Survey of Research Reports in Transportation Modelling

Technische Universiteit Delft
Bibliotheek Faculteit der Civiele Techniek
(Bezoekadres Stevinweg 1)
Postbus 5048
2600 GA DELFT

Faculty of Civil Engineering
Department of Transportation Planning and Highway Engineering

Delft University of Technology
1. Report number
   VK 3901.403 B

2. Title report
   Survey of Research reports in transportation modelling

3. Editor(s)
   Ir.ing. A. Nijsse
   S.J. Wamsteker-Andriessen

4. Institute
   Delft University of Technology
   Faculty of Civil Engineering
   Dep. of Transp. Planning
   P.O.Box 5048, 2600 GA Delft

5. Principal(s)

6. ISSN-number
   LVV-Report
   0920-0592

7. Subject
   Transportation Modelling

8. Research project
   Optimization of transportation models

9. Category report
   Survey or research reports

10. Date
    April 1993

11. Summary
    A survey of research reports in transportation modelling
    in two parts.
    Part one is divided in reports concerning economic development
    and car mobility, analyzing large transportation data files
    and transportation planning and spatial development.
    Part two consists of research reports concerning optimization
    of public transport networks and road networks, time dependent
    assignment in road networks, sub-divided in models and hard and
    software and road traffic informatics and telematics.

12. Contacts

13. Pages
    320

14. Price
    f 65,--
DELFT UNIVERSITY OF TECHNOLOGY

FACULTY OF CIVIL ENGINEERING

Department of Transportation Planning and Highway Engineering

Transportation Modelling Group

Scientific staff: Prof. dr. ir. R. Hamerslag
Ir. L.H. Immers

Research group: Ir. ing. A. Nijsse
Ir. E. de Romph
Ir. drs. J.E. Stada
Ir. L.A. Tavasszy
Ir. M. Westerman
Ir. N. v.d. Zijpp
Dr. ir. H.J.M. van Grol (Faculty of Applied Physics)

Computer group: A. Bos
P.C.H. Opstal

Secretary: S.J. Wamsteker-Andriessen

Stevinweg 1,
P.O. Box 5048,
The Netherlands
Telephone: +31 15 781681
Telefax: +31 15 786993

NL-2628 CN Delft
NL-2600 GA Delft
PART I

1 ECONOMIC DEVELOPMENT AND CAR MOBILITY

R. Hamerslag (1985) and W.H. Scheltes
Impact of fuel price- and income-changes on transportation. *62th annual meeting Transportation Research Forum*, Washington DC

R. Hamerslag (1986), L.H. Immers and W.H. Scheltes
The influence of economic recession and changing fuel prices on car mobility in the Netherlands. *Proc. of the Fourth World Conference on transport research*, Vancouver. The center for transportation studies University of British Columbia Vancouver, Canada VGT IWS

R. Hamerslag (1990)
Analyzing the Netherlands Travel Survey. Methods for analyzing and some findings. *5th DVWG Seminar Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft e.V*, Reihe B, ISSN 0418-9183


R. Hamerslag (1992) and M. Westerman
Quantitative relationship between traffic, infrastructure and environment. *Proceedings 25th ISATA, Road Transport Informatics/ Intelligent Vehicle Highway Systems*, Florence, Italy, pp 221- 228

2 ANALYZING LARGE TRANSPORTATION DATA FILES

R. Hamerslag (1987) and W.H. Scheltes
New algorithm for grouping observations from a large transportation database. In: *Transportation Research Record 1090, Statewide Data Collection and Management Systems*. Washington

R. Hamerslag (1988) and L.H. Immers

R. Hamerslag (1981)
Investigation into factors affecting the route choice in Rijnstreek-West with the aid of a disaggregate logit model. *Transportation* 10 pp. 373-391
J. van der Waard (1988)
The relative importance of public transport trip-time attributes in route choice.
PTRC annual meeting, Bath, England

R. Hamerslag (1982), J.P. Roos and M. Kwakernaak
Analysis of accidents in traffic situations by means of multiproportional weighted poisson model. Transportation Research Record 847 pp 29-36

3 TRANSPORTATION PLANNING AND SPATIAL DEVELOPMENT

R. Hamerslag (1980)

Hamerslag R. (1989) and B. H. Immers
The interaction between the transport system and the spatial development. Model specification and model characteristics. Delft university of Technology. Dept of Transportation Planning and Highway Engineering. VK 3302.202. ISSN 0920-0592

R. Hamerslag (1992), E.C van Berkum and M.A. Replogle
The influence of highway and railway construction on spatial development: a modelling approach. 6th World Conference on Transport Research, Lyon, France

R. Hamerslag (1993), E.C van Berkum and M.A. Replogle
A model to predict the influence of new railways and roads on land use development. 72nd TRB annual meeting, Washington D.C. Paper 930875
PART II

4 OPTIMIZATION OF PUBLIC TRANSPORT NETWORKS AND ROAD NETWORKS


Ben H. Immers (1989) and Peter H. Mijjer

Jan Willem Houtman (1987) and Ben H. Immers
A traffic Assignment Model to Reduce Noise Annoyance in Urban Networks. *Transportation Research Record 1143*

L.H. Immers (1991) and N.G.J. Oosterbaan

Ben Immers (1983), Jan-Jaap Schoone, Henk Voogd
The Planning of Motorway Alignments through Computer-Aided Design. *Conference of the Italian Society for Operations Research, Lavoro*

5 TIME DEPENDEND ASSIGNMENT IN ROAD NETWORKS

5.1 Models

R. Hamerslag (1989)

R. Hamerslag (1989)
Dynamic assignment in the three-dimensional timespace. Mathematical specification and algorithm. *Proc. Joint USA-Italy seminar on urban networks, dynamic control and flow equilibrium*. Capri, Italy

E. Kroes (1990) and R. Hamerslag
A practical dynamic network equilibrium model with elastic departure times. *Transportation Research Board. 69th annual meeting*. Washington D.C. (paper 890788)

E. de Romph (1992), H.J.M. van Grol and R. Hamerslag
A Dynamic traffic assignment model for short-term predictions. 2nd International Capri Seminar On Urban Traffic Networks. Capri, Italy
5.2 Hard and Software

H.J.M. van Grol van (1990) and A.F.Bakker
A special purpose parallel computer for traffic simulation. *70th TRB Annual Meeting*, Washington

H.J.M. van Grol (1992) and A.F.Bakker
Large-scale Network Assignment: Algorithms, Hardware and Implementation. *PTRC annual meeting*

6 ROAD TRAFFIC INFORMATICS AND TELEMATICS

R. Hamerslag (1989) and W.J. Kribbe
A proposal for a travel-time-prediction-system. *Proc. 5th World Conference on Transport Research*, Yokohama, Japan

Ben H. Immers (1992) and Laurent J.C. de Ben
Analogies in communication systems. *6th World conference on Transport Research*, Lyon, France

M. Westerman (1992), L.H.Immers and R. Hamerslag
Determination of real time travel times on congested motorways for short term applications. *2nd International Capri Seminar On Urban Traffic Networks*. Capri, Italy

R. Hamerslag (1991) and E.C. van Berkum

R. Hamerslag (1991) and E.C. van Berkum
R. Hamerslag (1990)
Analyzing the Netherlands Travel Survey. Methods for analyzing and some findings. *5th DVWG Seminar Schriftenreihe der Deutschen Verkehrswissenschaftlichen Gesellschaft e.V.*, Reihe B, ISSN 0418-9183

R. Hamerslag (1992) and M. Westerman
Quantitative relationship between traffic, infrastructure and environment. *Proceedings 25th ISATA, Road Transport Informatics/ Intelligent Vehicle Highway Systems*, Florence, Italy, pp 221- 228

R. Hamerslag (1987)
1988

Design of public transport networks.

R. van Nes, R. Hamerslag and L.H. Immers

1202 Transportation Research Record, National Research Council, Washington DC
Design of Public Transport Networks

ROB VAN NES, RUDI HAMERSLAG, AND BEN H. IMMERS

This paper describes the major features of an optimization model which can be used to design public transport networks. Design problems that can be solved with the model involve the redesign of either a part of a network or a complete network and the assignment of frequencies. The model consists of an additive procedure in which the decision to incorporate a route in the network or to increase the frequency of a route is based on an economic criterion which can also be regarded as an estimate of the Lagrange Multiplier of the optimization problem. A major advantage of the model is that the different design problems are solved with one single optimization process. Furthermore, the optimization process is kept understandable and the model is suited for use on a personal computer. Some results of the model are presented.

Due to the changing economic situation the financial constraints of public transport have become more and more important. The government is no longer willing to account for all deficits of the public transport companies. The policy has changed into granting a single subsidy with which the public transport companies have to offer service which can compete with other modes and transport facilities for those who cannot travel otherwise. Since the subsidy will be limited, the public transport companies will have to reconsider the service they are offering. Of course, it is also necessary to cut costs by improving the scheduling of personnel and vehicles, as well as the regularity of the service.

The design of the network deserves extra attention, as it is the network that determines the service offered. Moreover, the network is used as input for studies concerning other aspects such as timetables, scheduling, and regularity. Another reason for extra attention to the design of public transport networks is the fact that networks have often been adjusted and the assignment of frequencies. The model consists of an additive procedure in which the decision to incorporate a route in the network or to increase the frequency of a route is based on an economic criterion which can also be regarded as an estimate of the Lagrange Multiplier of the optimization problem. A major advantage of the model is that the different design problems are solved with one single optimization process. Furthermore, the optimization process is kept understandable and the model is suited for use on a personal computer. Some results of the model are presented.

The problem of the network design can be formulated as follows: which routes and which frequencies should be offered to fulfill the demand for public transport as well as possible, given a certain available budget.

EXISTING DESIGN METHODS

The type of model mostly used for network design is an evaluation model in which an origin-destination (OD) matrix is assigned to a network and with which all kinds of evaluation characteristics are calculated. Such an evaluation model enables a systematic comparison between alternative network designs. Although the use of evaluation models must be considered as a major improvement to the quality of the planning process, the disadvantage remains that only a few alternatives can be compared because of the effort involved. Also, these alternatives will often be biased towards the existing network and the implicit ideas of the planner, although in some cases this might be considered as an advantage.

The disadvantage of a limited number of alternatives, however, does not apply to models which design a network as well as evaluate it. These so-called optimization models use operations research techniques to find a feasible network. The name optimization model is misleading, however, because most models do not find an optimum solution and even then it is questionable whether an optimum of a model, which is a mathematical description of reality, will be an optimum in reality. Therefore, the importance of these models does not derive from the fact that they find a (near) optimum solution, but rather that they help to find new and feasible alternatives.

Despite the advantage of generating new alternatives, the use of these optimization models is very limited. This limited use can be explained by several reasons, one of them being the overall lack of experience in using models in public transport studies, but when we take a close look at the optimization models which have been developed, some other reasons can be found.

EXISTING OPTIMIZATION MODELS

In the last two decades all kinds of optimization models for the design of public transport networks have been developed. These models can roughly be divided into six categories:

1. Analytical models (e.g., Holroyd [2], Koceur and Hendrickson [3]). These models use simplified networks to derive optimum relations for parameters of the public transport system, for instance headway and route-spacing.
2. Models determining which links should be used to construct routes for a public transport network (e.g., Bilheimer and Gray [4], Rea [3]).
3. Models determining routes without considering the frequencies of the routes (e.g., Pierick and Wiegand [6], Simonis [7]).
4. Models assigning frequencies to a given set of routes...
(e.g., Scheele [8], Furth and Wilson [9], Hagberg and Hasselström [10]).

5. Models determining routes in a first and assigning frequencies in a second stage (e.g., Lampkin and Saalmans [11], Dubois et al. [12]).

6. Models determining routes and frequencies simultaneously (Hasselström [13]).

The first two categories determine neither routes nor frequencies and are therefore unsuited for the problem we have formulated. The third and fourth categories solve only part of our problem, either routes or frequencies. Actually, there are only two categories of models suited to our design problem, categories 5 and 6.

Determining Routes and Assigning Frequencies Separately

The models of category 5 solve the network design problem in two stages. In the first stage the routes of the network are determined. The objective is to transport a maximum number of passengers given a fixed OD-matrix. In this stage Lampkin and Saalmans [11] consider trips without transfers, while Dubois et al. [12] consider all trips. In the second stage frequencies are assigned to the generated set of routes. The objective is to minimize the total travel time given the OD-matrix and the available number of vehicles. In the calculation of the travel time Dubois et al. [12] introduced the possibility of walking instead of using public transport. All the methods used are clearly heuristic, but those of Dubois et al. [12] are more sophisticated. The major disadvantage of these models, however, is the fact that they solve the problem of routes and frequencies separately, while there is a distinct relation between these two components of the public transport system. Moreover, a fixed OD-matrix is used, so the relation between supply and demand for public transport services is not taken into account.

Determining Routes and Assigning Frequencies Simultaneously

The model developed by Hasselström [13] does not have these disadvantages. It solves the problem in three stages. First, the model considers a link network and eliminates links seldom or never used by passengers (compare the models of category 2). The result is a concentrated network which is used in the second stage to generate a large set of possible routes. Finally, the route of the network are selected by assigning frequencies using linear programming. The objective is to maximize the number of transfers saved by changing from a link network (transfers at every node) to a public transport network (transfers only at intersections). Instead of a known OD-matrix, Hasselström [13] suggests the use of a desire matrix (i.e., an OD-matrix for the situation in which an ideal public transport system exists) in order to lessen the bias towards the network with which the OD-matrix is determined. The disadvantage of the model is that although routes and frequencies are determined simultaneously, two different optimization problems are formulated.

Use of Optimization Models

All optimization models discussed have rarely been used in practice. Most models have been employed only in the projects they were designed for, or in the tests described in the presentations. The model of Hasselström [13] forms part of the VOLVO-package, which contains a variety of models for planning public transport networks (e.g., Andreason [4]), and has been used more often (Arntstrom [5], HTM [6] and Harris and Haywood [17]). The major disadvantages of all optimization models, however, are the complex structure (e.g., several optimization problems within one model) and the limited accessibility for planners as the models can be used only on a mainframe.

A NEW MODEL

The disparities between the capabilities of optimization models in the design process and the practical situation combined with the need to improve the design process are the reasons for developing a new model. If a model is to be used as a tool in the design process, it should fulfill the following requirements:

1. It should be suited for several design problems ranging from short-term analyses to long-term decisions, e.g., assigning frequencies, designing part of a network and designing a complete network.
2. It should be easily accessible and understandable for the user (i.e., the planner).

The model presented in this paper is an attempt to serve as such a model. It is suited for use on a personal computer and special attention is given to the interactive design process. Moreover, the optimization model will be included in a software package for the design of public transport networks. This package will also contain a model for the determination of an OD-matrix, an evaluation, model and interactive programs to arrange the necessary input. Activities for which the package can be used are as follows:

1. Evaluating a network,
2. Assigning frequencies,
3. Designing or redesigning part of a network,
4. Designing or redesigning a complete network.

For activities 2, 3 and 4, the optimization model can be used. The optimization process is structured to be simple and understandable.

OPTIMIZATION PROBLEM IN WORDS

The main objective is to design a network which can fulfil the demand for public transport as well as possible. It is obvious that this objective cannot be used in an optimization model, as it is unclear what is meant by "as well as possible." Does a network qualify as "good" if it offers services which can compete with other modes, or if it is especially suited to the needs of people who cannot travel otherwise? The decision on what is meant by "as well as possible" is a political one,
however, and should not be made within an optimization model.

An objective suited to an optimization model and for both interpretations of "as well as possible" is maximizing the number of passengers, given a certain budget. It is a well-known fact that transfers negatively affect the number of passengers. Recent research in the Netherlands shows a penalty of 6 minutes, not including the waiting time at the transfer point (Van der Waard et al. [28]). Therefore, maximizing the number of passengers is more or less equivalent to minimizing transfers, especially in middle-sized cities such as those in the Netherlands. Although maximizing transfers is a commonly used objective (see e.g., Hasselström [12]), it is preferable to maximize the number of direct trips. The objective of maximizing direct trips makes it possible to use a description in which the demand for public services depends on the quality of the services offered, while the objective of minimizing transfers requires a fixed OD-matrix. Therefore, we choose to maximize the number of direct trips.

The major constraint of the problem is the available budget. Since there is a strong relation between the available budget and the number of vehicles that can be put into operation, the optimization problem can be formulated as follows:

Maximize the number of direct trips given a certain fleet size.

A special aspect of the public transport system is the use of different vehicle types (e.g., bus, tram). As the vehicle type influences both generalized costs and total costs, this aspect will also be included in the optimization model. Of course it is possible that, by maximizing the number of direct trips, networks may be developed which offer very poor transfer facilities, resulting in far fewer passengers than the highest number desirable. Therefore, additional constraints, such as a maximum number of routes or a minimum frequency, may be necessary. The decision as to which constraints must be imposed depends on the characteristics of the demand pattern and the specific network.

RELATION BETWEEN SUPPLY AND DEMAND FOR PUBLIC TRANSPORT SERVICES

The formulated objective makes it necessary to describe the relation between supply and demand for public transport services. It is not possible to use elasticities which are based on empirical research of the behavior of passengers as a result of changes in the public transport system. Usually these elasticities have constant values and are time and place dependent. Therefore, a direct demand model is formulated, which is based on the simultaneous distribution-modal split model (see, e.g., Wilson [19]). The relation between supply and demand for public transport is described by the deterrence function.

The simultaneous distribution-modal split model can be formulated as (see, e.g., Wilson [19]):

\[ T_{ij} = r \cdot o_i \cdot d_j \cdot F_{ij} \quad \forall i,j \]  

where

- \( T_{ij} \) = number of trips between nodes \( i \) and \( j \),
- \( r \) = constant term
- \( o_i \) = factor for the generation of node \( i \),
- \( d_j \) = factor for the attraction of node \( j \),
- \( F_{ij} \) = value of the deterrence function for all modes for OD-pair \( i,j \).

Constrained by:

\[ \sum_j T_{ij} = D_i \quad \forall i \quad \text{and} \quad \sum_i T_{ij} = O_j \quad \forall j \]  

where

- \( D_i \) = arrivals at zone \( i \),
- \( O_j \) = departures from zone \( j \).

\( F_{ij} \) can be written as:

\[ F_{ij} = \sum_v \left( \frac{C_{ijv}}{C_{ij}} \right) \quad \forall i,j \]  

where

- \( F_v \) = the deterrence function for mode \( v \),
- \( C_{ijv} \) = the generalized costs for OD-pair \( i,j \) with mode \( v \).

Finally, the number of trips by public transport can be calculated with the following equation:

\[ T_{up} = T_{ij} \cdot \frac{F_{ij}}{F_v} = r \cdot o_i \cdot d_j \cdot F_{ij} / C_{ijv} \quad \forall i,j \]  

where

- \( T_{up} \) = number of trips by public transport between nodes \( i \) and \( j \),
- \( F_v \) = the deterrence function for public transport,
- \( C_{ijv} \) = the generalized costs for OD-pair \( i,j \) with public transport.

We will assume that a small change in the public transport system will only affect the number of trips by public transport, and will not affect the number of trips by other modes (\( T_{up} \)). This assumption is acceptable for situations where 10–20 percent of all trips are made by public transport (e.g., in the Netherlands). The values of \( o_i \) and \( d_j \) are known, so by using equation (4) it is possible to calculate the number of trips. The values of \( o_i \) and \( d_j \) have to be determined for a situation which is comparable with the new situation. In case of an existing public transport system, these values can be calculated with an observed OD-matrix and for instance a weighted Poisson model (e.g., Hamerslag et al. [20]). When large changes are expected to occur, such as new residential areas, a traffic forecasting model should be used to calculate the values of \( o_i \) and \( d_j \).

OBJECTIVE FUNCTION

Given is a set of possible routes \( Y \) with characteristics such as:

1. \( f_r \) = frequency of route \( r \),
2. \( s_r \) = vehicle type used on route \( r \) (e.g., bus, tram),
3. \( N_r \) = set of nodes connected by route \( r \),
4. \( t_v \) = in-vehicle times between the nodes of set \( N_r \).
Only \( f_s \) will be used as the decision variable in the optimization process; all other characteristics are assumed to be fixed for each route. For instance, if a route can be used by two vehicle types, two identical routes have to be included each with a different vehicle type.

The set of possible routes \( Y \) can be generated in several ways, for instance with the method described by Ceder and Wilson (21), or such a set can be developed manually, using the interactive programs that will be included in the package. The final package will contain a model for the generation of routes.

The objective is to maximize the number of public transport passengers who can travel without transfers:

\[
\max \sum_{y} \left\{ \sum_{i} \left( r \cdot a \cdot F_{s}(C_w) \right) \right\}
\]

The generalized costs for an OD-pair are determined by the set of routes \( S_i \), which offer a direct trip for the OD-pair \( i,j \). Therefore:

\[
F_{s}(C_w) = G(S_i)
\]

with \( S_i = \) a set of routes with \( i \in N_i, j \in N_j \), and \( f_s = 0 \) for \( \forall y \in S_i, S_0 \in Y \).

When equation (6) is substituted in (5) the objective can be written as:

\[
\max \sum_{s} \left\{ \sum_{r} \left( r \cdot a \cdot F_{s}(C_w) \right) \right\}
\]

The description of the public transport system results in a complicated analytical formulation of the objective. In order to derive a formulation which is more suitable for analytical analyses, a somewhat simplified description is used. For instance, let us assume an exponential function for \( f_{ij} \):

\[
f_{ij} = a \cdot \exp \left\{ -b \cdot \left( C_w + c \right) \right\}
\]

with \( a, b, \) and \( c \) as the coefficients.

The generalized costs can be written as:

\[
C_w = g_w + (60 \cdot h) \left( \sum_{r} (f_r) \right)
\]

with

\[
g_w = \text{a constant for OD-pair } i,j, \text{ determined by the time to access and to egress the system and the time spent in the vehicle},
\]

\[
h = \text{a parameter for the calculation of the waiting time (including the weight of the waiting time)}.
\]

Of course this description of the generalized costs is too simple in case there are several routes available for the OD-pair \( i,j \), but it is sufficient to illustrate the problem. Equation (7) can then be written as:

\[
\max \sum_{r} \left\{ \sum_{s} \left( r \cdot a \cdot d_s \cdot \exp \left\{ -b \cdot \left( g_w + (60 \cdot h) \left( \sum_{r} (f_r) \right) + c \right) \right\} \right) \right\}
\]

**CONSTRANTS**

The constraints of the problem are the available budget (S1) and the number of vehicles per vehicle type (S2). Furthermore, the possible frequencies are restricted to a limited set of integer values in order to make it easy for the passenger to memorize headways (S3), and of course only an integer number of vehicles can be assigned to a route (S4). These constraints can be written as:

\[
S1: \sum_{r} \left\{ k_s \cdot \left( \sum_{y} (n_{y} \cdot b_s) \right) \right\} \leq K
\]

with \( K = \text{the available budget,} \)

\( n_{y} = \text{the number of vehicles that is necessary for route } y, \)

\( b_s = \text{a binary variable that indicates whether route } y \text{ will be included in the summation (} b_s = \text{1 if } s = s; \text{ otherwise, } b_s = \text{0}), \)

\( k_s = \text{a factor for the costs of using a vehicle of type } s. \)

\[
S2: \sum_{y} (n_{y} \cdot b_s) \leq m_{y} \text{, } \forall \ s
\]

with \( m_{y} = \text{available number of vehicles of type } s. \)

\[
S3: f_s \in S \text{, } \forall \ y
\]

with \( f = \text{set of possible (integer) frequencies.} \)

\[
S4: n_{y} - 1 < (f_{y} - m_{y}) \leq n_{y} \text{, } \forall \ y
\]

with \( m_{y} = \text{the number of vehicles that is necessary for the frequency of one vehicle per hour on route } y. \)

**SOLUTION METHOD**

The formulated problem has a non-linear objective, linear constraints and a great number of integer variables. There are no efficient algorithms available to solve the problem without simplifying the formulation. For that reason Lampkin and Saalmans (11) and Dubois et al. (12) solve the problem in two stages; first, determine the routes and second, assign the frequencies. But, as there is a distinct relation between routes and frequencies, it would be better to determine them simultaneously. Therefore a new method has been developed. This method can be described as follows:

0. Set all frequencies equal to 0 and determine the elements of the sets \( S_y \).

1. Determine for each route \( y \) the efficiency \( r_y \) of an increase of the frequency by calculating the ratio of the number of extra passengers as a result of this increase and the necessary costs:

\[
r_y = \frac{\sum (r \cdot a \cdot d_s \cdot G(S_{w}))}{k_{y} \cdot m_{y} \cdot (f_{y} - f_{y})}
\]

\[
= \frac{\sum (r \cdot a \cdot d_s \cdot G(S_{w}))}{k_{y} \cdot m_{y} \cdot (f_{y} - f_{y})}
\]

(15)
with

\[ m, n \in N, \]
\[ f_1, f_2 \in f, \text{ and } f_1 < f_2. \]

\[ r_y \] = the efficiency of route \( y \).

\[ S_{m-n} \] = set of routes available for OD-pair \( m-n \).

\[ k_s \] = factor for the costs of using a vehicle of type \( s \) \( \left( s = s_y \right) \).

2. Select the route with the highest efficiency ratio and increase the frequency of that route.

3. Check the constraints \( S1 \) and \( S2 \) (eq. [11] and [12]); if they are no longer met the process stops; otherwise, continue with step 1.

A special feature of the method is the possibility of assigning some routes a fixed frequency, e.g., routes of other public transport companies, or routes of a vehicle type of which vehicles are no longer available. Because the optimization problem is limited to passengers who are being offered a direct trip, the values of the \( r_y \)s and the value of the objective function can quickly and easily be obtained. By restraining the set of possible routes \( Y \), the method can also be applied to other design problems; for example, if we are only interested in the assignment of frequencies, the set \( Y \) consists of the existing routes.

ANALYSIS OF THE METHOD

Lagrange Multiplier

Although the method is heuristic, the efficiency of a route \( (r_y) \) can be regarded as an estimate of the Lagrange Multiplier, which is introduced when the first constraint is included in the objective and the integer constraints are dropped. The Lagrange formulation can then be written as:

\[
\max \sum_y \left\{ \sum_i [a_i \cdot d_i \cdot G(S_i)] \right\}
- \mu \cdot \left\{ \sum_k \left[ k_s \left( \sum_y (f_y \cdot n_v f_y \cdot b_y) - K \right) \right] \right\} \tag{16}
\]

with \( \mu \) = the Lagrange Multiplier.

An optimum will be found when the Kuhn-Tucker conditions are fulfilled. Therefore it is required that:

\[
\frac{\partial}{\partial f_y} \left\{ \sum_y \left[ a_i \cdot d_i \cdot G(S_i) \right] \right\}
- \mu \cdot \left\{ \sum_k \left[ k_s \left( \sum_y (f_y \cdot n_v f_y \cdot b_y) - K \right) \right] \right\} = 0 \quad \forall y \tag{17}
\]

with \( m, n \in N, \)

\[
\mu \cdot \left\{ \sum_k \left[ k_s \left( \sum_y (f_y \cdot n_v f_y \cdot b_y) - K \right) \right] \right\} = 0 \tag{18}
\]

\[
\mu \equiv 0 \tag{19}
\]
Using equation (17) the Lagrange Multiplier \( \mu \) can be written as:

\[
\mu = \frac{\delta f}{\delta r} \left[ \sum S(S_{im}) \right]^{-1} \quad \forall y
\]

If we take the limit of equation (15) as \( \left(f_{x} - f_{y}\right) \) approaches zero, the resemblance between the equations (15) and (20) is obvious. From this point of view the method is based on minimizing the variance between the values of \( r_{y} \) by increasing the frequency of the route with the largest \( r_{x} \). The values of \( r_{y} \) will decrease gradually and finally converge to a solution in which they are more or less equal to each other and consequently equal \( \mu \).

Concavity

If the objective function is concave over the decision variables \( (f_{y}) \), the Kuhn-Tucker conditions are sufficient to determine the optimum. For an exponential, as well as for a lognormal reference function it can be shown that the objective is concave for frequencies greater than a certain value, depending on the coefficient being used (see Figure 1). The concavity of the objective function also guarantees that an increase of \( f_{x} \) will result in a decrease of \( r_{x} \), and consequently that the method converges to a solution.

Quality of the Solution

As the method can be used for several design problems, there are two aspects that have to be analyzed:

1. The assignment of frequencies.
2. The selection of routes.

Both aspects have been analyzed with the use of the simplified objective described with equation (10). When we restrict the problem to assigning frequencies only, we can derive an alternative solution technique. By introducing the first constraint (11) in the objective (10) the Lagrange equation is derived. In the optimum situation the Kuhn-Tucker conditions should be fulfilled. These conditions result in a set of non-linear equations, which because of the concavity of the objective can be solved using Newton-Raphson (see Simmons [22]). For these analyses a simple network (Figure 2 and Table 1) has been used.

For several sets of routes the frequencies were determined using Newton-Raphson as well as the new method with \( f_{x} = f_{y} = 0.1 \) and with \( f_{x} = f_{y} = 1 \). The results show that with a small stepsize the new method gives the same results as Newton-Raphson. If we use the integer stepsize, the results are quite satisfactory. The results are shown in Table 2.

The selection of routes is more difficult to analyze as the method used to analyze the assignment of frequencies cannot be used for the selection of routes. Therefore this has been done by comparing the first four selected routes with the results of every possible combination of four routes from the set \( Y \). For each combination the Newton-Raphson method was used to determine the frequencies and the value of the objective. This comparison showed that the selected four routes were the best combination. Moreover, this analysis showed

<table>
<thead>
<tr>
<th>Zone</th>
<th>( a_{i} )</th>
<th>( d_{i} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

**TABLE 2** COMPARISON OF CALCULATED FREQUENCIES FOR A TEST NETWORK

<table>
<thead>
<tr>
<th>Route</th>
<th>Frequency</th>
<th>( (f_{x} - f_{y}) = 0.1 )</th>
<th>( (f_{x} - f_{y}) = 1.0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2-6-5</td>
<td>1.9</td>
<td>1.9</td>
<td>2.0</td>
</tr>
<tr>
<td>1-2-6-7</td>
<td>2.3</td>
<td>2.3</td>
<td>2.0</td>
</tr>
<tr>
<td>4-3-7-6-5</td>
<td>1.4</td>
<td>1.4</td>
<td>1.0</td>
</tr>
<tr>
<td>4-3-2-6-8</td>
<td>2.4</td>
<td>2.4</td>
<td>3.0</td>
</tr>
<tr>
<td>No. of direct trips</td>
<td>606.5</td>
<td>606.2</td>
<td>692.9</td>
</tr>
</tbody>
</table>
that 3 percent of the combinations were near-optimal, i.e., within 2 percent difference from the optimum solution. Analyses with a more realistic network gave similar results.

ADDITIONAL FEATURES

Although the test showed good results for the method, there are indications that the purely additive nature of the method might have a negative effect on the optimal quality of the selected network. Therefore, an exchange routine to check the solution has been introduced. This routine is based on the interpretation of the efficiency $r$, as an estimate of the Lagrange Multiplier $\mu$ and checks whether the solution can be improved by replacing a selected route with another. If an optimum solution has been found it will not be possible to improve the solution in this way because the efficiency of the routes will be more or less equal. Moreover, an interactive routine is developed which can be used to analyze a solution by fixing frequencies, dropping routes or introducing extra routes, and to restart the optimization process, for example, to assign frequencies or to select alternative routes given the adapted solution. Therefore, the model does not present the solution, but allows the planner to play around with an optimized solution. Besides, the possibility of using different starting sets $Y$, developed manually or with the use of a route generation model, also offers different solutions from which the planner may choose.

INCORPORATION OF THE PUBLIC TRANSPORT SYSTEM IN THE MODEL

Another aspect which determines the quality of the model is the description of the public transport system. Some special features of this description will be discussed in this paragraph.

The area that is the subject of the study is divided into zones, which are located around the stops. For each zone an access- and egress-time is determined. Trips from or to the study area are supposed to enter or leave at fictional zones located at the major transfer points between the local and regional public transport system.

The generalized costs consist of the weighted sum of the time-elements of a trip by public transport, namely the in-vehicle-time, the access- and egress-time, the in-vehicle time and the waiting time. The in-vehicle time is weighted with a coefficient which depends on the vehicle type. The waiting time can be calculated with several formulas, so it is possible to account for the expected regularity of the route, for example.

A special situation occurs when several routes offer a direct trip for an OD-pair. In some models the frequencies of the routes are added, but this is clearly a very optimistic approach.

We will use an approach similar to that of Lampkin and Salmons (11), but instead of calculating an average waiting time we also take account of the possibility of bunched arrivals of vehicles. In this approach it is assumed that a passenger uses the first vehicle that arrives at the stop. This assumption has often been criticized (e.g., Marguer and Ceder [23]), but this criticism is not supported by empirical evidence.

All kinds of routes can be used in the set of possible routes: one-way and two-way routes, routes with loops, express routes and so on. Moreover, it is possible to use different deterrence functions to account for the different behavior of separate groups of travelers.

EXAMPLES

The model which has been described is suitable for a personal computer (Olivetti M24, MS-DOS, 640KB) and can be used for a network consisting of 250 nodes with a maximum of 150 zones, and a maximum of 750 possible routes.

Fictional Network

As an example of the design process using the optimization model the network of Figure 3 is used. A set of possible routes $Y$ was generated manually, and consists of 75 routes. Two alternative demand patterns are considered: a midday period and an evening peak hour. For the midday period a network is designed which offers 1205 passengers a direct trip, given a fleet size of 10 vehicles. This solution cannot be improved using the exchange routine. For the evening peak hour two networks were developed: a complete new network and a network which uses the midday network as a base network. A comparison of these two networks shows that adding new constraints, such as the use of a base network, results in less
optimal solutions. On the other hand, using a base network has the advantage that the network remains recognizable for the passenger. The major point, however, is that the optimization model can be used for both strategies. Results of the tests are shown in Figures 4, 5, and 6, and in Table 3.

Existing Network

The optimization model is also tested with data from the city of Groningen in the Netherlands (170,000 inhabitants). This network consists of 182 nodes and 115 zones (Figure 7). Three different starting sets were used. The first set consists of the existing routes run by the local public transport company. The second set was constructed by splitting the existing routes at the city center and connecting them in all possible ways. The third set was derived with the use of basic design principles. The shortest routes from the city center and the railway station to 14 termini were determined and these route segments were combined in such a way that each route passes the railway station and the city center. The optimization model was used to determine the best possible network based on the possibilities contained in sets 2 and 3, given the demand pattern for the morning peak hour. The first set, the existing routes, is used for comparison.

The results of these tests can be found in Table 4. They clearly indicate that the optimization method is suited for realistic situations. Sets 2 and 3 yield similar results for the number of direct trips, an increase of 300 trips. Set 3, however, is clearly the best solution when the total number of trips is included. This is due to the basic design principles used to construct set 3, because of which a network can be developed offering good transfer facilities.

These analyses show that the set Y is an additional constraint; set 2, which is determined by the network, yields a lower result compared to set 3, which is developed with fewer constraints. It is up to the planner to decide which constraints will have to be included in the set of possible routes. It should be noted that during these tests the number of routes was not used as a constraint. The method itself stopped selecting routes at 9, respectively 7 routes.

CONCLUSIONS

We have presented a new optimization model for the design of public transport networks. Special features of the optimization method are as follows:

1. The simultaneous selection of routes, assignment of frequencies and the determination of the number of passengers.

2. The single optimization process which can be used for several design problems, ranging from short-term analyses to long-term decisions.

3. The application on a personal computer, which together with the interactive approach and the inclusion in a software

---

TABLE 3 CHARACTERISTICS OF THE RESULTS WITH THE TEST NETWORK

<table>
<thead>
<tr>
<th>Period of Day</th>
<th>Base Network</th>
<th>Direct Trips</th>
<th>No. of Vehicles</th>
<th>No. of Routes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mid-day</td>
<td>No</td>
<td>1,205</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Peak</td>
<td>Yes</td>
<td>1,529</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>Peak</td>
<td>No</td>
<td>1,543</td>
<td>15</td>
<td>6</td>
</tr>
</tbody>
</table>
package, enables the use of the model by the planner independently.

4. The possibility of taking into account all kinds of additional constraints, such as a base network, existing routes, a maximum number of routes, etc.

The method has proved to give good results with test networks and with actual data. Further research will be carried out to develop a model to generate a proper set of possible routes which can be used as input for the optimization model.

ACKNOWLEDGMENTS

This research is supported by the Netherlands Technology Foundation (STW). Moreover, the authors would like to thank E.M. de Ruig for her assistance in preparing this paper.

REFERENCES


11. W. Lampkin and P. D. Saulmann, The Design of Routes, Service Frequencies and Schedules for a Municipal Bus Undertaking: A


Publication of this paper sponsored by Committee on Public Transportation Planning and Development.
Optimization of transport networks.

Ben H. Immers and Peter H. Mijjer

20th Int. Scientific Conference on Transport Planning and Traffic Engineering, Budapest
OPTIMIZATION OF TRANSPORT NETWORKS

Ben H. Immers
Peter H. Mijjer


Delft University of Technology,
Faculty of Civil Engineering,
Department of Transportation,
Planning and Highway Engineering,
PO Box 5048,
2600 GA Delft,
The Netherlands.
OPTIMIZATION OF TRANSPORT NETWORKS

Ir. L.H. Immers,
Department of Transportation Planning and Highway Engineering,
Faculty of Civil Engineering,
Delft University of Technology

Ir. P.H. Mijjer*,
Department of Transportation Planning and Highway Engineering,
Faculty of Civil Engineering,
Delft University of Technology

*is employed with the Hague Consulting Group, The Hague,
from 1 February 1989

SUMMARY

In practice there are often many possibilities to decrease the
phenomenon of congestions in a transport network by increasing the
capacity of that transport network, (e.g. by building a new road,
adding additional lanes to an existing road, etc.). Because of the many
possibilities which have to be investigated, the process of decision
making is extremely complex. To help the planner in finding a feasible
solution, a model has been developed which selects the optimal
combination of possibilities on the basis of minimum traveltime costs
and investment costs. This optimization model has been applied to
determine the optimal structure of the main road network desired in
north-eastern part of the province Brabant.
1. INTRODUCTION

With the planning of traffic infrastructure in the Netherlands, the government tries to pursue an integrated, balanced and well-considered policy by means of a complex set of schemes, procedures and plans. The often time-consuming procedures are necessary to give all aspects connected with such a non-reversible intervention a sufficient chance. In planning the new road connections, the government uses the analytical policy method. This method has four stages:
- problem identification and definition;
- developing of alternatives;
- determining the effects;
- comparing the alternatives and selection.

Important parts of the procedure to go through are among other things project studies, including a forecasting calculation with the help of which the coming demand pattern is determined. The pros and cons of the connection(s) planned come up in great detail in the Environment Effect Report that has to be drawn up. When going through this procedure, all parties involved (authorities, advisory bodies, civilians) will be fully consulted.

If it comes to an overloading of the (motorcar) network on a regional scale, then there are, although it has been decided to tackle the issue only by the extending the capacity of the road network, various possibilities to solve the congestion problem. Besides the building of new road connections (new alignments of motorways), it is possible to extend existing road sections (and junctions) by one or more traffic lanes. The optimality of an alternative is, to a considerable extent, determined by the future use of the road network. In a network on a regional scale with all kinds of criss-cross relations, it is extremely difficult to indicate where the extensions of the road network can be realized best. Various alternatives (combinations of new alignments of motorways and road-widenings) are interesting. In practice this problem is solved by defining a limited number of (different) alternatives and by calculating their effects. An evaluation of the effects shows which alternative is to be preferred.

In a regional network with dozens of road sections and just as much potential new alignments, the number of alternative network structures that can be defined is very large (in case of N road sections respectively new alignments, this number is \(2^{*N} - 1\)). It is definitely
impossible to calculate all these alternatives and to compare them. On the other hand, of course, it is highly questionable whether the 3 to 4 alternatives which (according to the current procedure) are drawn up, contain the optimal solution (or the almost optimal solution) for the problem noticed.

In this paper a method is described which can help the planner with the developing and the selecting of alternatives in this complex subject-matter. The method is based on a model, which on the basis of marginal costs and benefits, selects optimal extensions of the network. The initial impetus to this model has been given by Steenbrink [18] within the scope of the Integral Traffic and Transport Studies [16]. The model has been improved on a number of points and thereupon it has been applied as a test to a regional traffic problem, i.e. an investigation into the main road structure desired in north-eastern part of the province Brabant in the Netherlands.

This paper is built up as follows.

In chapter 2 a number of methods, which can be used for network optimization (including the method developed by Steenbrink), will be treated roughly. In chapter 3 the mathematical basis of the model and the adaptations/extensions proposed, will be worked out further, after which, in chapter 4, the application of the model for the calculation of the main road structure in north-east Brabant is described. The paper is concluded with chapter 5 in which some conclusions and recommendations are incorporated. For a detailed report of this investigation, the reader is referred to Mijjer [13].

2. OPTIMIZATION OF NETWORKS.

Optimization of networks, both the extent of the network and the use (exploitation) of it, is a frequent subject of investigation in traffic and transport planning. The complexity of the problem, caused, among other things, by the large number of alternatives that has to be investigated, has led to the developing of all kinds of techniques to support the decision making process, respectively to optimize it. Here it concerns both

1) discrete choice models (in case of combinatorial problems), and
2) models as a basis for strategic choice processes in which all kinds of uncertainties play a part (decision models, simulation models).
For the problem described in this paper (optimal extension of a network) especially the first category of models is important. Almost all models which are developed to solve this type of problem, can be reduced to a basic form (see Magnanti and Wong [12]), i.e.: optimize an objective function \( F \), in which the effects (costs, benefits, etc.) of investment alternatives have been included. Important constraints in this optimization process are:
- the conservation of traffic flows (no traffic can get lost),
- non-negativism of traffic flows,
- capacity limitations of network links (facultative).

