES is already the worst with  $\alpha = 0$ , unfortunately it will only get worse by increasing  $\alpha$ .

Table 2 displays the results for all measures other than OBJ for the four tests. Italic numbers do not pass the corresponding norms. Noticeable is that almost all results are good or excellent and at the finish line each test passed each norm. R<sup>2</sup>-RES, T3.5-RES, T4.5-RES, T5.5-RES, GEH5 and GEH10 show an improvement with each iteration for each test, from that it can be concluded that each of the methods result in a better restriction matching. R<sup>2</sup>-MAT, T3.5-TE, T4.5-TE and T5.5-TE show a downswing with each iteration, this implies that the O-D matrices increasingly move away from the a priori matrix. The difference between ORIGINAL and REVISED on the one hand and GRADIENT on the other hand is clear, the latter has a significant higher mark on the matrix and a significant lower mark on the restrictions. COMBI shows results that are in between, generally it is closer to the better method than to the worst. So COMBI is able to combine the qualities of the REVISED method with the qualities of the GRADIENT method.

If we omit the  $\alpha$  in OBJ-MAT, then we can calculate the squared errors of the matrix. Figure 18 shows these OBJ-MAT values for the four methods. GRADIENT has the lowest error which is not surprising since it does not manipulate the matrix aggressively. More remarkable is the bad performance of REVISED for this measure, it was also expected that it would perform better than ORIGINAL since this method changes the O-D pairs for each restriction.

It is possible to compare the a priori matrix  $\hat{g}$  with the a posteriori matrix  $g$  in a visual manner. The two corresponding values for an O-D pair can be plotted on the  $\mathbb{R}^2$  plane, this results in a scatter plot, with ideally each point near the  $x = y$  axis. Figure 19 shows these plots for each method. The results are quite interesting. A first impression shows that the GRADIENT method has the best results, for this method the points are closer to the centre than for the others. Besides that all points fit inside a cone originating from the centre, this shows that there is a bound on the maximum relative error for this method. ORIGINAL and REVISED have a larger spread of the points, COMBI is once more the median. A more remarkable result is that these three methods have a lot of points on the  $y$ -axis. This means that O-D pairs with a very small number of trips ( $< 0.3$  and  $> 0$ ) are exploded, these O-D pairs only have paths over restrictions with a very high factor. The methods do not take this excessiveness of the factors into account and change the value of the O-D pair. A possible explanation is the known problem of cut-through traffic<sup>41</sup>. If large volumes of traffic are registered on a rural road and the assignment only assigns some local O-D pairs to this road a misinterpretation occurs. The methods ORIGINAL and REVISED will increase the traffic for the local O-D pairs only. The GRADIENT method does not have these extreme relative increments.

### 7.1.3 Test on the $\alpha$ values for GRADIENT

Method GRADIENT needs an  $\alpha$  as its input, it determines the relative weight of the two parts of the objective function. It is not possible to suggest an  $\alpha$  that is good for each project, it is not even possible to give the best  $\alpha$  for a specific network. Different  $\alpha$ -values are tested for the Alphen network, as well the combination of GRADIENT and REVISED is tested. Each matrix estimation method is executed with 10 iterations in this test. There are three scenarios, firstly only GRADIENT once,

<sup>40</sup>Restrictions for which the factor  $X$  is very high or small, causing O-D pairs to change unnaturally

<sup>41</sup>'sluipverkeer' in Dutch.