The above optimization problem can be solved in different ways. The extent of the problem (network), the shape of the objective function (linear, convex, concave), as well as the nature of the constraints determine what techniques can be applied best. The objective function is linear if the capacity does not influence the traveltime (travel-costs), in other words there is an unlimited capacity. In all other cases the objective function is not linear (e.g. convex or concave).

It is only with small networks (30 nodes, 100 links) that the optimal solution of the formulated problem can be calculated. The techniques to be applied then are, among other things, the minimal spanning tree (Hu [9]), Benders decomposition-technique [2] and branch and bound (see e.g. Boyce et al [4] and Dionne and Florian [6]).

With larger networks it is not possible anymore (within acceptable time limits) to calculate the optimal solution in using existing techniques. In that case one has to turn to a heuristic approach, which implies that no guarantee can be given that the optimal solution will be found. However, with the help of "worst case analysis" or "empirical analysis" it can be indicated roughly how far one is removed from the optimal solution. Techniques, based on the adding, removing and, respectively, the changing of links, are often applied for problems with a linear objective function (see e.g. Billheimer and Gray [3] and Dionne and Florian [6]). One may speak of a convex objective function, when the traveltimes (travelcosts) as a result of congestion, increase disproportionately to the intensity. For this type of problem the decomposition technique of Benders can be applied as well (Cotes and Laughton [5]).

Haubrich [8], and Barbier [1] as well, use for their railnetwork studies a technique by which, out of the maximal defined network, links are continually left out. Steenbrink [10, 17, 18] also developed a
decomposition technique which is very suitable for solving the network optimization problem in question. In his method, the budget constraint has also been replaced by admitting a non-linear construction costs component in the objective function. In section 2.1 the method will be described in great detail.

In networks where "economies of scale" play a part, (public transport, goods traffic), we often have to deal with a concave objective function. The result of this is that the solution found is mostly suboptimal. Various techniques have been developed (e.g. the incremental improvement technique by Yaged [20]) to overcome this problem. For the sake of a public transport optimization problem, Van Nes et al [14] have developed a technique based on the application of Lagrange multipliers.

2.1. The method of Steenbrink

With this technique, a given (maximal) structure of a (road) network is taken for granted. By the choice of the dimensions (capacities) and the traffic intensities of/on the various road sections, an optimal structure of a road system (network) is generated on the basis of minimization of social costs (traveltime costs, investment costs, accident costs, etc.). These basic assumptions lead to a systems optimum. In this scope, the given structure of the network must be interpreted as the possibilities stated, from which a structure can be chosen. On the basis of the above-mentioned principles, an optimal structure is chosen by the model. In order to be able to determine the optimum, the optimization problem is divided into a number of subproblems and one main problem (decomposition). The number of subproblems is equal to the number of road sections, of which the dimension (capacity) can be changed. The solution of the subproblems leads to a function which conveys the optimal relation between the capacity and the traffic intensity of a road section. When the capacity is defined as a continuous quantity, the subproblems can simply be solved with the help of differential calculation.

Substitution of the solution of the sub-problems in the objective function gives a derived objective function, which is only dependent now on the traffic intensity. Then the main problem can be formulated as follows: in what way are the traffic flows to be assigned to the network, so that the (derived) objective function is minimized.
Steenbrink [18] proves that this problem is equivalent to the determination of the equilibrium between supply and demand in a network on the basis of the first principle of Wardrop (all routes used between two zones have a resistance, smaller, or equal to the resistance of routes not used.). When the derived objective function is convex, it can be proved that the objective function is minimized if between two zones that route is chosen with the smallest value for the marginal objective function. The marginal objective function is the function that arises from the differentiation of the derived objective function to the intensity. So, for convex functions the objective function is minimized, if for each O-D pair the route is chosen with the smallest marginal costs. To solve the main problem, the same techniques can be used as for the determination of equilibrium flows in a transport network. For this, Steenbrink uses a capacity-restraint assignment algorithm. With this algorithm the transport relations (origin-destination matrix) are assigned step by step to the network. However, the assignment-criterion, which is usually the traveltime, has now been replaced by the marginal costs (i.e. the value of the marginal objective function).

3. THE MODEL.

In developing the optimization model, the method, developed by Steenbrink, is taken for granted. The minimization of social costs is also aimed at, with which, however, the costs are limited to the sum of traveltime costs and investment costs (however, this objective function is to be adapted at will). This implies that from a given (maximal possible) structure of a network on the basis of minimization of traveltime costs and investment costs, the optimal structure has to be determined. The objective function in a formula-form is:

\[
\min \sum_{X,C} F(X,C) = \min \sum_{X,C} k*X*Z(X,C) + I(C) \quad \text{with,} \quad \sum_{X,C} \quad (1)
\]

- \( k \) = conversion factor for traveltime to traveltime costs
- \( X \) = intensity on a link
- \( C \) = capacity of a link
- \( Z(X,C) \) = delay-function for a link
- \( I(C) \) = investment function for a link
- \( s \) = link or road section
The delay function is a function with which the traveltime on a road section can be determined, dependent on the intensity. For this delay function the BPR-function has been taken as a starting-point (see e.g. Hamerslag [7]). In this function the traveltime on a road section is dependent on the capacity and the traffic intensity of/on that road section:

\[ Z(X,C) = T_0[1 + a(X/C)^b] \quad \text{with}, \quad (2) \]

- \( T_0 \) = the traveltime on an unloaded road section (i.e. the initial traveltime)
- \( a \) and \( b \) = coefficient and exponent, for which, the values 1 and 4 respectively, are generally assumed.

To calculate the amount of the investment costs for a road section, it must be taken into account that the extent of the investment costs is dependent on the point of departure. For it is often much cheaper to rebuild or to extend an existing road than to build a new road. Besides, there will be a sharp rise of the investment costs in proportion to the increase of the capacity, if, at the same time, the costs of junction design (dimensions), are taken into account. To calculate the investment costs, the following function has been taken as a starting-point:

\[ I(C) = [p*(C)^x + q]^y*L \quad \text{with}, \quad (3) \]

- \( C \) = capacity of a road section
- \( L \) = length of a road section
- \( q \) = constant
- \( p \) and \( r \) = respectively coefficient and exponent

By the variation of the constant, the coefficient and the exponent, various investment functions can be described. The sub-problems, the determination of the optimal relation between the capacity and the traffic intensity of/on a road section, are being solved with differential calculation. To solve the main problem, not the capacity-restraint assignment used by Steenbrink, but an equilibrium assignment technique is used. With the capacity-restraint assignment-algorithm, the trips are assigned according to the principle of the repeatedly loading and
unloading of a network. These loading and unloading percentages are
determined in advance. However, the extent of these percentages affects
the results, so that wrongly chosen percentages can lead to wrong
results. So attaining a network equilibrium (or: finding the optimum)
is not guaranteed with this algorithm. With the equilibrium assignment
technique the relations are also assigned to the network step by step.
However, with each iteration an optimal step-size is determined within
the algorithm. So, the attainment of equilibrium is better guaranteed
with this method, be it that the condition for this is that the
function describing the assignment criterion is convex, continuous and
non-decreasing (Jansen [11], NVI [15]).

3. 1. The assignment-criterion.

The substitution of (2) and (3) to (1) and the differentiating of (1)
to the capacity C (with a value of 1 and 4 for respectively a and b)
leads to the solution of the sub-problems:

\[
C = \left( \frac{4*K*T_0*X}{r*p*L} \right)^{1/7}
\]  

(4)

The substitution of (4) to (1) gives the derived objective function,
which is only dependent on the intensity. The differentiating of this
derived objective function (with a value 3 for the exponent r in the
investment function) leads to the marginal objective function:

\[
F_{\text{marg}} = k*T_0 + \frac{15}{7} M*X
\]  

(5)

Function (5) describes the marginal costs for a road section of which
the capacity can be adapted. In principle, three types or groups of
road sections can be distinguished:

1. existing road sections not to be adapted. The marginal costs for
   these road sections consist of traveltime costs only.
2. existing road sections that can be adapted. The function that
describes the marginal costs consists of three stages:
   - if the optimal capacity necessary is lower than the capacity
     present, the marginal costs consist of traveltime costs only;
   - if the optimal capacity necessary is between the capacity
     present and the maximal capacity allowed (maximal road-widening
of the road, e.g. 2 x 4 traffic lanes), the marginal costs consist of traveltime costs and investment costs;
- if the optimal capacity necessary is greater than the maximal capacity allowed, the marginal costs consist of traveltime costs only;

3. (new road sections not present. The function that describes the marginal costs consists of two stages:
- if the optimal capacity necessary is below the maximal capacity allowed, the marginal costs consist of traveltime costs and investment costs;
- if, during the optimization process, the capacity necessary exceeds the maximal capacity allowed, the marginal costs consist of traveltime costs only.

Figure 3.1 shows the marginal object function of an existing two-lanes trunk road, which can be rebuilt into a motorway with 2 x 2 up to and including 2 x 4 traffic lanes. In this figure the various stages to calculate the value of the marginal objective functions can clearly be distinguished.

Figure 3.1 : Marginal objective function of a two-lanes trunk road.

To get more uniformity in this calculation, the marginal objective function can be approximated by an estimated function. The use of an estimated function for the marginal objective function offers a number of advantages:
to solve the main problem with the help of an equilibrium assignment algorithm, the marginal objective function must be non-decreasing and convex. From figure 3.1 it appears clearly that the original function is locally not convex but concave.

3.2 The estimated function.

The estimated function \( G(X) \) must meet the following requirements:
- the function must be non-decreasing, continuous and convex. However, the function needs not to be strictly convex \( (G''(x) < 0) \);
- the function must fit in as much as possible with the characteristics of the marginal objective function.

As an approximation of the marginal objective function by an estimated function, the following function form has been chosen:

\[
G(X) = \frac{TE^*(X-QMIN)/QMAX}{QMAX} + \frac{TEX^*(X-QMAX)/QMAX}{QMAX} + GRT + \frac{TMIN + GRT}{LN(GRT)}
\]

with:
- \( K \) = conversion factor for traveltime to traveltime costs
- \( T_0 \) = initial traveltime on a road section
- \( X \) = traffic intensity on a road section
- \( LN \) = natural logarithm
- \( GRT \) = base; with this base the keenness of the changes between the various sections can be influenced.
- \( QMIN \) = the variable (motorvehicles) which causes the change of a section with traveltime costs only, to a section with both traveltime costs and investment costs.
- \( QMAX \) = the variable (motorvehicles) which causes the change of the section with both traveltime costs and investment costs to a section with traveltime costs only (i.e. the point where the maximal capacity has been reached);
- \( TMIN \) = coefficients with which the gradient of the first, the second and the third section respectively, can be influenced.
- \( TE \) and \( TEX \)
The marginal objective function belonging to the various types of roads can be approximated by the estimation of GRT, QMIN, QMAX, TMIN, TE and TEX.

For the same type of road as in figure 3.1, the estimated function belonging to that type of road has been represented in figure 3.2.

MARGINAL OBJECTIVE FUNCTION

Figure 3.2: Estimated marginal objective function.

In the light of a test network, it has been investigated what the consequences are of using an estimated function for the marginal objective function. For that purpose calculations have been carried out with both the estimated function and the (real) marginal objective function. In the calculations the demand pattern (and consequently flows on the road sections) has been raised continually. From the results it appears that when the transport flows are becoming high to such an extent that on some road sections the capacity is exceeded, there will arise differences in the results. Here the tendency is apparent that, when an estimated function has been made use of, less spreading of the traffic on alternative road sections will occur in case of high loads.

In case of high loads on a road section, the marginal costs are determined by traveltime costs only (for the maximal capacity allowed has been attained). In using the original marginal objective function, these traveltime costs are based on the BPR-function. The character of this function is that, when the capacity of a road section is exceeded,
the resistance will strongly increase and will soon come to "infinite". If the resistances in a traffic network are calculated with the help of the BPR-function, then alternative routes will soon be looked for when the capacity is exceeded. So, the tendency to assign the traffic to several routes will strongly occur. In practice the increase of the resistance of a certain route as a result of congestion, will not strongly be experienced by the traffic participant. Especially in the rush-hours, queueing of cars is accepted to a certain extent. Beside this, it is only possible to turn to alternative routes when the point of time at which, and the place where the queue will occur, are known in advance. So, in practice the load on overloaded roads will often be higher than the load calculated in a network model on the basis of the BPR-function.

In using an estimated function for the marginal objective function, the (marginal) traveltime costs, when reaching the maximal capacity, are described by a straight up going line (see figure 3.2). Thus, when using the estimated function, the traffic will flow off less rapidly from "overloaded" routes and so the reality is better described. Consequently, the use of an estimation for the marginal objective function does not necessarily lead to less useful results.

4. THE APPLICATION OF THE MODEL.

The optimization model has been applied to the main road system of the north-eastern part of the province Brabant. The network to be optimized is built up on the basis of various alternatives for the main road structure, as developed by the Provincial Department of Public Works of North-Brabant [19]. The network to be optimized contains 615 nodes and about 1800 links. Figure 4.1 shows a rough view of the network to be optimized.

Some results of the optimization which help to assess the quality of the solution found are:
- the value of the objective function during the iteration process,
- the calculated capacities of the road sections,
- the road section loads,
- the total kilometrage and the total traveltime (hours).
The number of iterations in which the optimum has to be found, has been put at 7. From the decrease of the value of the objective function during the iteration process it becomes evident that the process converges to a minimum (see table 4.2.). The most important results of the optimization process are among other things:
- the existing 2 x 1 trunk road between 's-Hertogenbosch and Helmond (RW 266) has to be extended to 2 x 2 traffic lanes;
- RW 266 has to be extended from Helmond to the A67;
- RW 55 between 's-Hertogenbosch and Oss should have a capacity above the 2 x 2 traffic lanes;
- RW 264 can do with a capacity of 2 x 2 traffic lanes, with the

1 RW = national highway
exception of the Poot of Metz, where a higher capacity is advisable;
enlargement and extension of RW 266, and (locally) capacity-
adaptations of RW 55 and RW 264, lead up to the fact that for the
motorway connection between Nijmegen, Veghel and Eindhoven, as
indicated in the plans of the Province of North-Brabant, a road
connection of a lower order will be sufficient.

Table 4.2: Value of the objective function during the iteration
process

<table>
<thead>
<tr>
<th>iteration</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>4466.5</td>
<td>3220.7</td>
<td>3155.7</td>
<td>3134.9</td>
<td>3126.4</td>
<td>3116.8</td>
<td>3111.5</td>
</tr>
<tr>
<td>objective function</td>
<td>4466.5</td>
<td>3220.7</td>
<td>3155.7</td>
<td>3134.9</td>
<td>3126.4</td>
<td>3116.8</td>
<td>3111.5</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS/RECOMMENDATIONS

The conclusion can be drawn that this improved optimization model
generates very useful results as a basis for taking decisions regarding
network investments on a regional level. Priorities with respect to
extensions of infrastructure are being determined in a regional context
(where the connection between network design and demand pattern is
taken into account). Application of the model asks for an extensive
initial effort (network definition, calculation of a demand pattern,
calibration, etc.), but once this being done, the effects of possible
network concepts can be quickly calculated. In principle, the model in
its present form can be applied to regional network studies. Neverthe­
less, it is advisable to examine whether it is possible to refine the
model on some points, i.e.:

1) in drawing up the network to be optimized for north-east Brabant,
variations for the main road structure have been used, as developed
by the Provincial Department of Public Works of North-Brabant. The
advantage of this is that the extension possibilities and the way
of carrying them out are definite and fit in with prevailing
policies. However, when there are no existing plans and policy
views available, then, in order to be able to apply the optimiza­
tion model, a choice has to be made in what way the extension
possibilities of existing roads and new roads to be built could be carried out. The results of the optimization process are also dependent on the (chosen) ways to extend the link capacity so that, indirectly, the final results are influenced. If, for a road section, there are more ways possible of extending link capacity (and so more types of marginal objective functions), then, during the optimization process, discontinuities may occur, causing the fact that also the convexity of the objective function cannot be guaranteed anymore. This problem can be overcome by using one (estimated) type of the marginal objective function for each road section.

2) As yet, the total social costs are defined as the sum of traveltime costs and investment costs. If, for the building of a road between two points, two different proposed routes are possible, and the sum of the traveltime costs and the investment costs for both these proposed routes is alike, then the model will not generate a clear choice between these two possibilities. So, in principle, the model will advise the building of both routes. Extension of the social costs, with, among other things, safety (accident) costs and costs for the damage to nature and environment, will not only solve this problem, but also yield more useful results.

3) In all model calculations a fixed demand pattern has been assumed. However, when new roads are being built in a region, respectively the capacity of the network is extended, then the demand pattern (O-D matrix) will react upon this. The nature and the extent of the changes in the demand pattern are, to a considerable extent, dependent on the nature and the extent of the changes introduced into the network. So, in principle, there is some discrepancy between the optimal network calculated and the demand pattern belonging to it. This problem can be solved by introducing a feedback mechanism. First of all, the altered demand pattern (belonging to the optimal network) is to be calculated. Then, with this altered demand pattern, the optimization process is passed through a second time. This process can be repeated several times. Another possibility is the method as described in van Nes et al [14]. In this model the demand pattern is directly connected with the quality of the network. Investment choices are (gradually) being
made on the basis of an optimization criterion such as the maximization of the number of relations served or the maximization of the transport performance. When the budget available is exceeded, the process will be stopped.

REFERENCES

[1] Barbier, M.  
Le futur reseau de transports en region de Paris,  
Cahiers de l'Institute d'Amenagement et d'Urbanisme de la Region Parisienne 4 (1966)


Network Design with Fixed and Variable Cost Elements,  
Transportation science 7 (1973), pg 49-74


[5] Cotes, G. & Laughton, M.  
Large-Scale Mixed Integer Programming Benders-Type Heuristics,  
Institut de recherche de l'hydro-Quebec, Varennes, Quebec, Can. 1981

Exact and Approximate Algorithms for Optimal Network Design,  
Networks 9 (1979), pg 37-59

[7] Hammerslag, R.  

[9] Hu, T.C.

[10] Jansen, G.R.M.


Network design and transportation planning, Transportation science, vol. 18, no. 1, February 1984

[13] Hijjer, P.H.


[15] NVI
Verkeerstoedeling binnenvaart - theoretische achtergronden en practische uitwerking, januari 1977 (Traffic-assignment inland navigation - theoretical backgrounds and practical elaboration, January 1977)

[16] NEI
[17] Steenbrink, P.A.
Optimalisatie van de infrastructuur, Verkeerstechniek, 1971, no:7
(*Optimization of the infrastructure, Traffic engineering, 1971, number 7)

[18] Steenbrink, P.A.
Optimization of Transport Networks, London, John Wiley & Sons Ltd, 1974

[19] Stuurgroep Hoofdwegestructuur Noord-Oost Brabant
Studie Hoofdwegestructuur Noord-Oost Brabant, deel 1 t/m 5
Rijkswaterstaat, Directie Noord-Brabant, 's-Hertogenbosch, september 1984
(*Steering Committee Main Road Structure north-east Brabant Main Road Structure Study of north-east Brabant, volume 1 up to and including 5 Department of Public Works, Board of North-Brabant, 's-Hertogenbosch, September 1984)

[20] Yaged Jr., B.
Minimum Cost Routing for Dynamic Network Models, Networks, 3 (1973), pg 193-224

*Footnote: literature not to be had in the English language!
A traffic Assignment Model to Reduce Noise Annoyance in Urban Networks.

Jan Willem Houtman and Ben H. Immers

Transportation Research Record 1143
A Traffic Assignment Model To Reduce Noise Annoyance in Urban Networks

Jan Willem Houtman and Ben H. Immer

The possibilities of reducing traffic noise annoyance in urban networks without reducing the total amount of automobile traffic are investigated. The basic idea is to reduce noise levels by influencing drivers' route choice. Possibilities to influence this choice were investigated by modifying an equilibrium assignment algorithm.

In the past few decades noise annoyance has become an increasing problem, especially in the Netherlands with its dense population. Therefore a law has come into force, Wet Geluidhinder, that specifies permissible noise levels under various circumstances. This study concerns road traffic (1). Investigations have shown that the noise annoyance problem is most severe within towns, which restricts the possibilities for traffic engineers to solve the problem: noise barriers cannot be applied in the inner cities, for instance.

This study seeks to determine whether modified route choice might help to solve the noise problem. On roads with few houses, or with few houses close to the road, traffic flow should be increased in order to reduce the flow on roads where noise annoyance occurs or can be expected. Thus a comprehensive rather than an ad hoc approach is provided. This study involves unmodified fixed travel demand and an unmodified travel mode choice and is restricted to motor traffic.

Because the noise level is a logarithmic function of the flow, it is expected that the best solution will be created when most traffic is concentrated on a small number of main routes. However, it is also possible that a concentration of traffic on several routes combined with a diversion of oversaturated flows to low-density roads will be a feasible solution as well. The model to be discussed appears to support this hypothesis.

HOW TO MEASURE TRAFFIC NOISE

Noise can be quantified objectively in various ways:

- Noise level (in decibels)
- Loudness (in sones)
- Loudness level (in phons)
- Frequency characteristics (in Hertz)
- Interval time (in seconds)

Noise level is the most instructive, especially when it is A-weighted. This means that the measures are adapted to the way humans observe different frequencies.

In the Netherlands the equivalent noise level \( L_{eq} \) in A-weighted decibels is the most commonly applied measure. It smooths a fluctuating noise as follows:

\[
L_{eq} = 10 \log \left( \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} \frac{p^2(t)}{p_0^2} dt \right)
\]

where

- \( p_{eff} \) = effective sound pressure,
- \( p_0 \) = reference pressure,
- \( t_2 - t_1 \) = observed time interval, and
- \( L_{eq} \) = duration-sensitive noise level.

It should be pointed out, however, that such a noise level gives less representative values for nighttime traffic, when isolated cars pass by, than in steady flows.

Basically, traffic noise is a function of flow, traffic composition and speed, and the distance between the facade of the houses and the heart of the road (hereafter called the facade distance). In accordance with the current legal standards, trucks are subdivided into medium-heavy and heavy traffic according to certain criteria. Buses belong to the heavy-traffic category.

Unlike noise levels, noise annoyance is subjective. All kinds of personal characteristics influence a person's sense of annoyance. Nevertheless, investigations have shown a remarkable correspondence when the number of strongly annoyed persons is determined as a function of noise level.

An inquiry in Amsterdam resulted in the following relationship:

Percentage of strongly annoyed persons = 0.0038 \* \exp(0.1143 \* L_{eq})

This means that a doubling of the traffic flow resulting in a 3-dB(A) increase of the noise level causes a 40 percent increase in the number of strongly annoyed people.

HOW TO MODEL THE NOISE PROBLEM

Background of the Problem

The model assumes the existence of an origin-destination (O-D) table for motor traffic, which means that it concentrates on route choice and assignment. Within optimization problems, a distinction can be made between user-optimizing and system-optimizing theories. Wardrop (2-4) formulated these two principles as follows:
1. Travel costs on all the routes actually used are equal, and less than those that would be experienced by a single vehicle on any unused route, and
2. At equilibrium, the average journey cost is minimal.

These two optimums do not usually coincide.

Although the system equilibrium creates the most efficient traffic pattern for the community, it should be observed that it is an idealized target that will not be observed in practice without some form of enforcement.

Beckmann et al. (5) proved that this is equivalent to a convex minimization problem. Using the network concept proposed by Florian (6), this can be written as

\[
\text{min } Z = \sum \int_0^{f_a} c_a(x) \, dx \tag{3}
\]

where

\begin{align*}
Z &= \text{objective function}, \\
f_a &= \text{flow on link } a, \text{ and} \\
c_a &= \text{average travel cost function for link } a \text{ (in this paper a travel-time function)}. 
\end{align*}

Furthermore, the objective function Z should meet such conditions as positiveness, monotonous increase, and convexity to guarantee the existence of a solution and to warrant that it is unique and stable.

Regarding these conditions and given the background of the two different optimization approaches, one should realize that

1. Although a system optimization would seem the obvious way to reduce noise levels and noise annoyance within a town, such a solution is too unrealistic to be practicable. At best it gives an idea of the most favorable situation that can be reached.
2. The logarithmic noise-level and noise-annoyance functions do not meet the above conditions, which are necessary for applying the existing optimization techniques.

Extension of the Set of Constraints

The arguments mentioned in the previous section led to the choice of the following approach: a user minimization of the travel time with the addition of an extra condition. This condition, giving the maximum flow \(X\) on a road as a function of facade distance, traffic composition and speed, and the noise standards to be met, is not a constraint in the traditional sense, because it is incorporated in the link travel-time functions as follows:

\[
C^* = X \quad \text{for } X < C \\
= C \quad \text{for } X > C 
\tag{4}
\]

where

\begin{align*}
C &= \text{link capacity in traffic theory}, \\
C^* &= \text{capacity to be used in the link travel-time function, and} \\
X &= \text{calculated maximum flow}. 
\end{align*}

This results in an unmodified travel-time function for \(X > C\) and a compressed function for \(X < C\).

As a consequence, all travel times will increase if \(X < C\). This can easily be understood, because the objective is to reduce high noise levels. Therefore flows on these links must be reduced. In the chosen user optimization this can only be realized by increasing travel time on these links. As a consequence, alternative routes that originally were longer become attractive. This is shown in Figure 1a for the Bureau of Public Roads (BPR) function (7). Figure 1b shows how the resulting increase in travel time would be realized in practice: it is easier to increase the free-flow travel time than to reduce the link capacity. The possibilities for these practical realizations have been investigated in a follow-up study (8).

The reason for the conversion presented above is that an extra constraint like \(f_a > X_a\), for all \(a\), might make a feasible solution impossible. Furthermore, equilibrium according to Wardrop's first principle might become impossible as well.

The maximum flow of each link is called the environmental capacity (EC). These environmental capacities have a minimum value of 245 vehicles/hr because the current legal standards only cover roads with a minimum flow of 2,450 vehicles per day.

It should be noted that a real noise optimum will not be obtained. The result is one of a set of feasible solutions. In the results section, an analysis will be presented for a moderately large Dutch town in 1995, for which year a population of 85,000 inhabitants is projected. The network contains 264 nodes—among them 57 centroids—and 766 links.

THE ASSIGNMENT MODEL

Description

Although it is not impossible for the environmental capacity to be exceeded, this should not occur. Traffic on oversaturated links should be redistributed over the network. It is important, therefore, to choose a good link travel-time function. Both Davidson's hyperbolic function (7, 9) and the BPR polynomial have proved to be good delay functions. The Davidson function was expected to result in a more pronounced redistribution because of the asymptote at capacity.

BPR:

\[
t = l_0 \cdot \left[ 1 + 0.15 \cdot \left( \frac{f}{C} \right)^4 \right]
\tag{5}
\]

Davidson:

\[
t = l_0 \cdot \left[ \frac{1 - 0.6 \left( \frac{f}{C} \right)}{1 - \left( \frac{f}{C} \right)} \right] (f < C)
\tag{6}
\]

For computational reasons the hyperbolic function is extended with a linear part for saturation degrees of 0.99 and over.

To test this assumption, the assignment was performed with both functions.

The Dutch legal standards offer two calculation methods (10, 11). One method is very exact and detailed, which makes it unsuitable for the calculation of noise levels on such a large
scale as is intended here. Therefore the other method was used, from which the environmental capacity can be derived as follows:

$$X = \frac{d}{P_1 \cdot 10^{y(1)} + P_m \cdot 10^{y(m)} + P_s \cdot 10^{y(z)}} \cdot 10^{L_{\max}/10}$$ (7)

where

- \(y(1) = 5.12 + 0.021 \cdot v - \log v\),
- \(y(m) = 6.84 + 0.009 \cdot v - \log v\),
- \(y(z) = 7.62 + 0.003 \cdot v - \log v\),
- \(P_1 = \) percentage of automobiles,
- \(P_m = \) percentage of medium-heavy traffic,
- \(P_s = \) percentage of heavy traffic,
- \(v = \) speed (km/hr),
- \(d = \) facade distance (m), and
- \(L_{\max} = \) noise standard [dB(A)].

### Input

The network was divided into different link types, each with its specific traffic composition, speed, and theoretical capacity. For roads within the built-up area \(P_1\) ranges from 0.94 to 1.00, \(P_m\) ranges from 0 to 0.004, and \(P_s\) from 0 to 0.02. Only the facade distance may differ for each separate link.

However, in a network description, a road consists of two directed links. Therefore, a final point to be dealt with was the fact that assignments are performed for each directed link separately, whereas the noise level on a road is dependent on the flow on both links together. In the model a 50-50 division of the total flow on a two-way road is assumed, corresponding to a maximum flow per link of half of the environmental capacity. This is the most unfavorable and therefore the safest assumption; a full utilization of the road capacity will seldom occur under these conditions.

### Sensitivity Analysis

To determine the trade-offs between the variables in the model, a simple sensitivity analysis of the noise-level calculation was carried out. The results are as follows:

1. A 1-dB(A) elevation of the noise standard corresponds to a 26 percent increase in the environmental capacity; every 3-dB(A) elevation results in a doubling of \(EC\).
became very large and hard to handle when full capacity is approached.

Alternatives 2 and 4 have been compared on the basis of the following five criteria:

1. The number of links in each noise bracket $i$ ($i = 1, \ldots, 5$). For noise levels beyond the noise standard of $60$ dB(A), noise brackets of $3$ dB(A) have been defined, because this bracket size corresponds to a doubling of the traffic flow.

2. A noise index value $IN_i$ for each bracket. By adding the lengths of all road sections within one bracket, weighted noise index values were obtained.

3. A total noise index value $INTOT$ for the whole network. For each bracket the noise index value $IN_i$ is multiplied by an annoyance factor $c_i$. The products, added over all brackets, give the $INTOT$ value. The annoyance factors were derived by Wardrop (1). The noise standard $L_{max} = 60$ dB(A) corresponds to a factor equal to 1; an excess of $x$ dB(A) results in a factor equal to $1\cdot e^{x/11.43}$ (Table 2).

4. The total travel performance in vehicle kilometers.

5. The saturation degree. To get an impression of the eventual oversaturation throughout the network, the average saturation degree for the busiest directions of all roads together and the quietest directions of all roads together are determined. Alternative 2 will always show larger values than Alternative 4 because most saturation degrees depend on the environmental capacity (which usually is smaller than the theoretical capacity).

### Analysis

At first the results were rather poor and disappointing. The number of road sections where the noise standard was exceeded had increased (rather than decreased) by 50 percent and the total travel performance increased by 64 to 75 percent. A closer observation of the plots, however, showed several locations where the input specifications required modifications. One such location was a highway north of the town with an important traffic function. A low environmental capacity for such a road is not realistic. In these cases it is better not to impose any restrictions on the flow and to install effective noise-reducing facilities if necessary.

Furthermore, the network structure was improved. As a consequence the number of links (including dummy links connecting centroids to the network) increased to 859. The results of the model at this stage are shown in Table 3. Redistribution of the traffic (in order to reduce the noise levels) now resulted in a 21 percent increase in travel performance versus a 6 percent improvement in the total index value $INTOT$.

It is obvious that noise levels higher than $66$ dB(A) are strongly reduced by incorporating the noise standard. Moreover, more than half of the noise levels in the $60$- to $63$-dB(A) noise bracket are less than $61$ dB(A). The fact that the total number of noise levels exceeded has increased follows from the equilibrium principle: alternative routes are used and travel distances increase.

Figure 2 shows these phenomena quite clearly. It shows the total length of the road sections where a certain noise level is exceeded as a function of this noise level.

The findings corroborate the earlier hypothesis that traffic flows on noise-sensitive roads are indeed reduced and flows on undersaturated roads are increased. The flow reductions, however, are not always sufficient to guarantee a noise level of $60$ dB(A) or less in the new situation.

### CONCLUSIONS AND RECOMMENDATIONS

On the basis of the comparison of the alternatives, the following conclusions were reached:

1. Mechanical application of the model may lead to wrong and unfavorable results. In particular, roads or road sections that should serve or maintain an important traffic function must be selected beforehand. Such roads have to be treated as described in the previous section for the highway north of the town.

2. The results and improvements that can be obtained are dependent on the size and structure of the network. When few

### TABLE 2 ANNOYANCE FACTORS BY NOISE LEVEL

<table>
<thead>
<tr>
<th>No.</th>
<th>Range [dB(A)]</th>
<th>Avg. Exceeding</th>
<th>Annoyance Factor $c_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>60–63</td>
<td>1.5</td>
<td>1.19</td>
</tr>
<tr>
<td>2</td>
<td>63–66</td>
<td>4.5</td>
<td>1.68</td>
</tr>
<tr>
<td>3</td>
<td>66–69</td>
<td>7.5</td>
<td>2.36</td>
</tr>
<tr>
<td>4</td>
<td>69–72</td>
<td>10.5</td>
<td>3.32</td>
</tr>
<tr>
<td>5</td>
<td>&gt;72</td>
<td>13.5*</td>
<td>4.68</td>
</tr>
</tbody>
</table>

\*Estimated.

### Table 3 Comparison of Results

| Criterion                        | BPR Function
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>With Extra Constraint</td>
</tr>
<tr>
<td>No. of road sections</td>
<td>60 &lt; $L$ $\leq$ 63 ($i = 1$)</td>
</tr>
<tr>
<td></td>
<td>63 &lt; $L$ $\leq$ 66 ($i = 2$)</td>
</tr>
<tr>
<td></td>
<td>66 &lt; $L$ $\leq$ 69 ($i = 3$)</td>
</tr>
<tr>
<td></td>
<td>69 &lt; $L$ $\leq$ 72 ($i = 4$)</td>
</tr>
<tr>
<td></td>
<td>$L &gt; 72$ ($i = 5$)</td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Noise index value $IN_i$ (m)</td>
<td>$IN_1 = \sum_{i=1}^{5} 1 \cdot 1,000$</td>
</tr>
<tr>
<td></td>
<td>$IN_2 = \sum_{i=1}^{5} 1 \cdot 1,000$</td>
</tr>
<tr>
<td></td>
<td>$IN_3 = \sum_{i=1}^{5} 1 \cdot 1,000$</td>
</tr>
<tr>
<td></td>
<td>$IN_4 = \sum_{i=1}^{5} 1 \cdot 1,000$</td>
</tr>
<tr>
<td></td>
<td>$IN_5 = \sum_{i=1}^{5} 1 \cdot 1,000$</td>
</tr>
<tr>
<td>Total noise index value (m), $INTOT = \sum_{i=1}^{5} c_i \cdot IN_i$</td>
<td>41,140</td>
</tr>
<tr>
<td>Travel performance</td>
<td></td>
</tr>
<tr>
<td>(vehicle-km)</td>
<td></td>
</tr>
<tr>
<td>Average /TC (%)</td>
<td></td>
</tr>
<tr>
<td>Quiet</td>
<td></td>
</tr>
</tbody>
</table>
alternative routes are available, the possible improvements will be moderate, but when many surrounding roads are taken into account, redistribution of traffic is possible but will result in a strong increase in the total travel performance.

3. One should be aware that the final result is not an optimum, but only one of a series of possible solutions that meet the given constraints as closely as possible.

4. It is always dangerous to weight the results because it may resemble manipulation. However, in this case a mere unweighted comparison of the results could have resulted in a serious misinterpretation. Of the different ways to weight, the simplest and most transparent one was chosen.

The redistribution of traffic in this test case led to obvious improvements for the inhabitants of the city. Instances of exceeding the noise standard by 6 dB(A) or more decreased by 58 percent, whereas 28 of the 55 instances in the 60- to 63-dB(A) noise bracket are less than 61 dB(A), an amount that cannot even be perceived by the human ear.

However, it is questionable whether similar improvements can be expected in every arbitrary network. Furthermore, the differences between actual travel time and desirable travel time need to be moderate to make practical realization possible.

The results, however, can be further improved by perfecting some aspects of the model. The following recommendations are made:

- Introduction of a separate O-D table for heavy traffic, because of the considerable influence of this category.
- Introduction of the number of houses or the number of inhabitants per link in order to get a more exact weighting of the results.
- Performance of a true minimization of the noise annoyance to get an impression of the maximum improvement that can be obtained.

ACKNOWLEDGMENTS

This study was performed as an engineering graduate project in the Traffic Division of the Department of Civil Engineering at the Delft University of Technology. It was guided by R. Hamerslag of the Transportation Research Laboratory and P. H. L. Bovy and G. R. M. Jansen, all from the Delft University of Technology.

REFERENCES


Publication of this paper sponsored by Committee on Transportation-Related Noise and Vibration.

L.H. Immers and N.G.J. Oosterbaan

70th TRB annual meeting, Washington
A MODEL FOR THE CALCULATION OF ENVIRONMENT-FRIENDLY TRAFFIC FLOWS IN URBAN NETWORKS

L.H. Immers
N.G.J. Oosterbaan


Delft University of Technology
Faculty of Civil Engineering
Department of Transportation Planning and Highway Engineering
Stevinweg 1,
2628 CN Delft

Delft, August 1990
SUMMARY

A Model for the Calculation of Environment-friendly Traffic Flows in Urban Networks

The environmental impact of traffic flows in urban networks is an increasingly serious problem. In this study the possibilities are investigated of modifying route choice of traffic in urban networks in order to meet the standards for noise annoyance and the emission of pollutants. Therefore for each link of the network an Environmental Capacity is defined, being the minimum capacity of a link resulting from the selected environmental standards. The maximum flow on a link is defined by the minimum of the Environmental Capacity and the free flow capacity. By assigning the traffic to the network using the equilibrium assignment technique, a desire pattern of flows is obtained which meets (as far as possible) the environmental standards.

In this paper the results of an application of the model are presented using the network of the Dutch town of Ede/Bennekom.
1. INTRODUCTION

One of the drawbacks of increasing automobility is its impact on the environment. The Dutch government has been giving this wide-ranging environmental issue a great deal of consideration, which has resulted in a set of policy measures. Firstly, extensive regulations have been laid down as to the qualification and quantification of environmental annoyance, e.g. the "Noise Act". On the one hand these specify the limiting values, the standard values and the preferred values within which particular environmental effects should be kept, on the other hand standard methods are given by which these effects can be assessed and mapped, e.g. "Technical Aspects of Air Quality Regulations", and the "Environmental Impact Map". Secondly, research is being carried out on the kind of measures that will decrease undesirable and polluting side-effects of various activities (for instance the introduction of petrol with low lead levels and the establishing of routes for the transport of noxious materials). From the mid-eighties, finally, the Dutch government has adhered to the so-called 'stand-still' principle. This implies that the levels of the various kinds of environmental pollution must (at least) not be raised.

The main subject of this paper is the environmental impact of road traffic. To illustrate its scale: in 65% of noise annoyance cases road traffic is the (chief) source; road traffic is responsible for more than half of the emission of carbon monoxide (70%), lead (80%) and nitrogen oxides (50%).

The predicted drastic increase in motor traffic [1] will lead to a more extensive use of the existing infrastructure, with all its consequences for the (surrounding) environs. Concern about the environment will result in the phased introduction of stricter standards with respect to environmental effects, making it obligatory (for local authorities) to issue reports and take particular measures, i.e. reorganising town and transportation planning in existing situations or else adjusting it to new situations.

In this paper we shall present a model which is especially well-suited for determining an assignment (circulation) of traffic in an area in such a way that a favourable situation will arise with respect to meeting the various environmental standards. This approach merits considerable attention for the following reasons:

- the model indicates to what extent desired environmental standards can be met and on which road segments measures to this end will be required;
- by modifying traffic flow circulation integral measures are taken at the source; by decreasing traffic at an environment-critical site, polluting factors will become less serious.
- particularly in situations where the greatest problems may be expected to occur (urban agglomerations) alternative approaches are often scarce (one cannot put up noise barriers everywhere).
- by taking into account the future development of traffic, policymakers will be able to deal with the present situation as well as anticipate expected developments.

Houtman and Immers [2,3] have recently developed a model by means
of which an environment-friendly traffic assignment can be established on the basis of permissible noise levels, as specified in the "Noise Act". In this paper we will show how this model was extended with a component for air quality requirements. The report is composed as follows: In chapter 2 the most important features of the previously developed model will be briefly explained. In chapter 3 we will discuss the theoretical background and consider possibilities for integrating air quality requirements into the existing model. In chapter 4 the results of some sensitivity analyses and an application of the model will be presented and analyzed, after which chapter 5, finally, will contain some conclusions and recommendations.

For an extensive report of the research project we refer to Oosterbaan [4].

2. THE MODEL WITH RESPECT TO NOISE ANNOYANCE

Traffic in residential areas may have considerable consequences for human well-being. Many local authorities are faced with the question which measures to take so as to meet a desired environmental quality. In order to answer this question efficiently a model has been developed at Delft University of Technology which optimizes noise annoyance in connection with accessibility.

Central to the model is the Environmental Capacity (Cx) per road segment. This Environmental Capacity is defined as the capacity resulting from the standards for the emission of noise.

\[
C_{x_{i}} = \left[ \frac{d}{Y_{i} p_{i} + Y_{m} p_{m} + Y_{z} p_{z}} \right] \cdot 10^{\frac{L_{\text{max}}}{10}} \text{ (veh/hr)} \quad (1)
\]

where

- \( Y_{i} = 5.12 + 0.021 u - \log u \)
- \( Y_{m} = 6.84 + 0.009 u - \log u \)
- \( Y_{z} = 7.62 + 0.003 u - \log u \)
- \( p_{i} = \) percentage of automobiles
- \( p_{m} = \) percentage of medium-heavy traffic
- \( p_{z} = \) percentage of heavy traffic
- \( u = \) average speed km/hr
- \( d = \) distance facade to road axis in m
- \( L_{\text{max}} = \) maximum noise level in dB(A).
Per link (road segment) the minimum is determined of the Environmental Capacity calculated from (1) and the free flow capacity (Cap).