Measure	Method	After iteration 1	After iteration 2	After iteration 3	After iteration 4	After iteration 5
R <sup>2</sup> -RES	ORIGINAL	0.99983	0.99992	0.99994	0.99995	0.99995
	REVISED	0.99948	0.99974	0.99985	0.99991	0.99993
	GRADIENT	0.99822	0.99899	0.99931	0.99948	0.99958
	COMBI	0.99822	0.99964	0.99972	0.99985	0.99987
R <sup>2</sup> -MAT	ORIGINAL	0.99259	0.99183	0.99154	0.99139	0.99131
	REVISED	0.99256	0.99170	0.99122	0.99090	0.99068
	GRADIENT	0.99439	0.99336	0.99286	0.99255	0.99237
	COMBI	0.99439	0.99246	0.99215	0.99176	0.99162
T3.5-RES	ORIGINAL	99.010	99.222	99.505	99.576	99.576
	REVISED	93.777	97.454	99.081	99.505	99.788
	GRADIENT	<i>69.378</i>	<i>75.035</i>	<i>79.066</i>	82.956	85.644
	COMBI	<i>69.378</i>	95.545	96.252	99.151	99.081
T4.5-RES	ORIGINAL	99.576	99.717	99.717	99.717	99.717
	REVISED	98.727	99.788	100.0	100.0	100.0
	GRADIENT	<i>84.512</i>	<i>88.967</i>	<i>92.504</i>	<i>94.130</i>	95.474
	COMBI	<i>84.512</i>	99.646	99.576	99.929	100.0
T5.5-RES	ORIGINAL	100.0	100.0	100.0	100.0	100.0
	REVISED	100.0	100.0	100.0	100.0	100.0
	GRADIENT	93.777	96.888	98.303	98.798	99.010
	COMBI	93.777	100.0	100.0	100.0	100.0
T3.5-TE	ORIGINAL	96.272	95.610	95.394	95.345	95.361
	REVISED	96.935	96.057	95.643	95.162	94.947
	GRADIENT	99.785	99.404	98.973	98.608	98.41
	COMBI	99.785	96.952	96.554	96.140	96.057
T4.5-TE	ORIGINAL	98.542	98.360	98.277	98.211	98.144
	REVISED	98.857	98.542	98.294	98.161	98.028
	GRADIENT	99.934	99.818	99.768	99.685	99.602
	COMBI	99.934	98.890	98.857	98.608	98.559
T5.5-TE	ORIGINAL	99.387	99.321	99.321	99.304	99.304
	REVISED	99.437	99.354	99.271	99.238	99.188
	GRADIENT	99.983	99.983	99.983	99.983	99.934
	COMBI	99.983	99.420	99.420	99.404	99.387
GEH5	ORIGINAL	98.303	98.868	99.151	99.293	99.293
	REVISED	91.584	96.252	98.444	98.798	99.293
	GRADIENT	63.579	69.590	73.621	77.016	80.622
	COMBI	63.579	94.059	93.635	98.091	98.303
GEH10	ORIGINAL	99.576	99.717	99.788	99.788	99.788
	REVISED	98.868	99.859	100.0	100.0	100.0
	GRADIENT	84.866	89.321	92.716	93.989	95.757
	COMBI	84.866	99.717	99.788	100.0	100.0