This minimum is taken as the capacity determining the travel time when assigning the traffic to the network. (see figures 1 and 2).

If the traffic is assigned to the network according to an equilibrium assignment technique (user-optimal travel time minimization with additional constraints as to the noise level) a desire pattern of traffic flows is obtained. This desire pattern shows an equilibrium situation as regards travel times and meets the legal standards of noise emission as closely as possible.

The suitability of the model was tested by applying it to a real site, i.e. Ede/Bennekom [2].

As mentioned in the introduction, the emission of noxious materials (gasses) is another environmental threat. In a similar way as for noise annoyance we investigated the extent to which air quality requirements may be incorporated into the Environmental Capacity (and so into the assignment process), so that with respect to this problem, too, the standards can be met as much as possible.

We shall go into this more extensively in chapter 3.

3. AIR QUALITY REQUIREMENTS

The Dutch government has both set the air quality standards and established the computation methodology of air pollution; see [5,6]. The model calculations will be limited to those for nitrogen oxide and carbon monoxide. Lead will be left aside since we may assume that the introduction of petrol with a reduced lead level will considerably decrease the emission of lead.

In the publication "Technical Aspects of Air Quality Regulations" [5,6], standards are set and specifications are given of the CAR-
model (Calculation of Air Pollution from Road Traffic), developed by TNO.

The contribution of traffic to air pollution as a result of carbon monoxide (CO) emission is calculated as follows:

\[ [\text{CO}] = N \times E_s \times \phi \times F_{\text{region}} \times F_b + \text{Ca.CO} \quad [\mu g/m^3] \]  
\[ (2) \]

in which:

\[ E_s = (1-p_v) \times E_p + p_v \times E_v \quad [\mu g/m.s] \]  
\[ (3) \]

\[ \text{Ca.CO} = R_{\text{CO}} + S_{\text{CO}} \times F_a \quad [\mu g/m^3] \]  
\[ (4) \]

where:

- \([\text{CO}]\) = resulting concentration of carbon monoxide \([\mu g/m^3]\)
- \(N\) = number of vehicles per 24 hours
- \(E_s\) = average emission \([\mu g/m.s]\)
- \(p\) = percentage of other traffic (non automobiles)
- \(E_p, E_v\) = emission parameter \([\mu g/m.s]\) for private cars and other traffic respectively, dependent on speed
- \(\phi\) = dilution factor \([s/m^2]\) dependent on the type of ground cover \(T\) and the distance \(s\) between kerb and road axis.
- \(F_{\text{region}}\) = meteo-correction factor with respect to regional differences in wind velocity.
- \(F_b\) = correction factor with respect to the type \(I_b\) of street in relation to the presence of trees
- \(\text{Ca.CO}\) = background concentration CO \([\mu g/m^3]\)
- \(F_a\) = distance between the road segment and the edge of the built-up area \((\text{km})\)
- \(R_{\text{CO}}, S_{\text{CO}}\) = constants for the measurement of the \(\text{Ca.CO}\) \([\mu g/m^3, \mu g/10^3 m^4]\).

In the direct exhaust emission of nitrogen oxides there is a large quantity of the rather harmless nitric oxide \(\text{NO}\) (there are no air quality requirements for \(\text{NO}\)) and a comparatively small quantity of the much more toxic nitrogen dioxide \((\text{NO}_2)\). However, by a chemical equilibrium reaction with ozone \((\text{O}_3)\) \(\text{NO}\) is fairly soon converted into \(\text{NO}_2\) according to

\[ \text{NO} + \text{O}_3 \rightarrow \text{NO}_2 + \text{O}_2 \]  
\[ (5) \]

On the basis of the emission of nitrogen oxides \((\text{NO}_x)\) the noxious concentration of \(\text{NO}_2\) can be calculated as follows:
\[
[\text{NO}_x] = P * N * E_s * \phi * F_{\text{region}} * F_b \quad [\mu g/m^3]
\]
\[
[\text{NO}_2] = \gamma *[NO_x] + \frac{\delta * \text{Ca.O}_3}{H + [NO_x]} + \text{Ca.NO}_2 \quad [\mu g/m^3]
\]

in which

\[
E_s = 1 - P_v * E_p + P_v * E_v \quad [\mu g/m.s]
\]
\[
\text{Ca.NO}_2 = R_{no2} + S_{no2} * F_s \quad [\mu g/m^3]
\]
\[
\text{Ca.O}_3 = R_{o3} + S_{o3} * F_s \quad [\mu g/m^3]
\]

where

\[
[\text{NO}_x] = \text{concentration nitrogen oxides due to direct emission} \quad [\mu g/m^3]
\]
\[
[\text{NO}_2] = \text{resulting concentration nitrogen dioxide} \quad [\mu g/m^3]
\]
\[
P = \text{coefficient in relation to the calculation of} \ [\text{NO}_x]
\]
\[
\gamma, \delta, H = \text{parameters dependent on the type of ground cover T}
\]
\[
\text{Ca.NO}_2 = \text{background concentration nitrogen dioxide} \quad [\mu g/m^3]
\]
\[
\text{Ca.O}_3 = \text{background concentration ozone} \quad [\mu g/m^3]
\]
\[
R_{no2}, S_{no2}, R_{o3}, S_{o3} = \text{constants for calculating background concentration NO}_2 \text{ and O}_3 \text{ respectively}
\]
\[
N, P_v, F_s = \text{the same variables as in the case of CO}
\]
\[
F_{\text{region}} = \text{the same correction factor as in the case of CO}
\]
\[
E_p, E_v, E_s = \text{emission parameters as in the case of CO but differing from these in numerical value}
\]
\[
\phi, F_b = \text{the same coefficients as in the case of CO, but in some cases for NO}_x \text{ different numerical values have to be applied.}
\]

Analogous to the way noise annoyance was dealt with, the maximum flow per link is established for which the various air polluting materials (gasses CO and NO\(_2\)) do not exceed the relevant air quality standard. From (2) the maximum day flow can be determined on the basis of the air quality standard for CO from:

\[
\text{MaxCO} = \frac{\text{Stand.CO} - \text{Ca.CO}}{E_s * \phi * F_{\text{region}} * F_b} \quad \text{(veh./day)}
\]
The environmental capacity \((C_{X_2})\) for the emission of CO can then be computed from:

\[
\frac{\text{MaxCO}}{10} = \text{for one-way roads} \tag{12}
\]

\[
\frac{\text{MaxCO}}{20} = \text{for two-way roads} \tag{13}
\]

It is assumed here that during peak hours the flow is 10% of the 24 hours flow.

For the maximum concentration CO (standard CO) a temporarily increased limiting value of 15000 \(\mu g/m^3\) was established. This value will be decreased in four stages down to 6000 \(\mu g/m^3\) in the year 2000.

The maximum day flow on the basis of the standard for NO\(_j\) can be calculated as follows:

\[
\text{MaxNO}_j = \frac{[\text{veh./day}]}{P \times E \times F \times F_{\text{region}} \times F_{\text{b}}} \tag{14}
\]

\([\text{NO}_x]_{\text{max}}\) can be calculated from (7) as follows:

\[
[\text{NO}_x]_{\text{max}} = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{15}
\]

where

\[
A = \gamma \tag{16}
\]

\[
B = \gamma \times H + \delta \times \text{Ca.O}_3 + \text{Ca.NO}_2 - \text{Stand.NO}_2 \tag{17}
\]

\[
C = H \times (\text{Ca.NO}_2 - \text{Stand.NO}_2) \tag{18}
\]

\[
\text{Ca.NO}_2 = R_{\text{NO}_2} + S_{\text{NO}_2} \times F_a \quad [\mu g/m^3] \tag{9}
\]

\[
\text{Ca.O}_3 = R_{\text{O}_3} + S_{\text{O}_3} \times F_a \quad [\mu g/m^3] \tag{10}
\]

For the maximum concentration NO\(_2\) (standard NO\(_j\)) a temporarily increased limiting value of 160 \(\mu g/m^3\) was established. This value will be decreased in two stages down to 135 \(\mu g/m^3\) in the year 2000.

The environmental capacity for the emission of nitrogen oxides can be calculated by dividing the maximum day flow by 10 (one-way road) and 20 (two-way road) respectively.

\[
\frac{\text{MaxNO}_j}{10} = \text{for one-way roads} \tag{19}
\]

\[
\frac{\text{MaxNO}_j}{20} = \text{for two-way roads} \tag{20}
\]
\[ Cx_3 = \frac{\text{MaxNO}_2}{20} \quad \text{(for two-way roads)} \quad (20) \]

- adjustment of the model

Incorporation of the air pollution standards in the model means we have to introduce or else compute a great number of additional data. Per road segment coefficients will have to be introduced as to:
- type of trees \( T_b \)
- type of ground cover \( T \)
- the weighting coefficient \( J \) with respect to the number of dwellings or residents per road segment.

Furthermore, for every road segment the distance up to the edge of the built-up area is to be measured. A number of additional nodes located at the edge of the built-up area were therefore defined. The distance of a road segment up to the edge of the built-up area can then be computed by means of the \( x \) and \( y \) coordinates from (see also figures 3 and 4):

\[ D_L = \text{Max} \left( D_{pl_1}, D_{pl_2} \right) \quad (21) \]

where

\[ D_{pl_1} = \text{Min} \left( d_n \right) \quad (22) \]

\[ D_L = \text{Max} \left( D_{pl_1}, D_{pl_2} \right) \]

Figure 3: Distance from a link to the edge of the built-up area

\[ D_{pl_1} = \text{Min} \left( d_n \right) \]

Figure 4: Distance from a node to the edge of the built-up area

As with the noise aspect, it is possible to omit capacity computations with relation to the air quality for particular road segments (links), if, for instance, these have a clear traffic flow function.

Per link the capacity may now be calculated, i.e. the minimum value of the free flow capacity (Cap) and the environmental...
capacities for, respectively, noise, air quality CO and air quality NO₂.

Per environmental aspect an indication of the capacity reduction of the entire network can be obtained, on the basis of the limits set by the environmental standards, from:

\[
\text{RESCAP}(I) = \left\{ \frac{\Sigma \text{ENVCAP}_j(I) \cdot \text{Cap}_j}{\text{NLINK}} \right\} \times 100 \%
\]

where

- \( \text{RESCAP}(I) \) = the remaining network capacity with respect to environmental aspect I in %
- \( \text{Cap}_j \) = the free flow capacity of link j
- \( \text{ENVCAP}_j(I) \) = the environmental capacity of link j with respect to environmental aspect I
- \( \text{NLINK} \) = the number of links.

For the total of environmental aspects considered the remaining network capacity can be calculated as follows:

\[
\text{RESCAP}(O) = \left\{ \frac{\Sigma \text{INFO2}_j \cdot \text{Cap}_j}{\text{NLINK}} \right\} \times 100 \%
\]

where

- \( \text{RESCAP}(O) \) = the remaining network capacity with respect to all the environmental aspects in %
- \( \text{INFO2}_j \) = the lowest environmental capacity of link j

4. ANALYSES

4.1. Sensitivity analysis

By means of a sensitivity analysis the effects of five parameters on the environmental capacity were investigated for CO and NO₂ respectively.

The following parameters were considered:
1. distance to edge of built-up area \( F_s \),
2. average speed \( u \),
3. percentage other (non automobiles) vehicles \( P_v \),
4. ground cover \( T \),
5. type of trees \( I_t \).
As to the values of the other parameters three situations were distinguished in the calculations:

with regard to air quality
- the other parameters have favourable values,
- the other parameters have average values,
- the other parameters have unfavourable values.

Some of the results of these sensitivity analyses are to be found in figures 5-8. They show the effects of modification of average speed and the percentage of other traffic.

Figures 5 and 6: relation between maximum day flow (N) (from the point of view of the standards for CO and NO₂ respectively) and the average speed (u).

Standard CO = 15000 µg/m³, standard NO₂ = 160 µg/m³, F_{region} = 1.05

a : F_a = 0 km, p_v = 0%, T = 1, I_b = 1
b : F_a = 5 km, p_v = 10%, T = 3A, I_b = 2
b' : F_a = 5 km, p_v = 10%, T = 4, I_b = 2
c : F_a = 10 km, p_v = 30%, T = 3B, I_b = 3
c' : F_a = 10 km, p_v = 30%, T = 4, I_b = 3
Figures 7 and 8: Relation between maximum day flow (N) (from the point of view of the standard for CO and NO₂ respectively) and the percentage of other traffic (pᵣ).

Standard CO = 15000 µg/m³, standard NO₂ = 160 µg/m³, F_region = 1.05

a : F_a = 0 km, u = u_a', T = 1, I_b = 1
b : F_a = 5 km, u = u_b', T = 3A, I_b = 2
b' : F_a = 5 km, u = u_b', T = 4, I_b = 2
c : F_a = 10 km, u = u_d', T = 3B, I_b = 3
c' : F_a = 10 km, u = u_d', T = 4, I_b = 3

The figures show that for CO the maximum flow (env. cap.) increases with average speed and with the percentage of other traffic; as for NO₂, on the other hand, the maximum flow decreases in both cases. In all cases the limiting value for NO₂ determines the outcome.
Other important results of the sensitivity analysis are:
- the sensitivity of the maximum flow for a particular parameter increases with the extent to which the values of the other parameters are favourable,
- the maximum flow decreases with increasing ground cover,
- the maximum flow decreases with increasing distance to the edge of the built-up area,
- the maximum flow decreases with an increasing number of trees along the road (this is a short-term effect. In the long term the presence of trees will have a positive effect on air quality).

4.2 Model calculations

In order to test the model the network and the OD-table of Ede/Benekom for the year 1995 were used, as drawn up by IWIS/TNO [7].

The network contains 264 nodes, comprising 57 centroids, and 859 one-way links (see fig. 9).

Figure 9: Network of Ede/Benekom, [7].
The categorization in types of road was made according to, among other things, road cross section, average flow and average speed. Per category of link a fixed free flow speed, free flow capacity and percentage of medium-heavy and heavy traffic was determined. The distance facade to road axis, ground cover, type of trees and distance road segment to the edge of the built-up area, vary for each link. The variables F-region (meteo-correction coefficient wind velocity), Standard CO and Standard NO₂ have the same value for the whole network.

For the area four calculations were carried out:
I  A traditional equilibrium assignment, not taking into account environmental standards.
II  An equilibrium assignment taking into account the noise standard (60 dB(A)).
IIIA An equilibrium assignment taking into account the noise standard and the temporarily increased limiting value with respect to the air quality standards (CO : 15000 μg/m³; NO₂ : 160 μg/m³).
IIIB An equilibrium assignment taking into account the noise standard and the strict requirements regarding the air quality standards (CO : 3000 μg/m³; NO₂ : 80 μg/m³).

An indication of the influence of the various quality standards on the traffic pattern can be obtained from table 1 where the remaining capacities (RESCAP(I)) for differing scenarios have been calculated.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>I</th>
<th>II</th>
<th>IIIA</th>
<th>IIIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise</td>
<td>100</td>
<td>55.8</td>
<td>55.8</td>
<td>55.8</td>
</tr>
<tr>
<td>CO</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>95.1</td>
</tr>
<tr>
<td>NO₂</td>
<td>100</td>
<td>100</td>
<td>99.9</td>
<td>55.5</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>55.8</td>
<td>55.8</td>
<td>50.8</td>
</tr>
</tbody>
</table>

It turns out that in the situation Ede/Bennekom the influence of the air quality standards does not become noticeable until the strict requirements are applied (scenario IIIB).
Table 2 shows the traffic flow effects and the environmental effects resulting from the various assignments.

Table 2: Summary of results

<table>
<thead>
<tr>
<th></th>
<th>L norm = 60 dB(A)</th>
<th>L norm = 60 dB(A)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Standard CO = 15000 µg/m³</td>
<td>Standard CO = 3000 µg/m³</td>
</tr>
<tr>
<td></td>
<td>Standard NO₂ = 160 µg/m³</td>
<td>Standard NO₂ = 80 µg/m³</td>
</tr>
</tbody>
</table>

| Noise | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 | Category 1 |
|-------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|------------|
| CO    | Category 1 | 30         | 24         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         | 24         | 30         |
|       | Category 1 | 3          | 0          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          | 3          | 0          |
|       | Category 1 | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          | 0          |
| Index 1 | 0          | 0          | 0          | 4850       | 0          | 0          | 113059     | 0          | 0          | 4850       | 0          | 0          | 113059     | 0          | 0          | 4850       | 0          | 0          | 113059     | 0          | 0          | 4850       | 0          | 0          | 113059     |
| Index 2 | 0          | 0          | 0          | 1826488    | 0          | 0          | 626967     | 0          | 0          | 1826488    | 0          | 0          | 626967     | 0          | 0          | 1826488    | 0          | 0          | 1826488    | 0          | 0          | 1826488    |
| Index 3 | 0          | 0          | 0          | 42442      | 0          | 0          | 71318      | 0          | 0          | 42442      | 0          | 0          | 71318      | 0          | 0          | 42442      | 0          | 0          | 42442      | 0          | 0          | 42442      |
| NO₂    | Category 1 | 0          | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         | 0          | 0          | 40         |
|       | Category 1 | 0          | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         | 0          | 0          | 12         |
|       | Category 1 | 0          | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         | 0          | 0          | 18         |
|       | Category 1 | 0          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          | 0          | 0          | 7          |
|       | Category 1 | 0          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          | 0          | 0          | 1          |
| Index 1 | 0          | 0          | 0          | 21310      | 0          | 0          | 19050      | 0          | 0          | 21310      | 0          | 0          | 19050      | 0          | 0          | 21310      | 0          | 0          | 21310      | 0          | 0          | 21310      | 0          | 0          | 21310      |
| Index 2 | 0          | 0          | 0          | 290734     | 0          | 0          | 167672     | 0          | 0          | 290734     | 0          | 0          | 167672     | 0          | 0          | 290734     | 0          | 0          | 290734     | 0          | 0          | 290734     |
| Index 3 | 0          | 0          | 0          | 49130      | 0          | 0          | 23326      | 0          | 0          | 49130      | 0          | 0          | 23326      | 0          | 0          | 49130      | 0          | 0          | 49130      | 0          | 0          | 49130      |

Kilometrage (km) | 89750 | 108717 | 108717 | 89750 | 108717 | 113503 |
Satur. degree (%) | 20.5 | 44.8 | 44.8 | 20.3 | 44.8 | 52.5 |

1) total length, 2) idem with noise weighted 3) summed number of dwellings with noise weighted.
For the assessment of the environmental effects class boundaries were drawn up as shown in table 3.

Table 3: Class boundaries per environmental aspect

<table>
<thead>
<tr>
<th>Class</th>
<th>Noise [dB(A)]</th>
<th>CO [μg/m³]</th>
<th>NO₂ [μg/m³]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>L_{norm} -&gt; L_{norm}+3 Stand.CO -&gt; Stand.CO+1000 Stand.NO₂ -&gt; Stand.NO₂+10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>L_{norm}+3 -&gt; L_{norm}+6 Stand.CO+1000 -&gt; Stand.CO+2000 Stand.NO₂+10 -&gt; Stand.NO₂+20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>L_{norm}+6 -&gt; L_{norm}+9 Stand.CO+2000 -&gt; Stand.CO+3000 Stand.NO₂+20 -&gt; Stand.NO₂+30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>L_{norm}+9 -&gt; L_{norm}+12 Stand.CO+3000 -&gt; Stand.CO+4000 Stand.NO₂+30 -&gt; Stand.NO₂+40</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>&gt;L_{norm}+12 Stand.CO+4000 Stand.NO₂+40</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The results of the calculations indicate the following:

- Air quality standards are of little or no significance when applying the (increased) limiting values. Scenario IIIA shows a picture that is identical to scenario II. Introduction of the noise standard does have a distinct influence on the traffic flow.

- When applying the 'strict' requirements with respect to air quality standards it appears that a qualitative improvement and a quantitative deterioration occur (compare II to IIIB). The total number of road segments and the summed length of the road segments where one or more air quality standards are exceeded, increase, but the magnitude of these excesses decreases. These occurrences are caused by a redistribution of the traffic from road segments (far) above the standard, to road segments with a flow smaller than the environmental capacity. Remarkably, this redistribution manifests itself more strongly in the case of the less critical CO.

- When only the noise standard is applied the air quality is also drastically improved (compare scenarios I and II, strict standards). The added incorporation of the air quality standards has a varying effect on the assessments of the acoustic quality (somewhat more annoyance but fewer road segments where standards are strongly exceeded). These findings seem to indicate the need for compromises between the various quality requirements.

- Incorporation of the strict requirements with respect to air quality standards leads to an increase in the total vehicle mileage. In order to prevent environmental standards being exceeded, a detour is made. The average number of vehicles (related to the free flow capacity) also increases. This increase is partly caused by the increase in mileage, partly by the assignment of the traffic to alternative routes with a low free flow capacity.
5. CONCLUSIONS AND RECOMMENDATIONS

On the basis of the tests performed we may conclude the following (we refer to [2] for the conclusions of previous analyses):

- Assuming (temporarily) increased limiting values, air quality standards do not play any part in assessing environmental annoyance in the Ede/Bennekom network. (In a 'compact' town this may be entirely different).

- Application of strict standards clearly shows that incorporation of environmental annoyance with respect to air pollution influences the results of the assignment process.

- The effects of the various environmental aspects on the traffic assignment process may be graded as follows: 1) noise, 2) air quality NO₂, 3) air quality CO.

- If different environmental standards are incorporated into the assignment process two effects may be observed:
  Firstly: In many cases adjusting the assignment process for one environmental aspect leads to an improvement of other environmental aspects which were not incorporated into the assignment process.
  Secondly: by adding an environmental aspect the improvements in the traffic circulation calculated for another environmental aspect may be partly nullified.

Besides the recommendations with respect to noise annoyance as stated in [2], it may, for a greater understanding of the problems of air quality and the applicability of the relevant software, be interesting to investigate the following points:

- Application of the model in a situation more problematic as regards air pollution, e.g. a medium-size, compact town (e.g. Delft).

- Research into the magnitude of the annoyance as a function of the extent to which the air quality standards are exceeded. (In the present model the excess itself is taken as a measure of the annoyance experienced).

- Incorporation of the (increasing) travel time in calculations of environmental effects.
REFERENCES


3. Houtman, Jan Willem "Research into the effects of speed reducing measures on travel time in networks" Delft University of Technology, Department of Transportation Planning and Highway Engineering, Research Report, February 1985


The Planning of Motorway Alignments through Computer-Aided Design.

Ben Immers, Jan-Jaap Schoone, Henk Voogd

Conference of the Italian Society for Operations Research, Lavoro
THE PLANNING OF MOTORWAY ALIGNMENTS THROUGH
COMPUTER-AIDED DESIGN

Ben Immers
Jan-Jaap schoone
Henk Voogd

Paper presented at the conference of the Italian Society for

Delft University of Technology,
Faculty of Civil Engineering,
Department of Transportation Planning and Highway Engineering,
P.O. Box 5048,
2600 GA Delft, The Netherlands.
SAMENVATTING

Dit paper is gewijd aan een toepassing van de multikriteria benadering bij het ontwerpen van een tracé voor een nieuwe autoweg. Enkele recente ontwikkeling op het gebied van de kwalitatieve multikriteria analyse worden behandeld in relatie tot het probleem van het systematisch lokaliseren van een nieuwe verkeersroute. Het paper heeft de volgende structuur: in paragraaf twee wordt de globale struktuur van het model behandeld. Paragraaf drie is gewijd aan een uitwerking van enkele belangrijke modelcomponenten, namelijk het gebruik van boogstralen, de behandeling van "unieke" waarden en de behandeling van kwalitatieve kriteriumgewichten. De werking van het model wordt geïllustreerd in paragraaf vier aan de hand van een praktische toepassing op een gebied in Midden Limburg, waar rijksweg 73 is geprogrammeerd.

SUMMARY

This paper is devoted to an application of the multicriteria approach to the design of a new motorway alignment. Some recent developments in the field of qualitative multicriteria analysis will be treated in relation to the problem of systematic route location. The paper has the following structure: in section two the general principle of the model is outlined. Section three is devoted to an elaboration of some important components, viz. the use of the curve radius, the treatment of "peak" values and the treatment of qualitative criterion weights. The model is illustrated in section four by means of an empirical application to a motorway location problem in the Dutch province of Limburg.
1. Introduction

Faced as we are with the present economic circumstances, more and more attention and thought must be given to the decisions to be made in public planning. Especially in a period of a seemingly increasing shortage of financial resources, more emphasis should be given to the effectiveness of policy proposals. Policy decisions which involve many important socio-economic and environmental consequences, like the location of new infrastructures, need careful consideration (see also Klaassen et al., 1981). For example, an unbalanced location of a new motorway alignment may not only lead to undesired physical and economic side-effects, but also to an unnecessary deterioration of environmental qualities and/or to an inefficient land use.

Evidently, many factors should be taken into account when new transportation infrastructure is planned. In practice, it is not unusual that in defining the relevant factors the operating agency's angle of incidence dominates. This often implies and emphasis on system costs, which are mostly estimated in terms of capital costs (i.e. acquisition and clearance costs, guideway or track costs, parking space costs, construction costs), operating costs and return on investment. Besides, some attention may be paid to travel costs, safety costs and costs of noise pollution.

However, such a cost-oriented approach of planning invited rather harsh criticism in practice. There are often considerable difficulties in physically tracing out and measuring in money units all of the consequences of a project. Many politically important impacts, viz. landscape-ecological and environmental effects, can hardly be taken into explicit consideration in a cost-benefit approach. Due to these important limitations, the attention is more and more focussed on approaches which enable a more comprehensive judgement. Especially the multicriteria approach can be mentioned in the respect (see for some reviews, inter alia, Nijkamp, 1979; Rietveld, 1980; and Voogd, 1983a).

This paper will be devoted to an application of the multicriteria approach to the design of a new road. Some recent developments in the field of multicriteria evaluation is treated in relation to the problem of systematic route location through computer-aided design (see also Schoone, 1982; Immers and Schoone, 1982; Immers et al., 1983).
Already in the early sixties some important work in the field of road location has been done by MIT in the USA, where Alexander and Manheim (1962) developed a composite-mapping procedure for highway planning. An approach which stresses the cost aspects has been presented a decade later by Athabassoulis and Galogero (1973). Much research has also been done by Turner (1978), who created a computer-based method which shows the desirability of a highway location based on engineering feasibility, improving social and economic conditions and environmental impact. Related approaches have been presented by, inter alia, Van Staalduine and Voogd (1975), Arnold et al. (1977) and Koeppel (1979).

In this paper an approach will be elaborated which incorporates some interesting new features, such as the possibility to cope with qualitative weighting schemes and the ability to deal with "peak" criterion values. Practice shows that especially environmental peak values are extremely important issues in transportation policy making (see also Voogd, 1983b).

The structure of the paper is as follows: in section two, the general principle of the route location model is outlined. Section three is devoted to an elaboration of some important components of the model, viz. the relationship between zoning system and the design procedure, the treatment of "peak" values and the treatment of qualitative criterion weights. Next, the model will be illustrated in section four through an empirical application to a road location problem in the Dutch province of Limburg. The paper is closed by means of some summarizing conclusions.
2. The general model structure

In order to describe the spatial characteristics of the study region properly, a zoning system must be postulated. For reasons of simplicity a grid zoning system will be used, i.e. the region under consideration is subdivided into \( g \) square grids. The purpose of the model is to find the best possible route from one particular grid in the zoning system (the origin grid) to another a priori designated grid (the destination grid), given a number of explicitly defined criteria and criterion weights.

Each route can be seen as composed of a number of related links \( z_n \); in other words, a route \( L \) can be defined as a set of links, or:

\[
(2.1) \quad L = \{ z_n \mid z_n \text{ is connected to } z_{n-1} \text{ for each } n > 1 \};
\]

where \( z_1 \) starts in the origin grid. Obviously, a large number of sets \( L \) can be distinguished for a given location problem. The model aims at finding the optimal set given the appraisal score \( a_z \) attached to each link (to avoid a cumbersome notation the link index \( n \) will further be dropped). This can be formally written as:

\[
(2.2) \quad \min \phi = \sum_{z \in L} a_z .
\]

It is assumed in (2.2) that a lower appraisal score implies that the particular link is more suitable to be included in the route.

Each link will pass two or more grids. Hence, for each link a set \( q_z \) can be distinguished which includes the grids \( g \) which are directly affected by that particular link. The appraisal score is now defined as a function of the length of the link (to be denoted as \( s_z \) ) and the multicriteria valuation \( v_g \) given to grid \( g \), or:

\[
(2.3) \quad a_z = f(s_z, v_g) \quad g \in q_z
\]

In order to arrive at a multicriteria valuation the weighted summation rule is used:
\[ V_g = \sum_j w_j (e_{jg} + c_{jg}) \]

where \( w_j \) is the weight attached to criterion \( j \) (\( j=1,\ldots,J \)); \( e_{jg} \) is the score attached to grid \( g \) with respect to criterion \( j \) and \( c_{jg} \) is variable which receives a high value if an important physical barrier has to be taken (e.g. a river). Thus,

\[
\begin{align*}
    c_{jg} &= 0 \text{ if no barrier is passed} \\
    &= \text{real number} > 0 \text{ if barrier is passed}
\end{align*}
\]

For some criteria (e.g. environmental or hydrological effects) it is necessary to consider not only the criterion score of each individual grid, but to include also possible spatial spillover effects. In the model this is done by relating a grid score to the surrounding area defined by 9 x 9 grids. The following formula is used to adapt the criterion score \( e_{jg} \) to the quality of the surrounding grids:

\[ e_{jg}^* = \sum_{g'} (e_{g'} - d_{gg'} \cdot e_{jg}) \]

where \( \gamma \) is a negative value attached to the power \( \gamma \) and \( d_{gg'} \) is the distance between grid \( g \) and other grids \( g' \) included in the 9 x 9 zoning subsystem.

Model (2.2) is solved as follows. First, each grid will be appraised by means of the various design criteria which can be distinguished (e.g. cost-criteria, environmental criteria, physical planning criteria). Evidently, the criterion scores \( e_{jg} \) must be standardized in order to be comparable (see for an overview and discussion of various standardisation methods: Voogd, 1983a) In addition, particular barriers in the study area must be made explicit by means of so called "cost lines". In case a route crosses a cost line, the particular criterion score must be increased with an appropriate value (see (2.5)). Next, an appraisal score can be determined for each link through formula (2.3). In the next section it will be shown that this may be done in several ways. By applying a conventional shortest route algorithm, an optimal route can now be found as the set of links in which the sum of the appraisal scores is minimal. In this case the three-builder algorithm of Dijkstra (1959) is used because of its flexibility and efficiency.
3. Some important elaborations

The model outlined in the preceding section can be elaborated and refined in several ways. In this section three different issues are dealt with in more detail, viz. the curve radius, peak values and qualitative weights.

3.1. Curve radius

To carry out a detailed design of a modern road it is not sufficient merely to know the width and cross-section it must have. Knowing a speed likely to be used on a road, it is obvious that, in order that vehicles shall not leave the road, nor need to slow down, the radius of the curves on the road should be tied to the speed. In case of a highway the curve radius should at least be 2000 meter, which implies that when two links are allowed to make an angle of 45 degrees, each grid zone should have a size of 1.5 x 1.5 km or more. However, if the change of direction of a route is about 22.5 degree, the minimum size of a grid will be 400 x 400 meter.

Given the curve radius, the search directions will be limited to three, as is shown in Figure 3-1. The straight lines represent feasible search directions, whereas the dotted lines are search directions which are not feasible because of a too large angle. The left part of Figure 3-1 concerns the situation in case the length of a grid is 1.5 km or more; the right part holds for grids which have a smaller size. Notice that the direction of the "arrival" link is of influence on the-search direction.

![Figure 3-1. Feasible search directions](image-url)
It is now possible to illustrate the various ways in which the appraisal score $a$ for each link can be made explicit (see formula (2.3)). Five different situations can be distinguished from Figure 3-1, viz. direction A and B for the left part and direction C, D and E for the right part. If the length of a link is expressed in the unit "grid length", the five elaborations of formula (2.3) become:

(3.1) \[ a_A = 1.0000 \cdot \frac{v_3 + v_4}{2} \]

(3.2) \[ a_B = 1.4142 \cdot \frac{v_3 + v_2}{2} \]

(3.3) \[ a_C = 2.0000 \cdot \frac{v_7 + 2v_8 + v_9}{4} \]

(3.4) \[ a_D = 2.2361 \cdot \frac{v_7 + v_8 + v_5 + v_6}{4} \]

(3.5) \[ a_E = 2.8284 \cdot \frac{v_7 + 2v_5 + v_3}{4} \]

Formulae (3.1) - (3.5) represent more or less the "average appraisal score" of a link, weighted by its length.

3.2. Peak values

In practice the weighted summation rule (formula (2.4)) is often criticized by the fact that by simply adding numbers important information might be lost. Peak values, for instance high environmental values, can thus be easily "overlooked" by adding them to other low values. In order to deal with this problem, a so-called scaling parameter may be added to formula (2.4) in the following way:

(3.6) \[ v_g = \left[ \sum_j \left( w_j (e_{jg} + c_{jg}) \right)^a \right]^{1/a} \quad (a>1) \]

whereby $a$ denotes the scaling parameter, which has to be selected a priori by the analyst. Obviously, in case $a$ approaches infinity
The formula (3.6) becomes:

\[
V_g = \lim_{\alpha \to \infty} \left[ \sum_j \left( w_j (e_{jg} + c_{jg}) \right)^{\alpha} \right]^{1/\alpha}
\]

\[
= \max_j w_j (e_{jg} + c_{jg})
\]

By using a scaling parameter \( \alpha > 1 \) high values can be stressed in the analysis, thus increasing the possibility that the particular grid will not be affected by a route. This concept is especially helpful as a kind of "sensitivity analysis" showing the user what will happen if a certain spatial issue gains in importance.

3.3. Qualitative weights

There are several ways to deal with qualitative weighting schemes in multicriteria evaluations (e.g., see Nijkamp et al., 1983). Because of its simplicity, a very attractive approach to deal with qualitative criterion weights \( w_j \) is by using the probabilistic method developed by Rietveld (1982). This method may be illustrated by means of a brief example.

Suppose we have three criteria for which holds that their weights add up to unity, i.e.:

\[
\sum_j w_j = 1
\]

In Figure 3-2 it is visualized that all metric weights which fulfil condition (3.8) can be situated in a triangle. If only a qualitative ordering of the weights is known, the attention can be focussed on a limited area of this triangle, viz. the shaded area in Figure 3-2. The corner points of this area are the so-called extreme weight sets which are in accordance to the postulated ordering. For example, if the following weight set is expressed: \( w_1 \geq w_2 \geq w_3 \), the following extreme metric weights can be distinguished: \((1,0,0)\), \((\frac{1}{2},\frac{1}{2},0)\) and \((\frac{1}{3},\frac{1}{3},\frac{1}{3})\). However, in reality the weights will be situated within the shaded area. By assuming that the "real" metric weights are uniformly distributed over this area, one can find analytical expressions for the most probable metric weight values. It appears that the expected values are a convex function of the corresponding ranks.
Figure 3-2. The boundaries of metric weights attached to three criteria

This means that there is only a small number of relatively important criteria and a large number of relatively unimportant criteria. The expected values of the weights can be calculated as follows:

\[
\begin{align*}
E(w_1) &= \frac{1}{J^2} \\
E(w_2) &= \frac{1}{J^2} + \frac{1}{J(J-1)} \\
E(w_{j-1}) &= \frac{1}{J^2} + \frac{1}{J(J-1)} + \ldots + \frac{1}{J.2} \\
E(w_j) &= \frac{1}{J^2} + \frac{1}{J(J-1)} + \ldots + \frac{1}{J.2} + \frac{1}{J.1}
\end{align*}
\]

By means of (3.8) qualitative weighting schemes can be transformed into numerical values, which can be substituted in formula (2.4).
4. Empirical Illustration

The model elaborated in this paper has been applied to a road location problem in the province of Limburg in The Netherlands. It concerns the design of a new motorway alignment between junction A73/A77 near Boxmeer and the A2 motorway near Maasbracht/Echt. The proposed alignment, called "rijksweg 73", is still in debate, although the necessity of such a connection ("the backbone of the province of Limburg") is never questioned. The differences in opinion focus on the part between the city of Venlo and Maasbracht, which is therefore taken as subject for further consideration in this empirical illustration of the road location model. In Figure 4-1 the study area is given, which includes a relatively large number of...
villages and small towns. There are also several places with valuable environmental and landscape qualities. Besides, some important land use functions are allocated in this area, e.g., a water control area. The complexity of the design problem is also increased by the river the Meuse, which flows almost across the study area.

The area is subdivided in grids of 500 x 500 meters. The characteristics of each grid has been assessed by means of eight criteria, which are based on a report published by the Ministry of Public Works (1979). The following criteria are distinguished:

1. Construction and acquisition costs
2. Landscape and environmental attractiveness
3. Agriculture attractiveness (soil quality, scale, reconstruction plans)
4. Urbanization plans
5. Drinkwater protection areas
6. Physical pollution (noise nuisance, air pollution)
7. Combination possibilities with other kind of infrastructures
8. Existence of valuable objects (monuments).

A "cost line" has been defined for the river Meuse, i.e. an additional value of 10 times the maximum cost value of criterion 1 has been included in (2.5) because of the costs involved in building a bridge. The scores for each criterion are standardized between 0 (very suitable) and 10 (unsuitable for a road). Hence, for each criterion a map can be created to stimulate discussions and to assist in the interpretation of the outcomes of the analysis. An example of such a map made by a line-printer is given in Figure 4-2.

The results of the analysis are included in Table 2. The present computer program computes both for each criterion and for combinations of criteria (through formula (2.4)) an optimal alignment. These combinations are based on the qualitative weighting schemes, which are represented in Table 1. The various outcomes in Table 2 show that the total appraisal scores (i.e. $a$ in formula (2.2)) differ considerably for each weight set. This may be explained by the very complex structure and many conflicting issues in the study area. Some results of the design procedure are visualized in Figure 4-3.
Table 1. Weighting Schemes

<table>
<thead>
<tr>
<th>SCENARIO</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Costs</td>
<td>.230</td>
<td>.076</td>
<td>.154</td>
<td>.076</td>
<td>.076</td>
<td>.076</td>
<td>.154</td>
<td>.154</td>
</tr>
<tr>
<td>B. Landscape and environment</td>
<td>.059</td>
<td>.176</td>
<td>.059</td>
<td>.059</td>
<td>.176</td>
<td>.176</td>
<td>.118</td>
<td>.176</td>
</tr>
<tr>
<td>C. Agriculture</td>
<td>.066</td>
<td>.133</td>
<td>.199</td>
<td>.066</td>
<td>.133</td>
<td>.133</td>
<td>.199</td>
<td>.066</td>
</tr>
<tr>
<td>D. Traffic Attraction</td>
<td>.083</td>
<td>.083</td>
<td>.083</td>
<td>.249</td>
<td>.083</td>
<td>.083</td>
<td>.249</td>
<td>.083</td>
</tr>
<tr>
<td>E. Population of Limburg</td>
<td>.059</td>
<td>.176</td>
<td>.118</td>
<td>.059</td>
<td>.118</td>
<td>.176</td>
<td>.118</td>
<td>.176</td>
</tr>
<tr>
<td>F. Ministry of Transport</td>
<td>.187</td>
<td>.125</td>
<td>.062</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
<td>.125</td>
</tr>
<tr>
<td>G. Ordinal Priorities</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>5</td>
<td>7</td>
<td>2</td>
<td>6</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 2. Some results

<table>
<thead>
<tr>
<th>scenario</th>
<th>α = 1</th>
<th>α = 20</th>
</tr>
</thead>
<tbody>
<tr>
<td>A costs</td>
<td>248.64</td>
<td>93.06</td>
</tr>
<tr>
<td>B landscape and environment</td>
<td>201.76</td>
<td>67.33</td>
</tr>
<tr>
<td>C agriculture</td>
<td>215.64</td>
<td>66.40</td>
</tr>
<tr>
<td>D traffic attraction</td>
<td>296.06</td>
<td>164.82</td>
</tr>
<tr>
<td>E population of Limburg</td>
<td>208.66</td>
<td>66.50</td>
</tr>
<tr>
<td>F Ministry of transport</td>
<td>269.14</td>
<td>108.18</td>
</tr>
<tr>
<td>G ordinal</td>
<td>215.46</td>
<td>76.20</td>
</tr>
</tbody>
</table>
Figure 4.2. An example of a map of criterion scores.

Dark = high grid value (not suited for road construction)
Light = low grid value (suitable for road construction)
Figure 4-3. Some possible road alignments
5. Some summarizing conclusions

In this paper a model has been outlined which can be used to arrive at systematically designed and, above all, accountable and debatable motorway alignments. The emphasis in this approach on specifying the exact design criteria and the priorities (weights) may induce a greater and better public and political involvement, provided that the results of an analysis are clearly published. The multicriteria core of the model enables the designer to highlight the real issues to be resolved. A very comprehensive judgement is possible, although by its very nature, this road location model does not necessary lead to clear-cut solutions for a design problem. However, this seeming disadvantage can be one major benefit: there is a general mistrust of any procedure which appears to offer a simple, straightforward answer to difficult and delicate decisions which involve a variety of subjective considerations. An approach with different weighting schemes, which can be represented in a qualitative way, is more likely to be acceptable to decision makers and citizens than more restrictive design procedures based on single cost criteria or the expertise of one single designer.
References

Alexander, C. and Manheim, M., The Use of Diagrams in Highway Route Location, M.I.T., Civil Engineering Systems Laboratory (1962)
Ministry of Public Works, Nota Rijksweg 73, de ruggegraat van Limburg, tekst, figuren, kaartenmap tracés (1979)
Nijkamp, P., Multidimensional Spatial Data and Decision Analysis, Wiley, Chichester (1979)
Rietveld, P., Multiple Objective Decision Methods and Regional Planning, North-Holland, Amsterdam (1980)
Schoone, J.J., Route Optimalisatie met het model ARM, Delft University of Technology, Dept of Civil Engineering (mimeographed) (1982)
Voogd, H., Multicriteria Evaluation for Urban and Regional Planning, Pion, London (1983a)
1989

Dynamic Assignment in Three-Dimensional Time Space.

R. Hamerslag

1220 Transportation Research Record. National Research Council, Washington, DC
Dynamic Assignment in Three-Dimensional Time Space

Rudi Hamerslag

In traditional assignment models, cars are assigned to a route and are therefore present on all links on that route simultaneously. Calculations from this type of model give few positive results. If the assignment is done in space with time as a third dimension, this problem can be overcome. The first part of the paper gives a simple example of the equilibrium assignment model showing that, in some parts of the network, congestion is unrealistically calculated as a consequence of bottlenecks upstream. The second part of the paper gives a description of the three-dimensional assignment models. The proposed algorithm conforms with existing two-dimensional assignment models, although details in the algorithm are different. The effect of improving the capacity of bottlenecks on congestion downstream is shown. A computer model of the assignment model works under MS-DOS on a microcomputer.

In traditional assignment models two-dimensional (2-D) origin-destination (O-D) matrices are assigned to two-dimensional networks. Cars between each O-D pair are assigned to the links belonging to a certain route. Because these links do not have a time dimension, the implicit assumption is made that cars are present on all links at the same time. So cars that in reality are caught in a particular bottleneck can also be considered in the calculation the cause of congestion downstream. To improve the assignment process, a time dimension is added to the traditional 2-D assigned space. A three-dimensional (3-D) O-D matrix is assigned to a three-dimensional network.

The following are discussed in this paper:
1. The problems with the 2-D assignment.
2. The principles of dynamic assignment in 3-D time space.
3. The principles of the algorithm used.
4. The increase in capacity downstream from the bottleneck, and
5. A few remarks about computing.

TRADITIONAL ASSIGNMENT MODELS

In traditional 2-D assignment models [e.g., that of van Vliet (2)], networks are defined by links. These links connect two models (e.g., j and k). Each node j (1, 2, 3, . . .) and k (1, 2, 3, . . .) has coordinates x_jt and x_kt. Each link has a certain length (z_{jk}) with a distance, time, or generalized time dimension. In this paper, time will be used as a dimension. The shortest routes are calculated between each O-D pair. In the all-or-nothing assignment program, all cars between each O-D pair are assigned to the shortest route.

The equilibrium method (2, 3) can be used if there are overloaded links in a network. The time on every link \( jk \) \( (z_{jk}) \) is calculated by using a delay function:

\[
    z_{jk} = F(q_{jk}, C_{jk}, z_{jk0})
\]

where

\[
    q_{jk} = \text{the traffic flow on link } jk,
\]
\[
    C_{jk} = \text{the capacity of link } jk,
\]
\[
    z_{jk0} = \text{the time of a link } jk \text{ in an unloaded network}, \text{ and}
\]
\[
    z_{jk} = \text{the time of link } jk \text{ in a loaded network}.
\]

See Brandston's overview (4).

The value of \( q_{jk} \) is calculated by an iterative process. Equilibrium will be reached when the flow on all routes in use is equal and when there are no more unused links (Wardrop's principle). To reach equilibrium, the linear approximation method can be used (3). The flow in iteration \( i \) \( (q_{jk}^i) \) is calculated as a linear combination of \( q_{jk}^i \) and \( q_{jk}^{i-1} \). The value \( q_{jk}^i \) is the assigned traffic to the shortest routes in the network with \( z_{jk}^{i-1} = F(q_{jk}^{i-1}, C_{jk}) \).

The next example was inspired by the traffic system southwest of Rotterdam where a bridge limits the traffic crossing the river. The O-D matrix in Table 1 was assigned to the network in Figure 1. The traffic flows run from right to left. Figure 1 shows an all-or-nothing assignment and an equilibrium assignment. The equilibrium model shows that part of the cars are assigned to routes 4-10-9 and 2-1-8. This assignment is made because of congestion on links 3-2 and 5-8. In reality this congestion does not appear because the buses are held in bottleneck 7-6. The equilibrium assignment model gives fewer satisfactory results in this example.

ASSIGNMENT IN TIME SPACE

The main problem of the assignment models is that traffic is assigned to a network without a time dimension. To improve these methods, a time dimension has been added. Links are defined by nodes \( jk \) and period \( p \).

Instead of time on a link \( jk \), the time on link \( jk \) is introduced during period \( p \). A period capacity is used instead of an hour capacity. The traffic flows are also defined by nodes \( jk \) and period \( p \). The routes are calculated on the surface and in space, so a 3-D time space is used.

If a link is overloaded, then the path (a) will switch to a route along other nodes as in the 2-D space, (b) will switch to a route in a later period, or (c) both.

The delay on the links is also determined in time space and
<table>
<thead>
<tr>
<th>From</th>
<th>4</th>
<th>7</th>
<th>sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>To</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1250</td>
<td>3750</td>
<td>5000</td>
</tr>
<tr>
<td>9</td>
<td>1250</td>
<td>3750</td>
<td>5000</td>
</tr>
<tr>
<td>sum</td>
<td>2500</td>
<td>7500</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 1 ORIGIN-DESTINATION MATRIX**

Flows in overloaded links

<table>
<thead>
<tr>
<th>link</th>
<th>Flow</th>
<th>capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>7 - 6</td>
<td>7500</td>
<td>4000</td>
</tr>
<tr>
<td>6 - 5</td>
<td>7500</td>
<td>4000</td>
</tr>
<tr>
<td>3 - 2</td>
<td>5000</td>
<td>4000</td>
</tr>
<tr>
<td>2 - 9</td>
<td>5000</td>
<td>4000</td>
</tr>
<tr>
<td>8 - 5</td>
<td>5000</td>
<td>4000</td>
</tr>
</tbody>
</table>

**FIGURE 1 Assignment of O-D matrix in Table 1 to a network with the all-or-nothing method (top) and the equilibrium method (bottom). (In the following figures, links loaded between 85 and 95 percent of capacity are lightly shaded, and links that are loaded more than 95 percent of capacity are shaded more darkly.)**

works. Overloaded links lead to an overflow to the links in another period.

Kroes et al. (9) mentioned a method called “equilibrium assignment in the timespace,” which was also developed in the Netherlands. In addition to the road network, a network with shadow links was made that represented the alternative of driving at other than peak-period times. With a 2-D equilibrium model, part of the traffic is assigned to this network. Although the name is similar, this method is different from the method presented in this paper.

In the method proposed by Ben-Akiva et al. (10), an equilibrium method is used to change the departure times and link times. The method can be used only for a very small hypothetical network. The 3-D assignment method in this paper uses 3-D O-D matrices and networks with departure times that are not affected by congestion. A study has been started to integrate our method with those of Kroes et al. (9) and Ben-Akiva et al. (10).

**THE ALGORITHM**

Three-dimensional assignment can be formulated as a 3-D equilibrium model. The algorithm consists of the following steps:

1. Read a 2-D network.
2. Determine the 3-D O-D matrix.
3. Determine the period capacity of the links.
4. Calculate the delay in the links.
5. Calculate the shortest routes in 3-D space.
6. Assign the 3-D O-D matrix to the shortest routes.
FIGURE 2 Example of an assignment in time space of the O-D matrix of Table 1.
7. Load the network.
8. If the stop criterion has not been reached, return to Step 4.

Although the 3-D algorithm, generally speaking, is similar to the algorithm in 2-D space, there are some important differences on a more detailed level:

1. Read the 2-D O-D matrix and the 2-D network. Existing 2-D networks can be used in the 3-D calculations; consequently there is no need for conversion or extra input of data. This factor is a practical advantage of the method.

2. Determine the 3-D O-D matrix. The 3-D matrix is determined by splitting up the 2-D O-D matrix into periods defined by the departure time fractions. This system is a good way to approximate the peak periods. For longer periods (e.g., holiday traffic), more complicated methods should be used to determine the 3-D matrix.

3. Determine the period capacity. The period capacity of the link can be determined as a fraction of the hourly capacity.

4. Determine the link delay. A 3-D delay function is used to determine the delay in 3-D links. This function is similar to that in 2-D space.

\[ z_{ijk} = F(q_{ijk}, Q_{ijk}, C_{ijk}, z'_{ijk}) \]  

where

- \( q_{ijk} \) = the number of cars on link \( kj \) during period \( p \),
- \( Q_{ijk} \) = the number of waiting cars from previous periods,
- \( C_{ijk} \) = the capacity (cars per unit) during period \( p \),
- \( z_{ijk} \) = the time on the loaded link \( kj \) during period \( p \),
- \( z'_{ijk} \) = the time on the unloaded link \( kj \) during period \( p \).

In general, \( z_{ijk} \) will have different values for the various periods. The overloaded links in previous periods influence the delay in the later periods.

5. Determine the shortest routes in 3-D space. Figure 3 shows a 3-D network with a string of links. Link 2-3 has a lower capacity than the other links. The Y-axis is the time scale. The links of the successive time periods are shown. The dashed lines are the 3-D paths of the first and last cars in each period. The last car in the first period is the same as the first car in the second period, and so on.

The departure time of the first car in the first period equals zero. This car uses links in period 1. The departure time of the last car in the first period equals 60 min. It is also possible to reduce some of the period capacities for delays caused by highway construction.

6. Assign the 3-D O-D matrix to the shortest 3-D routes. The important difference is that in the 2-D space all car trips of an O-D pair are assigned to all links along the shortest route. Because in the 3-D space links are used during different periods, the cars must be assigned to different periods. The cars that depart in the first period use link 3-4 in the first and second periods. The ratio of the car trips that are assigned to link 3-4 in the first and second periods is proportionate to the areas marked 1 and 2 (Figure 3). The car trips of the third period are partly assigned to link 2-3 in the third and fourth periods. The ratio is proportionate to the areas marked 3 and 4.

7. Load the network. Loading the network is done by part of the all-or-nothing assignment flows just calculated and with flows from the previous iteration. As in the 2-D space, it can be done in various ways.

Two methods have been tested. The first method is similar to the linear approximation method of the equilibrium method. The first experience with this method was not very successful, as was reported at the UTSG conference in London (12) in 1988. However, the method is being improved, so linear approximation may be useful after all. Some research is still required to make this suitable for publication.

The second algorithm uses the equation

\[ q'_{ijp} = q_{ijp} \cdot g' + q_{ijp} \]  

The value of \( g' \) depends on the number of iterations \( i \) and will also be chosen in such a way that there are no overloaded links.

\[ g' = \min[[1/(i + 1), (q'_{ijp} - C_{ijp})/q_{ijp}] \]

**EXAMPLE OF UPSTREAM CONGESTION**

It is possible to gain insight into the problem of new congestion that appears after improvements upstream. The example in
The calculation time necessary will be about 100 times a can give better insight into the ability of this method to improve stream congestion can be prevented by delaying the traffic feeding onto links upstream. The 3-D assignment technique necks on downstream congestion. It seems possible that downstream congestion can be prevented by delaying the traffic feeding onto links upstream. The 3-D assignment technique can give better insight into the ability of this method to improve the working of the traffic system.

COMPUTING
The system runs as part of the TFT? workbench on OLIVETTI M21 and PC-AT and PC-386 with EGA cards for small networks (13). The assignment of large networks is also possible. However, 3-D calculation needs more computer time than 2-D assignment. The calculation time is the product of

• The number of iterations,
• The number of time periods, and
• The time necessary for the calculation of an all-or-nothing assignment.

The calculation time necessary will be about 100 times a 2-D all-or-nothing assignment or 10 times an equilibrium assignment.

To improve the calculation speed, a special processor is being developed so the system can be used for very large networks. The first prototype of this processor is about 200 times faster than a Micravox. An even faster execution is possible (14). Because of these improvements it is expected that a longer calculation time will not be required for very large networks.

CONCLUDING REMARKS
Since the traditional 2-D assignment methods have some shortcomings, a time dimension has been introduced to improve this method. The algorithm, generally speaking, is similar to the 2-D variant. However, on a detailed level there are some differences that cannot be neglected: the method can be used for large networks; the existing 2-D networks can be used as input for the calculations; the calculation time is longer; and the development of computer hardware makes the method suitable for very large networks.

In conclusion, the dynamic assignment in 3-D time space can be used for the following purposes:

• A more realistic assignment of traffic on congested networks;
• Acquisition of new insights into new downstream congestion after improving capacity upstream;
• Ability to calculate, based on downstream congestion, the influence of decreases in capacity caused by such factors as road construction, road maintenance, and accidents;
• Ability to calculate the areawide effect of feeding cars into a network system on certain strategic chosen links; and
• Ability to use the program as part of a delay warning system during road congestion.

REFERENCES
1989

Dynamic assignment in the three-dimensional timespace. Mathematical specification and algorithm.

R. Hamerslag

Proc. Joint USA-Italy seminar on urban networks, dynamic control and flow equilibrium. Capri, Italy
DYNAMIC ASSIGNMENT
IN THE
THREE-DIMENSIONAL TIMESPACE.
Mathematical specification and algorithm.

by Dr Rudi Hamerslag (*)

Paper prepared for the joint USA-Italy seminar on transportation networks, Naples & Capri, June 1989

Professor in Transportation Engineering of the Delft University of Technology. Faculty of Civil Engineering(also partime attached to the Faculty of Technical Mathematics and Informatics), Stevinweg 1, 2600 GN Delft, The Netherlands.
Phone (...)15 784609(University); (...)3430 12809(at home)
DYNAMIC ASSIGNMENT IN THE THREE-DIMENSIONAL TIMESPACE.
Mathematical specification and 3d-algorithm.

by Dr Ir Rudi Hamerslag (*)

* Professor in Transportation Engineering of the Delft University of Technology, Faculty of Civil Engineering (also part-time attached to the Faculty of Technical Mathematics and Informatics), Stevinweg 1, 2600 GN Delft, The Netherlands. Phone (...)3430 12809 & (...)15 784609

ABSTRACT

In traditional assignment models, cars are assigned to a route and so are simultaneously present on all links on that route. However, cars which are delayed in bottlenecks cannot appear downstream at the same time. This can give less good outcomes of the calculations.

If the assignment is done in space with time as a third dimension, this problem can be overcome. A simple example of the equilibrium assignment model shows, that in some parts of the network, congestion is unrealistically calculated as a consequence of bottlenecks upstream. This problem is solved by adding a third time dimension to the traditional two-dimensional equilibrium model. The algorithm is conform to existing two-dimensional assignment models, although it deviates in detail. The effect of improving the capacity of bottlenecks on congestion downstream is shown. A computer model of the assignment model works under MS-DOS on a microcomputer.
1 INTRODUCTION

To study the traffic in road networks a lot of models has been used. There are three types of models:

- the all-or-nothing method, that is often used.
- the probabilistic models, which should be used in networks without or with less overloaded links of the logit type (Dial(1969)) or the probit type (Burrel(1968))
- the equilibrium model, that should be used in networks with congestions.

All these models have in common that the calculations are made in the two-dimensional space, that represents the surface of the earth. In the next paragraph some of the problems with the two-dimensional equilibrium assignment method will be illustrated.

There are two types of models which take a time dimension in account:

- the assignment models. Examples of these kind of models are e.g SATURN and CONTRAM;
- the models which specify the influence of congestion on the departure/arrival times.

The characteristic of the SATURN (eg van Vliet(1982)) method is that time dependent OD matrices are assigned to independent 2-D networks for different time periods.

In the CONTRAM assignment method (Leonard et al (1978)), a limited number of cars are also assigned sequentially to 2-dimensional independent networks. Overloaded links lead to an overflow of links in a next time period.

Kroes et al(1987) mentioned the method called "Equilibrium assignment in the timespace". Additional to the road network a network with shadow links was made, that represents the alternative of driving out of the peak period. With a two dimensional(2-D) equilibrium model a part of the traffic is assigned to this network. So the number of cars that avoid the peak period is calculated. This method developed in the Netherlands, deviates from the method presented in this paper, but the name is quite similar.

In the method proposed by Ben Akiva et al(1984) an equilibrium method is used to change the departure times and the link times. This method can only be used for a very small hypothetic network without overlapping links. Vythoulkas (1988) recently used larger networks.

The method from this paper is of the dynamic type and an addition to the traditional to the 2-D assignment method. To improve the assignment process, a time dimension is added to the two-dimensional assigned space. A three-dimensional OD matrix is assigned to a three dimensional(3-D) network. The method can be applied to very large networks.

The idea for this method is from the mid-seventies. The improvement of the computer hardware made it possible to elaborate the idea. There have been some publications about this subject on the "Vervoersplanologisch Colloquium" in The Hague 1987, on the UTSG Conference in London 1988 and the TRB Annual Meeting in Washington 1989. (Hamerslag(1987,1988a,1989))

Because the reader will not have followed these publications, we will give a brief general description of the method that is roughly speaking, similar to previous publications.

Additional to previous publications, more details are given of the mathematical formulation (par3) and
of the used algorithm (par 4). To integrate the model with distribution, model choice and departure time models, more attention is given to the calculation of the travel times in the three-dimensional space between OD pairs. Some more information is given about the convergence of the iteration process (par 5). A less artificial larger network is used to illustrate the reability of the method (par 6). A few remarks about computing are made in the last part of this contribution.

The three-dimensional method in this paper uses 3-D OD matrices and networks with departure times which are not influenced by congestion in the network. A study to integrate the present method with changing departure time caused by congestion is started together with Kroes and Ben-Akiva. The results will be published elsewhere.

2. TRADITIONAL ASSIGNMENT MODELS

In traditional assignment models two-dimensional Origin and Destination matrices (OD-matrices) are assigned to two-dimensional networks. Cars between each OD pair are assigned to the links belonging to a certain route. In these models networks are defined by links. These links connect two models (e.g. j and k). Each node j (1,2,3,...) and k (1,2,3,...) has coordinates x,y and x,y.

Each link has a certain length (Z) with a distance, time or generalized time dimension. The shortest routes are calculated between each OD pair. In the all-or-nothing assignment program all cars between each OD pair are assigned to the shortest route.

The equilibrium method will be used if there are overloaded links in a network. The time on every link jk (Z) is calculated by using a delay function:

\[ z_j = F(q_{jk}, c_{jk}, z_{jk}) \]

where

- \( q_{jk} \) is the traffic flow on link jk
- \( c_{jk} \) is the capacity of link jk
- \( z_{jk} \) is the time of a link jk in an unloaded network
- \( z_{jk} \) is the time of link jk in a loaded network.

E.g. Brandston (1976) gives an overview of the form of used functions.

The value of \( q_{jk} \) is calculated by an iterative process. An equilibrium will have been reached when all routes in use are equal and when unused links are longer (Waldrop (1952)).

Different algorithms are used to calculate the equilibrium. A lot of research has been performed to find a fast algorithm to solve the equilibrium e.g. Murchland (1969), Ruiter (1974), Florian and Nguyen (1974), LeBlanc et al (1974, 1985) etc.

All the traditional assignment methods have in common that the links do not have a time dimension. So the implicit assumption is made that cars are present on all links at the same time. So cars which in reality are held in a certain bottleneck, can also cause a congestion downstream in the calculation.
This will be illustrated by an example. The network (figure 1) exists of a motorway system (100 km/h) and a secondary road system (50 km/h). The traffic flows run from the left side of the figure to destinations in the middle and on the right side. The traffic flows which are calculated with the all-or-nothing assignment method, follow the motorways. The links which are loaded more than .95 of the capacity, are shaded.

The flows which are calculated with the equilibrium assignment show a different view. The congestion on the motorway caused a delay, so a part of the traffic is reassigned to the roads of the secondary system.

However the traffic is in reality held on the upstream bottlenecks. The downstream bottlenecks do not exist in the middle and left of the figure. The equilibrium model reassigns traffic caused by congestions which do not exist, because the cars are held in the upstream bottleneck. The equilibrium assignment model gives less satisfactory results in this example.

3. ASSIGNMENT IN THE TIME SPACE

Cars which are delayed in bottlenecks cannot appear downstream at the same time. This problem arises because the traffic is assigned to a network without a time dimension. To improve these methods a time dimension has been added. Links are defined by nodes jk and period p. The length of the periods are e.g. 10-15 minutes. Longer periods may also be chosen if this is relevant for the problem.

Instead of time on a link jk, the time \( z_{jk} \) on link jk is introduced during period p. A period capacity \( C_{jk} \) is used instead of an hour capacity. The traffic \( q_{jk} \) flows are also defined by nodes jk and period p.

The routes are calculated in surface and in space. So a 3-d time space is used. If a link is overloaded then the path:
- will switch to a route along other nodes as in the 2-d space
- or will switch to a route in a later period
- or both

The delay on the links is also determined in the time space and may be different from period to period.
At the end a 3-d OD matrix \( T_{odp'} \) is assigned to a 3-d network of trip departing from o to d in period \( p' \).

This problem can be formulated as an equilibrium assignment analogical to the 2-d method.

\[ q_{odp'} = \sum_{od} \sum_{p'} T_{odp'} \delta_{p'od} \text{ for } p' < p \]  

The link flows are equal to the sum of all flows between od pair \( T_{odp'} \) along the 3-d route \( r \) 

\[ T_{odp'} = \sum_{p'} T_{odp'} \]  

The cost between od in period \( p' \) \( z_{odp'} \) is the sum of all linkcosts along the routes

\[ T_{odp'} > = 0 \]

The equilibrium is determined by minimizing

\[ \min \Sigma_{p} \text{ integral} \{ z_{p} \times dx \} \text{ between the boundaries 0 and } q_{mp} \]

The 3-D algorithm can be formulated as a 3-D equilibrium model. The algorithm exists of the following steps:

I. Read a 2-d network
II. Determine the 3-d OD matrix
III. Determine the period capacity of the links
IV. Calculate the delay in the links
V. Calculate the shortest routes in the 3-d space
VI. Assign the 3-d OD matrix to the shortest routes.
VII. Load the network
VIII. If the stop criterion has not been reached then return to step four

The 3-d algorithm is, generally speaking, similar to the algorithm in the 2-D space. There are, however, some important differences on a more detailed level.

I. The 2-d OD matrix and the 2-d network.
   Existing 2-D networks can be used in the 3-D calculations, with the consequence, there is no need for conversion or extra input of data. This is a practical advantage of the method. 
   The freeflow linkcost is determined by

\[ z'_{mp} = z_{m} \]

II. The 3-d OD matrix.
The 3-d matrix is determined by splitting up the 2-OD matrix in periods defined by the departure time fractions \( f_p \)

\[
T_{odp} = T_{od} \cdot f_p
\]

This may be a good approximation for the peak periods. For longer periods e.g. holiday traffic more complicated methods should be used to determine the 3-d matrix.

### III. The period capacity.

The period capacity of the link can be determined as a fraction of the hour capacity. The capacity is multiplied by the ratio of period length \( z_p \) and 60 minutes

\[
C_{zp} = \frac{C_p}{60} \cdot z_p
\]

It is also possible to reduce some of the period capacities to take into account the delay caused by roadworks.

### IV. The link delay.

A 3-d delay function is used for the determination of the delay in the 3-d links. This function is analogical to that in the 2-d space.

\[
z_{zp} = F(q_{zp}, Q_{zp}, C_{zp}, z'_{zp})
\]

where

- \( q_{zp} \) is the number of cars on link \( kj \) during period \( p \).
- \( Q_{zp} \) is the number of waiting cars from previous periods.
- \( C_{zp} \) is the capacity(cars per unit) during period \( p \)
- \( z'_{zp} \) is the time on the unloaded link \( jk \) during period \( p \).
- \( z_{zp} \) is the time on the loaded link \( jk \) during period \( p \). In general, \( z_{zp} \) will have different values for different time periods.

The difference with the delay function in the 2-d space is that the overloaded links in previous periods \( Q_{zp} \) influence the delay in later periods. The 3-D delay function used for the calculation of this contribution is:

\[
\text{If } (q_{zp} + Q_{zp}) < 1.3 \cdot C_{zp} \text{ then}
\]

\[
z_{zp} = z'_{zp} + 0.105 \cdot \frac{(q_{zp} + Q_{zp})}{C_{zp}}, z_p
\]

else

\[
z_{zp} = z'_{zp} + (q_{zp} + Q_{zp})/C_{zp} - 1 \cdot z_p
\]

If \( z_{zp} > z_p \) then the routes in a later period are used and \( q_{zp} + Q_{zp} > C_{zp} \)
which means that there are cars waiting in the queue.

V. The shortest routes in the 3-d space.

Figure 2 shows a 3-D network with a string of links. The link 2-3 has a lower capacity than the other links.

The Y-axis is the time scale. The links of the successive time periods are shown in the figure.

The dashed lines are the 3-D paths of the first and last car in each period. The last car in the first period is the same as the first car in the second period and so on.

The departure time of the first car in the first period equals zero. This car uses links in period 1. The departure time of the last car in the first period equals 10. This car uses the links in the 2nd period. Some of the cars which depart between the first and last car are using links during the first and the second period.

The departure time of the first car of period 3 is 20 min. There is a delay in node 2-3 caused by congestion and so the car arrives in node 3 more than 10 min later. This car also uses links in different time periods.

So the departure times are equal to 0, 10, 20, 30 etc, instead of zero as in the 2-d assignment. The point of time the nodes are passed, depends on delays, which may be different from period to period.

The cars use links in different time periods.

The routes from the origin are determined in the 3-d space similar to the 2-d space. It is possible to use the Moore or the Dijkstra algorithms. The difference is this is done for all the time periods instead of one time period as in the 2-d space.

A rather important difference with 2-d assignment methods is, that the routes in the 3-d space are found by comparing space paths and time paths simultaneously. This enables a comparison of routes between origin and destination.

VI. The 3-D all or nothing assignment.

The 3-D OD all-or-nothing assignment matrix is loaded to the shortest 3-D routes. The important difference is, in the 2-D space all cartrips of an OD pair are assigned to all links along the shortest route.

Because in the 3-D space links are used during different time periods, the cars must be assigned to different time periods.

The cars which depart in the first period use link 3-4 in the 1st and 2nd period. The ratio of the car
trips which are assigned to the link 3-4 in the 1st and 2nd period is proportional to the areas marked by (italic) 1 and 2.
The car trips of the 3rd period are partly assigned to the link 2-3 in the 3rd and 4th period. The ratio is proportional to the areas marked by (italic) 3 and 4.

VII. Loading of the network.
As in the 2-D space the loading of the network is possible in two ways:
- Successive loading of parts of the matrix.

\[ q'_{jp} = q^{*}_{jp} \cdot g' + q''_{jp} \]

The first experiments with the 3-D algorithm performed been carried on in this way. (Hamerslag(1987,1988))

- The flow in iteration \( i \) (\( q'_{jp} \)) is calculated as a linear combination of \( q^{*}_{jp} \) and \( q''_{jp} \). The value \( q^{*}_{jp} \) is the assigned traffic to the shortest routes in 6.

\[ q'_{jp} = q^{*}_{jp} \cdot g' + q''_{jp} \cdot (1 - g') \]

The second method has the advantage that the calculated times between OD-pairs can be used for interrelated calculation of distribution models, model choice models or arrival/departure time models.

The loading in this paper has been done with this method. A fixed and a decreasing weighting factor are applied. A fixed weighting factor (as used in eg MICROTRIPS) is

\[ g' = 0.20 \]

It is also possible to use a weighting factor that decreases with the number of iterations \( i \).

\[ g' = \frac{1}{i + 0.5} \]

Powell and Sheffi (1982) proved, a decreasing predetermined weighting factor can be used to reach the 2-D equilibrium. Horowitz(1989) has recently used a decreasing weighting factor in a case study with in the 2-D space with reasonable results.

5 CONVERGENCE

In figure 3 the total time in the system as a function of the number of iterations is given. The fixed and variable weighting factor \( g' \) is used. Both methods converge to the same limits. The RMSE of the linktime changes is also given. The variable weighting factor gives less changes in the RMSE. So the variable weighting factor gives a better result here.
The differences in traffic loads between the 10th and the 25th iteration are less. The largest differences are found in the 6th and 7th period on the links 5-8 and 8-3. The differences with the fixed weighting factor are larger than those with the variable factor. The differences are acceptable in calculations for practical purposes.

6 THREE-DIMENSIONAL EXAMPLE

The OD matrix and network of the example of Figure 1 are used to illustrate the working of the method in the three-dimensional space. Figure 4 gives the flows during the successive 8 periods. The traffic is held in the upstream bottleneck, links 14-11, 6-5 and 13-2. There is no congestion in links 14-9-10-3, 11-12-17 downstream, like in the 2-d space. Less traffic has been assigned to the secondary roads (11-20-19-23-18);

It seems also to be possible to prevent downstream congestions by delaying traffic by feeding on links upstream. The 3-D assignment technique can give a better insight into it.

4 COMPUTING

The system runs as part of the TFTP workbench on OLIVETTI and PC-AT and PC-386 with EGA cards for small networks. Hamerslag(1988b)

The assignment of large networks is also possible. However, the calculation needs more computer time than the 2-D assignment.
Figure 4. Example of an assignment in the timespace. Note the bottlenecks in 14-11 and 6-5. Some of the bottlenecks downstream disappear.
The calculation time concerns the product of:
- the number of iterations
- the number of time periods
- the time necessary for the calculation of an all-or-nothing assignment.

The calculation time will be about 100 times a 2-d all-or-nothing assignment or 10 times an equilibrium assignment.

The longer calculation time will not be a problem if e.g. parallel computing techniques are used. The reported improvement by Chen and Boyce (1989) and by Mouskos and Mahassani (1989) are promising. More than 20 times faster calculations are possible.

To improve the calculation speed a special processor is being developed. The first prototype of this processor is about 200 times faster than a microvax. Even a faster (perhaps 1200 times) execution is possible (Grol 1988).

Because of these improvements it is expected, longer calculation time will not be required for very large networks.

8 CONCLUDING REMARKS

The traditional 2-d assignment methods have some shortcomings. A time dimension has been introduced to improve this method.

The algorithm, generally speaking, is similar to the 2-D variant. However, more detailed, there are some differences which cannot be neglected.

It can be used for large networks. The existing 2-D networks can be used as input for the calculations. The calculation time is longer. The development of computer hardware makes it suitable for very large networks.

The dynamic assignment in the 3-d time space can be used for the following purposes:

- A more realistic assignment of traffic in congested networks. It is possible to get insight into the new downstream congestions by improvement of the capacity upstream.
- The influence of a certain decrease in capacity caused by road construction, road maintenance, accidents, etc. can be calculated on downstream congestion.
- It will be possible to calculate the area wide effect of feeding cars into a network system on certain strategic chosen links.
- It may be possible to use the algorithm for road guiding systems
- The program can be used as part of a delay warning system caused by congestion.

ACKNOWLEDGEMENT

I thank Erik P. Kroes and Moshe Ben-Akiva for encouraging me to give more attention to a more
accurate calculation of the travel times between OD pairs; so the method can be used interrelated with new departure/arrival time models.

LITERATURE


Huye-Kuo Chen(1989) and D.E.Boyce Code optimisation for non linear transportation networks model: A case study TRB 68th annual meeting, Washington


Hamerslag, R (1989) Dynamic assignment in the time space 86th TRB annual meeting, Washington (880596)


Mouskos, K.C (1989) and H.S.Mahmassani Guidelines and computational results for vector processing of network assignment codes on supercomputers TRB 68th annual meeting, Washington (preprint 880615)

Murchland, J.D.(1969) Road network traffic distribution in equilibrium. Mathematisches Forschungsinstitut, Oberwolfach


A practical dynamic network equilibrium model with elastic departure times.

E. Kroes and R. Hamerslag

Transportation Research Board. 69th annual meeting. Washington D.C. (paper 890788)
A PRACTICAL DYNAMIC NETWORK EQUILIBRIUM MODEL WITH ELASTIC DEPARTURE TIMES

Eric Kroes
Hague Consulting Group

Rudi Hamerslag
Delft University of Technology

Summary

The paper describes a practical dynamic equilibrium model for simulating car traffic in congested highway networks. The model consists of two main components: an assignment model and a departure time choice model. The assignment model is an improved version of the 3-dimensional assignment in the time-space by Hamerslag, now able to perform equilibrium assignments. The departure time choice model has been adapted from a model by Ben-Akiva, and has been entered into the equilibrium loop.

The paper describes the main characteristics and innovations to both model components, and presents some results of the simulation model. The following topics are covered:

- fixed versus elastic departure time choice
- different weighting factors in the assignment
- different parameter values in the departure time choice model

Finally some conclusions are given regarding the development and application of the model.
1. INTRODUCTION

Traffic simulation models are tools used to predict and evaluate how car traffic will react to network improvements. But unfortunately most existing traffic simulation models do not deal adequately with the time dimension. They ignore the fact that car drivers tend to avoid peak hour congestion by rescheduling their departure times (see e.g. Kroes et al., 1987). They also disregard the fact that traffic which is held up at one bottleneck cannot be present at other bottlenecks downstream at the same time. Ignoring these phenomena does not cause problems in situations of moderate congestion. But for networks with severe congestion the existing traffic simulation models become inadequate, and lead to unrealistic predictions.

Therefore, for situations with significant congestion, new traffic simulation models are required, which incorporate the time dimension and which are able to predict the time-dynamic impacts of congestion on car traffic patterns. This paper describes a research project, in which such a model was developed. The work was commissioned by the Dienst Verkeerskunde of the Netherlands Ministry of Transport (Rijkswaterstaat). It was carried out in collaboration between Hague Consulting Group and the Technical University of Delft.

The objective of the project was to develop and apply a prototype version of a practical traffic simulation model which incorporates and integrates in an equilibrium framework the following two major time dynamic features:

1. The explicit representation of the time dimension in the route choice simulation (assignment) procedure.
2. The explicit representation of the process of departure time choice.

The first feature was included by using a 3-dimensional assignment method, as originally developed by Hamerslag (1988), and modifying it into an equilibrium procedure. The second feature was included by adapting a departure time choice model by Ben-Akiva et al. (1984), combining it with the assignment model and embedding this inside the equilibrium loop.

In this paper we describe some characteristics of the two main model components, and explain what modifications were made in order to achieve equilibrium. Then we show some overall simulation results, focussing on the following issues:

- fixed versus elastic departure time choice
- different weighting procedures in the assignment
- different parameter values in departure time choice.

Finally, we formulate some conclusions on the development and use of the model.
The time dimension was included in the assignment part of the simulation model by using a 3-dimensional (3-D) assignment procedure, which allows choice of routes in both space and time. A first practical algorithm for doing this has been developed by Hamerslag (1988), and we refer to his paper for the description of it. Hamerslag's first algorithm used an incremental loading approach to achieve a feasible solution, which was not an equilibrium solution.

In order to move towards a 3-D equilibrium assignment we decided to modify the original algorithm. At the same time this modification allowed us to obtain better estimates of the mean journey time for departures in each time band, which were required as an input for the departure time choice model. The modified assignment method differs in two main respects from the original method, which we will discuss in turn:

1. It calculates the journey times for the vehicles in each time band in a different way.
2. It uses a different procedure to load the vehicles to the network, which makes it possible to achieve equilibrium.

The journey times are calculated in the new algorithm as follows. For each group of vehicles leaving in a certain time band the algorithm keeps track of the exact location in the network of the first vehicle and the last vehicle of that group. So the vehicles are followed on a continuous timescale. This is illustrated in Figure 1. Here the y-axis is the time scale, and the x-axis represent the string of links. Link 2-3 has a lower capacity than the other links. The dashed lines are the 3-D paths (shortest routes in the time-space) for the first and the last vehicle in each time band. The last vehicle in band 1 has the same path as the first vehicle in band 2, and so on.

The departure time for the first vehicle in time band 1 equals 0. This vehicle passes all links during the first time band. The departure time for the last vehicle in time band 1 equals 10. This vehicle passes all links in the second time band. Some of the vehicles departing between this first and the last vehicle are passing some links during the first, and some the second time band. The departure time for the first vehicle of time band 3 is 20. There is a delay now on link 2-3 caused by congestion due to other vehicles that arrived previously. This is reflected in the time-space line, which is steeper than for the previous vehicles. After period 4 this congestion has disappeared, and the first vehicle of time band 5 (departure time equals 40) travels again at the same speed on link 2-3 as the vehicles leaving in the first two time bands.

In the loading stage each link needs to be loaded with the number of vehicles that use that link during each time band. This is different from the traditional 2-D assignment, where all vehicles are loaded to all links along the shortest route. So the vehicles
travelling on the continuous timescale illustrated in Figure 1 need to be apportioned to the limited number of discrete time bands that are distinguished for the loading stage. This is done in the following way. Many vehicles which depart in the first time band use link 3-4 partially in time band 1, and partially in time band 2. So a fraction of these vehicles needs to be loaded to this link in time band 1, and the remaining vehicles to the same link in time band 2. The algorithm distributes these vehicles over the time bands in proportion to the ratio of areas 1 and 2 shown in the figure. The sum of these two areas is the total vehicle kilometer hours on link 3-4 by the traffic stream leaving the origin during the first time band.

Then the second modification. In order to achieve an equilibrium solution the new algorithm uses an iterative procedure, which works as follows. In the first iteration all shortest paths are sought (for each time band!) in an unloaded network. So all speeds are free-flow, and the network is assumed to be empty. Then all vehicles are loaded onto their shortest routes, and all link speeds are adjusted using a speed-flow relationship. Then the next iteration starts. In this iteration a new set of shortest paths is sought, now within the loaded network of the previous iteration. This takes account of the reduced speeds on the (over)loaded links, and the amount of overloading which may be present at each link. Then all vehicles are loaded onto their new set of shortest routes, and the algorithm calculates the new flows as a weighted average of the loadings calculated in the previous iteration and the all-or-nothing loadings calculated in the current iteration. So the flows (on each link for each time band) effectively become:

\[ Q_i = g_i \cdot Q_i^a + (1 - g_i) \cdot Q_{i-1} \]

(1)
This process is repeated until some stopcriterion is reached, or a predefined number of iterations has been completed.

For a more detailed, step-by-step description of the algorithm, we refer to Hamerslag (1989).

3 THE DEPARTURE TIME CHOICE PART OF THE MODEL

The representation of the departure time choice process within the simulation model was included by adding an explicit departure time choice model. This predicts departure time choice as a function of desired departure time and expected delays in the network. The model is based upon a model used by Ben-Akiva et al. (1984). It considers travellers who go through one or more bottlenecks, and who may be able to adjust their departure times to avoid slow traffic. Consider, for example, a commuter who has to be at his/her work place at an official work start time. This desired arrival time may be during the peak period and the road that the commuter takes is congested at that time. This commuter may have a choice between on time arrival with a long travel time, and a late or early arrival with a shorter travel time. The difference between actual and desired arrival time is the schedule delay incurred by the traveller who trades it off against travel time (Kraft and Wohl, 1967). Cosslett (1977), Small (1982) and Abkowitz (1980) developed econometric demand models of work trip scheduling that are based on this trade-off between travel time and schedule delay. They estimated logit models of the choice of departure or arrival time bands with different specifications of the utility function.

The model predicts the probability that a driver chooses a departure time band $t$, out of the set of all possible departure time bands $T$, which is given by the following logit equation:

\[
p(t) = \frac{\exp (U_t)}{\sum_{t \in T} \exp (U_t)}
\] (2)
The two key variables that vary among alternative departure time bands are the travel time $Z$ from origin to destination (including all delays) and the schedule delay. The utility $U_c$ can therefore be expressed as:

$$U_c = \beta_1 [ (Z_p - Z_c) + \beta_2 \text{abs}(T_p - T_c) ]$$

with: $Z_p - Z_c$ - the difference in total travel time between the desired departure time band $p$ and time band $t$  
$T_p - T_c$ - the time difference between desired departure time band $p$ and departure time band $t$  
$eta_1, \beta_2$ - parameters

The parameter $\beta_1$ is a scale parameter, determining the overall sensitivity of the choice probabilities to deviations from the desired departure time band, and the level of randomness in this process. The greater $\beta_1$, the less variance there is around the desired departure times. Parameter $\beta_2$ sets the trade-off ratio between journey time delays and earlier or later departure times. When $\beta_2$ is 2.0 it means that the perceived disutility of leaving 10 minutes early or late is valued the same as the disutility of 20 minutes journey time delay.

Of course the specification of (3) is an extremely simple one, assuming equal disutility for earlier and later departures, and not allowing for any non-linearities in the disutility. Although more complex utility expressions for departure time choice models have been estimated in the US (Small 1982) we have retained the simple model form for this study, because no empirical evidence was available yet on the true form of the utility expression for the Netherlands. Currently research is carried out by Hague Consulting Group to derive parameters for departure time choice models in The Netherlands in the context of increased congestion, so we hope to be able to use a better founded model in the near future.