Table 2: The results for several measures and methods for the Alphen network

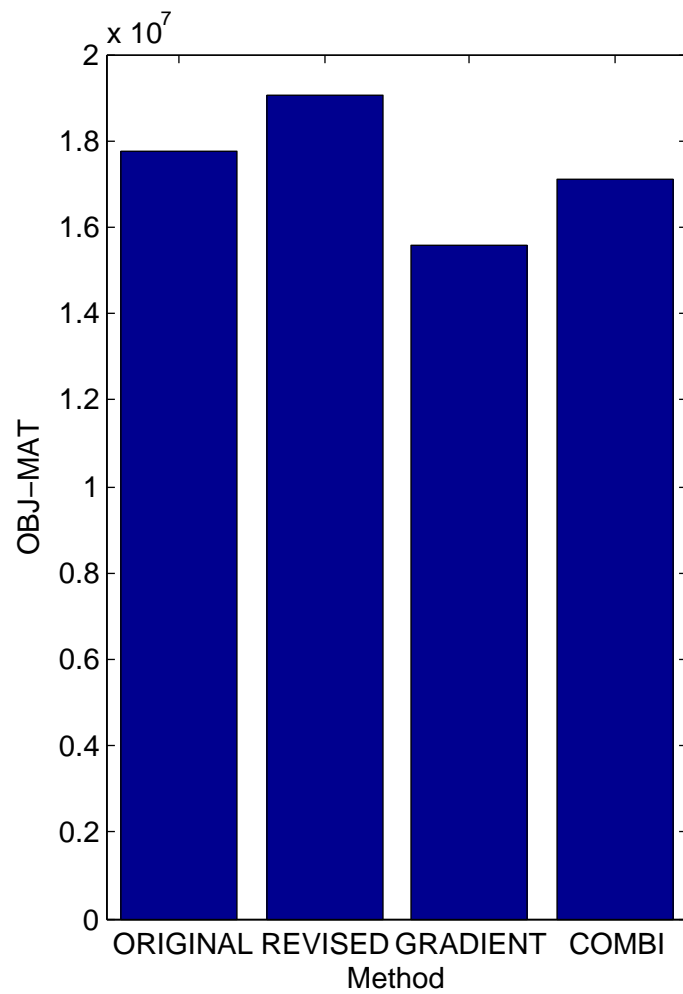


Figure 18: OBJ-MAT values for four methods on the Alphen network



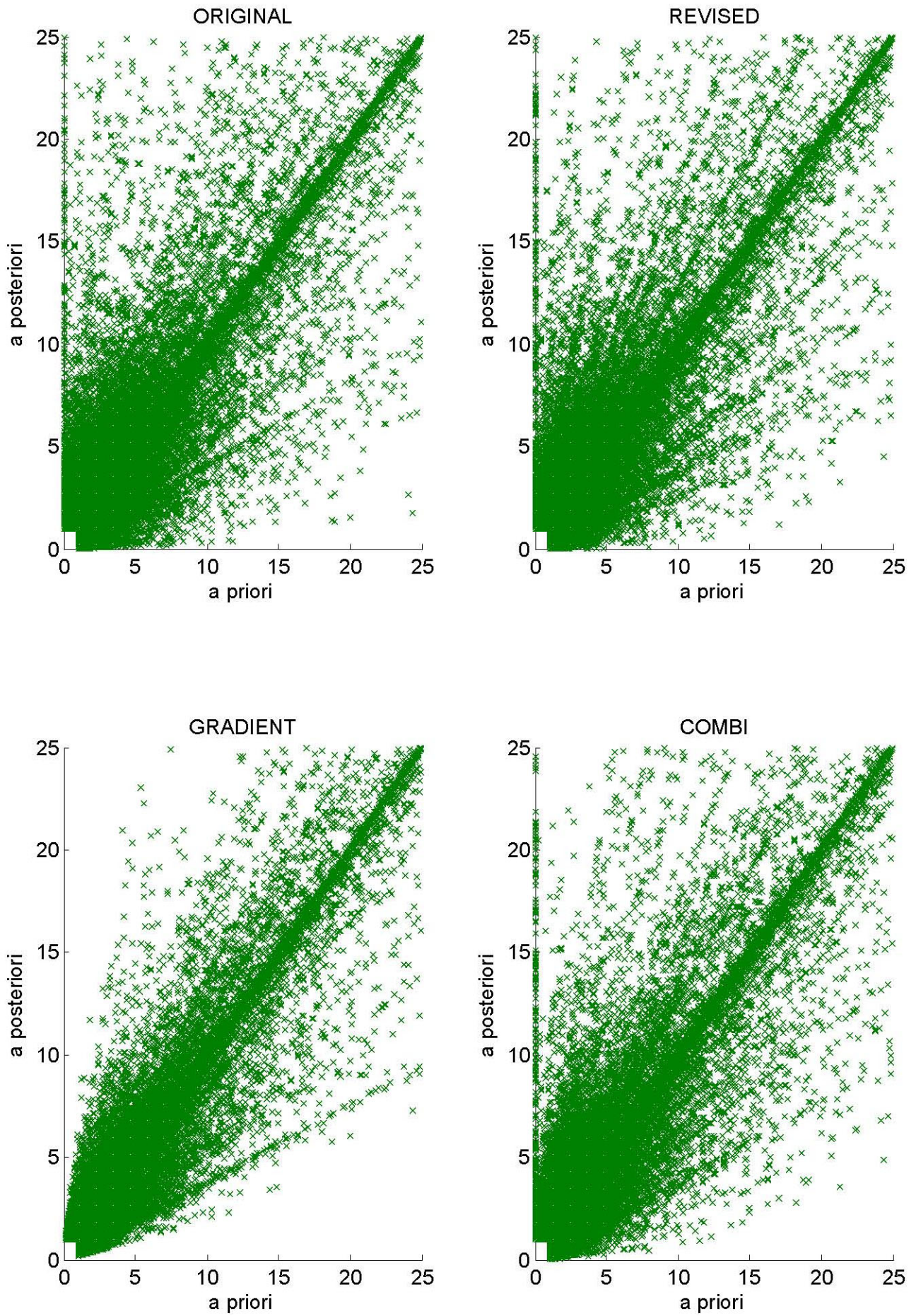


Figure 19: Scatter plots of the a priori and a posteriori O-D pairs for several methods on the Alphen network.

secondly GRADIENT and successively REVISED (COMBI1) and thirdly REVISED and successively GRADIENT (COMBI2). The following  $\alpha$ 's are tested in each scenario: 0.01, 0.1, 0.2, 0.5, 0.9, 0.98, 0.99, 0.998 and 0.999. There is no feedback with assignments in this test.

The results are shown in table 3, italic values did not meet the prevailing norm. An interpretation of the results is in fact straightforward and in line with the expectations. The higher the  $\alpha$ , the better the a priori matrix is preserved and the lower the  $\alpha$  the better the restrictions are matched. Furthermore it turned out that ten iterations of GRADIENT are not enough to meet the T-value norm for restrictions. Alternating with REVISED does have a positive effect on the matching of the restrictions, the corresponding measures improve significantly. The results for the matrix range from very good to good, so the improvement is at low cost.

The behaviour of the GRADIENT method for  $\alpha \leq 0.5$  is remarkable. There is little or no improvement for the quality measures regarding the restrictions if  $\alpha$  is decreased. But the other quality measures show a clear negative effect on the matrix. Apparently there is a critical  $\alpha$  value, underneath that point no more significant improvement is achieved for the restrictions. An plausible explanation can be found in the working of the GRADIENT method, the search directions are dominantly determined by the right term of the objective function for all  $\alpha$ 's smaller than 0.5. So the iterations are almost equal for each  $\alpha \leq 0.5$  which results in similar outcomes.

#### 7.1.4 Dynamic Tactical Traffic Model Amsterdam

GRADIENT is successfully applied to the Amsterdam network. GRADIENT is called five times interlaced with assignments, each GRADIENT call had 50 iterations.<sup>42</sup> After some adjustments it was decided to take  $\alpha = \frac{20}{21}$ .

---

<sup>42</sup>Such a high amount of iterations was easily attainable considering the small network.

Method	R <sup>2</sup> -RES	R <sup>2</sup> -MAT	T3.5-RES	T4.5-RES	T5.5-RES	T3.5-TF	T4.5-TF	T5.5-TF	GEH5	GEH10	
GRADIENT	$\alpha = 0.01$	0.99822	0.99442	<i>69.378</i>	<i>84.512</i>	93.777	99.785	99.934	99.983	63.579	84.866
	$\alpha = 0.1$	0.99822	0.99472	<i>69.378</i>	<i>84.441</i>	93.777	99.785	99.934	99.983	63.579	84.866
	$\alpha = 0.2$	0.99821	0.99509	<i>69.378</i>	<i>84.441</i>	93.777	99.785	99.934	99.983	63.579	84.795
	$\alpha = 0.5$	0.99818	0.99650	<i>69.236</i>	<i>84.371</i>	93.777	99.785	99.950	100.0	63.508	84.724
	$\alpha = 0.98$	0.99659	0.99964	<i>61.881</i>	<i>78.571</i>	90.665	99.983	100.0	100.0	55.587	78.996
	$\alpha = 0.99$	0.99541	0.99982	<i>58.982</i>	<i>75.955</i>	88.543	100.0	100.0	100.0	53.465	76.521
	$\alpha = 0.998$	0.99218	0.99998	<i>55.446</i>	<i>71.570</i>	85.573	100.0	100.0	100.0	50.849	72.631
	$\alpha = 0.999$	0.99121	0.99999	<i>54.526</i>	<i>71.004</i>	85.149	100.0	100.0	100.0	49.646	71.994
	REVISED	0.99948	0.99256	93.777	98.727	100.0	96.935	98.857	99.437	91.584	98.868
	COMB1	$\alpha = 0.01$	0.99964	0.99279	95.262	99.717	100.0	97.084	98.890	99.453	93.494
$\alpha = 0.1$		0.99964	0.99301	95.120	99.717	100.0	97.084	98.890	99.453	93.423	99.788
$\alpha = 0.2$		0.99964	0.99328	95.120	99.717	100.0	97.084	98.890	99.453	93.423	99.788
$\alpha = 0.5$		0.99964	0.99440	95.120	99.717	100.0	97.084	98.890	99.453	93.423	99.788
$\alpha = 0.9$		0.99962	0.99655	94.979	99.576	100.0	97.084	98.923	99.470	93.140	99.717
$\alpha = 0.98$		0.99957	0.99622	94.272	99.364	100.0	97.018	98.907	99.470	92.504	99.505
$\alpha = 0.99$		0.99954	0.99541	94.059	99.010	100.0	96.952	98.873	99.437	92.079	99.222
$\alpha = 0.998$		0.99950	0.99355	93.847	98.939	100.0	96.952	98.857	99.437	91.584	99.010
$\alpha = 0.999$		0.99949	0.99310	93.847	98.798	100.0	96.952	98.857	99.437	91.584	99.010
COMB2		$\alpha = 0.01$	0.99969	0.99244	95.191	99.646	100.0	96.670	98.807	99.437	93.069
	$\alpha = 0.1$	0.99970	0.99291	95.191	99.717	100.0	96.653	98.807	99.437	93.140	99.788
	$\alpha = 0.2$	0.99969	0.99345	95.120	99.717	100.0	96.670	98.824	99.437	93.069	99.788
	$\alpha = 0.5$	0.99948	0.99256	93.777	98.727	100.0	96.935	98.857	99.437	91.584	98.868
	$\alpha = 0.9$	0.99956	0.99794	94.413	99.364	100.0	96.869	98.940	99.453	92.008	99.576
	$\alpha = 0.98$	0.99924	0.99911	92.574	98.515	100.0	97.034	98.956	99.453	90.240	98.727
	$\alpha = 0.99$	0.99899	0.99928	91.867	98.020	99.788	97.101	98.956	99.453	89.533	98.586
	$\alpha = 0.998$	0.99844	0.99943	91.018	97.313	99.364	97.200	98.956	99.453	88.685	97.666
	$\alpha = 0.999$	0.99830	0.99945	91.089	97.313	99.293	97.184	98.956	99.453	88.543	97.595