4 THE INTEGRATED SIMULATION MODEL

When the improved 3-D equilibrium assignment in the time-space is combined with the departure time choice model described in the previous section, a very powerful integrated simulation model is obtained. In Figure 2 we have summarised some key differences between the traditional 2-D equilibrium assignment, the already more advanced SATURN and CONTRAM models, Hamerslags' original assignment algorithm and the newly developed integrated simulation model.

The differences are clear. The traditional equilibrium assignment operates entirely in the 2-dimensional space, assuming all flows and demand to be constant in time. SATURN and CONTRAM move into pseudo-dynamics, allowing the OD-matrix to vary over time, and adding certain dynamic elements to the loading stage (for instance taking account of accumulated queues in SATURN). But the
Figure 2. Comparison between various assignment models on dynamic capabilities.

<table>
<thead>
<tr>
<th></th>
<th>2-D Equilibrium</th>
<th>SATURN/CONTRAM</th>
<th>3-D Hammerslag</th>
<th>Integrated sim. model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shortest path finding</td>
<td>2-D</td>
<td>2-D</td>
<td>3-D</td>
<td>3-D</td>
</tr>
<tr>
<td>Link loading</td>
<td>2-D</td>
<td>2/3-D</td>
<td>3-D</td>
<td>3-D</td>
</tr>
<tr>
<td>OD matrix</td>
<td>2-D</td>
<td>3-D</td>
<td>3-D</td>
<td>3-D</td>
</tr>
<tr>
<td>Elastic OD matrix</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Equilibrium procedure</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
</tbody>
</table>

Figure 3 Graphic representation of road networks used for the shortest path finding in various simulation methods.

shortest path finding is still in 2-dimensional space, although again an element of dynamics is added by using repeated 2-dimensional networks. Hamerslag's original 3-D algorithm is the first practical method explicitly simulating the time dimension in the shortest path finding, in loading and in the OD matrix.
The differences in the path finding stage are illustrated in Figure 3. Hamerslag's method, however, does not have the possibility to reach an equilibrium solution. And demand is still assumed to be inelastic, i.e. independent of the delays on the network. The integrated simulation model, presented earlier in this section, adds these two missing elements.

5. PRACTICAL TESTS OF THE MODEL

The newly developed integrated simulation model has been operational since early 1989. In the first half of this year we have performed a large number of simulations in order to test and illustrate how the model works. These simulations were done using a simplified network for the Rotterdam area, together with an OD matrix estimated from traffic counts and an assumed desired departure time profile.

It is useful to point out that the network specification required for the dynamic assignment is exactly the same as for a traditional (static) assignment. In other words, no additional inputs are required to describe the network. So all existing networks can in principle be used without modification. The same holds for the OD matrix: the basic input required is a two-dimensional OD-matrix, which is the same as the input to a conventional static assignment. The transformation into a three-dimensional matrix is obtained by applying the departure time choice model. The simulation model can also accept as an input a time dependent OD matrix, if one is available.

In order to predict the departure time choice probabilities the model requires a desired departure time distribution. For this we used the profile given in Figure 4, one profile for all origin zones. It can be noticed that this distribution has an unusually sharp peak, which is bound to cause congestion. However, note that this is the assumed distribution of the desired departure times (input to the model), not the actual departure time distribution (output of the model). We have chosen this (artificially) sharply peaked distribution of desired departure times for two purposes. Firstly to demonstrate how traffic congestion arises, due to fact that (too) many people want to travel at approximately the same time. Secondly to clearly illustrate how the dynamic equilibrium assignment interacts with the departure time choice model.

In the next sections we present some of the results obtained with the integrated simulation model. We start by showing some results of the improved 3-D equilibrium assignment, where demand is kept fixed. Then we show what happens if the departure time choice model is added, and demand is allowed to interact with congestion. This is repeated with various alternative values for the key model parameters, to investigate the sensitivity of the model results. Since space is limited we present only overall results, no detailed link flow evaluations. For more detailed
results, including comparisons of the obtained link flows, we refer to the original study report (Hague Consulting Group 1989).

6 SOME SIMULATION RESULTS WITH FIXED AND ELASTIC DEMAND

The results of a simulation with the integrated model, where the predicted departure times have been constrained to the desired profile given in Figure 4, are summarised in Table 1 (all tables are given at the end of the paper). A number of observations can be made regarding the simulation and the equilibrium process in this simulation:

- the departure times are kept fixed for the different iterations, but the arrival times vary significantly, especially in the first few iterations; the first iteration assumes free-flow speeds, resulting in a very short mean journey time (11 min.) and a nice even arrival profile; however, this is totally neglecting any delays in the network; in the second iteration full congestion is taken into account, leading to a sharp increase in mean journey time (from 11 to 29 min.) and long delays in the arrival profile; from the third iteration onwards the arrival time profile does not change much any more, and the mean journey time sets at a level of around 21 min.

- the Root-Mean-Squared-Error (RMSE) between all linktimes in subsequent iterations is a useful indicator to monitor the stability of the equilibrium solution; from Table 1 it is clear that after iteration 3 the RMSE decreases steadily to a value below 1 after iteration 7.

Now what happens when demand is allowed to interact with congestion? Table 2 summarises the results of a simulation with
elastic demand. The following can be observed:

- from the first iteration onwards the departure times spread out, so that there are now also cars leaving in the first two time bands and the last two time bands; the departure time profile varies significantly in the first three iterations; the first iteration assumes free-flow speeds, resulting in a completely symmetric departure profile; this profile is already less peaked than the desired departure profile, due to random deviations from the desired departure times; in the second iteration full congestion is taken into account, and as a consequence departure times are spread out further; from the third iteration onwards the departure time profile does not change much any more; note, however, that the profile becomes slightly asymmetric, with somewhat more early departures than late departures (see Figure 5); this is because the delays on the network increase the mean journey times after the peak more than before the peak; therefore it is more attractive to depart 10 minutes earlier than 10 minutes later.
- the mean journey time sets at a level of around 15 minutes, significantly less than the 21 minutes found for the previous simulation with fixed departure times; clearly the assumed additional flexibility to avoid congestion works very well to reduce the delays in the network.
- the size of the Root-Mean-Squared-Error (RMSE) between all linktimes in subsequent iterations is slightly higher for the first few iterations, but then becomes similar to the value found in the simulation with fixed departure times.

Figure 5 Comparison between predicted departure time profiles with fixed and with elastic demand

![Graph showing comparison between fixed and elastic demand](image-url)

Overall it can be concluded that adding elastic demand to the simulation procedure does not cause any problems as far as con-
vergence of the simulation model is concerned, and the simulation results (at an aggregate level) appear plausible. The predicted mean journey time is significantly lower than for the simulation with a fixed (and highly peaked) departure time profile.

7 ALTERNATIVE WEIGHTING FACTORS IN THE ASSIGNMENT.

We have tried two different weighting factors for the iterative equilibrium assignment procedure:

1. A fixed weighting factor \( g = 0.20 \),
2. A decreasing weighting factor: \( g = \frac{1}{(i + 0.50)} \).

In order to monitor convergence of the model the number of iterations was extended to 100, and some overall results are given in Figures 6 and 7. From the results the following can be concluded regarding the impact of the different weighting factors:

- when the decreasing weighting factor is used the mean journey time shows clear fluctuations during the first ten iterations; when the number of iterations increases the mean journey time settles at a stable level; when the fixed weighting factor is used the mean journey time shows less fluctuations in the first few iterations, but it retains some variation throughout all subsequent iterations, however many they are
- the RMSE between all linktime changes in subsequent iterations develops in a similar way: for the decreasing weighting factor we find large values in the first few iterations,

Figure 6 Comparison between mean journey times with fixed and decreasing weighting factor

![Diagram showing mean journey time in minutes for fixed and decreasing weighting factors over iterations.](image)
while these are approximating zero for the last iterations; the fixed weighting factor simulation shows smaller RMSE in the first few iterations, but retains values around 0.5 for the last iterations.

when the levels of the mean journey times are compared the fixed weighting factor seems to lead to a more variable, but on average slightly lower value than the decreasing weighting factor; this suggests a slightly better solution. Although there are some differences in the results produced by the two different weighting factors there are no clear indications on which one performs better. When the resulting link flows are compared the differences are extremely small. At a practical level the more stable result provided by the decreasing weighting factor seems more attractive than the somewhat variable result provided by the fixed factor, even though the latter may produce lower mean journey times.

8 ALTERNATIVE PARAMETER VALUES IN THE DEPARTURE TIME CHOICE MODEL

In order to investigate the sensitivity of the simulation model to the size of the scale parameter in the departure time choice model, different simulations were carried out with the following values for the scale parameter: 0.10, 0.20, 0.30 and 9.0. The last value was deliberately chosen very high, to investigate whether or not the model approached the desired departure time profile, as it should. The other values were chosen within the range of values that was expected to be reasonable. The resulting departure time profiles of the simulations, all with a value -2.00 for the trade-off parameter, are summarised in Figure 8. From these results the following can be concluded:
- when the scale parameter is set at 9.0 the departure time choice model exactly reproduces the desired time profile
- when lower, more realistic values are used for the scale parameter the departure time choice model allows deviations from the desired times; as the scaling factor is chosen lower the spread across the different time bands gets larger, as expected (ultimately, when the scaling parameter approaches zero the departure time choice gets completely random, resulting in a uniform distribution of the departures across time)
- the model seems fairly sensitive to value of the scaling parameter; a value of 0.30 gives results which are still very close to the desired departure time profile, which means very little spread; at 0.20 there is already some more spread, and a value of 0.10 appears to spread even more
- again it can be observed that the predicted departure time profile is somewhat asymmetric, with slightly more departures switching to earlier time bands
- the predicted mean journey time varies with the size of the scale parameter: lower mean journey times are obtained for smaller values of the scale parameter
- the RMSE reduces to similar values in all simulations.

It can be concluded that the departure time choice model appears to be fairly sensitive to the value of the scaling parameter. High values (0.30 or higher) lead to predicted departure time profiles which are hardly different from the desired departure time profile. Lower values increase the spread, and a value of 0.10 leads to results which seem reasonable (as there is no empirical evidence available this is the only way to evaluate the value of the scaling parameter).

Figure 8 Comparison between predicted departure time profiles for different scaling parameter values

![Graph showing predicted departure time profiles for different scaling parameter values](image-url)
The second main parameter in the departure time choice model is the trade-off parameter. In order to investigate the sensitivity of the model to the size of this parameter three simulations were compared, with the following values for this parameter: -0.50, -1.00 and -2.00. A priori it was not clear what the size of this parameter should be: some people might prefer to leave 10 minutes earlier to avoid 5 minutes of delay, but others might prefer 10 minutes of delay to leaving home 5 minutes earlier. Much depends on individual preferences and personal constraints. But it is clear that the sign should be negative (as deviations from the desired departure time cannot reasonably generate any benefits, only disbenefits).

The results of the simulations, all with a scale parameter value of 0.10, are given in Figure 9. The following can be concluded:

- the lower the value of the trade-off parameter, the less the deviations from the desired departure times: a value of -2.00 gives much less spread than a value of -0.50; this is in line with the expectations, as a lower value represents a stronger disutility associated with leaving earlier or later
- the predicted departure time profile appears to be quite sensitive to this trade-off parameter: the predicted number of departures during time band 5 (the absolute peak) varies between 3500 for -2.00 and less than 2000 for -0.50; this means that about 50% of all departures which would desire to leave in time band 5 are predicted to switch to another time band when the trade-off parameter equals -0.50
- the predicted mean journey time varies with the size of the trade-off parameter: lower mean journey times are obtained for higher values of the parameter
- the RMSE reduces to similar values in all simulations.

Figure 9 Comparison between predicted departure time profiles for different trade-off parameter values
It can be concluded that the departure model appears to be also quite sensitive to the value of the trade-off parameter. Low values (-2.00 or lower) lead to predicted departure time profiles which are not very different from the desired departure time profile. Higher values (-1.00 or higher) increase the spread significantly, and a value of -0.50 moves more than 50% of all peak departures away from their desired departure time band. Again there is no empirical evidence available on which result is best. We have guessed that a value of -1.00 might be a realistic one.

9. CONCLUSIONS

We have taken two existing model components, adapted and modified these and combined them into one integrated simulation model. We have briefly explained how the key parts of the model work, and we have carried out a large number of simulations to try it out and to test how it performs. Only a few of these could be presented in this paper. At this stage we feel the following conclusions can be drawn about the model in its current state:

- the results obtained with the model sofar show that it is capable of doing what it was designed for: simulating traffic in congested networks with an explicit representation of the time dynamic characteristics of that traffic with respect to route choice and departure time choice, and leading to an equilibrium solution
- the model provides a much more detailed output than conventional static 2-D assignments, explicitly showing the traffic moving through the network and demonstrating the build-up and breakdown of congestion at bottlenecks over time
- the model does not require much more input data than the conventional static model: similar 2-D networks and OD matrices can be used; however, when more detailed information is available, for instance a time-differentiated OD matrix, this can also be entered directly into the model
- the sensitivity tests performed sofar suggest that the model is well-behaved, producing results which are plausible and seem realistic.

Although the model has been shown to work, and to provide the sorts of predictions which we wanted it to provide, it is still a prototype. Further work is required before it can be fully recommended as a general method at the practical application level.

It is clear that real life testing of the developed model is the major requirement. As the proof of the pudding can only be in eating it, applications of the method for actual studies would provide the best insights into its usefulness to address the current and future transport problems of congested networks, and to assist in finding feasible solutions for these problems. Before such real life testing can be done some additional research is
required. Specifically, the following areas would benefit from further research:

1. The development of the model into a professional and practical computer program needs further work.
2. The assignment part of the model would benefit from more work investigating the basic properties of the improved threedimensional equilibrium assignment procedure.
3. The departure choice part of the model would benefit from an empirically based utility function specification (for instance using results from the DVK SP Congestion study).
4. A further issue for research is the equilibrium process of the simulation method. A formal mathematical investigation of the properties of the method would be useful.
5. And finally the hardware on which the program is used merits attention. There are now possibilities to develop dedicated computer systems using parallel processors, which would make it possible to perform very detailed simulations within a reasonable timescale.

We expect to be able to do this further development work and research within the next year, so that real life applications of the model can soon be made.

10. ACKNOWLEDGEMENTS

We want to acknowledge the substantial contribution by Moshe Ben-Akiva (MIT/Hague Consulting Group) to the success of this project, and the stimulating comments we obtained from Peter de Wolf (Rijkswaterstaat DVK).

11. REFERENCES


Cosslett S. (1977) Demand Model Estimation and Validation (Edited by D. Mc.Fadden et. al.) Urban Travel Demand Forecasting Project, Institute of Transport Studies, University of California, Berkeley.


12. TABLES Table 1 Simulation results with fixed demand

\[ g = \frac{1}{i+0.5} \]
\[ \beta_1 = \text{not relevant: demand fixed} \]
\[ \beta_2 = \text{not relevant: demand fixed} \]

<table>
<thead>
<tr>
<th>Iteration: 1</th>
<th>Mean journeytime: 11.2 min</th>
<th>RMSE linktime: 3.9 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 940 1940 3880 2120 1060 60</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 2</th>
<th>Mean journeytime: 29.1 min</th>
<th>RMSE linktime: 9.7 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 673 953 1613 1400 2540 2820</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 3</th>
<th>Mean journeytime: 20.2 min</th>
<th>RMSE linktime: 17.7 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 780 820 1544 1400 2644 2540</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 4</th>
<th>Mean journeytime: 22.3 min</th>
<th>RMSE linktime: 8.5 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 757 831 1377 1949 2300 2786</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 5</th>
<th>Mean journeytime: 21.0 min</th>
<th>RMSE linktime: 3.4 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 744 838 1391 1942 2482 2602</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 6</th>
<th>Mean journeytime: 21.6 min</th>
<th>RMSE linktime: 1.6 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 736 842 1400 1938 2424 2660</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration: 7</th>
<th>Mean journeytime: 21.3 min</th>
<th>RMSE linktime: 1.0 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Departures</td>
<td>0 0 1000 2000 4000 2000 1000 0 0</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 0 0 731 832 1418 1935 2531 2552</td>
<td></td>
</tr>
</tbody>
</table>
Table 2  Simulation results with elastic demand (scale parameter 0.10 and trade-off parameter -1.00)

\[ g = \frac{1}{i + 0.5} \]
\[ \delta_1 = 0.10 \]
\[ \delta_2 = -1.00 \]

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Mean journeytime</th>
<th>RMSE linktime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.1 min</td>
<td>3.9 min</td>
</tr>
<tr>
<td>Departures</td>
<td>153 417 1133 1953 2687 1953 1133 417 153</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 144 401 1090 1903 2643 1997 1182 638</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>18.3 min</td>
<td>7.4 min</td>
</tr>
<tr>
<td>Departures</td>
<td>273 713 1375 1672 2111 1791 1333 536 197</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 260 692 939 1081 1733 1576 2195 1523</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>13.1 min</td>
<td>15.3 min</td>
</tr>
<tr>
<td>Departures</td>
<td>237 615 1318 1771 2263 1814 1287 508 187</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 226 551 962 1221 1849 1861 2000 1330</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>15.3 min</td>
<td>7.7 min</td>
</tr>
<tr>
<td>Departures</td>
<td>235 608 1339 1809 2256 1774 1281 510 188</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 223 557 909 1354 1827 1758 1968 1403</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>15.3 min</td>
<td>3.3 min</td>
</tr>
<tr>
<td>Departures</td>
<td>235 607 1342 1817 2238 1750 1292 526 193</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 224 562 977 1423 1743 1760 1937 1474</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15.5 min</td>
<td>1.3 min</td>
</tr>
<tr>
<td>Departures</td>
<td>236 608 1339 1826 2240 1746 1290 522 192</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 224 568 856 1460 1745 1715 1932 1499</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>14.9 min</td>
<td>1.0 min</td>
</tr>
<tr>
<td>Departures</td>
<td>234 601 1328 1833 2238 1747 1297 528 194</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 223 541 865 1477 1704 1803 1914 1474</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>15.5 min</td>
<td>0.9 min</td>
</tr>
<tr>
<td>Departures</td>
<td>235 602 1327 1843 2232 1739 1291 534 196</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 223 548 845 1505 1715 1761 1906 1496</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>15.4 min</td>
<td>0.5 min</td>
</tr>
<tr>
<td>Departures</td>
<td>235 603 1325 1844 2228 1735 1294 537 198</td>
<td></td>
</tr>
<tr>
<td>Arrivals</td>
<td>0 223 553 830 1524 1683 1762 1949 1476</td>
<td></td>
</tr>
</tbody>
</table>
A Dynamic traffic assignment model for short-term predictions.

E. de Romph, H.J.M. van Grol and R. Hamerslag

2nd International Capri Seminar On Urban Traffic Networks. Capri, Italy
1. INTRODUCTION

The growth of traffic in the last decade coupled with environmental constraints has led to several developments in traffic management to solve urban traffic congestion. Among these developments is Dynamic Traffic Management (DTM). The point of DTM is to optimize the throughput of the network, given the present supply and demand for traffic. The instruments, which DTM can use to accomplish its task are:

- rerouting and route guidance
- ramp metering
- incident management
- tidal flow
- car pooling, etc.

To operate and control these instruments safely and efficiently, models are needed to gain insight in the short term development of traffic.

Several requirements can be formulated for these models, which are either new models, or traditional models, adjusted to this new task.

A short term traffic model should:

- be able to respond to DTM measures
- be able to simulate the traffic realistically, especially the modelling of traffic jams
- be suitable for networks, not just a short stretch of road
- be able to be used in a real-time situation

There are several models describing the traffic conditions in a network. They can be roughly divided into two groups:

- simulation models
- assignment models

This categorization is based on the way the end-result is obtained. Simulation models will define the traffic conditions at time \( t + \Delta t \) by extrapolating the conditions at time \( t \). Therefore the result is obtained through time-propagation. Assignment models will calculate the traffic conditions using route-choice and (iterative) assignment techniques.

Simulation models:

There are two different basic concepts for simulation models: Microscopic simulation techniques (Gazis [5]) and macroscopic simulation techniques (Lighthill & Withham [13]). The large number of parameters and the small time-step makes microscopic models unsuitable for network-wide, on-line applications (see e.g. Schwerdtfeger [17]). Macroscopic models are usually used for short motorway sections with several on- and off-ramps (see e.g. Papageorgiou [15][16], and Michalopoulos [14]).

Assignment models:

The main interest for DTM lies in dynamic assignment models. Several approaches are described in recent papers. Some approaches are based on the state-equation (relation between the increase of traffic on a link \( a \) at time \( t \) to the difference between the instantaneous in- and out-flow) See e.g Bin Ran [1], Boyce [2], Vythoulkas [19] and Byung-Wook Wie [3]. Other models are developed as an improvement over static assignment models. See e.g. Hamerslag [8][9][10], Janson [11] and Cascetta [4].
CONTRAM (Leonard [12]) is an assignment model which uses packets for its assignment.

To meet the requirements for DTM we developed a model called: Dynamic Traffic Assignment. The model is based on the model Hamerslag described in "A three dimensional assignment in timespace" [8], and other papers [9][10]. The model has been adjusted to work in real-time, and improvements have been made towards its dynamic aspects.

The model is mainly focused on freeway systems. In a real-time situation it gives a prediction of the traffic situation for a short time period (± 1 hour). The model gives the average development of traffic flow on each link of the network for a number of consecutive time periods; for example 12 periods of 5 minutes. A link could be from a few hundred metres to several kilometres long and a period could be 1 to 10 minutes. The traffic conditions are assumed to be homogeneous on a link in a period.

In an off-line situation the model can be used to gain insight in the performance of dynamic traffic management instruments.

Although this paper describes the model for short-term modelling, there are also models needed for the longer term. For example dynamic route guidance for international traffic in Europe (e.g tourist and trucks) needs models for the longer term (several hours). Also the traditional models for long-term prognoses need improvement due to the increasing effects of congestion.

We believe that the principle of the model described here is suitable for different applications. The user-demands for information will however be different and the information required by the model shifts from actual (on-line) information to historic information when the total time span increases.

2. DYNAMIC TRAFFIC ASSIGNMENT

Like most other dynamic assignment models, the simulation period is divided into intervals of equal length, to which we will refer as periods. For each period the traffic conditions in the network are separately defined. The time-varying travel demand is presumed given.

The following assumptions are taken:
- all travellers are completely informed and taking into account the future congestions
- all travellers choose their minimum cost path
- the traffic conditions on a link for a period are homogeneous
- the traffic conditions on a link for a period are subject to a relation between density, intensity and speed. The nature of this dependency depends on the link, and the period.

The model calculates, in an iterative process, the densities on the network for every link in every period. In every iteration the routes from every origin in every departure period
are calculated. According to these "routes in time" the OD-matrix is loaded on the network. The iteration process starts with initial (free-flow) travel times. For the next iteration new travel times are calculated based on the densities created in the previous iteration. The relation between density and traveltime is determined with speed-density functions.

The iteration process is the same as used for static equilibrium or capacity restraint assignment methods. In each single iteration however the shortest-path finding and the assignment is performed in time. We will call one iteration an "All or Nothing Assignment in Time". One iteration however has no practical meaning due to temporal discontinuities (as Janson calls it [11]). Even if there is only one path in the network several iterations are needed to eliminate the discontinuity between the traveltime used for the route choice and the traveltime resulting from the loaded density.

If the difference is equal to zero a stable solution is reached. Whether this solution provides the dynamic user equilibrium cannot be proven yet.

The "All or Nothing Assignment in Time" ( = 1 iteration) forms the basis of the iteration process and is explained in the following figure:

After one "All or Nothing Assignment in Time" new travel times are calculated based on the resulting densities. These travel times are used to define the shortest path tree for the next iteration. The first iteration is performed with initial (free-flow) travel times. In case of an equilibrium assignment, the subsequent iterations are weighted using:

\[ \rho_{a,p} = (1-\lambda) \cdot \rho_{a,p}^{old} + \lambda \cdot \rho_{a,p}^{new} \]  

Where \( \rho_{a,p}^{old} \) is the result from all previous iterations and \( \rho_{a,p}^{new} \) the result of the last iteration. Until now \( \lambda \) is fixed (e.g. 0.2) or \( \lambda = 1 / (i+0.5) \) where \( i \) is the iteration number.
The iteration process is shown as a flow chart in the next figure:

The process starts with an empty network. The calculated traveltimes are thus free flow travel times. One "All or Nothing Assignment in Time" is performed and the results of this iteration are combined with the results of the previous iteration according to formula one ($\lambda = 1$ in the first iteration). If the stop criterium is not reached, new traveltimes are calculated for all links in the network for every period, based on the densities calculated in the combination at the end of the previous iteration.

The stop criterium is reached if the travel times of the last iteration are the same as the travel times used in the previous iteration.

The iteration process is in principle the same as for static assignment methods.

No mathematical proof of uniqueness or convergence is given. However several experiments with different networks and OD-matrices have shown a convergence in all cases after 20 to 40 iterations. The most iterations were needed for networks with heavy congestion. We believe the iteration process can still be improved.

2.1. NETWORK

The network is defined as a directed graph $G(N,A)$, with node set $N$ and link set $A$. Some of the attributes of a link are defined for each period separate. For example the speed or the traveltime of a link can be different for each time period. The length of the link is the same for each period. Parameters of a link, such as capacity, density, traveltime and speed, are considered to be homogeneous over the length of the link and over the duration the period.

No assumptions are made concerning the link-length or the difference in link-length.

2.2. OD-MATRIX

The travel demand is specified in an OD matrix. The accuracy of the prediction naturally
depends on the quality of the OD-matrix. There are basically two ways to construct the OD-matrix. In an ideal situation the OD-matrix is specified for each period separately. In a less ideal situation the OD-matrix is defined for the complete time-span and is spread over the periods with a departure time distribution.

A compromise can be made by calculating a departure time distribution for each period. Based on the estimated traveltime from each origin to its destinations and an assumed arrival time, the departure time for each origin can be estimated. Based on this estimation a departure time distribution for each origin is calculated.

### 2.3. TRAVELTIME FUNCTIONS

The traveltime for each link in each period is estimated by an iterative process, based on the traffic density on a link. The relationship between intensity, density and speed (traveltime) is described by:

\[ q(t) = p(t) \cdot v(t) \]  

(2)

Usually the traveltime is determined with a traveltime-function or delay-function. This function describes the relation between intensity and traveltime (e.g. BPR function). In a dynamic assignment model a traveltime-intensity function causes some inconsistencies. The function shows an increasing traveltime with growing intensity. However when we observe the intensity, density and speed on a road section in reality, we find a different relation. In non-congested situations traffic will flow freely and with growing density and intensity the traveltime will indeed increase. In congested situations however the speed will drop rapidly and the intensity will decrease while the traveltime still increases.

Therefore, because the intensity does not decrease, a traveltime-intensity function is not suitable to describe a congested situation. A true traveltime-intensity function would show a duality for the traveltime in relation to the intensity (see figure).

The relation between density and traveltime does not show this duality and is therefore more suitable for dynamic assignment. Instead of using the intensity as explanatory parameter the density is used.
The true form of this traveltime-density function is subject to further research. The function used in the model at the moment is a speed-density function used by Smulders [18]:

\[
v(p) = \begin{cases} 
  v_{\text{max}} \cdot \left[1 - \frac{p}{p_{\text{max}}}\right], & 0 \leq p \leq p_{\text{crit}} \\
  \phi \cdot \left[\frac{1}{p} - \frac{1}{p_{\text{max}}}\right], & p_{\text{crit}} \leq p \leq p_{\text{max}}
\end{cases}
\]

In which \(v_{\text{max}}\) is the free-flow speed and \(\phi\) is chosen to make the function continuous at \(p_{\text{crit}}\).

In accordance with the density on a link, the matching speed (traveltime) is found using this speed-density function.

2.4. PATH FINDING

The path finding is done according to the arrival time in each node of the path. The passing time in a node is calculated with the consecutive traveltimes of the preceding links in the path. The traveltime on a link can be different for each period.

The method is best illustrated with the following example:

![Pathfinding with trajectories](image)

The path in time is drawn as a trajectory (distance-time) graph. The x-axis shows the path. The path starts in node A (Origin) and ends in node G (Destination). The y-axis shows the time. The time is divided in 7 periods.

The traveltimes for link A-B is equal to one period. So the traffic leaving node A in the first period, arrive at node B at the beginning of the second period. The traveltime to traverse link B-C is equal to 2½ period. So the traffic arrives at node C halfway during
It takes three consecutive periods to traverse link B-C, the second, the third and the fourth period. Because they experience a different traveltime in every period, the total time to traverse link B-C is the sum of these three traveltimes, taking into account the distance travelled. The three traveltimes added up results in a total time to traverse B-C equal to $2\sqrt{2}$.

In general, the time to traverse link $a$ from node $i$ to $j$ is calculated by calculating the arrival time in node $j$:

- $d_i$ = departure time from node $i$,
- $d_j$ = arrival time in node $j$,
- $T_{per}$ = period length,
- $p_i$ = period number of departure from node $i$,
- $p_j$ = period number of arrival in node $j$,
- $a$ = link between $i$ and $j$.

$d_j$ can be calculated as follows:

$$d_j = \text{time of arrival in period } p_j + \text{traveltime in period } p_j$$

The "time of arrival in period $p_j$" and the "traveltime in period $p_j$" on link $a$ can be calculated as follows:

In each period $p$ a certain distance is travelled on link $a$. This distance, expressed as a fraction of the link-length covered in period $p$, is equal to: $\alpha_p$

$$\alpha_p = \frac{T_{per}}{t_p}$$

In which $t_p$ is the traveltime on link $a$ in period $p$.

Because the moment of departure from node $i$ can be at any time during a period, the fraction of distance travelled in the first period ($p_i$) could be less than $\alpha_p$. The fraction of distance travelled is equal to: $\sigma_i$

$$\sigma_i = \left[ p_i - \frac{d_i}{T_{per}} \right] \cdot \alpha_p$$

The same occurs in the period ($p_j$) in which node $j$ is reached. The distance travelled on link $a$ in that period can be less than $\alpha_p$. Because the end of the link has been reached.

$$\sigma_j = \left[ \frac{d_j}{T_{per}} - (p_j-1) \right] \cdot \alpha_p$$
The distance travelled, expressed as a fraction of the link-length is equal to: $\sigma_j$
So if the period of arrival in node $j$ ($p_j$) is known, the actual arrival time is (following from (4) and (7)):

$$d_j = (p_j - 1) \cdot T_{per} + \sigma_j \cdot t_{p_j} \quad (8)$$

Because the sum of the fractions travelled on link $a$ has to be 1, the period of arrival in node $j$ ($p_j$) can be calculated with (following from 5, 6 & 7):

$$\sigma_j + \left( \sum_{p \neq p_j} \sigma_p \right) + \sigma_j = 1$$

The arrival times in the nodes can be used to define the shortest path tree in the same way as other (static) shortest path finding algorithms.

2.5. ASSIGNMENT

The OD-matrix is assigned for each origin and each departure period according to the calculated trajectories. To calculate the contribution of a traveller to the density on link $a$ in period $p$, the duration of presence on link $a$ in period $p$ is calculated.

The assignment method assumes that all traffic, departing in a period, leaves the origin uniformly spread over the duration of a period. The traffic is assigned according to the surfaces between the trajectory the first car follows and the trajectory the last car follows.

This is best illustrated with the following example:

**Assignment with Trajectories**

The first car leaving from node A in period 1 arrives at node B at the end of period 1. The last car leaving in period 1 arrives at the end of period 2 at node B. All the cars leaving in this period are evenly spread between these two trajectories. So the traffic is
assigned according to the surfaces between the two trajectories. In the first period only 50% is assigned to link A-B. In the second period 50% is assigned to A-B and 50% to B-C. In the third period 100% is assigned to B-C. Since a density is assigned, the total amount of cars present in each period should be 100%, except for the period of departure and the period of arrival. The link C-D shows that there can be a significant difference in traveltime between the last car and the first car. So the distance between cars can increase or decrease.

This method gives an accurate assignment of traffic in time. Test results however showed that after several iterations almost the same distribution of traffic is reached with a simplification of this method. The "trajectory assignment" method. With this method it is assumed that all traffic leaves at the beginning of the period. This method is computationally less demanding, and the iteration process takes care of the spreading of the cars.

The method is illustrated with the following example:

**ASSIGNMENT WITH TRAJECTORIES**

<table>
<thead>
<tr>
<th>Periods</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

The traffic departing from node A in the first period with destination G (e.g.: OD\textsubscript{AG1} = 100), follow the trajectory calculated with the pathfinding. During the first period they pass the link A-B. In the first period the 100 cars are all assigned to the link A-B. In the second period they are present on link B-C. In the third period they are still present on link B-C. In the fourth period they are partly present on link B-C and partly on C-D and D-E. So in the fourth period they are assigned according to each part. Resulting in ±50% to B-C, ±20% to C-D, and ±30% to D-E. Etc..

**In general**: two situations can occur:

1. **Several links are covered in one period.** In this case travellers are only part of a period present on the link, and therefore should only be assigned for this part.

2. **One link is covered in several periods.** The travellers are present on the link during multiple periods and should be assigned entirely for each individual period.
The density on a link \( a \) in period \( p' \) is a summation of the contribution of all OD-pairs \((o,d)\) from all departure periods \( p \) to the link \( a \).

\[
\rho_{a,p'} = \sum_{o,d} \rho_{a,p',o,d,p}
\]

where \( \rho_{a,p',o,d,p} \) which represents the contribution of the OD-pair \( OD_{o,d,p} \) to the density on link \( a \) in period \( p' \) and is defined as:

\[
\rho_{a,p',o,d,p} = \gamma_{a,p',o,d,p} \cdot OD_{o,d,p} \cdot \delta_{a,o,d,p}
\]

where \( \delta_{a,o,d,p} \) denotes the presence of link \( a \) in the shortest path for OD-pair \( OD_{o,d,p} \). The value \( \delta_{a,o,d,p} \) can take is either 0 or 1.

And \( \gamma_{a,p',o,d,p} \) is the fraction of \( OD_{o,d,p} \), contributing to link \( a \) in period \( p' \).

\[
\gamma_{a,p',o,d,p} = \frac{d_{i,p'} - d_{i,p}}{T_{per}}
\]

where \( d_{i,p'} \) and \( d_{i,p} \):

\[
d_{i,p'} = \text{maximum} \left[ d_{i} \cdot (p' - 1) \cdot T_{per} \right]
\]

\[
d_{j,p'} = \text{minimum} \left[ d_{j} \cdot p' \cdot T_{per} \right]
\]

are respectively the departure and arrival time travelling between \( i \) and \( j \) in period \( p' \).

In case the arrival period is equal to the departure period: \( d_{i,p'} - d_{i,p} \) gives the duration of time in which the traffic is present on link \( a \) in period \( p' \). In case the traffic is present on link \( a \) during the complete period; \( d_{i,p'} - d_{i,p} \) is equal to \( T_{per} \) and \( \gamma_{a,p',o,d,p} \) is 1.

Using densities instead of intensities offers, next to a more realistic representation of the delay function, the possibility to check the conservation of traffic present in the system. The amount of traffic present on the network should be the same for each period. Except in the period of departure from the origin and in the period the traffic reaches its destination.

2.6. ADDITIONAL FEATURES OF THE MODEL

For a realistic simulation of traffic some extra features are needed in the model.

Jam building upstream

The density on a link has a maximum. Maximum density represents a no-motion traffic-jam. Once the maximum density has been reached on a link, no more traffic can be assigned to it.

In one iteration a higher density than maximum density is allowed, in the next
iteration however no more traffic is assigned to the overflown link. The traffic is assigned to the preceding link in the path in the same period. With this method the density on the overflown link will decrease after every iteration until maximum density is reached again. Eventually no overflow will occur. Upstream the overflown arc, due to the re-assigned traffic, the traveltimes will increase. Due to the longer traveltimes the density will increase. Depending on the demand this effect will proceed upstream and a queuing effect, which simulates the jam-building upstream, is achieved. Due to the assignment of traffic to a preceding link a discontinuity in flow will appear. However in the final iterations the amount of traffic assigned to preceding links is minimal. The continuity of flow is maintained this way.

**On-ramp link dependency**

At sections of the freeway with on-ramps a less efficient throughput takes place. The more cars enter the freeway at the on-ramp, the more delay occurs. The delay on the freeway is influenced by the ratio of the traffic on the on-ramp and the traffic on the main lanes. If the merging process is without delay the capacity of the freeway would not be influenced. However when the intensity on the on-ramp gets larger the capacity of the freeway decreases, due to an inefficient merging process. By making the capacity of "the section of the freeway with the on-ramp" dependent of this ratio a more realistic behaviour is achieved.

This can be explained with the following figure:

The straight line shows the capacity in a non-delayed merging process, the curved line shows the dependency used in the model. If the demand on the on-ramp close to 0% and the demand on the main road close to 100%, the influence on the capacity is minimal. However when there is 50% demand on the on-ramp and 50% demand on the main road the capacity of the section is lowest. The true form of the dependency relation needs further research.

**Rerouting**

Rerouting measures are implemented as fixed routes between two nodes for certain groups of OD-pairs. The percentage of the traffic to be rerouted can be specified.

**Ramp-metering**

Ramp-metering is implemented as a limited capacity of an on-ramp. Since ramp-
metering shows a more efficient merging process, the on-ramp link dependency explained above can be made less strict.

**Speed control measures**

The speed on certain links in the network can be controlled per period. Also the relation between speed and density can be changed.

**Tidal flow**

By changing link-capacities (to zero) for certain periods, the effects of tidal flow measures can be investigated.

### 3. ON-LINE DATA USAGE

For on-line usage the model needs to adapt its predictions with on-line data.

In the Netherlands a system is planned for monitoring the freeway system. With hundreds of induction-loops in the road, information of the situation on the network will be centrally available, including on- and off-ramps. Every minute, intensity and speed are measured.

The on-line corrections the model can make are divided in three parts:

**Day specific situations**

Every day the OD-matrix and the attributes of the links have to be corrected to the specific day. With the weather the capacity of a link changes. Roadwork can change routes and capacity. Special events can attract more traffic, etc.

**Short term adaptions**

The correctness of the prediction depends on: "the correct delays, the correct route-choice, a correct number of cars on the freeway, and a correct number of cars entering the freeway in the future periods". Based on the information of the monitoring system you can easily adapt the speed-density functions to specific situations. With information of on- and off-ramps the number of cars present on the freeway system can be corrected. Whenever an incident occurs the attributes of the links involved need correction. The new capacity is easily measured.

**Long term adaptions**

On longer term the model can be tuned by correcting the OD-matrix and route-choice. This is however an off-line process and needs different models.

This part of the model still needs a lot of research. At the moment the research is mainly focused on realistic modelling of traffic-flow on networks. At this time parallel research is carried out for on-line estimation of OD-matrices at Delft University of Technology.
4. IMPLEMENTATION

The model is implemented on a Silicon Graphics workstation in a window environment. Most of the model's input and output is presented with graphics. This makes the model very user friendly. The program has shown to be a very powerful tool to investigate and validate the model.

4.1. SOFTWARE

The program is written in C and works with a window environment. The model can display several windows with output of the model. Several calculations can be saved and shown simultaneously. Model parameters and attributes of links can be changed for each calculation. This makes the program very suitable for studying the influences of different parameters.

The main window shows the network. Different roadtypes are shown in different colours. Links can be selected and attributes changed.

This window shows a small test network, which is the freeway system in the south-west of the Netherlands (Randstad).

The results of the calculations are presented in different windows:
- The density on the network is shown in a film like manner in which the densities are shown in different colours and/or numbers.
- For each Origin Destination pair the trajectories can be displayed. Changes over iterations or changes over periods are shown.

The window shows for a given OD-relation the path in time. The x-axis shows the path. The y-axis shows the time in periods.

- Each link can be viewed separate in a line graph. Three different results can be shown in the same graph. Different attributes (e.g. density, intensity, speed, traveltime) can be chosen with popup menu's.

The window shows the situation on one link. The x-axis shows the time and the y-axis the chosen attribute (speed, density, traveltime or intensity).
The input of the model is also shown in graphs and can be changed interactively.
- The departure time distribution can be changed in a bar chart
- The speed-density function can be changed in a line graph, and different types of functions can be chosen for different roadtypes.

4.2. HARDWARE

Simultaneously with the development of the model, research is done towards the design and construction of a special purpose parallel computer for assignment calculations. In case the on-line implementation of the model will be too demanding on computation time this research can form a cost-effective answer [6][7].

5. TESTS: COMPARISON WITH A MICRO SIMULATION MODEL

The best way to validate a model is to test it in a real-life situation. However the model can be tested in some simple situations first. If the model does not perform well in these simple situations, simulations of large networks are useless.

Some simple basic tests have been done to validate the traveltime estimation of the model, which is important for a good behaviour of the model. No tests on route choice are included due to lack of comparison material.

Results of the following tests are reported:
- a 10 km roadway with a bottleneck
- a 10 km roadway with an on-ramp

The model is in principle not intended to reproduce each kilometre roadway in detail, but it should be able to simulate a correct in- and outflow and traveltime for the entire roadway.

These tests are chosen for two reasons:
1) They can be found in any network and are crucial for the traffic behaviour in a network. In case a bottleneck situation is included in a larger network the model is capable to simulate the delay of traffic due to this bottleneck and the effects up- and downstream in the network.
2) Test results of these situations are well known, which makes them easy to validate. We could have included test results of the network at page 13 or even larger networks, but no comparison material is available for these networks.

The test results were compared with a (validated) micro-simulation model, FOSIM. The results of the micro simulation were aggregated to 1 minute data. The tests pushed the model a bit beyond its possibilities, because an appropriate time period for network applications would be about 5 minutes.