Table 3: Results of test for several  $\alpha$  values

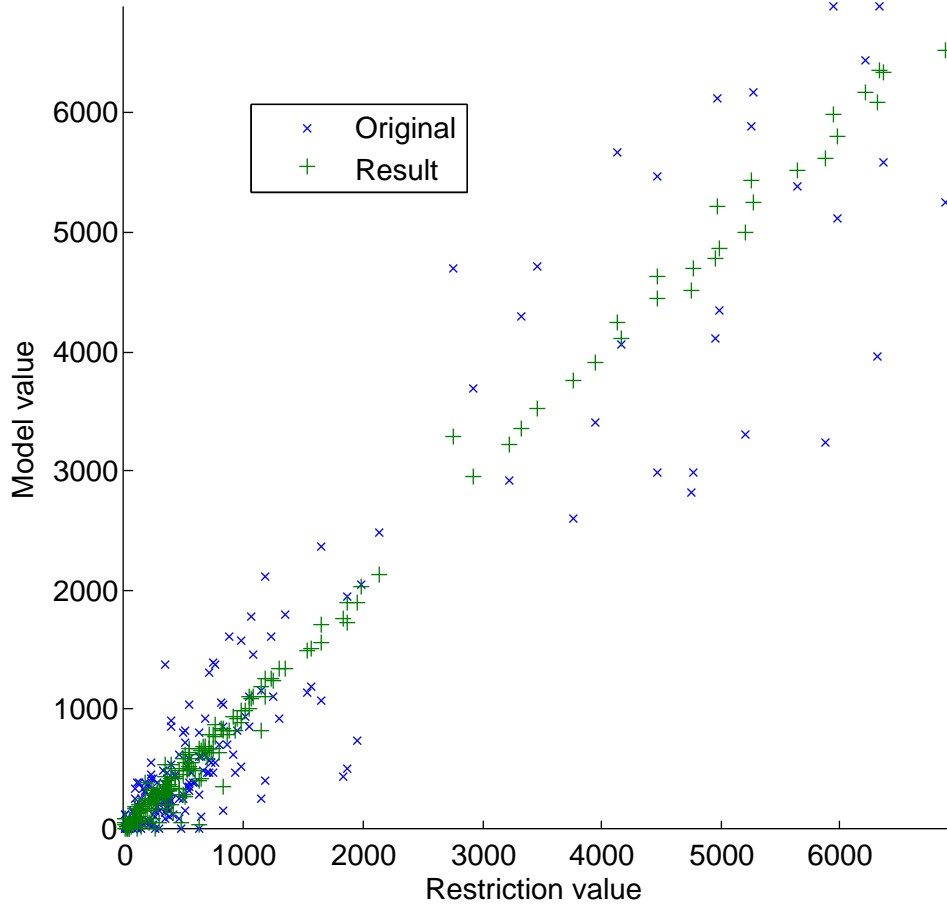


Figure 20: Scatter plot of the a priori and a posteriori situation for the restrictions in the Amsterdam network

The final result had the following score on the measures:

Measure	value
$R^2$ -RES	0.99465
$R^2$ -MAT	0.97727
T3.5-RES	84.804
T4.5-RES	90.686
T5.5-RES	96.569
T3.5-TE	79.087
T4.5-TE	89.354
T5.5-TE	95.627
GEH5	78.431
GEH10	90.686

So the norm for T4.5-RES and T3.5-TE were not achieved, nevertheless the results were judged as good. Figure 20 shows a scatter plot of the a priori and a posteriori restriction matchings. It shows that the points are less spread in the a posteriori situation than in the a priori situation. Furthermore there are no outliers. Figure 21 shows the a priori - a posteriori O-D matrix scatter plot for Amsterdam. The two lines are boundaries for an absolute error of 40, all points are between these boundaries. This means that the trips for an O-D pair does not change by more than 40, which is an acceptable result.



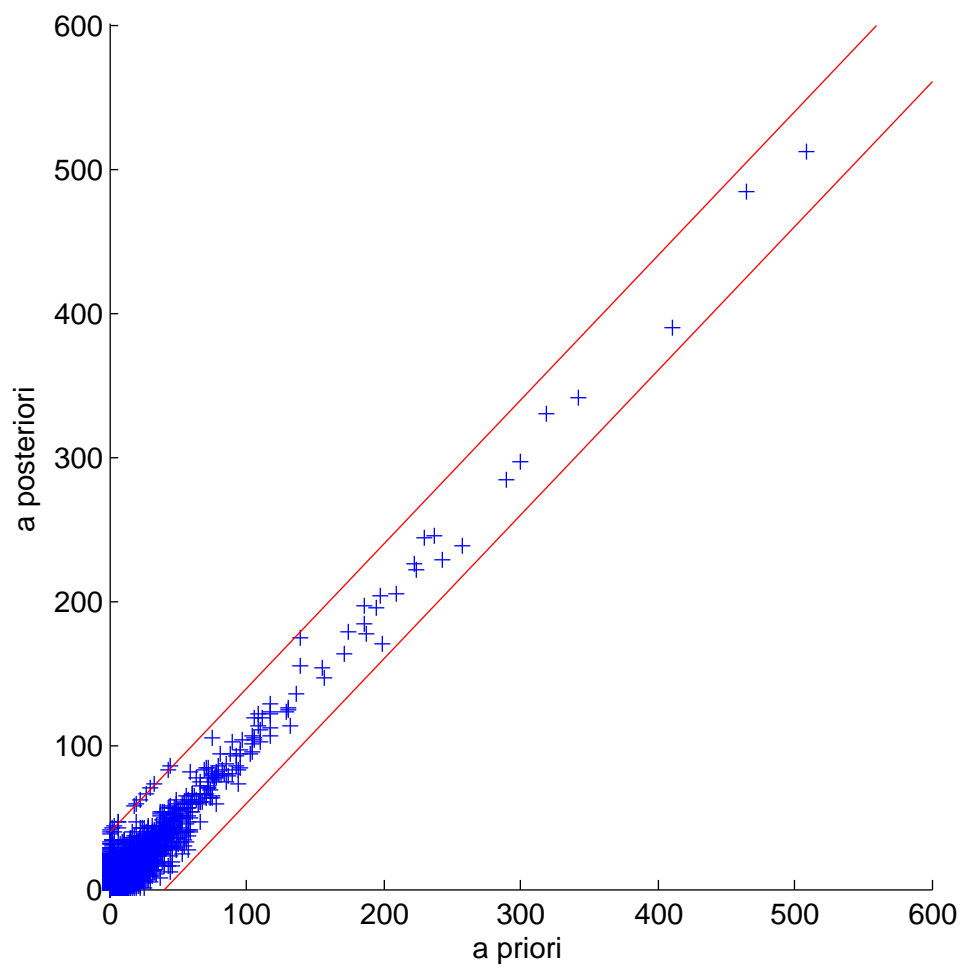


Figure 21: Scatter plot of the a priori and a posteriori O-D matrix in the Amsterdam network

## 7.2 Memory reduction

The bottleneck for memory usage is the storage of the routes and route choice. There are several ways to store this information, full path information will give the user and algorithms the most freedom. However this approach will need the storage of a tremendous amount of paths. Namely the number of routes per O-D pair, times the number of O-D pairs, times the number of dimension. The latter comes from the fact that routes and route choice is not equivalent for different dimensions. In all the approaches of this study the  $\mathbf{P}$  values are used to store route choices, for each dimension they effectively consist of a matrix. In OmniTRANS this will be the sceneline matrix in most cases. The number of elements of such a matrix is the squared number of zones. A lot of the entries will be zero, since only for some O-D pairs there will be a path that is significant for the restriction. In general it can be said that the elements of the matrices are between one and zero.<sup>43</sup> To store these matrices as a regular matrix (i.e. a double array) will be highly inefficient.

### 7.2.1 8-bit unsigned integer

In general a matrix is stored as a double array of doubles<sup>44</sup>, usually two 32 or 64 bit integers. One bit is used for the sign (+ or -) and 32-bit integer ranges from  $-2^{31} - 1$  to  $+2^{31} - 1$ , that is -2147483647 to 2147483647. In a screenline matrix all numbers are between 0 and 1, so it highly inefficient to store it as a regular matrix. The use of a double array of unsigned 8-bit integers can give a memory use reduction from 87.5 to 93.75 percent.

An instance of an unsigned 8-bit integer can be one of the integers between 0 and 255. These numbers can be mapped to the interval  $[0, 1]$  by dividing them by 255. An implementation of this mapping was made in Matlab to store the matrices (see next paragraph) as double arrays of 8-bit unsigned integers. Although this gave significant lower memory usage it was not used due the rounding errors and the fact that there was a better way to store the matrices.

### 7.2.2 Sparse matrix structure

The majority of elements of the matrices is zero, until now these zeroes are stored in the memory. The sparsity of a matrix is determined by the fraction of zero elements. In the case of the  $\mathbf{P}$  values the matrix is so sparse that it is more efficient to store them in a sparse matrix structure. The only elements that are stored in a sparse matrix structure are the non-zero elements. For each non-zero element of the matrix the value and the position indices are stored.