The tests show the capability of the model to simulate some simple basic test situations. The model however is intended for large networks and should not be confused with micro-simulation models and its purposes. Whether the model behaves correct for large networks is part of further research. Tests at the freeway system of Amsterdam are planned (starting in August 1992).
A bottleneck
A ten kilometre road with three lanes running into two lanes. The road is made with ten links of one kilometre. The simulated time is twenty minutes, with 20 periods of 1 minute.

The situation:

The traffic is travelling from left to right. Every minute a number of cars are entering the section according to the following table:

<table>
<thead>
<tr>
<th>period</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>#cars</td>
<td>40</td>
<td>56</td>
<td>72</td>
<td>88</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>104</td>
<td>88</td>
<td>72</td>
</tr>
</tbody>
</table>

The next graphs show the intensity (veh/hr) on a link in time on 6 of the ten links. Along the x-axis the periods in time are displayed. The fifth graph is the first link with two lanes. The solid line displays FOSIM results and the dashed (grey) line displays our results.
The next six graphs show the matching speeds (km/hr) on the same links.

Comments:

At 1000m the graphs show the intensity and the speed of the traffic entering the network. At 3000m the same shape shifted in time is observed. Due to a queuing in the bottleneck (at 5000m) the intensity at 4000m drops at the 12th minute. The matching speed decreases to 20 km/hr. Up to the 14th minute the number of cars entering the bottleneck decreases and the intensity increases again. After the 14th minute the number of cars entering the bottleneck decreases further and the intensity decreases too, while the speed further increases. At 4500m the same process shifted in time is observed. Downstream the bottleneck a steady intensity of approximately 5000 veh/hr is observed. This intensity is shifted in time on the last graph at 7500m.

The behaviour at the bottleneck is not reproduced in detail by the model, but downstream the bottleneck the results matches good. So the delay due to the bottleneck is reproduced. The queuing effects at 4000 m are with FOSIM larger as with our model.
An on-ramp
A ten kilometre road with two lanes with an on-ramp at the 5th kilometre. The road is made of ten links of one kilometre, the on-ramp starts at the 3rd kilometre and merges at the fifth kilometre. The simulated time is one hour, with 60 periods of 1 minute.

Situation:

The traffic is going from left to right. Every minute a number of cars are entering the section according to the following table:

<table>
<thead>
<tr>
<th>period</th>
<th>main</th>
<th>ramp</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>43 43 43 43 43 43 43 43 43 43 43 43 43 43 43 43</td>
<td>12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12</td>
</tr>
<tr>
<td></td>
<td>18 26 34 39 38 34 38 34 29 23 23 14 2</td>
<td>16 43 29</td>
</tr>
<tr>
<td></td>
<td>17 18 43 43 23 23</td>
<td>19 20</td>
</tr>
<tr>
<td></td>
<td>14 2</td>
<td></td>
</tr>
</tbody>
</table>

The next graphs shows the intensity (veh/hr) on a link in time on 6 of the ten links. At 3500 meter and 4500 meter the intensity at the on ramp is in the second lower line. The solid line displays FOSIM results and the dashed (grey) line displays our results.
The next six graphs show the matching speeds (km/hr) on the same links.

Comments:

The first graph at 1500m shows the traffic entering the network. For the main road a steady intensity of 3000 veh/hr is entering the network for a total time of 20 minutes. At 3500m the same intensity is observed shifted in time. The traffic entering at the on-ramp is shown in the second line in these graphs. A steady intensity of 800 veh/hr enters the on-ramp for 10 minutes, the next 4 minutes the intensity increases to 2200 veh/hr and then it decreases to zero. At 4500m the traffic does not merge yet and the same intensity shifted in time on the main road can be observed. At the on-ramp however a queuing is observed at 4500m. Speeds are very low. At 5500m the traffic has merged and a intensity of approximately 4000 veh/hr is observed. The same intensity, shifted in time, with a slowly increasing speed is observed at 6500m and 8500m.

The results at the on-ramp are not reproduced in detail. FOSIM has more queuing at the on-ramp then at the main road. Our model queues the same on the main road as at the on-ramp. This can be good observed at 4500 m. The total effect of an on-ramp is fairly good described.
6. CONCLUSION

The model "Dynamic Traffic Assignment" can predict a realistic flow in a network. There are no constraints on the network definition, or the time period length. The continuity of flow and the conservation of traffic is maintained.

The model can simulate traffic jams and the effects up- and downstream. The relation between link occupancy and traveltime is modelled with speed-density functions, instead with intensity- traveltime functions. A situation with congestion, where a low intensity and a high traveltime is valid is included this way.

The effects of Dynamic Traffic Management instruments, such as: "rerouting and (network coordinated) ramp-metering, speed control and tidal-flow can be investigated.

The model is intended to work with large networks (see e.g. figure on page 13). To validate the model, it has been tested on a few small basic situations. The test results showed that the model can reproduce the situations satisfactory. The situations were not reproduced in detail, but that was not expected.

The results on the basic situations, which occur in every network, makes the correct working on a large network plausible.

Further validation of the model with data from the freeway system of Amsterdam is planned. With the research towards special purpose hardware and on-line updating, possibilities for Real-Time Traffic Management are within reach.
REFERENCES

A special purpose parallel computer for traffic simulation.

H.J.M. van Grol and A.F. Bakker

70th TRB annual meeting, Washington
A SPECIAL PURPOSE PARALLEL COMPUTER
FOR TRAFFIC SIMULATION

H.J.M. van Grol
A.F. Bakker

Delft University of Technology
Faculty of Applied Physics
Physics Informatics/Computational Physics
Lorentzweg 1, 2628 CJ Delft, The Netherlands
Fax: (31) -15- 786081

July 1990
Revised
December 1990

Prepared in response to the call for papers by the Transportation Supply Committee (A1C05) Transportation Research Board for the 70th Annual TRB Meeting.
ABSTRACT
A SPECIAL PURPOSE PARALLEL COMPUTER
FOR TRAFFIC SIMULATION

H.J.M. van Grol
A.F. Bakker

Delft University of Technology
Faculty of Applied Physics
Physics Informatics/Computational Physics
December 1990

Traffic simulations are widely used for long and medium term forecasting of traffic. Now-a-days, with the growing problem of queues during rush hours, the demand arises for dynamic traffic management and, therefore, short term forecasting. Apart from the need for new, adapted dynamic assignment models the second important part in this new development is the required computational power. Most commercially available computers are unable to produce an accurate forecasting for the next 5 to 30 minutes within the desired time and budget.

Analysis of existing assignment models and their most time consuming part, shortest path finding, has shown that the main structure of the models can be parallelized. The use of parallelism thus seems apparent. Several general purpose parallel computers, such as N-cubes, are commercially available. However, apart from being more expensive, they loose a large part of their expected performance by the amount of 'necessary' inter-processor communications. Additionally, the programming of such computers has turned out to be more difficult than expected.

We propose a 'simple' linear array of typically 16 processors, the so called Linear Processor Array (LPA). This one-dimensional parallel computer with highspeed buffered interconnections between each pair of neighboring processor boards, parallel accessible by both a control board and a general purpose host computer, forms a transparent concept for the programmer. The optimally configured boards together with the high speed intercommunication allow a cost/performance improvement of a factor 100 compared to a minisupercomputer like the Convex C1.

For future developments of on-line traffic control, routeguidance, ramp metering etc. the LPA should be a powerful tool.
1. INTRODUCTION

Many transportation planning and control activities involve traffic simulations for medium and long term forecasting. The existing models are efficient enough to do this within the desired time. With larger networks the simulation time increases rapidly. Furthermore, with the growing importance of dynamic traffic management (since building new road is considered not to be a desired solution anymore), the demand for short-term forecasting grows. Existing models do not meet these demands, neither in time nor in accuracy.

Apart from the development of new adapted dynamic assignment models, the time-constraints require us to think about the necessary computing power. With most commercially available computers it is not possible to produce an accurate forecasting for the next 5 to 30 minutes within the desired time and budget. The use of special purpose computers will open ways to a desired solution.

In this paper we will try to give a cost-effective answer to this problem. To start with, we will investigate the existing assignment-models and look into the development of the new dynamic models. In chapter two this will lead to a simplified representation of assignment models in which the algorithm is analyzed to find possible optimizations. We will show that the main structure of the models can be parallelized. After a short discussion about the use of general purpose computers and an introduction to special purpose computers, we will discuss the use of special purpose hardware for traffic routing problems, in chapter three. This will lead to the proposed Linear Processor Array which will be explained in detail. In chapter four we will present some preliminary results and we conclude with the expected performance improvement.
2. TRAFFIC SIMULATION MODELS

Traffic simulation models, or, more specifically, network assignment models, have evolved from the simple, static assignment models (all-or-nothing assignment) to the more complex dynamic assignment models (threedimensional assignment model).

In the past the models have mainly been used for long and medium term forecasting and play an important role in the transportation development schemes. Long computations are manageable in these cases. The computation time, however, will grow rapidly with more complex models, larger networks etc.

A third important objective, short-term forecasting, has recently risen. Short-term forecasting is used for dynamic traffic management, which deals with on-line network-wide traffic control, routeguidance, ramp metering etc. Short-term forecasting, in the range of 5-30 minutes, will impose a time-constraint on the simulations.

All together there exist enough reasons to justify a search for methods to speed up the calculations. Improvements can be made on both the algorithm side and the computer side. In the remainder of this chapter we will concentrate on possible improvements on the algorithm side.

First we take a look at the collection of assignment models, and find a generalized form. Although the reader will be familiar with the models, we will summarize them.


The simplest model is the all-or-nothing model. This model, however, does not take into account the multiple routes different travelers, from one origin-destination pair, take in reality; nor, does it take into account the load-dependence of traveltimes or time-dependence in general. The remaining models can be split into three categories, based on their resolution of these three deficiencies. The corresponding algorithms are, respectively, the stochastic, the equilibrium and the dynamic assignment algorithms. We will now list a few examples of each category.

**Stochastic Assignment**

The principle of stochastic assignment is to simulate the route-choice by assuming the traveler does not have perfect knowledge about his route.

There are two essentially different stochastic assignment methods. The first one will iteratively perform all-or-nothing assignments, based on randomly adjusted arc-traveltimes. The assignment at each iteration is a fraction (1/number of iterations) of the total assignment. The maximum number of possible routes between one origin and destination is equal to the number of iterations. The second method consists of only one or two iterations but results in a large number of taken routes.

Both methods have been described in detail in Van Vliet [5]. These methods are mainly used in non-congested networks.

**Equilibrium Assignment**

In this category the traveltimes depend on the trafficload. Multiple routes are found at equilibrium. The equilibrium is based on the principle of Wardrop [6]:

*The traveltimes of all used routes between an origin and a destination are equal to or smaller than the traveltimes of the unused routes.*

*No traveler can shorten his journey by switching to another route.*
There are several ways to reach the equilibrium: methods with a theoretical background and more ad hoc methods. The practical difference is the number of iterations necessary to reach the equilibrium and the amount of work in each iteration. At the end of each iteration travel times are recalculated. Methods include:

- **Equilibrium:**
  The equilibrium assignment is computed using optimization techniques. At each iteration the old assignment is compared with a new all-or-nothing assignment. The resulting assignment is the optimal weighting between these assignments. This is done until no significant improvement can be made.

- **Capacity Restraint:**
  The network is iteratively loaded and unloaded, according to a predefined recipe.

- **Fixed Demand Incremental:**
  The network is iteratively loaded with a decreasing amount $1/(i+1)$, where $i$ is the iteration number.

**Dynamic Assignment**

Instead of observing the assignment in one time-interval, assuming the total amount of traffic on a route to be evenly spread along the route, we take into account the time aspects. In other words, by dividing the observed interval into a number of periods, travelers are only contributing to the load of an arc in the period in which they are really using it. Since travelers do not start their journey all at once, the OD-matrix is defined for each period separately.

One dynamic assignment method is described by Hamerslag [2][3]. For each period an all-or-nothing assignment is computed, resulting in an all-or-nothing assignment in three dimensions (time and space). Techniques from the previous categories are used to find a realistic assignment. The number of iterations, required in the former algorithms, is now multiplied by the number of simulated periods. This method can realistically simulate traffic in congested networks. It is possible to simulate temporary decreasing capacities caused by, for instance, accidents or ramp metering.

**Resume**

Summarizing, we can conclude that all algorithms are repetitive all-or-nothing assignments. The workload (computational load) comes almost entirely down to the all-or-nothing assignment. To speed up the computation, hence, we must concentrate on the optimization of the all-or-nothing assignment.

**2.2. The All-Or-Nothing Assignment.**

The all-or-nothing assignment, as the name suggests, assigns all traffic to a single shortest route and none to the others. It will, therefore, compute the shortest path between each O(origin)-D(estination)-pair and assign the associated amount of traffic to each consecutive arc along that path.

```c
# Compute the all-or-nothing assignment
for each OD-pair in the network {
    Compute the shortest path between Origin and Destination
    Assign the path
}
```
It is more efficient to calculate shortest path trees (spt's), because the corresponding parts of the shortest paths from one origin to several destinations can be long. Thus:

```python
# Compute the all-or-nothing assignment
for each origin in the network {
    Compute the shortest path tree
    Assign the spt
}
```

We will now examine the algorithm in more detail.

### Shortest Path Finding Algorithms

There are four main techniques of finding shortest paths:

- Heuristic technique: one-to-one
- Algebraic technique: all-to-all
- Combinatorial technique: one-to-all
- Optimizing technique: one-to-all

For our purpose the heuristic technique is not suitable (one-to-one) and the algebraic technique is inefficient and uses much memory. This leaves the combinatorial and the reoptimizing techniques. The reoptimizing technique is the fastest technique and is based on the reoptimization of an existing spt for origin \( u \) to a new spt for origin \( v \). As it is a more complex algorithm, consumes a large amount of memory, and is only about two times faster, one might prefer to use the combinatorial technique.

The algorithms using the combinatorial technique can be split into two groups, the labelcorrecting and the labelsetting algorithms. The labels are the values associated with each node in the network representing the cost of traveling coming from the present origin-node (root of the spt).

The following algorithm is a general shortest path finding algorithm:

```python
# Compute the shortest path tree
Initialize \( d_v = \infty, \ v \in N - \{r\}, \ d_r = 0 \)
Init_Q \( Q = \{r\} \)
While ( \( Q \neq \emptyset \) ) {
    Select_node select \( u \in Q; \ Q = Q - \{u\} \)
    for each \( (u,v) \in FS(u) \) {
        if \( d_u + c_{u,v} < d_v \) {
            \( p_v = u \)
            \( d_v = d_u + c_{u,v} \)
            Update_Q \( Q = Q \cup \{v\} \)
        }
    }
}
```

For an explanation of the symbols used, see appendix.
The tree is built from the root; therefore, these algorithms are also called treebuilder algorithms. The set Q contains the nodes that need to be examined. All algorithms are similar except for the way Q is maintained in Init_Q, Select_node and Update_Q. The efficiency of the algorithm depends on the way this is done. The minimum computation time is reached when Init_Q, Select_node and Update_Q cost minimal time.

Well known algorithms are Moore [12], which is a labelcorrecting algorithm and Dijkstra [8], which is a typical labelsetting algorithm.

A more detailed description can be found in [7,9,10,11]. A simple labelsetting algorithm is one that simply sorts the entries in Q, (S-ord). It has a time complexity of \( O(m,n) \), where \( m \) is the number of arcs in the network and \( n \) the number of nodes. The fastest algorithm using the combinatorial technique is a threshold algorithm (T-calc) which combines the good qualities of both the labeling methods. Although, in principal, the computational complexity of this algorithm is \( O(n.2^n) \), and thus about the worst possible, in practice it behaves like \( O(n) \) and is robust. For the maintenance of \( Q \) it uses a combination of methods: address calculation and a lifo-procedure (last-in-first-out) (see van Grol & Bakker [11]). In this way the algorithm dependent parts (Init_Q, Select_node and Update_Q) cost minimal time.

The differences between the methods increase with growing network size.

Having reduced the time complexity of the spt to \( O(n) \) we will now examine the time complexity of the assignment part.

Assignment

The simplest way of assignment is to follow the path between Origin and Destination and assign the associated amount. The following is the procedure to assign one shortest path tree \( (q_a, a \in M \text{ contains previously assigned loads}) \):

```plaintext
# Assign the spt for Origin
for each Destination in the network {
    u = Destination
    while u is not equal to Origin {
        \[ q_{p,a} = q_{p,a} + OD(Origin, Destination) \]
        u = pu
    }
}
```

In this way the arcs near to the origin are assigned many times. The time complexity is \( O(n.l_r) = O(n.n) \), where \( l_r \) is the mean number of arcs in a path.

A more efficient method can be obtained by simultaneously assigning several OD elements to an arc. Supposing that \( l_u \) is the level in the tree (the number of nodes counted from the origin), we can sort all nodes according to their level. Starting at the highest level, we can execute the following procedure.
# Assign the spt for Origin
\[ K = \{1 \ldots n\}, \quad S_u = 0, \ u \in N \]
# Sort the nodes in set \( K \) top-level down

\textbf{Sort} \( K \)

for each node \( u \) from \( K \) {
\[ q_{p^*u} = q_{p^*u} + OD(\text{Origin},u) + S_u \]
\[ S_{p^*} = S_{p^*} + OD(\text{Origin},u) + S_u \]
}

The set \( K \) contains all nodes with their levels. This reduces the complexity of the assignment to \( O(n) \) but sorting \( K \) has a complexity of \( O(n \log n) \). By using an addressable array, \( L(.) \), to sort the nodes by their level we can reduce the total complexity to \( O(n) \). The algorithm then becomes:

# Sort levels of spt for Origin

\[ L(.) = 0 \]

for each node \( u \) {
\[ L(l_u) = L(l_u) \cup \{u\} \]
}

# Assign the spt for Origin

\[ S_u = 0, \ u \in N \]

for each level top down {
for each node \( u \) in the set \( L(\text{level}) \) {
\[ q_{p^*u} = q_{p^*u} + OD(\text{Origin},u) + S_u \]
\[ S_{p^*} = S_{p^*} + OD(\text{Origin},u) + S_u \]
}
}

\textbf{Resume}

A complete description of the algorithm can be found in the appendix. The overall computational complexity (path finding & assignment) for one spt is now \( O(n) \). This means that a number of operations is executed on each node and \( O(n) \) is thus the minimum complexity. Only the number of operations can now be minimized. The computational complexity of the all-or-nothing assignment is \( O(n^2) \). A good implementation of this algorithm is the best we can do. A major improvement could only be achieved by an implementation in assembler.

When we take a distant look at the algorithm we can make two observations: first of all we can see that the computations of the spt’s are independent. This allows the use of parallelism. The second observation is that the all-or-nothing assignment is mainly a dataflow problem. All network data will flow through the algorithm several times, while the number of operations on the data is minimal. Bearing this in mind we will discuss the possible use of special purpose hardware in the next chapter.
3. SPECIAL PURPOSE HARDWARE

First we will discuss the use of general purpose computers in traffic simulations in paragraph one, since their limitations motivated the current research of using special purpose hardware. The principles of the special purpose computers are described in paragraph two. The third paragraph will focus on dedicated architectures for traffic simulations. After an introduction to the proposed Linear Processor Array (LPA) in paragraph four, we will conclude with a more detailed description of the LPA, dedicated for traffic simulations, in paragraph five.

3.1. General Purpose Computers

Standard traffic simulations usually run on general purpose computers, such as PC's, workstations and mainframes. For small networks and simple algorithms the turn around time of the simulations is satisfactory for most applications. Moving from PC to mainframe or minisuper workstations significantly improves performance and visualization, but the demand for larger networks and more complex assignment algorithms requires supercomputer power. However, the cost and the limited availability of supercomputers eliminate this option.

In general, commercially available computers were designed to solve all problems, and are not tailored to efficiently solve a typical problem such as found in traffic simulations. To improve the cost/performance ratio, or to bypass hardware limitations of general purpose computers, one can design and build a special computer, the architecture of which maps perfectly on the problem or algorithm involved. This approach can be considered a low cost alternative for supercomputers.

3.2. Special Purpose Computers.

Special purpose computers are designed to efficiently carry out a particular task at supercomputer speed. In general, they cannot handle any other task, or, if they can, the performance will be poor. However, a design can cover a class of problems, and thus can be used for a wide range of applications without performance penalty. The design of a special purpose computer is based on the problem(s) or algorithm(s) to be used, and allows an architecture that explores parallel and pipelined operations wherever applicable. It allows problem dependent memory organization, problem adapted basic instruction set, etc., with the purpose to improve the total performance of the computer.

Special purpose computers range from (a) single purpose to (b) multi-purpose computers and they differ in the flexibility of programming them.

(a) Single Purpose Computers.

Single purpose computers have a basic instruction-set that will only cover the operations required for the task it has been designed for. This approach allows the algorithm to be hard-wired, which guarantees an optimal speed. Fixed-wired parallel and/or pipelined architectures restrict the flexibility to modify the algorithm, but leave open the possibility to vary enough parameters to motivate the effort to design such a machine. The cost/performance ratio of this type of computer is low (factor 100 better than supercomputers) and they are 24 hours/day available for the computer experimentalist. A variety of single purpose computers
have been successfully exploited in signal processing and computational physics [13,14,15,16].

(b) **Multi-purpose computers**

Multi-purpose computers are designed to efficiently solve a class of problems rather than a single problem. These architectures reflect the common property of the algorithms involved and can be programmed to solve a problem from the class of problems it was designed for. High-level languages are used (C and F77) to program the computer.

The speed is obtained by using many processor nodes interconnected by a communication network that is suitable for the class of algorithms involved. The architecture of the nodes is kept simple but effective to allow the construction of efficient compilers. The choice of processor, memory structure and size, interconnection network and wordlength characterize the multi-purpose computer.

They can be shared memory or distributed memory machines running in Single-Instruction Multiple-Data (SIMD) or Multiple-Instruction Multiple-Data (MIMD) mode. Flexibility and programmability of these computers are traded for ultra-speed as in single purpose computers. However, the newest commercially available microprocessors are fast, the architecture is scalable, and can result in a cost/performance ratio improvement by two orders of magnitude compared to supercomputers.

So-called "general purpose" parallel computers, such as N-cubes, may use fast processor nodes, but their memory architectures and their slow interconnection networks do decrease the overall performance dramatically. Only between 10 and 20% of the advertised peak performance is reached by careful programming. Existing parallelizing highlevel language compilers are still far from ideal. Automatic decomposition of sequential program flows of problems that are often parallel by nature is not an efficient way to obtain fast codes for parallel computers.

In practice the user has to choose a network topology and program the nodes to use that network efficiently. Often topologies are chosen to be ring structures, so that the programmer can implement his algorithm without being distracted by more exotic topologies. Still, the node interconnections are slow because of their all-topology structure.

3.3. A Special Purpose Computer For Traffic Simulations.

In order to select or design a computer for traffic simulations, the algorithm has to be examined for possible parallel and/or pipelined operations. As the algorithms involved are still in development, thus demanding flexibility, we will not consider a single-purpose computer.

Decomposition of the total problem into coupled parallel processes is a way to find a scalable parallel architecture. The node architecture, memory size and interconnection channel will then determine the final computer.

As we showed in chapter two, a time critical part of the algorithms used in traffic simulations is the calculation of \( n \) shortest path trees (spt's), where \( n \) is the number of nodes in the network. We concluded that the spt's can be independently calculated, thus allowing us to decompose the problem into \( n \) problems of one spt. Using \( n \)
processor boards, we can, hence, calculate all the spt's in parallel. This will improve the simulation speed by order \( n \). In a parallel computer, with \( P \) processor boards, we can calculate \( n/P \) spt's on each board, which gives a speed-up factor of \( P \). The latter solution is preferable for reasons of scalability, especially for large \( n \).

Each processor board will need all network information, thus \( P \) times the amount of memory needed to store the network is the minimal total memory size. Minimizing \( P \) to keep memory costs down is compensated by making the processor boards as fast as possible, and keeps the overall performance high.

In the assignment schemes the above decomposition holds too. However, to find the total load per arc, we must accumulate the partial arc loads. We need, therefore, to efficiently interconnect the processor boards. Here we can use a pipeline structure where each processor board is one pipe stage of the whole pipe. Consequently, we order the processor boards in one string to construct this pipeline. One highspeed data channel from each board to its neighbor board is sufficient to obtain a fast pipeline. The accumulated arc loads are collected in the last pipe stage (last processor board in the string). To start the next iteration the updated network has to be broadcasted to all processor boards. When all processors are connected to a common bus, broadcasting can be accomplished using one talker and \( P \) listeners on this bus, see figure 3.1.

Before going into details of the design, we first discuss a more generalized architecture that will cover the above architecture ideas for traffic routing simulations, but can be used for other purposes as well (multi-purpose computer).

![Fig 3.1. Data-flow in the traffic assignment problem mapped on an array of P processor boards.](image)

### 3.4. The Linear Processor Array (LPA) Architecture.

A LPA is a one dimensional array of identical processor boards, each of which is connected to its two neighboring boards only by a data bus. In addition, the boards share a common data-, address- and control-bus which is also interfaced to a general purpose host computer. One processor board is configured as a control board, which supervises the chain of processor boards through a special control bus, see figure 3.2.

A large class of problems in computational physics (both authors work in this field) can be solved using this parallel architecture. A natural domain decomposition allows the mapping of different subdomains on different processors. All subdomains can be processed simultaneously. In general the calculations in a subdomain need data from the other subdomains, but when "local environment problems" are involved, such as in finite difference calculations, only the directly neighboring subdomains are contributing to the results. For this class of problems we can decompose the domain by slicing...
in just one direction. Each domain slice is mapped onto one processor, and each processor communicates only with its two neighboring processors. The host passes data and programs to the LPA and collects results of the calculations via the common bus. The processor architecture, the local memory and the interface to the two neighboring processor boards are optimized to tackle the problem they are designed for.

3.5. The LPA For Traffic Simulations.

The LPA architecture is well suited for the traffic routing problem. Programs and network data, which are (mostly) identical on the different boards, can be broadcasted to the processor boards. Each processor node requires enough memory to contain all network information and the locally calculated results. At the start of every iteration (all-or-nothing assignment), the network information on all the processor nodes is identical. Then, every node starts to calculate its share of the total number of shortest paths instructed by the control board. Clearly, this calculation is intensive. A fast processor is needed for all floating point calculations. In addition, since a lot of data is involved, the memory interface is important. Bus contention should be avoided at all times. This is partly solved by the parallel approach taken, with independent nodes and distributed memory, but a fast memory interface, tuned for this particular problem is still called for.

Figure 3.2. Global architecture of the Linear Processor Array.

The processor needs to be selected on grounds of floating point calculation speed primarily. The fastest processor available today is the Intel i860 Microprocessor. It contains a core unit, floating point unit and Instruction & Data caches on one chip. Theoretical speeds are 40 MIPS and 80 MFLOPS. Practical speeds using high level languages such as C and Fortran are in the order of 15 MFLOPS. Hand coded assembly is capable of performing between 30 and 80 MFLOPS, depending on the algorithm, and the amount of vectorization and pipelining that can be employed. Furthermore, the Intel i860 can execute integer and floating point operations in parallel and has a 64 bit wide memory bus.

Because of the speed at which the processor processes data, the memory interface has become the only bottleneck. The rate at which data can be retrieved from and stored
into memory determines the overall processing speed. The network information is typically scattered through memory, thus rendering the on-chip (small) data cache practically useless. Therefore, the interface to the dynamic memory (use of static memory only would be too expensive) is vital.

At the end of each iteration, every board has calculated part of the load for every arc. These partial loads can be accumulated in a pipelined fashion using the connections to the neighbor boards: every board receives the partial loads from its right neighbor, adds its own partial loads, and hands the results to its left neighbor. Finally, the accumulated partial loads are handed to the control board (which is the leftmost board), which can start the next iteration by broadcasting the updated network.

The connections between the boards are realized by First-In First-Out buffers (FIFO’s), which are 64 bit wide memory components that move data in receiving order to the neighbor board (size of buffer is several kByte). The FIFO’s are used to automatically synchronize the asynchronous processors, and buffer data sent between them. This allows the processor nodes to act autonomously and send data whenever ready. The control board is not needed to synchronize nodes or buffer data. All processor nodes can use their FIFO’s concurrently, allowing for maximum throughput.

**4. RESULTS**

In chapter two we have shown how the performance of traffic simulation programs can be improved and in chapter three how we can use special purpose hardware to reduce the computing time and the cost involved. In this chapter we will define the improvements expected by using special purpose hardware.

We will first consider the improvement by parallelization. Using an architecture by the LPA-concept as we propose, with 16 i860-based processor boards, we can expect to gain an improvement of a factor 16 compared to one i860-based processor board. Although this seems very obvious, most computers with parallel processors are unable to improve their performance by the number of processors they use (compared to one processor), see [1,4]. The improvement is justified by the negligible overhead in inter-board communication. The inter-board communication, required to accumulate the arc-loads and to broadcast network data, does not increase with the number of processor boards and is small compared to the total computation.

An operating system running on a computer, allows the users to use all kinds of facilities, such as file-support, I/O in general, multi-tasking, scheduling, timing etc. Using such an operating system will decrease the overall performance of the system. By avoiding most of these facilities we can gain some performance improvement. The operating systems running on the LPA-nodes will allow the minimum amount of facilities to run the problem efficiently.

Secondly we will compare the expected performance of the LPA with some general purpose computers that are commonly used. The i860 is a fast processor as explained in the previous chapter. The processor alone is already in competition with several fast general purpose computers. We have executed some test runs on the Intel Microprocessor Software Development system: STAR860, which is an AT-386 with an i860 CPU based add-on card. This board is not optimally configured, and we can thus improve on the performance in the final design.
As comparison we have run tests on a Convex CI, several differently configured Silicon Graphics (SG, R3000, 33-25-20 MHz, 64kByte cache), a MicroVAXIII, a MicroDutch (68020), an Hewlett Packard Workstation (HP, 68030) and a Personal Computer (AT-386, 25Mhz, 64kByte cache). The Convex CI is a vector processor with an architecture resembling a Cray supercomputer. Though not the fastest, it is a widely used computer.

The test-programs where programmed in C. We have used two shortest path tree algorithms, S-ord and T-calc, see paragraph 2.2. The assignment was done in two different ways, the 'simple' assignment and using levels, also described in paragraph 2.2. For the implementation of the algorithms we can use pointers or indexing. The calculations can be done in integers or floating point notation.

The resulting computing times are given in tables 4.1-4.3.

**Table 4.1. Computing times in seconds on a number of computers.** The program, either T-calc or S-ord is implemented with pointers or indexing and the calculations in either integer or floating point notation. The assignment is implemented with the use of levels. The network size $N=3347$. HP is Hewlett Packard, SG is Silicon Graphics.

<table>
<thead>
<tr>
<th>data-struct</th>
<th>floating point</th>
<th>integer</th>
<th>floating point</th>
<th>integer</th>
</tr>
</thead>
<tbody>
<tr>
<td>spt-alg</td>
<td>T-calc</td>
<td>S-ord</td>
<td>T-calc</td>
<td>S-ord</td>
</tr>
<tr>
<td>Microdutch</td>
<td>4307</td>
<td>6155</td>
<td>2626</td>
<td>2989</td>
</tr>
<tr>
<td>AT-386</td>
<td>1485</td>
<td>2137</td>
<td>733</td>
<td>674</td>
</tr>
<tr>
<td>MicroVAXIII</td>
<td>1180</td>
<td>1712</td>
<td>1119</td>
<td>1324</td>
</tr>
<tr>
<td>HP</td>
<td>1149</td>
<td>1562</td>
<td>789</td>
<td>764</td>
</tr>
<tr>
<td>Convex CI</td>
<td>411</td>
<td>535</td>
<td>315</td>
<td>382</td>
</tr>
<tr>
<td>SG (20 MHz)</td>
<td>266</td>
<td>315</td>
<td>237</td>
<td>273</td>
</tr>
<tr>
<td>SG (25 MHz)</td>
<td>260</td>
<td>302</td>
<td>238</td>
<td>270</td>
</tr>
<tr>
<td>SG (33 MHz)</td>
<td>161</td>
<td>189</td>
<td>139</td>
<td>164</td>
</tr>
<tr>
<td>STAR860</td>
<td>190</td>
<td>202</td>
<td>185</td>
<td>192</td>
</tr>
</tbody>
</table>

**Table 4.2. The difference in computing times between using a simple assignment method and one using levels.** The times are given for two spt-algorithms and on several computers. The network size $N=3347$. SG is Silicon Graphics.

<table>
<thead>
<tr>
<th>data-struct</th>
<th>pointer</th>
</tr>
</thead>
<tbody>
<tr>
<td>assignment</td>
<td>T-calc</td>
</tr>
<tr>
<td></td>
<td>levels</td>
</tr>
<tr>
<td>SG (20 MHz)</td>
<td>266</td>
</tr>
<tr>
<td>SG (25 MHz)</td>
<td>260</td>
</tr>
<tr>
<td>SG (33 MHz)</td>
<td>161</td>
</tr>
<tr>
<td>STAR860</td>
<td>190</td>
</tr>
</tbody>
</table>

On all computers available optimizers were used. The network used contained 3347 nodes, 9394 arcs and an arc-node ratio 2.8. As comparison, a network of 17931 nodes was also used. The computing times given are the all-to-all times; an all-or-nothing assignment with each node as origin and destination.
Table 4.3. The difference in performance of the computers on networks with different network sizes. The times given are in case of \( N=3347 \), all-to-all, in case of \( N=17931 \), 1000 spt's. The spt-algorithm used was T.calc and the assignment uses levels. SG is Silicon Graphics.

<table>
<thead>
<tr>
<th>size</th>
<th>3347</th>
<th>17931</th>
</tr>
</thead>
<tbody>
<tr>
<td>Convex C1</td>
<td>411</td>
<td>683</td>
</tr>
<tr>
<td>SG (25 MHz)</td>
<td>260</td>
<td>581</td>
</tr>
<tr>
<td>SG (33 MHz)</td>
<td>161</td>
<td>417</td>
</tr>
<tr>
<td>STAR860</td>
<td>190</td>
<td>319</td>
</tr>
</tbody>
</table>

The i860 is not the fastest processor in table 4.1 because of its memory configuration. The Silicon Graphics (33MHz) has an advantage of having a 64kByte cache. This advantage disappears when a larger network is used, see table 4.3. With integer indexing the i860 is always superior. Next to an upgrade of the i860 from 33 MHz to 40 MHz, the processor board can be improved by using a pipelined multibank memory instead of the single bank memory implementation on the STAR860.

5. CONCLUSIONS

The use of special purpose hardware is only legitimate when the task to be performed is time critical and the budget is limited. In chapter 4, we have shown that the STAR860 is two times as fast as the Convex C1. The performance of a single i860-based processor board, in comparison to the STAR860, can be improved by a factor of about 2, using a faster version of the i860 and a better memory architecture.

The price of a single i860-based processor board is mainly determined by the memory-cost and, depending on the amount of memory, is estimated to be between $5,000 and $10,000. A 16 board LPA together with a general purpose host, of about $40,000, and additional costs of manufacturing of about $15,000 amounts to a total cost of less than $0.3 million. The price of a Convex C1, however, is about $0.6 million, and thus leads to another factor 2 in cost/performance improvement.

Hence, a 16 board LPA will give a cost/performance improvement of a factor 100 compared to a Convex C1.

For future developments of on-line traffic control, routeguidance, ramp metering etc the LPA should be a powerful tool.

6. REFERENCES

Assignment:


Transportation Research Record 1220, 1989.


Shortest path finding:


Special Purpose Computers in Computational Physics:


ISBN 0-12-049260-1


APPENDIX: THE ALL-OR-NOTHING ASSIGNMENT.

The following is a description of the algorithm used to perform the all-or-nothing assignment. This is an optimal algorithm depending on the way Init_Q, Select_node and Update_Q are implemented. The symbols used are defined as follows:

- $N, M$: set of all nodes, arcs in the network
- $n, m$: number of nodes, arcs in the network
- $FS(v)$: the forward-star representation of node $v$; defines the network
- $OD(r,v)$: Origin-Destination matrix; number of travelers going from node $r$ to $v$
- $c_{a}, c_{u,v}$: used arc-length (arc-traveltime) for arc $a$ from node $u$ to node $v$
- $d_v$: calculated distance (traveltime) from the origin-node to node $v$
- $p_v$: previous node from node $v$ in the shortest path tree (spt), defines the spt
- $l_v$: level of node $v$ in the shortest path tree.
- $q_{a}, q_{u,v}$: calculated load on the arc $a$, from node $u$ to node $v$
- $r$: current origin-node (root)
- $Q$: temporary set; contains nodes to be examined
- $L(n)$: set containing all nodes from spt on level $n$
- $S_v$: temporary; trafficload going up to node $v$

In short the all-or-nothing assignment looks as follows:

```plaintext
# Compute the all-or-nothing assignment
for each origin in the network {
    Compute the shortest path tree
    Sort levels
    Assign the spt
}
```
The subroutines Init_Q, Select_node en Update_Q are not defined here. Although crucial to the efficiency of the algorithm, the functionality remains the same. The calculated loads, from preceding assignments are kept in $q_a, a \in M$.

# Compute the shortest path tree for origin $r$

Initialize $d_v = \infty, \forall v \in N - \{r\}, d_r = 0, l_r = 1$

Init_Q $Q = \{r\}$

While ($Q \neq \emptyset$) {
    Select_node select $u \in Q; Q = Q - \{u\}$
    for each $(u,v) \in FS(u)$ {
        if ($d_u + c_{u,v} < d_v$) {
            $p_v = u$
            $l_v = l_u + 1$
            $d_v = d_u + c_{u,v}$
            Update_Q $Q = Q \cup \{v\}$
        }
    }
}

#Sort levels

Initialize $L(.) = 0$

for each node $u$ {
    $L(l_u) = L(l_u) \cup \{u\}$
}

#Assign the shortest path tree

Initialize $S_u = 0, u \in N$

for each level top down {
    for each node $u$ in the set $L(level)$ {
        $q_{p_u} = q_{p_u} + OD(r,u) + S_u$
        $S_{p_u} = S_{p_u} + OD(r,u) + S_u$
    }
}
1992

Large-scale Network Assignment: Algorithms, Hardware and Implementation.

H.J.M. van Grol and A.F. Bakker

PTRC annual meeting
1. INTRODUCTION

Network assignment models are frequently used for medium and long term transportation studies. For these studies mainly the static assignment models are used. Assignment models are nowadays also used for short term applications. For these applications dynamic assignment models are used. Whether static or dynamic assignment algorithms are considered, the required computing time increases with the square of the network-size (expressed in the number of nodes).

For large networks the computing time to obtain results may take too long, especially when they are used for real-time applications. Reducing the computing time is the subject of this study. The required computing time is influenced by: the scale of the problem, the desired accuracy, the model, the algorithm, the implementation and the chosen computer. This paper is focused on the last three issues.

Algorithms: Assignment models use an iterative procedure to define the traffic-load and the travel-time on each arc in a network. In each iteration the Origin-Destination matrix (OD-matrix) is assigned to the network. Each OD-pair is assigned to one route. For the determination of the routes, that are generally based on minimum travel-time or -cost, shortest path finding algorithms are used. They consume most of the computing time. In chapter 2 a few of the most efficient and robust algorithms will be reviewed.

Implementation: A good algorithm will only show its quality when it is well implemented. For a good implementation aspects such as data-structures, accuracy, programming techniques, etc. must be considered. The implementation (sometimes even the algorithm) must be adapted to obtain maximum performance on a specific architecture. Maximum use should be made from the available compilers/optimizers. These issues are discussed in chapter 3.

Hardware: When an efficient algorithm is properly implemented, the only way to improve the performance is by using a faster computer. The alternatives range from fast sequential or pipelined supercomputers through parallel computers to special purpose computers. The choice depends on the required performance, the characteristics of the problem and the available budget. The complexity of this choice will be discussed in chapter 4.
Analysis of the traffic assignment problem has lead to the design and construction of a special purpose computer, see van Grol and Bakker [6]. This computer is called the Linear Processor Array (LPA). The architecture and other aspect concerning the design and construction of a computer will be discussed in chapter 5.

The LPA is still in the construction phase (last stages) and therefore the performance cannot be tested. The estimated performance of the LPA will be compared to the performance on a Cray Y-MP and a CONVEX. This comparison is presented in chapter 6. Finally in chapter 7 some conclusions will be given.

2. ALGORITHMS

Assignment models are used to define the traffic-load-distribution in a network based on the traffic-demand. The traffic-demand is given by an Origin-Destination matrix (OD-matrix). Assignment models are also called route-choice models because the trips made between two points in the network are distributed across all feasible routes between these two points. The distribution is often based on the equilibrium principle of Wardrop [16] and it is determined by an iterative procedure.

Traditionally, static assignment models were used. In static assignment models the traffic-demand is considered to be time-invariant and the traffic is simply assigned to the complete route traveled. These and a few other inconsistencies, see Hamerslag [8], Hungerink [10], etc. have led to the development of dynamic assignment models.

In dynamic assignment models the traffic demand is considered to be time-variant and the time-interval, for which the assignment is carried out, is discretized into periods. During each period the traffic-conditions are presumed to be constant. A dynamic assignment model, which in structure strongly resembles the static assignment model, is introduced by van Grol and de Romph [7]. The algorithms for this model are computationally more demanding than for static assignment, because an assignment is executed for each period in which traffic departs.