#### EXAMPLE 5 (Sparse matrix)

Consider the following matrix:

$$M = \begin{pmatrix} 0 & 0 & .75 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

If it is stored as a normal matrix it will use  $4 \times 4 = 16$  doubles in the memory. With a sparse matrix structure  $M$  will be stored as

$$\begin{array}{ll} (1, 3) & .75 \\ (3, 2) & 0.5 \end{array}$$

using 4 integers (for the indices) and two doubles (for the values) in the memory which is significantly less.

---

<sup>43</sup>This is not necessary since a route can pass the same screenline twice, but then the screenline is not properly defined

<sup>44</sup>A double is used to store real numbers, it is an integer that stores the integer part and an integer that stores the decimal part of the real number

A drawback to this structure is that it is not possible to directly obtain the value at an arbitrary position. In the normal structure this can be directly read from the memory, with the sparse structure the list of indices must be searched. Fortunately the operators used with these matrices in this study appear to be very effective with the sparse matrix structure. The value on a specific index is effectively never needed. Although it might not be clear directly from the three algorithms, the only operators used in the implementation of the algorithms are the (of course element-wise) matrix addition and element-wise multiplication. For these operators the program can walk through the list of non-zero elements since the result with zero elements are obvious. This will be made clear via an example.

**EXAMPLE 5 (CONTINUED)**

Consider the following matrix as well

$$N = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

To evaluate  $N + M$  the only thing needed is to add the non-zero elements of  $M$  to  $N$ , which are perfectly stored in the sparse structure. To evaluate  $N \cdot M$  (element-wise product) the only non-zero elements in the result are the non-zero elements of  $M$ , which are again perfectly stored in the sparse structure. In this case only the indices (1,3) and (3,2) are considered to change  $N$ , in the case of the addition the other elements are unchanged and in the case of the multiplication the other elements become zero. So both operators are even faster if a sparse matrix structure is used. The application of the operators on two sparse structured matrices are also straightforward.

It is not surprising that this sparse matrix structure is eventually used in the Matlab implementation to store the  $\mathbf{P}$  values. A very significant memory reduction was achieved by this. The total memory usage of Matlab in Windows dropped from 1.3 GB to 300 MB with the introduction of the sparse matrix structure for the Alphen model. The memory consumption of Matlab in idle state is already 270 MB, so the improvement is tremendous.

## 8 Conclusions & Recommendations

### 8.1 Conclusions

The study delivered two new matrix estimation methods for OmniTRANS. It is proven that REVISED overtake some shortcomings of ORIGINAL. The results do not depend on the order of the input and restrictions treated later in the algorithm are not favoured. The results of the REVISED method (that applies an averaged factor for each O-D pair) can be explained and justified easier than the ORIGINAL method. This is a potential expansion for Omnitrans.

GRADIENT is a story apart. The method has its roots in a mathematical sound problem formulation contrary to ORIGINAL and REVISED. The convergence is nevertheless disappointing, all quality measures regarding the restrictions stay far behind. Because the gradient descent method causes the O-D matrix to change over one search direction in each iteration the differences that are made are small. A large advantage of GRADIENT is the conservation of the a priori matrix structure, the scatter plots show that deviations are bounded (at least visually). ORIGINAL and REVISED on the other hand show unwished deviations, unpredictable, unbounded changes do occur in the performed tests. The parameter  $\alpha$  in GRADIENT is a steering wheel for modellers to obtain the desired result, but it is hard to determine which  $\alpha$  value is 'good'. Sometimes this parameter should be calibrated into the fourth or fifth decimal.

Combinations of the methods (REVISED and GRADIENT in particular) lead to predictable outcomes, but are not able to combine the largest advantages of both methods. After application of ORIGINAL or REVISED the matrix drastically changed, GRADIENT is not able to restore the structure. GRADIENT only succeeds in softening the sharp edges in that case. Furthermore it is hard to justify such combinations of methods when used in commercial projects.

Michael Florian, a professor from Montreal, Canada and an authority on the subject has this opinion about matrix estimation: *Matrix Estimation is rather an art than science*. This study does agree with that. The approaches are so numerous and diverse that it is not possible to create an unequivocal 'right solution method'. Modellers and users must have an understanding of where they work with and what the caveats of these methods are. A good advise is to build and investigate scatterplots and identify outliers for example.

This study showed that sparse matrix structures significantly reduce the memory usage while it gives arithmetic advantages as well.

### 8.2 Recommendations

#### 8.2.1 Integration of methods

The current practice considers the assignment and matrix estimation as two distinct methods. To achieve an optimal solution these methods are iterated, but within both methods iterations take place as well. It should be possible to integrate these methods into a single one since they use almost the same data sets and variables. In several assignment methods like volume averaging and equilibrium assignment a new route is calculated in each iteration step and a fraction of the demand is assigned to it. Directly after this change in the assignment a step of the GRADIENT method can be performed. The advantage comes from the assignment, because it holds<sup>45</sup> while doing matrix estimation it does not have to build equilibrium from scratch again. So the previous proportions are still in the memory and the assignment continues with calculating new shortest paths. Additional research is needed to examine the feasibility of this approach. The update of the loads on the network can be a caveat in this approach.

---

<sup>45</sup>It keeps the variable space and thus the results up till then in the memory

## 8.2.2 Other restrictions

The currently available restriction are traffic counts, screenlines, blocks and trip ends. With the generalization to restrictions it is possible to introduce new types. This is especially demanded in public transport, an overview of potential restrictions with explanation of the  $\mathbf{P}$  values is given:

- **Turning movement count:** Traffic volumes that pass a particular turning moving are sometimes registered and available. It is a good information source for matrix estimation. The definition is analogous to that of traffic counts, the difference lies in the network element where the information is stored. The  $\mathbf{P}$  values are the proportions of each O-D pair and relevant dimension that makes the particular turning movement.
- **Transit link count:** The traffic volume passing a transit link can be treated analogous to regular traffic counts. For the  $\mathbf{P}$  values the sum over all lines passing the location should be taken
- **Total boarding/alighting<sup>46</sup> for a transit line:** The total number of boarding passengers on a transit line can be measured, together with the frequency of the line a counted flow can be derived. This can act as a restriction for matrix estimation. The particular  $\mathbf{P}$  values should be the fraction of demand for each O-D pair and each relevant dimension for which the line is boarded. This has similarities with a screenline, although the locations are stations. One should be aware that not every boarding movement on a station is on that particular transit line, the assignment method has to distinguish different flows for different lines while deriving the  $\mathbf{P}$  values. This is similar for the number of alightings at a transit line.
- **Total boarding/alighting at a transit stop:** The number of boarding passengers a on specific line and specific transit stop can be measured. This is the traffic count equivalent where the latter restriction was the screenline equivalent. So the  $\mathbf{P}$  values should be fraction of demand for each O-D pair and each relevant dimension for which the line is boarded at the particular stop. The same caution as above holds for the generation. This is similar for the number of alightings at a transit stop.
- **Total boarding/alighting at a transit station:** The number of passengers entering a station can be measured. This restriction can be defined as a screenline containing all non-transit incoming links of a station. The  $\mathbf{P}$  values are as well analogous to the  $\mathbf{P}$  values for a screenline. This is similar for the number of alightings at a transit station.

---

<sup>46</sup>'Uitstap' in Dutch.

## References

- [Abrahamsson (1998)] Abrahamsson, T., 1998. *Estimation of origin-destination matrices using traffic counts  $\tilde{U}$  a literature survey*. Technical Report, International Institute for Applied Systems Analysis, Laxenburg, Austria.
- [Barceló (1997)] Barceló, J., 1997. A survey of some mathematical programming models in transportation. *TOP. Journal of the Spanish Society of Statistics and Operations Research* 5, pp. 1-40.
- [Bell (1991)] Bell, M., 1991. The estimation of origin-destination matrices by constrained generalized least squares. *Transportation Research B* 25, pp. 13-22.
- [Cascetta & Nguyen (1988)] Cascetta, E. & Nguyen, S., 1988. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B* 22 (6), pp. 437-455.
- [Cascetta (2001)] Cascetta, E., 2001. *Transportation Systems Engineering: Theory and Methods*. Kluwer Academic Publishers, Dordrecht.
- [Chen, et al. (2009, in press)] Chen, A., et al., 2009. Norm approximation method for handling traffic count inconsistencies in path flow estimator. *Transportation Research Part B (in press)*, doi:10.1016/j.trb.2009.02.007
- [Chen & Florian (1996)] Chen, Y & Florian, M., 1996. O-D demand adjustment problem with congestion: Part I. Model analysis and optimality conditions. In Bianco, L. & Toth, P. (Eds.), *Advanced methods in transportation analysis*. Springer-Verlag, Berlin, pp. 1-27.
- [Doblas & Benitez (2005)] Doblas, J. & Benitez, F.G., 2005. An approach to estimating and updating origin-destination matrices based upon traffic counts preserving the prior structure of a survey matrix. *Transportation Research Part B* 39, pp. 565-591.
- [Goudappel Coffeng website] <http://www.goudappel.nl/>
- [Omnitrans International website] <http://www.omnitrans-international.com/>
- [Schilpzand] Schilpzand, M. *OmniTRANS technical document - Matrix Estimations*
- [Spiess (1987)] Spiess H., 1987. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research B* 21, pp. 395-412.
- [Spiess (1990)] Spiess, H., 1990. A gradient approach for the O-D matrix adjustment problem. *Centre de Recherche sur les Transports de Montréal*, Publication 693.