Figure 1. A flow-diagram for assignment models. The dotted lines of the in-lay are unique to the dynamic assignment models.
One iteration in any assignment model consists of three parts. The main part is an all-or-nothing (AON) assignment based on the current travel-times. At the end of each iteration the resulting traffic-load distribution is combined with the traffic-load distribution from the previous iterations and at the start of the iteration the travel-times are updated based on the current traffic-load distribution. In the AON assignment the trips are assigned to the current shortest paths (nothing to other routes). The iteration process for both static and dynamic assignment models is visualized in figure 1.

The AON assignment itself consists of the determination of the shortest paths and the assignment of trips to them. The most efficient way to do this is by calculating and assigning shortest path trees. A shortest path tree specifies all the shortest paths from one node to all the other nodes.

The most time-consuming part of the algorithms is the determination of the shortest paths trees. A large number of different shortest path finding algorithms exist [4][15][5]. Their run-time performance ranges from \(O(n^2)\) to \(O(n^3)\), where \(n\) is the number of nodes in the network. The most efficient algorithms use the so called combinatorial technique or the re-optimizing technique and solve the one-to-all problem. The combinatorial technique constructs a shortest path tree, while the re-optimizing technique uses an existing shortest path tree and re-optimizes it to a tree rooted at another node. In figure 2 the different techniques are categorized, together with two for assignment purposes less efficient approaches. The re-optimizing technique, see Florian [3] and Gallo et al. [4], cannot be used with dynamic assignment models.

![Figure 2. Classification of the shortest-path-finding techniques.](image)

The combinatorial technique operates by satisfying the following optimality condition for all neighboring pairs of nodes in the network.

Optimality condition: a tree is a shortest path tree if and only if:

\[
d_i + t_{ij} - d_j \geq 0, \quad \forall (i,j) \in A
\]

Where \(d_i\) is the length of the path from the origin to node \(i\), \(A\) is the complete set of arcs in the network, and \(t_{ij}\) the arc-length from node \(i\) to node \(j\). This means that when no path can be shortened by rerouting it, the resulting tree is a shortest path tree, see figure 3.

The order in which the node-pairs (arcs) are investigated determines the efficiency of the algorithm. Mainly two types of algorithms exist; the label-correcting and the label-setting algorithms. The "label" of a node is the path-length from the origin to this node. Instead of selecting arcs these algorithms select a node and then investigate all arcs connected to it. A shortest path tree is determined by examining
Figure 3. Illustration of the optimality condition. If the optimality conditions for nodes $i$ and $j$ are not satisfied, the path from $r$ to $j$ can be shortened by rerouting it through $i$. Then after setting $d_j = d_i + t_{ij}$ the new optimality conditions for nodes $i$ and $j$ become $d_i + t_{ij} - d_j = 0$ and $d_j + t_{ji} - d_i = t_{ij} + t_{ji} > 0$.

marked nodes, starting with the origin, which initially is the only marked node. When a shorter path is found to a node, the path is changed and the node is marked to be examined. In label-correcting algorithms the processing order of the marked nodes is determined by the order of marking. This generally results in the correction of the labels and thus the reexamination of nodes. In label-setting algorithms the marked nodes are selected in the order of increasing distance to the origin. This assures that the label is correct once the node is being examined. The so called threshold algorithms are a hybrid form of the correcting and setting algorithms.

In figure 4 an overview is given of the algorithms.

Figure 4. Classification of the algorithms using the combinatorial technique.

The algorithms differ in the way the marked nodes are processed. The marked nodes are placed:

- C-stack —on a stack
- C-queue —in a queue
- C-deque —(double ended queue) conditionally at one of both ends of a queue
- C-dlque —(double linked queue) conditionally in one of two queues
- S-ord —in a sorted linked list
- S-heap —in sorted heap
- S-calc —in an addressable array (label-setting)
- T-calc —in an addressable array (strictly spoken label-correcting)
A more detailed discussion of these algorithms and their performance has been given in van Grol et al. [5]. The best algorithms are T-calc and C-dlque, see next chapter.

**T-calc:** This threshold algorithm, see van Grol et al. [5], uses the label-value of a node to determine the place of it in an addressable pointer-array. Each pointer leads to a linked list of nodes with approximately the same label-value. Strictly, this method is a label-correcting method, because the range of label-values in one linked list can be larger than the minimum arc-length and the nodes are not selected based on their labels. The worst-case computational complexities of this algorithm is thus \( O(n \cdot 2^n) \).

**C-dlque:** This label-correcting algorithm, see Pallottino [13] uses two queues to manage the marked nodes. Depending on whether the nodes are marked for the first time or not, they are placed in one of both queues. The worst-case computational complexities of this algorithm is \( O(m \cdot n^2) \).

These worst-case complexities serve mainly as a performance guarantee. The actual (average) performance is generally much better.

### 3. IMPLEMENTATION

A good algorithm will only show its qualities when it is well implemented, as was already indicated in the introduction.

Quite often the importance of the implementation is overlooked. The implementation depends on the computer that is used, the available tools and the quality of information about it. An extensive discussion about implementations is therefore not in place. However, for a good implementation at least the following must be considered:

- well chosen data-structures
- whether the implementation is meant for a particular computer
- the required precision
- optimum use of the compilers / optimizers

**Data-structures:** The way in which data (the network) is organized has a major influence on the efficiency of the algorithm. The best data-structure for the assignment algorithms is the so called forward star representation. The forward star representation of a node \( u \) is an arc-list of arcs that connect \( u \) to its neighbors. In this case, when during the determination of the shortest path tree the arcs connected to a node must be investigated, they are directly available.

The best data-structure for storing the marked nodes, differs for each algorithm as was shown in the previous chapter.

**Computer:** When a particular computer is used, the implementation (and often the algorithm) must be adapted to obtain the best results on it. A vector-processor obviously has different requirements than a parallel processor. When an algorithm cannot be adapted to the particular requirements, the performance will not benefit from the available hardware. The algorithm presented in the previous chapter are not suitable for vectorization. They are, however, suitable for parallelization.

**Precision:** Whether to calculate results in integers, single-precision floating point or
double-precision floating point depends on the desired accuracy. Choosing a higher accuracy than required may reduce the performance considerably. This depends on the particular hardware of the computer that is used.

**Compilers:** To obtain maximum performance, the compiler- and optimizer- capabilities should be studied. In case of vector processors or parallel processors, additional directives can be added to aid the compiler in obtaining the best object code.

**Performance of the assignment algorithms**

The algorithms from the previous chapter have been tested on a number of 'real' road-networks. A few results are shown here. The tests were carried out on a Silicon Graphics workstation 4d/35, which contains a 33 MHz Mips R3000 RISC microprocessor and has a peak performance of 33 MIPS and 16.7 Mflops.

The computational complexity of T-calc was \( O(n \cdot 2^n) \), but in practice it is expected to have a running time of \( O(n) \). The reason for this is that each node is examined only once for one shortest path tree, while the examination itself is independent of the size of the network. Also the management of the marked nodes is independent of the size of the network.

The expected relation between running time for the all-to-all problem and the size of the network, \( n \) is thus \( O(n^2) \). Figure 5a shows the running-times of T-calc and C-dlque. The relation is not exactly followed but close enough for estimations.

![Graph](image)

**Figure 5ab. The all-to-all running times for a number of network from T-calc (circles) and C-dlque (plus-signs). (a) Shortest path finding and (b) all-or-nothing assignment**

In figure 5b also the assignment is shown. These algorithms, especially T-calc are robust. The network characteristic —apart from the size— do not influence the performance of these algorithms very much.
4. HARDWARE

Since the beginning of commercial computing the performance of computers has increased rapidly. Nowadays computers out-perform computers from the 50's by more than seven orders of magnitude. Still, demand grows with supply and even in these days we find ourselves in need for even more computing power. Several traffic assignment-programs are available for the PC, and with a 386 or 486 PC-AT reasonably sized networks can be processed. However, when it comes to large networks, time critical calculations or more complex calculations the computing power offered by PC's is not enough.

Workstations, minicomputers and mainframes can handle large problems with more ease, partly because of their higher processor performance, but mainly because of their more balanced architecture. Large memories, separate IO-processors, etc. allow the processing of more and larger jobs. However, possibly only supercomputers can deliver the desired performance. Because of the state-of-the-art technology they use, reaching for what is physically possible, supercomputer are very expensive. Calculating on a supercomputer is a costly affair; buying a supercomputer even more so.

With a limited budget only few options remain; comply with the long computing times, use supercomputers but not as frequent as desired or use a special purpose computer.

Special purpose computers

A special purpose computer is a computer build for one particular task, which it can process with maximum performance. Because other tasks are not supported, opposed to general purpose computers, savings can be made in hardware and design, and more attention can be given to the effective part of the computer.

Special purpose computers can be divided into two groups, single purpose computers and multi purpose computers. In the single purpose computers the 'special hardware' is applied at such a low level (the algorithm implemented in hardware) that only one task can be performed. In the multi purpose computer the specialization is used at a higher level. Characteristics of a problem are identified and the architecture is designed to process problems that share these characteristics.

Architectural improvements

To get some idea of the possible changes to the classic uniprocessor architecture from Von Neumann that improve the performance, a short overview will be given. The performance improvements due to developments in circuitry —the main cause for the enormously improvements since the 50's— are not included in this overview because they do not form an architectural change.

Almost all architectural improvements are introductions of some sort of parallelism; instead of one execution unit doing all the work several execution units work simultaneously. Parallelism is not limited to entire processors but can be identified in many other places in a computer. Even the 8-bits of a byte that are processed at once is a form of parallelism. Johnson and Durham [11] suggest a division in terms of data level and control level parallelism. Data level parallelism is spoken of when a single instruction causes the concerted movement or processing of many bits of data. Control level parallelism is spoken of when there are multiple processors, each of which
independently executes a sequence of instructions that belong to one program. Data level parallelism does not interfere with the sequential execution of a program, while control level parallelism does.

**Data level parallelism:** Architectural improvements in this category are:

- **Instruction pipelining:** Each instruction is divided into several stages. By processing these stages in separate execution units the rate of execution can be improved. RISC (Reduced Instruction Set Computers) architectures are based on these changes. By simplifying the instruction set (each instruction fits into one Word), by using separate instruction- and data-busses, by disconnecting the memory interface from the processor, etc. RISC processors can execute one instruction each clock cycle.

- **Vector pipelining:** Often the same operation is executed on different data. By collecting the data in a vector the operation on the vectors can be pipelined winning a factor equal to the amount of stages in the pipeline. Supercomputers such as the Cray- and NEC-series, but also mini-supercomputers such as the Convex-series are examples of machines using vector pipelining.

**Control level parallelism:** Architectural changes in this category are:

- **Few powerful processors / mainly simple processors:** Which of both is the most efficient depends on the characteristics of the problem; is the parallelism in the problem fine-grained, or coarse grained, what is the data-interdependency, etc.

- **Shared/distributed memory:** Processors can only work with shared memory when their memory activities are relatively low or when each processor has some private memory (cache). When there are not too many data-dependencies but memory activity is high, distributed memory can be used.

- **Network topologies:** In distributed memory systems the processors must be connected some way to exchange information. Networks can be hierarchical / fully connected, static / dynamic. Each of these have advantages and drawbacks.

Other classifications have been given in the past. The categorization by Flynn in SISD (single instruction stream, single data stream), SIMD (single instruction stream, multiple data stream), MISD and MIMD is possible the most well known, see Hockney et al. [9].

**General purpose versus special purpose computers**

In commercially available (super) computers all of the above techniques are used. Many computers that are called general purpose computers show a much higher efficiency to some problems than to others. Vector processors need independent data vectors, while parallel computers prefer a large degree of independence in code. Clearly, differentiating between general purpose computers and special purpose computers is not simple, but it is of no major importance here. When the argument is limited to maximum performance, the computer choice limits to general purpose supercomputers and special purpose computers. Supercomputers are expensive because of the used techniques. To cover the cost, many different people must use them and to satisfy all of them the computer must be efficient enough for 'all' possible problems. A special purpose computer can be cheap because it is only used by people with the same type of problem and it therefore doesn't have to solve any other problems.
The decision to buy or construct a special purpose computer is viable to a number of considerations:

- **Budget**: With a plentiful budget a supercomputer could be bought, or at least enough supercomputer time. Only when the budget is limited the effort of building or buying special purpose hardware should be considered.

- **Duration of problem**: The problem-field for which the special purpose computer is considered, must be a *long lasting supply of problems* (large jobs and/or real-time applications), because otherwise the effort of developing the computer becomes pointless.

- **Development time**: The development time should not be too long, preferable less than a year, because otherwise the problem could already have been solved on general purpose computers. Moreover the technology catches up with the design (1 magnitude every 5-6 years).

Other factors such as *maintenance* and *programmability* must also be considered.

In the next paragraph a possible special purpose computer for traffic assignment will be investigated.

5. A SPECIAL PURPOSE COMPUTER FOR TRAFFIC ASSIGNMENT

5.1. Possible special purpose architectures

In this paragraph possible single- and multi-purpose solutions will be briefly discussed. The requirements for a multi purpose solution will be specified.

**Single Purpose Solution:**

Suppose only shortest path finding algorithms are considered. A single purpose solution can than be sought for algorithms based on the algebraic technique and for algorithms based on the combinatorial technique.

Matrix manipulation methods based on the algebraic technique are good candidates for a single purpose solution because they can be parallelized easily. One particular algorithm —the 'doubling algorithm' (see Bertsekas [2])— obtains a minimum computing time for an all-to-all problem in \( O((\log n)(\log m^*)) \) time, where \( n \) is the number of nodes and \( m^* \) is the maximum number of arcs of all shortest path in the network. This performance, however, is obtained on a hypercube with \( n^3 \) processors. For relatively small networks (300 nodes) the amount of processors is already large (9,000,000). Thus, for large networks, such a computer is not feasible.

In the algorithms based on the combinatorial technique the innermost loop consists of the examination of the neighbors of each node. The examination of the neighbors could be parallelized. However, the amount of neighbors is relatively small. It varies from 2 to more than 8, with an average of about 3. The performance improvement could be a factor 3, but for the algorithm as a whole this factor is smaller.

For both approaches only the shortest path finding algorithms were considered. When also the assignment is included, for which the above solutions are not efficient, the improvements will become smaller. A single purpose solutions for traffic assignment algorithms —although feasible— is not effective.
Multi Purpose Solution:

In this case the problem must be studied in a wider perspective. The relatively simple structure of the assignment problem does not allow many points of view and it almost immediately leads towards an obvious solution. Parallelism can be found at a number of levels in the assignment problem, but the determination and the assignment of shortest path trees is the form of parallelism considered here.

Parallelizing the assignment:

The assignment can be parallelized by distributing the assignment for all the origins over a number of parallel processing units. In principle, the maximum speedup would equal the amount of parallel processing units used. The parallelization is however not without penalty. The maximum speedup by parallelization is inversely proportional to the sequential part of the problem (the part that cannot be parallelized). This is often referred to as 'Amdahl's Law'. In 1967 Amdahl [1] stated that "the effort expended on achieving high parallel processing rates is wasted unless it is accompanied by achievements in sequential processing rates of very nearly the same magnitude". The sequential part should be as small as possible. A closer look at the assignment problem is thus required.

A sequential part is first of all represented by 'update' and 'combine', see figure 1. Another sequential part is caused by parallelization problems.

Traffic assignment is a memory intensive problem. Connecting all (more than 10) processing units to a shared memory —although possibly feasible— is not effective; too much time would be spend waiting for memory. Thus a distributed memory system is required. Each processing unit will have its own private memory with the entire network. After all trees are assigned, the resulting traffic-load distribution must be determined. However, the results are spread across the processing units. To collect the traffic-load distribution, all separate sub-distributions have to be transported to the execution unit, which will than calculate the result. Likewise, the new updated travel-times must be distributed to all the processing units. The time consumed by the data-collection and distribution contributes to the sequential part of the process.

Using the notation from Hockney [9], the serial execution time, $T_1$, is expressed as:

$$T_1 = t_{ser} + t_{par}$$

Where $t_{ser}$ represents the above mentioned sequential parts and $t_{par}$ the parallelizable part. When the algorithm is parallelized an extra term emerges, which represents the summation and distribution. This term, $t_{distr}$ represents the time to perform the summation and distribution between 2 processing units. Because there are $p$ processing units it will have to be multiplied by $p$. The parallel execution time, $T_{par}$, becomes:

$$T_{par} = t_{ser} + \frac{t_{par}}{p} + p \cdot t_{distr}$$

Obviously, unless $t_{distr}$ is very small, the multiplication by $p$ will undermine the efficiency of the computer. Preferably, the update travel-time should be broadcasted to the processing units, such that it does not depend on the size of $p$. Similarly the summation time should also be independent of $p$. When this is achieved, the following
parallel execution time is found;  

\[ T_{par} = t_{ser} + \frac{t_{par}}{p} + t_{distr} \]  

(4)

By determining the sizes of \( t_{ser} \), \( t_{distr} \) and \( t_{par} \), the speedup and efficiency for a parallel computer can be estimated, according to equation 3 and 4. The running time of the algorithms is \( O(n^2) \), while the distribution time is \( O(n) \). Therefore, the size of the network will alter the efficiency of the computer. To determine reasonable limits, the running-time of \( T\text{-}calc \) for an all-to-all problem (\( n \approx 250 \)) and a more-to-all problem (\( n > 30000 \)) have been measured. Based on these running-times and estimated times for the distribution (based on the time required to copy the distribution to another place in memory) the minimum and maximum speedup and efficiency can be determined. In figures 6 and 7 the speedup and efficiency of both options is shown. Clearly, the broadcasting and summation have a crucial effect on the resulting performance.

\[ \text{Figure 6. Performance aspects according to equation 4 of a parallelized program. The two dotted lines show the estimated boundaries. (a) The speedup, } S, \text{ compared to sequential execution as a function of the number of parallel processors. The solid line is the ideal performance with } t_{ser} = 0. \text{ (b) The efficiency, } E, \text{ as a function of the number of processors.} \]

Requirements:

Based on the results from the last paragraph, the requirements for a multi purpose computer can be specified:

- a number of independently working processing units with fast processors and sufficiently fast and large memory.
- fast communication channels between the processing units

They must
- be sufficiently fast
- enable broadcasting/summation in an efficient way

Naturally there is a balance between the processor speed and the required speed for the communications.
Figure 7. Performance aspects according to equation 3 of a parallelized program. The two dotted lines show the estimated boundaries. (a) The speedup, $S$, compared to sequential execution as a function of the number of parallel processors. The solid line is the ideal performance with $t_{ser} = 0$. (b) The efficiency, $E$, as a function of the number of processors.

5.2. The Linear Processor Array (LPA)

In this paragraph the architecture of a special purpose computer, which satisfies the requirements presented in the previous paragraph, will be presented. The architecture concept of this special purpose computer is called the Linear Processor Array (LPA). This architecture is a linear array of independently operating processor boards. Each board has a fast processor and a local memory. Apart from the linear network by which the processor-nodes are connected, they are also connected by a common bus. This bus also connects the computer to the outside world, which is represented by a host computer.

The linear network is a simple architecture concept. It is a network with a low connectivity (2), and is thus often dismissed. Most parallel computers have an interconnection network with connectivity higher than 2, thus forming a mesh, a hypercube, etc. In practice, however, these computers are often configured as a linear array, because the simple concept of a linear array allows the programmer to relate to his problem more easily.

The Linear Processor Array (LPA), see figure 8 consists of $P$ processor boards and 1 control board, the inter-processor bus and the global bus. The LPA is connected (by the global bus) by an interface and the host bus with a general purpose host computer.

The requirements formulated in the previous paragraph form the basis of the LPA architecture. The first requirement is satisfied by the $P$ processor-boards of which the LPA consists. The specific architecture of the processor-boards, with a fast processor and a large memory, will be discussed later. The second requirement leads to the linear network, represented by the inter-processor bus, and the global bus.

The inter-processor bus is actually a sequence of $P$ independent busses, one between each pair of neighboring boards. The inter-processor bus can be used to set...
Figure 8. Overview of the architecture of the Linear Processor Array. The ip-bus is the inter-processor bus for the transportation of data between neighboring boards.

up a pipeline. Using this pipeline, the flow-distribution, present on all boards, can be summarized in about the same time as it would take to summarize the flow-distribution on one board and its neighbor. Similarly the broadcasting of travel-times can be established. The time required for summation and broadcasting is now practically independent of the number of processor boards. Thus the speed of the LPA grows linearly with the number of boards.

A broadcast could also be established by the global-bus, which enables the transfer of data to all boards simultaneously. The summation, although possible, would encounter the disadvantage, that the communication time would again depend on the size of $P$, as described in the previous paragraph.

Although the LPA forms a conceptual solution to the communication problem, the speed of the communication remains important to keep the sequential part small. This is because of Amdahl’s law, see paragraph 5.1.

Discussing the full details of the LPA would go too far. Only the architecture of one processor-board will be briefly discussed.

The LPA processor (control) board

The control board and the processor boards have the same design. Functionally the only difference is that the processor-board may not become bus-master.

The requirements for the processor board are that it must be fast, it must have a large memory and it must support fast inter-processor communication. The choice has been made to use a classic design, for reasons of flexibility, programmability and because the revenue of designing a special purpose architecture would be too small.

The design of the processor-board is based on the Von Neumann Uniprocessor architecture with a few extensions. Figure 9 shows the functional design of the processor board. It consists of a processor, the memory system, a bus interface, the inter-processor bus and the local bus connecting the former.

The processor used is a RISC micro-processor, the Intel i860 (40 MHz), which
Figure 9. Functional design of the processor board. It consists of a processor, a memory system, a bus-interface, an inter-processor bus and a local bus.

has a peak performance of 40 MIPS and 80 Mflop. A number of features characterize the i860. It has a so called dual instruction mode in which the processor initiates and executes an integer and a floating point instruction simultaneously. Furthermore, it knows the dual operation mode which allows two floating point operations simultaneously. It has a 64-bit databus and 2 on-chip caches; one 4Kbyte instruction cache and one 8Kbyte data cache.

The memory system consists of dynamic memory, with a maximum size of 128 Mbyte, and static memory, with a size of 512 Kbyte.

The inter-processor bus can buffer 2 Kbyte of data and forms an asynchronous link to other neighbor processor boards. The data-paths are 64 bit wide. With a clock-rate of 40 MHz a maximum speed of 320 Mbyte could be achieved with the inter-processor bus.

5.3. Programming the Linear Processor Array

A computer is a symbiosis of hard- and software. In an efficient computer they will receive a balanced amount of attention. In a special purpose computer the tendency is to underestimate the importance of the software because the problem fits naturally on the hardware. This might be so in single-purpose computers, but in a multi-purpose computers such as the LPA, the software should be well supported.

The software for the LPA can be grouped in three categories:
- user software
- (parallel) programming support software
- operating system software

The user software is based on the interface provided by the programming support
software and the operating system. The programming support software in its turn is based on the interface provided by the operating system. Finally, the operating system itself is based on the interface provided by the hardware.

This software-system, as it is presented above, resembles an onion in which each layer provides a set of functions dependent only on the layers within it. This is illustrated in figure 10.

The LPA will be equipped with an operating system. Operating systems, —present on most commercial computers— provide an working-environment, which is much more user-friendly than the bare hardware. Several user-services are taken care of and the user can focus on his or her own applications. The operating system does however produce a certain overhead. Processor-time is used for the operating system, which is not effectively used by the user to solve problems. The operating system for the LPA has been tuned to produce as little as possible overhead.

**Operating system**

The operating system (kernel) on the LPA consist of a number of independently operating, identical kernels, one on each of the boards of the LPA. This is the consequence of the fact that the operating system is an adaption of an existing operating system. It is, however, not a disadvantage, as it only supports the extensibility of the system as a whole.

The operating system on each board supports (originally):

- Multitasking
- Virtual memory
- System calls
- Full IEEE Floating Point Exception Handling

and added for the LPA are:

- Enhanced signal and interrupt handling
- message passing —in functionality (UNIX system V) not in performance
- special memory-map system call

**Programming support software**

A single board of the LPA can now be programmed as a 'normal' computer. However, programming one single processor-board is quite different from programming the LPA as a whole. Programming support software helps bridging this gap.

Ideally the user would write a single program, which would then be automatically parallelized to run on the LPA. Such facilities are not available and the parallelization will remain the task for the user. However, the programming support software will help to start up the application on the LPA, to initialize communication between subprograms, etc.
H.J.M. van Grol & A.F. Bakker

The functions provided by the support software are thus two-fold:

- initialization of the LPA
- setting up communication-channels between the processes on the LPA and the host and maintaining them during the session. The programming support software thus provides an easy programming environment for the LPA.

Another major task for the support software is to provide the possibility to simulate the LPA on a general purpose machine. That enables the user to test and debug LPA-applications while other applications are running. Additionally, when the LPA is/was not available it allows LPA-programs to be developed.

The software package developed for this purpose is called the *LPA-simulator*. The LPA-simulator is a set of functions (in C) combined in a library, which must be linked with the user-program. Two different libraries are available, one to run on the LPA and one to run on the Host or another multitasking computer.

The LPA-simulator provides the user with several communication services:

- command handling —sending commands from the host or control-board to the processor-boards
- message passing —high performance message passing, for data-transport to and from the processor-boards
- neighbor communication —support for using the inter-processor bus

The message passing service mentioned above is not the same as the message passing service provided by the kernel. The functionality is more restricted, but the performance is much better.

6. RESULTS

In this chapter the expected performance of traffic assignment algorithms on the LPA will be compared to the performance of these algorithms on two widely used computers, a supercomputer from the Cray-series and a mini-supercomputer from the CONVEX-series.

The LPA will also be compared to a set of Silicon Graphics workstations. Previous test-results on an i860-board build in a 386-PC have shown that the performance of one LPA processor-board is (at the same clock-rate) just as fast as a Silicon Graphics workstation. The set of workstations therefore make a good comparison. To obtain the test-results for the i860, optimizers were used, but the special features of the i860 —dual instruction, dual operation, etc— were not used at all. The final performance may be improved by more than a factor 2.

6.1. Performance on the Cray Y-MP4 and the CONVEX C2

The Cray Y-MP is a super computer based on vector-processing. Depending on its configuration, it has from 1 up to 8 parallel vector-processors. The experiments that follow have been carried out on one vector-processor. The Cray Y-MP has a clock-cycle of 6 nanoseconds. Based on this alone, the performance would have to be a factor 5 better than the Silicon Graphics workstations. However, the Cray Y-MP has an architecture which is optimized for vector processing and the most efficient assignment
algorithms are not suitable for vectorization. Consequently the performance on the Cray Y-MP is not much better than on a Silicon Graphics workstation.

The CONVEX C2 is a mini-super computer also based on vector-processing. The same argument as above are valid for the Convex. The performance for large networks is slightly less than a Silicon Graphics workstation.

The algorithms have been implemented using the available compilers and optimizers. Local vectorizations of the assignment algorithms, such as the examination of the neighbor-nodes, have led to an increase in computing time instead of a reduction. The results are shown in table 1. The best results were obtained with just scalar optimizations.

Table 1. Run-time performance of an assignment algorithm on one processor of the Cray Y-MP and the CONVEX C2 for several network sizes. The performance on a Silicon Graphics workstation is denoted by SG.

<table>
<thead>
<tr>
<th>network</th>
<th>Cray Y-MP</th>
<th>CONVEX C2</th>
<th>SG</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>optimization</td>
<td>optimization</td>
<td></td>
</tr>
<tr>
<td>none</td>
<td>scalar</td>
<td>none</td>
<td>scalar</td>
</tr>
<tr>
<td>1104</td>
<td>19</td>
<td>29</td>
<td>19</td>
</tr>
<tr>
<td>3348</td>
<td>48</td>
<td>75</td>
<td>52</td>
</tr>
<tr>
<td>17931</td>
<td>256</td>
<td>404</td>
<td>290</td>
</tr>
</tbody>
</table>

To execute assignment problems on a vector-processor more suitable algorithms should be used. The matrix-manipulation method is suitable for vectorization, but as mentioned earlier, this algorithm is not efficient. Some time ago, better results have been obtained with the vectorization of the assignment problem. Mouskos et al. [12] obtained an improvement of a factor 5 by vectorization and changes to the code. Unknown is, however, what kind of algorithm they exactly used and how it was implemented.

The Cray Y-MP that was used has 4 processors. An additional speed-up of at most a factor 4 can thus be obtained on this computer. The Convex has two vector processor. Here an additional speed-up of factor 2 can thus be obtained.

6.2. Performance on a set of Silicon Graphics workstations

As an example of a parallel implementation without the inter-processor bus, a set of workstations, connected by ethernet (local area network) is used to test the effects of equation 3. The effects are shown in figure 11.

The results show that for some applications a series of workstations can produce a satisfactory speed-up. However, for larger number of parallel processing units this configuration will loose its efficiency.

It does show however that, with commercial computers, performances close to the performance of the LPA can be obtained. Only the missing inter-processor bus undermines their efficiency (apart from optimizations of the processors and operating system).
7. CONCLUSIONS

When assignment models are used, with large networks and possibly in time-critical situations the computing time for the model must be reduced.

A good algorithm is essential, since their performance ranges from $O(n^2)$ to $O(n^3)$. The algorithms with the best performance are T-calc and C-queue. They both use the combinatorial technique. The first algorithm uses a threshold method, the second a label-correcting method. The relation between running-time and network-size (in number of nodes) is almost reduced to $O(n^2)$. The assignment itself also has a running-time of $O(n^2)$ and it reduces the total performance by about a factor 2.

For the implementation of the algorithm, the best data-structure is the forward star representation. Often calculations are done with maximum accuracy. Reducing the accuracy to what is required, may improve the performance considerably.

Depending on the required performance, the duration of the problem, the required flexibility and most importantly the available budget, the choice for a computer must be made. Maximum performance is expected from supercomputers, but these are expensive. The same performance can be expected by special purpose computers but at a much lower price.

A special purpose computer is proposed, the Linear Processor Array which is an cost-efficient alternative for supercomputers. The LPA is linearly scalable. The speed of a 16 board LPA for traffic assignment applications is about twice the speed of a Cray Y-MP4 (4 vector-processors) for the same applications. The Cray Y-MP/4 costs about $10$ Million, see Supercomputer [14], and the LPA about $250,000$. Allowing some room for error and mis-judgement, it can be concluded that the cost-performance ratio is improved by a factor 40 (maintenance, energy and personnel costs have not been included, but would only support the argument.

The computational bottleneck solved in the LPA is the efficient collection and
broadcasting of data by the fast asynchronic inter-processor connections. When computer-manufacturers would introduce such connections on their processor boards, commercial buildings blocks would be created which are efficient for a large range of problems. An LPA could then be constructed more easily and possibly at lower cost.

8. ACKNOWLEDGEMENTS

This work was sponsored by the Stichting Nationale Computerfaciliteiten (National Computing Facilities Foundation, NCF) for the use of supercomputer facilities, with financial support from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (Netherlands Organization for Scientific Research, NWO).

9. REFERENCES

[14] Supercomputer 40, (1990), volume VII, nr. 6, pp. 5, ISSN 0168-7875
Network assignment problems take a computing time that is at least proportional to the square of the network size (expressed in the number of nodes). For large scale networks this can lead to undesirable delays in obtaining useful results. Reducing the computing time is the subject of this study. The computing time is influenced by: the scale of the problem, the desired accuracy, the model, the algorithm, the implementation and the chosen computer. This study focuses on the last three issues.

Algorithms. A time consuming part of the assignment models is shortest path finding. The algorithms for shortest path finding have been tested many times in the past and were studied thoroughly. A few of the most efficient and robust algorithms are a threshold-method, T-calc (address calculation), and a label-correcting method, C-dlque (double linked queue).

Implementation. The good implementation of the algorithms is just as important as their inherent quality. It is obvious that the programmers ability is crucial and that the efficiency of the implementation depends on the architecture of the computer. Also the quality of the (optimizing) compiler must be considered. The best data-structure for traffic assignment algorithms is the so called forward-star representation.

Hardware. When an efficient algorithm is properly implemented, the only way to improve the performance is by using a faster computer. The alternatives range from fast sequential or pipelined supercomputers through parallel computers to special purpose computers. The choice depends on the required performance, the characteristics of the problem and the available budget. When the problem is long-lasting, the algorithm is suitable and the available budget is limited the construction of a special purpose computer can be considered. Buying a special purpose computer has a lower threshold, but still requires some effort to obtain optimal results.

Efficient assignment algorithms consecutively compute and assign the shortest path tree for each of the $n$ origins in the network. The computations for different
origins are independent and can thus be carried out in parallel. The ultimate way to improve the performance of a computer — with a given technology — is also by parallelization. A \( P \) parallel computer with distributed memory can be used. A shared memory system is not efficient, because assignment algorithms are memory-intensive computations. The local memory must be large enough to contain the total problem (the network and part of the Origin-Destination matrix).

The main penalty of this way of parallelization is that the problem must be distributed and that the results have to be collected. The optimal solution would be one in which the distribution and collection would be independent of the amount of parallel processing units. Such a system would be linearly scalable (\( P \) times faster, \( P \) times more expensive).

The development of a special purpose computer is considered. A linear processor array (LPA) comprising \( P \) parallel processor-boards (with scalable memory-size) that are linearly interconnected by high-speed buffers is designed and is partly constructed. The linearly interconnected boards allow the creation of a \( P \)-stage pipeline thus making the distribution and collection virtually independent of \( P \) \((P \ll n)\).

The speed of a 16 board LPA for traffic assignment applications is about twice the speed of a Cray Y-MP with 4 processors for the same application. By comparing the prices of the Cray Y-MP and the LPA, it may be concluded (leaving some room for error) that the cost-performance ratio is improved by a factor 40.

A computer, such as the LPA is commercially not available. When computer manufacturers would introduce the above mentioned inter-connections on their processor boards, commercial buildings blocks could be created, that are efficient for a large range of problems. An LPA could then be constructed more easily and possibly at even lower cost.
1989

A proposal for a travel-time-prediction-system.

R. Hamerslag and W.J. Kribbe

*Proc. 5th World Conference on Transport Research, Yokohama, Japan*
A PROPOSAL FOR A TRAVEL-TIME-PREDICTION-SYSTEM

dr.ir. R. Hamerslag\textsuperscript{1} and dr. W.J. Kribbe\textsuperscript{2}

1) Professor of Traffic Engineering, Faculty of Civil Engineering (also part-time attached to the Faculty of Mathematics and Informatics)

2) Associate Professor of Information Systems, Faculty and Informatics.

Paper prepared for the 5th World Conference on Transport Research, to be held in Yokohama, Japan, July 10-14, 1989.

March 15, 1989
A PROPOSAL FOR A TRAVEL-TIME-PREDICTION-SYSTEM

dr.ir. R. Hamerslag and dr. W.J. Kribbe

Delft University of Technology
P.O. Box 356
2600 AJ Delft, The Netherlands

1) Professor of Traffic engineering, Faculty of Civil Engineering (also part-time attached to the Faculty of Mathematics and Informatics)

2) Associate Professor of Information Systems, Faculty of Mathematics and Informatics

SUMMARY

In this article a system is proposed with which it would be possible to observe queues and to measure them in terms of "expected delay" instead of the usual length we are accustomed to. This information would enable the road users to adapt their "behavior" (or their "planning") provided that this information could be given "in time". Then the disadvantageous consequences of queues could be kept limited to a minimum.

The system proposed now is not meant to replace other systems that are being developed. They should be seen as an addition to these systems.

Comparatively, this system needs little additional investment in new infrastructure (computers and network). It can also be applied to road sections which up to now stood little chance of getting queues but where occasionally unexpected congestion arise.

Adaptations in the behavior of road users are expected to decrease queues.

1. INTRODUCTION

In the Netherlands the traffic, measured in the number of car kilometers, is still increasing. It is expected that the of "peak hours" will increase more and more. Moreover, congestions are not only restricted to the big cities. The queues will spread over a large part of the country. Through the increase of the traffic, the system becomes more and more sensitive to disturbances. These disturbances may for example arise from road works, accidents, defects of vehicles and events like fairs and exhibitions.

Besides being confronted with the normal congestions at peak hours, the road user has increasingly to face congestions at a different time of day. This latter type of congestion is extra annoying, because it is disregarded.

At present about a fifth part of the congestions result from works in the Netherlands. According to reports half of the congestions in California should result from unexpected events. Because the congestions beyond the peak hours are not predictable people will be late for meetings, they will be late at the office, etc. It will be clear that the picking up and the delivering of goods cannot always take place according to plan, which may result in considerable financial damage.

Not only in the Netherlands the queue problem is an important social issue, but also in other parts of Europe, e.g. in America and Japan, this issue is, broadly speaking, problematical.
2. INFORMATION FOR THE ROAD USER

With systems that are in development (see chapter IV), the information is passed on during the trip. This brings about that only the route of the trip can be changed. Further most systems use equilibrium models (see [13],[11]) and suppose that car drivers have a complete knowledge about the congestion in the network. Also systems in which the time of leaving the origin is influenced by congestion (see [1]) make the same supposition.

The system that will be proposed here is intended to pass on the information beforehand. It is to be expected that reliable information in the long run will lead to congestion-avoiding behavior. Depending on the circumstances one will decide on choosing a different:
- route
- time of leaving
- way of transport
- location to do some shopping or to hold meetings
or one will not go at all.

3. MODELS FOR PREDICTING CONGESTION

3.1. Observing congestions and giving advice

Existing systems to observe congestions are based on manual procedures. Surveying will take place visually by the police and the road-maintenance service who will pass on their findings to the traffic police by radiotelephone. This detection method, the so called moving observer method is labor-intensive and consequently expensive. Financially it is not possible to apply this method with sufficient reliability (an every-10-minutes frequency) in the total Dutch motor way system.

Today, the warning of the traffic system takes place in two ways. The most usual way is the passing on of radio messages on the length of the queues. Besides, warnings are passed on via "matrix" signs giving the speed-advice. The meaning of this speed-advice is to prevent accidents.

3.2. The predicting of congestions and delays

A congestion occurs when the intensity (Q) on a road section surpasses the capacity (C). When the stochastic research is neglected, the extent of the delay (D) in time units can be calculated as a function of these quantities as

\[ D = \frac{(Q-C)}{Q} \times t/2 \]

if \( Q > C \)

The way of calculation is illustrated in figure 1.

To predict the congestion, we need estimates of these quantities. However, a problem in making these estimates is that both the intensity and the capacity differ from hour to hour, from day to day and from season to season (see 3.4). A congestion occurs as a resultant of changing (stochastic) events. This situation is shown in figure 2. A consequence will be that, one day, amidst apparently similar circumstances, a considerable delay takes place whereas, the next day, this will not be the case.

To predict the delay, "delay functions" are used in existing transport-flow-models, which co-ordinate the travel time, the capacity
and the intensity on a road section. An example of such a function is the BPR-function (see [15] en 1):

\[ T = T_{\text{std}} \left( 1 + 0.15 \times (Q/C)^4 \right) \]

so

\[ D = 0.15 \times (Q/C)^4 \]

1 BPR-functie

On behalf of the calculating of networks, this function is usable for both strategic and control level. However, the stochastic element in the capacity and the intensity is neglected here. For the benefit of a sound prediction, the calculation of the delay at a special time and a special road section will have to be done far more accurate than is now required for studies at strategic level. The analysis of the flow capacity will have to take place per road section.

3.3. The measuring of delay instead of capacities

The capacity of a road is no fixed quantity. The capacity is determined by the average time drivers think acceptable between two cars, the pendel headway (see [2]). This headway will be greater when the circumstances are less favorable. Differences in distances of .2 seconds lead to a difference of capacity of 10%. If dark or when it is raining people take less risks than on a bright day.

When there are road works going on or when unusual events are taking place alongside the road, the effect on the average times between cars driving one after another will be the same (see [14]). Summarizing, it may be said that the capacity is influenced by:
- the type of road
- the width of the roadside
- road works
- weather
- bright/dusk/dark
- illumination

So when predicting capacity, the above mentioned points have to be considered too. Because of the large spreading in the distance between the vehicles and the relatively small spreading in the average pendel headway, it is difficult to determine the quantities of the influence mentioned here on the capacity.

In the Netherlands information on intensities is gathered automatically at counting points. The spreading the information by counting point appears to be much smaller than the total spreading. Therefore, essentially, it should be possible to improve the prediction of the delay. The new system proposes to use these data.
3.4. Measuring and predicting intensity

It has to be investigated to what extent the spreading in the observations can be reduced by classifying on periods of time. Basically, a model for predicting intensity may have a shape as mentioned in 2

\[ Q(a,t) = \sum_{i=0}^{t-1} c_i \cdot Q(a,i) \]

where

- \( Q(a,t) \) indicates the relevant points in time (for example every month, day, day of the week, etc....)
- \( c_i \) are the coefficients to estimate the prediction of intensity
- \( i \) is the intensity on link a at point of time t

2 prediction of intensity

3.5. Alternative methods for observing congestions and delays.

A different method to determine the delay could be following one (or more) individual vehicles on a route and, in that way, record the decrease and the increase of the speed. However, it is by no means certain that there are technical means available which enable us to follow the large numbers of cars individually with sufficient reliability. Installation of such advanced apparatus (if available) would be too expensive. Furthermore, the observation is disturbed when the followed car leaves the motor way.

Making use of the information already gathered at the counting points, alternatives are also possible. The first alternative is based on the use and the comparison of the cumulative distribution on the successive counting points. The cars are added up cumulatively at the counting points. The time difference is found in comparing this information at two counting points. From this a possible delay can be deduced. In figure 3 an example of such a method is given.

However, this comparison is only possible when all vehicles are observed. If a car is not or wrongly detected, this will lead to a mistake in the determination of time. It might be possible to overcome this by carrying out the calculation for a number of car models and by determining the time differences. However, the problem that all vehicles have to be observed at all approaches and exits, continues to exist. Of course this requires an
extra investment.

A different alternative is based on observing divisions of frequency of different car models in a group. Small groups of cars (for example 30) are sorted out to model. At the next counting point the structure of the group of cars will have changed. The possibilities are for instance that some cars have overtaken other cars, some cars have left the road and other cars will have come on the road after passing the counter point. From the results of monte carlo simulation (see appendix 1), it appears that such an approach is realizable as long as the distance between the counting points is no more than 10 to 20 kilometers. In this simulation a percentage up to 15% of vehicles coming and going has been reckoned with. Before it is possible to draw a final conclusion, the simulation, however, will have to be repeated with real data.

An approach similar to this last alternative has been followed by Pfannerstill. Experiments on 5 km motor way between Aken and Heerlen without on and off traffic have demonstrated that groups of cars can be recognized by using pattern recognition techniques. In the Netherlands such a system could be realized with a limited extension of the current points of measurement, so that the cost, in proportion to the cost of proposed solutions, would be relatively low.

3.6. The influence of the delay on a road section somewhere else in the network

In the "route-guidance" or "traffic-co-ordination" systems equilibrium models are used to send the traffic to the various routes possible (see [10]). These kinds of models could be used with delay functions that are tuned to real situations. However, a problem is that these models do not sufficiently reckon with the dynamics of the system. The influence of a congestion - upstream - on the rest of the network is being neglected. This is not realistic and deserves extra attention at short term predictions. When, instead of a traditional two-dimensional model, a three-dimensional model is applied in the space time period (see [9]), then this problem has been removed. In figure 4, figure 5, figure 6, figure 7 an example has been given of such an application.

figure 4

expected MEAV at 7.05 hour

figure 5

expected MEAV at 0.30 hour
The prediction is made for every 10 minutes beforehand. The prediction has been made for various parts of the network in situations of good and bad weather. The figures have been made with the TFTP system (see [8]). A dark shade means a greater risk of delay. By means of the procedure described here, it can at the same time be determined how it is possible that a delay can influence the traffic regulation downstream, which can be important to the police for the measures to take.

A. CONCEPTS OF SYSTEMS IN DEVELOPMENT

At present the technical development of traffic systems (see [3], [16]) proceeds for the greater part in the direction of systems which can be installed on roads or in cars. In Europe examples of this are CARIN of Philips and ARI of BLAUPUNKT (see [4]). In America ERGS is being developed and so are the systems CAGS and AMTICS in Japan (see respectively [5], [17], [12]). Broadly speaking, the bases of these systems are that the time of departure cannot be changed. However, in many cases this will be possible, though.

A second principle of the systems mentioned above is that there is a network of alternative routes, so that network models of transport can be applied (see for example [3]). Although this will mostly be the case in urban building, the question remains whether it should always be desirable to divert the traffic via secondary routes outside the cities.

5. CONCEPT OF A NEW SYSTEM

Moreover, with the technological advancement it is possible to store large quantities of "situation-dependent" information in databases. Developments in the field of artificial intelligence make it likely that the information can be supplied in such a way that explanation will be given with regard to the style of argumentation used, so that there will be a choice: either to follow the advice or to give one's own interpretation of this information.

The measurements of intensity have to be stored together with the information on delays and the circumstances among which these delays have taken place. Then, from this, rules can be deduced with respect to the delays expected when these circumstances will happen again. Besides, the system will have to contain rules on the reaction of car drivers. A closer investigation will have to be made.
The system proposed contains:
- a data acquisition consisting of intensities and delays per counting point;
- a system with rules on the circumstances under which the changes in the capacities, the intensities and the delays will appear;
- a component which passes the predictions on to the road users with which, at the same time, an explanation is given about the way in which a prediction is established;
- a component with which the influence of the warning on the intensity can be calculated;
- a component with which the influence of additional rules, to be drawn up by the traffic controle, regarding events and such, can be calculated.

6. CONCLUDING REMARKS

The prediction of delays (resulting from queues) before a road user starts driving, can be more effective than information given on the length of queues during the drive. The providing with proper information, in good time, enables the road user to take the decision "how to avoid the congestion" on his own.

The information is only useful when it is also reliable. The calculations of delay with the well-known equilibrium models are too rough to be used on short term predictions. It is suggested to decrease the spreading in the data by specifying the delay to road section, point of time and circumstance.

It is not possible to determine the delay per road section directly. The delay can be deduced from traffic counts and speed information. Simulation shows that further investigation can be useful here.

Delays in networks can be calculated with equilibrium models if specified functions as to place, time and circumstance are used. The adding of a time dimension to these models can enlarge their sense of reality.

The approach proposed requires an information system. The amount of data in this system becomes, because of the differentiation in data and the frequency with which these data have to be gathered, very large. The possibility of using the well-known styles of argumentation from artificial intelligence to shorten the calculation time, has to be investigated.

The purpose of this system is to have road users react on warnings, which will cause a decrease of part of the queues. The information system will implicitly measure the reactions on the warnings and will consequently take these into account at the next warning.

REFERENCES


conference on transportation for the future, Sodertalje, Sweden


1992

Analogies in communication systems.

Ben H. Immers and Laurent J.C. de Ben

6th World Conference on Transport Research, Lyon, France
1. INTRODUCTION

The policy document 'Telematics in the Transport Sector' of the Dutch Ministry of Transport holds out the prospect of various new developments in the field of traffic and transport control in connection with telematics.

Telematics here is understood to mean the linking of telecommunication systems and information systems. In traffic and transport it has found particular application in the fields of:

- route planning (passenger transport)
- traffic guidance (passenger and freight transport)
- enforcement and surveillance (passenger and freight transport)
- monitoring and handling of goods (freight transport)
- market mediation and transactions (freight transport)
- substitution (passenger transport)

For this range of applications intensive communication is required between the system and the system manager (monitoring), between the system manager and the users (route information etc.) and between the users themselves (EDI etc.). Therefore new developments and investments are called for with regard to the infrastructure of telematics.

In this process of innovation and development three questions will have to be answered:

- What new developments are likely to emerge in the field of telecommunications?
- To what extent can we quantify the future demand for telecommunications (type and size); are there any interactions with other forms of communication, and
- What is the best way of shaping the telecommunications infrastructure (besides road, water, air, rail and pipelines).

Some of the basic (more clearly defined) subjects of research related to the above may be formulated as follows:

1. Research into the interactions between communication, traffic and transport on the one hand, and land use and land-use planning on the other hand.
2. Research into highly local networks (internal, within a company, intramural).
3. Research into the impedance sensitivity in the use of the various telecommunication techniques and related to this the network structure.

Three departments at Delft University are involved in exploring the above themes: the Department of Transportation Planning and Highway Engineering, the Department of Housing, Land-use Planning and Pollution Control, and the Department of Telecommunication and Traffic Guidance Systems.
The third subject of research (the impedance sensitivity of telecommunication techniques) will be dealt with in this paper. The investigation will be focused in particular on quantifying nature and extent of the demand for communication (between various agents). In this investigation two questions are very important:
- To what extent will existing and new communication techniques be used, and
- To what extent will new communication techniques have a substituting or else a generating effect on the use of existing communication techniques (physical movements, telephone, etc.)

In seeking to answer the above questions, the complexity of the message to be transferred, and related to this, the quality of the data transmission (communication) play a significant part. Thus telephone connections that are sensitive to disturbances are unsuited for transmitting complex information where for instance facial expression is essential. As for the desired network structure, besides speed and capacity also the distribution pattern of the data transmission is important. In this distribution process three determining factors are to be distinguished:
- Data transmission involves costs and time. These are partly dependent on the distance.
- To a large extent the exchange of information is kept down by the circle of acquaintances and relations that, for various reasons (e.g. desired face-to-face contacts), live in the immediate surroundings.
- A great many messages and information exchanged by means of telecommunication techniques lead to or have been preceded by physical movement. Therefore, constraint factors relating to physical travel behaviour also affect the distribution of information through telecommunication techniques.

In the study the following steps will be taken:
- The impedance will be measured for the various forms of communication; to what extent do threshold values arise with the various media.
- Which factors determine the level of the contacts per medium; are they mainly socio-geographical variables or economic ones too (income), or else do autonomous trends occur.
- To what extent does interaction between the various forms of communication occur; can these interactions be translated into substituting or else generating effects.

On the basis of the results of the three above-mentioned partial studies, it should be possible to predict the nature and amount of the communication (contacts) per separate medium (face to face, telephone, infrared, satellite, etc.). This knowledge is relevant to some other studies which are being carried out within the framework of the present project, viz.:
- the setting up of the telecommunications infrastructure.
- the calculation of trade-offs between travel behaviour and telecommunications.

2. FORMULATING THE PROBLEM IN DETAIL

Communication arises as a result of contacts between people. The aim of communication is transferring information from origin (person who sends) to destination (person who receives). In general communication leads to a reduction of uncertainties or things unknown. Thus it helps to provide
people with a great many primary necessities of life, such as food, clothes and education. Not surprisingly then, there is a great diversity in both scale and spatial manifestation.

There are a lot of means by which information can be transmitted (communications media). A balancing of costs and benefits (speed, capacity, quality) may lead to either a barrier or an acceptance with regard to using a particular communications medium.

On the other hand the expected use of the communications medium determines to a large extent the desired network structure. In developing existing and potential alternative communications media, the opportunities for investment of both manager and user will be of great importance.

Meeting supply and demand plays a significant part in the development and expansion of communications network structures. In this respect data-communication systems for transmitting immaterial information flows (radio, television, telephone, telex, face-to-face contacts, etc.) are no different from traffic and transport infrastructure networks for the transfer of material, physical information flows (passenger traffic, postal traffic).

For the planning and management of communications networks the factors determining to what extent they will be used need to be known. It is useful here to distinguish three lines of approach:

- **Extent of the demand for communication**

In the past few years a great many studies have been carried out about potential factors determining a specific demand for communication. Within companies and households one might think of: position (function), form of organisation, degree of automation. So far most of the studies found in the literature have been limited to the more qualitative lines of approach (cf. Maggi, Maggi & Hengevoss and others).

If we look at communication on a larger scale then factors of influence might be found in

A) local and regional socio-economic characteristics; the composition of the household, the number of residents, individual spending budgets, nature and quantity of employment.

B) developments in society as a whole; e.g. technological and socio-cultural developments (emancipation, individualization, increase in scale).

- **Spatial distribution pattern**

The characteristics of individuals' interaction patterns will be somehow related to the actual spatial situation and any developments taking place in it. The distribution of activity spaces, the tariff, but also the introduction of mobile agents (origins and destinations, e.g. the car phone) are some of the developments influencing this interaction pattern.

- **Interaction between communications media**

If individual choice is taken as the starting-point as regards the use of communications media then this implies that generation and substitution effects may arise. Of particular interest is the phenomenon where a transfer of information by physical movement is replaced by an immaterial, electronic transfer of information e.g. teleworking, teletraining and teleshopping. However, electronic transfer of information may also generate additional physical transfers (or other interactions) and vice ver-
sa. In the past, statements have been made about the influence of generating and substituting effects of the immaterial flows of information on mobility in future situations (e.g. the Swiss MANTO-Project). The policy document 'Telematics, Traffic and Transport' states that commuter traffic will be reduced to a maximum of 15%, shopping and business traffic to a maximum of 20%.

This study deals only with communications on interregional relations. A model representation is sought which will show more clearly the interaction between different communications media. The paper contains a number of ideas and a first (rough) analysis. Attention is focused on the distribution pattern of three communications media, namely personal business traffic, postal traffic and telephone traffic. Chapter 3 describes the relation between costs and use from a theoretical point of view. The theory is tested by means of practical observations (Chapter 4 and 5). Chapter 6, finally, contains some comments with regard to the possibilities of model development, taking into account the interaction between the various communications media.

3. BARRIERS TO COMMUNICATION

Theoretically, if time and costs budgets are irrelevant, the quantity of information to be transferred will only depend on the capacity of the information medium. Basically, in case of sufficient capacity any message is transferable. In practice however, time and costs budgets are limited; they constitute therefore a barrier (impedance) to the use of an information medium. A reduction in the number of contacts will be related to an observable increase in the impedance.

With respect to communication two factors may determine the value of the impedance:
- the costs (length of time) of bringing about a contact; these depend on the information medium chosen, and on the distance between the individuals involved in the contact.
- the costs (length of time) of transmitting a message; these depend on the information medium chosen, the distance between the individuals involved and the nature of the message to be transmitted (complex or simple, professional or private).

Table 1 shows the practicability of three communications media on the basis of impedance. The terms optimal, good and bad indicate the mutual relations with respect to potential use.

Table 1: influence of the impedance for three communications media

<table>
<thead>
<tr>
<th></th>
<th>short distance message</th>
<th>short distance complex message</th>
<th>long distance simple message</th>
<th>long distance complex message</th>
</tr>
</thead>
<tbody>
<tr>
<td>post</td>
<td>bad</td>
<td>bad</td>
<td>good</td>
<td>optimal</td>
</tr>
<tr>
<td>telephone</td>
<td>optimal</td>
<td>good</td>
<td>optimal</td>
<td>good</td>
</tr>
<tr>
<td>face to face</td>
<td>good</td>
<td>optimal</td>
<td>bad</td>
<td>bad</td>
</tr>
</tbody>
</table>
For postal traffic the transfer time and the transfer costs are independent of the transfer distance. For contacts across relatively short distances postal traffic is a relatively expensive communications medium; for contacts across relatively long distances it is a comparatively cheap medium.

Just as for postal traffic, transfer time and transfer costs for telephone traffic are negligible. The length of the call is related to the complexity of the message. Optimum use of the telephone takes place in case of comparatively simple messages, with the emphasis on speed and the possibility of entering into a direct dialogue. Telephone calls of this kind are usually rather routine affairs, characterized by short contacts at non-prearranged moments.

Face-to-face contacts will mainly take place on relations with a relatively short distance. Transfer costs in the form of travel times will increase (too) rapidly with increasing distance. The directness of the contact enables one to transfer (highly) complex messages. Quite often a face-to-face contact may take place following a telephone contact (for instance due to the complexity of the message).

Fixed costs consist in, among other things, investments connected with installation and management of the infrastructure and investments in know-how. These costs are mainly paid for by the manager of the communications medium; the user is only indirectly involved. However, the user, too, will have to invest time and money (experience and costs of telephone, car, bicycle) before being able to use the medium. The latter may in some cases lead to an (initial) barrier.

For the time being the influence of the nature of the message on the length of the conversation for the various communications media will not be considered. If this simplifies reality, it also provides a clearer picture of the possible effects of the other factors of influence, namely the communications medium and the distance between the individuals involved. A diagram of the user's distribution pattern of communication is shown in Figure 1.

The flow of information to be transferred (I: y-axis) is plotted against the distance between the individuals involved in the contact (d: x-axis). The horizontal line I_{max} is a rough estimate of the need for communication. It is assumed that the various communications media cannot entirely meet the total demand.

The amount of telephone traffic is more or less independent of distance. Postal traffic contributes relatively little on the shorter distances;
the contribution of face-to-face contacts on the shorter distances is relatively large. The enclosing continuous line indicates the sensitivity to distance if the communications media are considered in relation to each other. The shaded part indicates the region in which only telephone and post contribute in meeting the demand for communication.

It is interesting to investigate next to what extent this distribution pattern corresponds to the distribution pattern that can be deduced from observations concerning the use of the various media. The results of these analyses will be presented in Chapters 4 and 5.

4. METHOD

For the year 1983 the influence of distance as a measure of spatial diffusion was investigated for three communications media, viz.
- separate post 1983 (averaged in kg per week)
- telephone calls 1983 (in Erlang, at peak hours)
- passenger traffic 1983 (number of trips)

The unity of Erlang refers to the amount of telephone traffic, related to the duration of the call. One Erlang corresponds to e.g. 60 calls of 1 minute each or one call of 60 minutes. It was furthermore investigated to what extent for telephone traffic regional characteristics affect the distribution pattern. Finally the stability of the distribution pattern in time was determined. To this end data were used of telephone calls during the years 1968, 1974 and 1983.

The analysis techniques employed are the same as the ones commonly used in traffic engineering (e.g. the multiproportional poisson estimate, cf. Hamerslag et al.). The data concerning the use of the various media were supplemented by relevant socio-economic data from the same years (source: Dutch Telecom, Dutch Central Statistical Office). The zones into which the country was divided (cf. Fig. 2) correspond to the 22 telephone districts in the Netherlands (the main telephone exchanges of Dutch Telecom). As to the amount of intrazonal traffic assumptions were made.

Fig. 2: Division of The Netherlands into zones
5. ANALYTICAL FINDINGS

5.1. The influence of distance

Figure 3 is a diagram of the distribution pattern for the three communications media. The Weighted Poisson Estimator was used as the estimation method.

\[ F(Z(i,j),t) = a \exp(-b(\ln(cZ(i,j)+d))) \]

\( Z(i,j) \) is the impedance (e.g. distance or generalized times). By means of regression analysis the coefficients \( a \) (level component), \( b, c, \) and \( d \) (indicating the degree of sensitivity to changes in the distance) are estimated.

In Table 2, analogous with model approaches in traffic engineering, the...
results of the regression analysis are given for the three communications media. Since the magnitude of the flow of information (I) is expressed in different units, not much significance is to be attached to the value of the level component (a). The table shows clearly noticeable differences in the distribution pattern (compare the b-coefficients). The values of the coefficients c and d in (***) have been calculated, starting from the b-value of the car function. It is possible now to roughly calculate the ratio of the three communication media with regard to distance sensitivity. This is (approximately):

\[
\text{car traffic} : \text{telephone traffic} : \text{separate post} = 1 : 0.10 : 0.07.
\]

Table 2 Coefficients of the lognormal deterrence function

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>separate post (1983)</td>
<td>216025</td>
<td>-0.046</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>telephone traffic (1983)</td>
<td>-</td>
<td>-0.395</td>
<td>0.07</td>
<td>0</td>
</tr>
<tr>
<td>person traffic (1983)</td>
<td>1207</td>
<td>-0.149</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>87282204</td>
<td>-0.395</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

5.2. The influence of regional characteristics

From the lognormal distribution estimates can be made of the amount of telephone traffic generated by each zone. These estimates can be compared to the observed values, on which the estimates are based. Table 3 shows for 16 zones to what extent the estimated amount of generated telephone traffic deviates from the actual value. The deviation in percentage shown in Table 3 has been calculated as follows:

\[
\text{Dev} = \sum \left( \frac{\hat{y}(i,j) - y(i,j)}{\sqrt{v(i)}} \right)
\]

Table 3: Deviations in percentages between estimated and actual generation per district; year: 1983.

<table>
<thead>
<tr>
<th>Zone</th>
<th>asd</th>
<th>bd</th>
<th>dv</th>
<th>ehv</th>
<th>gv</th>
<th>gn</th>
<th>hlm</th>
<th>hgl</th>
<th>ht</th>
<th>lw</th>
<th>mt</th>
<th>nm</th>
<th>rt</th>
<th>tb</th>
<th>ut</th>
<th>zl</th>
</tr>
</thead>
<tbody>
<tr>
<td>Afw</td>
<td>-5.6</td>
<td>-6.4</td>
<td>7.0</td>
<td>2.9</td>
<td>0.1</td>
<td>7.7</td>
<td>-1.1</td>
<td>9.0</td>
<td>-4.7</td>
<td>3.1</td>
<td>-0.1</td>
<td>-1.0</td>
<td>-12.5</td>
<td>-1.3</td>
<td>-2.8</td>
<td>0.2</td>
</tr>
</tbody>
</table>

asd = Amsterdam, gd = The Hague, hlm = Haarlem, hgl = Hengelo, hv = Groningen, Inw = Leeuwarden, tb = Tilburg;
bd = Breda, gn = Groningen, lw = Leeuwarden, tb = Tilburg;
ehv = Eindhoven, hlm = Haarlem, mt = Maastricht, ut = Utrecht;

The north-eastern districts show a slight overestimation (Groningen 7.7%, Leeuwarden 3.1%, Deventer 7.0%, Hengelo 9.0%). In the 'Randstad' (the urban conglomeration in the west of the Netherlands) there is in many cases an underestimation (Amsterdam -5.6%, Rotterdam -12.5%, Utrecht -2.8%). Apparently telephone traffic from or to the northeast is less widely distributed than telephone traffic to or from the Randstad area (fewer interregional calls).

8
5.3. Stability in time

In the above analyses of the distribution pattern, data from the past were used. In order to make correct predictions, however, it is important to have insight into expected developments as to the distribution pattern in future situations. The estimated values of the distribution function for telephone traffic for three years are given in Table 4. A diagram is shown in Figure 4.

Table 4: The estimated values of the deterrence function $F(Z(i,j))$ and the values of the level component $\delta(t)$.

<table>
<thead>
<tr>
<th>crow flight distance(km)</th>
<th>value deterrence function $F(Z(i,j),1968)$</th>
<th>level-comp. $\delta(1968)$</th>
<th>level-comp. $\delta(1974)$</th>
<th>level-comp. $\delta(1983)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>816.42</td>
<td>1.00</td>
<td>1.90</td>
<td>2.76</td>
</tr>
<tr>
<td>41</td>
<td>567.60</td>
<td>1.00</td>
<td>1.86</td>
<td>2.83</td>
</tr>
<tr>
<td>57</td>
<td>359.70</td>
<td>1.00</td>
<td>1.84</td>
<td>2.69</td>
</tr>
<tr>
<td>70</td>
<td>310.58</td>
<td>1.00</td>
<td>1.80</td>
<td>2.61</td>
</tr>
<tr>
<td>97</td>
<td>194.77</td>
<td>1.00</td>
<td>1.81</td>
<td>2.65</td>
</tr>
<tr>
<td>130</td>
<td>141.79</td>
<td>1.00</td>
<td>1.82</td>
<td>2.59</td>
</tr>
<tr>
<td>165</td>
<td>105.95</td>
<td>1.00</td>
<td>1.86</td>
<td>2.62</td>
</tr>
<tr>
<td>214</td>
<td>84.10</td>
<td>1.00</td>
<td>1.69</td>
<td>2.23</td>
</tr>
<tr>
<td>320</td>
<td>33.03</td>
<td>1.00</td>
<td>1.32</td>
<td>2.05</td>
</tr>
</tbody>
</table>

Figure 4: Distribution functions of telephone calls; years: 1968, 1974, 1983
The level component $\delta(t)$ shows an increase in the course of the years. This increase is more or less the same on all relations (on relatively short distances as well as on relatively long ones). From this it may be concluded that although the total number of telephone calls has increased over the past few years, the distribution pattern (degree of sensitivity to changes in the distance) has hardly changed.

6. CONCLUSIONS

On the basis of theoretical assumptions it was assumed that the distribution pattern of interregional telephone contacts was hardly affected by either the factor of distance or regional characteristics. It so happens that there is only one tariff zone in the Netherlands besides the tariff for local telephone calls. Yet the analytical findings of empirical data indicate that there is a distinct relation between the distance (district) and the number of telephone calls, contradicting assumptions based on the sacrifice/utility principle.

Some of the factors that might determine the sensitivity of the number of telephone calls to changes in the distance follow below (see the Introduction):

- To a large extent the exchange of information is kept down by the circle of acquaintances and relations that, for various reasons (e.g. desired face-to-face contacts), live in the immediate surroundings.

- A large number of messages and exchanges of information are linked to a physical movement later on (or have been preceded by one). This is one reason why the distance factor which strongly influences the distribution pattern of physical travel behaviour, also affects the distribution pattern of the transfer of information by means of telecommunication.

A comparison of the distribution patterns of the three communications media studied shows that there are some clear differences in sensitivity to distance.

Both findings of the study underline the necessity of investigating the interaction that has been shown to exist between the use of (new) telecommunication techniques and physical travel behaviour, as well as the interaction between telecommunication techniques themselves. To a great extent they determine both the level of use (the occurrence of congestion) and the distribution pattern (the network structure to be chosen).

The above-mentioned interaction may even lead to a situation where the desired effects of the introduction of new communication techniques fail to come about or (still worse) reverse effects occur.

The distribution pattern of telephone calls does show a high degree of stability in time. If this property also holds true for other telecommunication techniques, then, taking into account the interaction mentioned above, it may well be possible to make predictions about the future use of telecommunications media.
ARNBAK, J.C.;
The two economies of telecommunication networks; In: Informatie en informatiebeleid, vijfde jaargang, 1987 no.4.

CBS;
OVG-enquêtebestand (Dutch National Travel Survey); Documentation for users data 1983; Heerlen.

CBS;
Werkzame personen naar geslacht, positie in het bedrijf en bedrijfstak per provincie (Work force per province according to gender, employment status and branch; jaarlijkse arbeids-krachtentelling, Heerlen.

Ministry of Transport and Public Works;
Nota telematica in Verkeer en Vervoer (White paper on Telematics in the Transport Sector); The Hague, February 1990.

DIETVORST, A.G.J.;
Telephone traffic and the economic structure of the Netherlands; application of a field theory; Katholieke Universiteit Nijmegen, Geografisch planologisch instituut, Nijmegen, November 1979.

DREWE, P.;
Telematics, the state of art and where do we go from here; In: Proceedings Colloquium Vervoersplanologisch Speurwerk, The Hague, November 1988.

FISHER, J.C. en R.H. PRY;
A simple substitution model of technological change; Technological forecasting and social change 3, pp.75-88, 1971.

HAMERSLAG, R en A.W. DERSJANT;
A multiproportional estimation method for the simultaneous determination of the value of the deterrence function by mode; PTRC annual meeting, 1976; Warwick.

DE JONG, M.A. en M.F.A.M. VAN MAARSEVEEN;

MAGGI, R.;
Towards an economic theory of barriers to communication; Papers of the regional science association, voll. 66, 1989, pp 131-141.

MAGGI, R. en T. HENGEVOSS;
Raumwiderstand bei der telecommunication; Mitteilungen des Arbeitskreises für neue Methoden in de Regionalforschung, Regional Science and Urban Economics, 17 109-21.

PTT;
Thoughts, Methods and Investigations; The Hague; December 1988.
Telecommunication in combination with information systems (telematics) is bound to play an important role in future transportation systems. Possible fields for application of telematics in the transport sector are:
- route guidance, route planning
- traffic control
- enforcement and surveillance
- handling of goods
- transactions of goods
- substitution of trips

These new developments in the transport sector require intensive communication between a.o. the system and system manager (e.g. monitoring), between the system manager and users of the system (e.g. route information) and between the users themselves (e.g. electronic data interchange). This intensive communication necessitates new developments/investments in the telematics infrastructure.

In mastering this development process some questions which have to be addressed are:
- Which new developments will emerge in the field of telecommunications/telematics
- Is it possible to quantify the demand for telecommunication facilities and are there any interactions between the different forms of communication
- What is the optimal form of the future telecommunication infrastructure.

This paper deals with the results of a study in which especially the future demand for telecommunication facilities has been investigated. The following subjects will be addressed:
- Quantification of the impedance of different forms of communication as well as determination of possible barriers to the use of communication media
- Determination of factors affecting the level of communication
- Determination of possible interactions between the different forms of communication, resulting in substitution or generation effects.

After a general introduction to the subject, referring to studies performed by a.o. Maggi, Dietvorst, Hengevoss and Drewe, the analogies in communication patterns are investigated by comparing origin-destination data for personal business traffic, post traffic and telephone calls. For each communication medium polarities (potentiality for generation/attraction of interactions) and deterrence (impedance) functions are calculated and compared with each other.

As for telephone calls, data are available for three different years (1968, 1974, 1983). Data of these three different years are used to investigate the stability in time of the distribution pattern.

The results of the study clearly indicate that there are analogies in the use of different communication media. These analogies refer to the level of interaction as well as the distribution pattern of the interactions. Furthermore the results of the study indicate that the introduction of new telecommunication techniques will not only generate substitution effects but also generation effects with respect to the daily trip pattern of people.
1992

Determination of real time travel times on congested motorways for short term applications.

M. Westerman, L.H. Immers and R. Hamerslag

2nd International Capri Seminar On Urban Traffic Networks. Capri, Italy
Determination of real time travel times on congested motorways for short term applications

M. Westerman, L.H. Immers, R. Hamerslag
Delft University of Technology, Faculty of Civil Engineering
Dept. of Transportation Planning and Highway Engineering

ABSTRACT
The paper starts with a distinction between various target groups, where each target group has its own applications and its own demands concerning actuality and accuracy of real time travel times.

Measuring real-time travel times can take place by means of several types of hardware and software. First, a list of the most common types of hardware is given including the advantages and disadvantages for application within the boundaries of the limiting conditions. Second, a summary of existing automated systems to record travel times is presented. None of these systems appears to be able to meet the system demands. A new method to measure real time travel times is the main focus of this paper.

This new method is based on the comparison of absolute vehicle counts by means of induction-loops at successive measuring sites along the motorway. To compensate for detection-errors corrections are made based on the correlation of intensity-fluctuations at the successive measuring sites.

We have implemented this method on a computer and tested it with real traffic data. The results show that the model can produce reliable (mean) travel times on road segments with lengths ranging from five to ten kilometres. In general, the presence of a traffic-jam between the measuring sites presents no problem for the reliability and the accuracy of the method.

1. INTRODUCTION
The determination of speeds on motorways related to volumes, traditionally takes place by means of induction loop detectors. Using volume, speed and distance the travel time is calculated using a delay-function.

This method leads to acceptable results for traffic planning. For dynamic traffic management and traffic control the travel time calculated in this way is not sufficiently accurate. Especially in congested situations irregular speeds occur, resulting in travel times with a very low accuracy. Furthermore, the relation between volumes and speeds - as indicated by the delay-function - only gives an estimation of the travel times. This relation, and therefor also this estimation, is not beyond all doubt.

Therefor it is necessary to develop a better method for determining real time travel times, which can be applied for short term applications.
This method, which we will describe in this paper, is new with respect to its approach of travel times. It does not use a relation between several elementary data to calculate the travel times, but directly determines the travel times.

2. INFORMATION REQUIREMENTS

We can distinguish various target groups for information about real-time travel times. Each of these target groups has its own applications and its own demands concerning actuality (for instance each minute, every 5 minutes, every hour, etc.) and accuracy. It is important to make this distinction, because this brings about significant limiting conditions.

Below, we listed the possible groups of applications for information about real-time travel times ([1]-[5]).

**Advanced traffic management and traffic control systems**

Short term management, like advanced traffic management systems and traffic control systems, is concerned with tactic decisions and purposes like incident handling and variable signing.

Automatic Incident Detection based on travel times on short road segments from 500 metres to 1 kilometre. For this application the dispersion in the travel times should not exceed 25% or \( \frac{1}{2} \) minute.

For variable signing based on drops of the travel time the maximal allowed dispersion amounts circa 20% or 5 minutes and depends on length of road segment. No actions should be taken in the case the drops are shorter.

**Traffic and transportation planning and policies**

Long term management, like traffic and transportation planning and policies, is concerned with strategic decisions and purposes like policy and plan making. The travel times needed for this application do not need to be very accurate because the concerned time period is long, circa 25-35% dispersion is allowed.

Also road maintenance authorities can use information about real-time travel times as basic information for actions to be taken by operational services like public works and police. Because only short road segments are concerned this information should be accurate up to 5-15 minutes.

**Advanced traveller information systems**

Advanced traveller information systems make use of real-time travel times. Car drivers for instance, can get information about real-time travel times in their cars. This information can be given directly or can be converted into other information like alternative routes. It should be very accurate, because car drivers will not react according to unreliable information. The maximum allowed dispersion is circa 10%. This means in urban areas about 2-3 minutes and in sub-urban areas about 5 minutes.
Fleet management

For planning routes of trucks or for dynamic route planning systems (home based) for long range road transfer and haulage it is important for the transporters to have information over travel times on the planned routes. Although the needed accuracy highly depends on the planned journey the dispersion should not exceed 10-15% or 20-30 minutes.

For distribution centres and couriers, which apply their planning for the short range, also a dispersion of 15% is allowed which in this case means about 5-10 minutes.

The car industry and car electronic industry is affected by information about real-time travel times because the apparatus they are constructing will only function if the right information is available. The apparatus are only reliable if the information is correct, so a dispersion of more than 15-20% is not allowed.

To test research concerned with the above mentioned applications researchers may want to use real-time travel times. Dispersions of about 15% are acceptable in this case.

3. MEASURING EQUIPMENT

For measuring real time traffic data different types of hardware can be used. Each type of hardware has its own advantages and disadvantages. It strongly depends on the application and the demands which type of hardware is preferred.

We will shortly describe the most common types of hardware and give their advantages (+) and disadvantages (-) for measuring real time travel times.

1. Induction loops

A vehicle passing an induction loop in the road surface causes an induction power which to a certain degree is characteristic for this vehicle.

+ popular, tested, well used, inexpensive, no need for on-board unit, bad weather conditions present no problems

- only one-way communication, high failure, high maintenance, low transmission rate

2. Infrared systems

Infrared beacons use a beam of energy in the infrared band between a transmitter and a receiver. A traversing vehicle breaks this beam and generates an impulse.

+ possibility of two-way communication with on-board unit, not very expensive, bad weather conditions present no problems, high transmission rate, reasonably tested

- only used to detect speeds of cars and distance between cars or to exchange messages, small beam, short distances
3. Microwave systems
Between a beacon along the road and a antenna on the car messages can be exchanged.
+ possibility of two-way communication with on-board unit, not very expensive, bad weather conditions present no problems, high transmission rate, good vehicle location precision
- can be used to measure speeds of cars without on board unit (Doppler effect) but need for on board unit to communicate, small beam, short distances

4. Radio beacons
By using radio signals of at least three (radio)broadcasting stations and calculating the phase difference between them (due to differences in the distance between receiver and transmitters) the position of the vehicle can be gained with an accuracy of 10 centimetres.
+ good vehicle location precision,
- need for on board unit,

5. Cameras
Traffic flow is continuously registered by means of cameras linked to a computer. By using image processing vehicles can be distinguished.
+ no need for on board unit, very precise vehicle location and maybe even identification
- only in test-phase, expensive, no two-way communication, not insensible to weather conditions

6. Cellular radio
Coverage area is divided into small cells with their own basic radio stations using a certain radio frequency. The same frequencies can be re-allocated in non-adjacent cells.
+ Pan-European, two-way communication, bad weather conditions present no problems, reasonable on board unit costs
- vehicle location only cell-size precision, low transmission rate

7. Satellites
Communication between satellite and vehicle and satellite-groundstation-vehicle is possible.
+ covers large area, long distances, continuously vehicle location information, bad weather conditions present no problems
- need for expensive on board unit, not very precise vehicle location, used for billing purposes or for giving information to users

Evaluation
The system making use of satellites needs an expensive on-board unit and has a low location precision. Also the current system making use of the cellular telephone offers a low location precision. Infrared and microwave systems both need on-board units and are not tested very thoroughly [22].
The use of two-way systems like beacons (infrared or microwave) would make it possible to construct a very elegant system for measuring real-time travel times. A disadvantage is that a large number of vehicles has to be occupied with an on-board unit to get sufficient reliable and accurate travel times. For getting reliable intensity measurements when using beacons, (almost) all vehicles have to be occupied with on-board units. As this will not be a realistic situation for some years, induction loops are still necessary. A combined system, composed of induction loops and beacons, would have realistic and practical possibilities.

At the current state of the art, for measuring real-time travel times on the complete motorway system, induction loops would be the best alternative. It is not very expensive, it is well tested, it has no need for an on-board unit and bad weather conditions present no problems. The fact that only one-way communication is possible and that the transmission rate is low, is not necessarily a problem. The high failure rate of the induction loops should be given due consideration.

In the next we will describe some existing methods which may be used for determining real-time travel times on motorways, based on elementary data gathered by means of induction loops.

4. EXISTING METHODS FOR MEASURING TRAVEL TIMES

In the past some research has been carried out concerning real time travel times on motorways. In this chapter we will give a survey of this research and give the advantages and disadvantages. As the elementary data has to be gathered by means of induction loops and the mutual distance between two loops should be sufficiently large to be applied on the complete Dutch motorway network (i.e. road segments with lengths of 5 to 10 kilometres), these are taken as limiting conditions.

1. Recognition of individual vehicles by means of footprints

At Delft University of Technology, the Department of Electrical Engineering, Telecommunications and Traffic-control, as part of a study into Automatic Incident Detection, is paying attention to the recognition of vehicles. The developed method can also be used for determining travel times ([6], [7] and [8]). This method uses "induction-loop signatures" or "footprints" (a vehicle passing an induction loop in the road surface causes an induction power which to a certain degree is characteristic for this vehicle). By comparing the footprints of successive measuring sites travel times can be obtained.

A disadvantage of the method is the required detailed analysis of the footprints which are necessary and which strongly depend on the composition of the signal. Changes in the signal (due to replacement of the hardware or software or changes in car size) have an impact on the method of analyzing the signals. Another disadvantage is that the signals are distorted when vehicles pass the loops very slowly or halfway (e.g. in case of traffic jams or overtaking manoeuvres). The "re-identification" of footprints is - due to the above factors - only reliable on short distances.

Consequently this method is not applicable for determining travel times on road segments with lengths of 5 to 10 kilometres.
2. Recognition of groups of vehicles by means of footprints

In the Federal Republic of Germany research is carried out by AVE Verkehrs und Informationstechnik GmbH in Aachen ([9], [10]), concerning automatic monitoring of traffic conditions by re-identification of vehicles.

For a number of vehicles passing the first measuring site, e.g. 50, the most striking characteristics are taken together and considered as a whole. The footprints of vehicles passing the second measuring site are continuously, same group size and shifting per vehicle, compared with the characteristics measured at the first site. When the correlation between these two groups is maximal, both groups are considered to be the same. In this way, absolute recognition of individual vehicles is not necessarily.

The maximum allowed distance between two successive measuring sites for this method is about 5 kilometres.

A problem, when using this method, could arise near ramps, because here the composition of the traffic flow can change considerably. Differences in speed and vehicles changing traffic lanes or overtaking each other also disturb the composition of the traffic flow. The effect of the disturbances on the accuracy of the travel times has not been studied yet. Other disadvantages of the method are the high demand on the available hardware and software as well as the already mentioned disadvantages with respect to the intensive analysis of the footprints (see 1.). This would make it too expensive for nationwide use.

3. The measuring of speeds and intensities

This method for determining travel times is tested by Grontmij [11] as part of a system for variable signing, currently under development at the Transportation and Traffic Research Division of Rijkswaterstaat.

\[
\begin{align*}
I_k(t_i), &\quad v_k(t_i) &\quad I_{k+1}(t_i), &\quad v_{k+1}(t_i) &\quad I_{k+2}(t_i), &\quad v_{k+2}(t_i) \\
N_k(t_i) &\quad | &\quad N_{k+1}(t_i) &\quad | &\quad N_{k+2}(t_i) \\
&\quad |\quad \to L_k &\quad | \\
\end{align*}
\]

\[I_k = \text{number of vehicles passing site } k \text{ per period } \delta t \text{ (intensity)} \]
\[v_k = \text{(harmonic) mean speed (m/s) in site } k \text{ over period } \delta t \]
\[L_k = \text{length of road segment (m)} \]
\[N_k = \text{number of vehicles in segment} \]
\[t_i = \text{measuring time} \]
\[\delta t = t_i - t_{i-1} \text{ (aggregation period)} \]

Using the above draft two methods for calculating travel times were developed.

**Method I:**

In the case of uncongested traffic the mean travel time on a road segment can roughly be calculated using the incoming and outgoing speed in that section, assuming that the speed of the traffic flow has the incoming speed until halfway the section and that it has the outgoing speed in the second half of the section.
In this way, the travel time $T_k$ between sites $k$ and $k+1$, in the case of uncongested traffic, is

$$T(t_i) = \frac{1}{2} \frac{L_k}{v_k(t_i)} + \frac{1}{2} \frac{L_{k+1}}{v_{k+1}(t_i)}.$$ 

The study showed [12] that even if applied for traffic moving in a jam, the results are sufficiently accurate, provided that $\delta t \leq 2$ minutes and $L_k \leq 1$ kilometres.

**Method 2:**

An intuitively better method is using the ratio of the intensity and the incoming speed for obtaining a valuation of the local traffic density. The same is done for the outgoing traffic. The traffic is assumed to be uncongested and the measured figures (speed and intensity) are assumed to be representative for the section as a whole.

The mean of these two values is multiplied by the length of the section thus resulting in an initial valuation (in case of uncongested traffic) of the total number of vehicles in the section

$$N_k(t_i) = \frac{1}{2} L_k \cdot \left\{ \frac{I_k(t_i)}{v_k(t_i)} + \frac{I_{k+1}(t_i)}{v_{k+1}(t_i)} \right\}$$

$N_k(t_i)$ being known, is updated (also in the case of congestion) by adding the difference between the incoming and outgoing intensity:

$$N_k(t_{i+1}) = N_k(t_i) + \left\{ I_k(t_{i+1}) - I_{k+1}(t_{i+1}) \right\} \cdot \delta t$$

The travel time on the given section is:

$$T(t_i) = N_k(t_i) / I_k(t_i).$$

When the distance between the successive measuring sites does not exceed 500 metres the described methodology is accurate (method 1 slightly better than method 2 [12]), applied to distances up to 1 kilometre the results of the methodology are still reasonably accurate but when the distance is 1 kilometre or more the quality of the calculated travel times severely decreases (both methods).

When the traffic is monitored using measuring sites on a mutual distance of 5 to 10 kilometres - as is here the case - none of the described methods is applicable.

4. Correlation between intensity-flows

This method is based on the assumption that intensity is a continuously fluctuating process and that fluctuations, measured at location A, for some rate can be measured again at a location B downstream.

The method was developed by ITP/TNO using the software package PRIMAL [13].

Information about the number of vehicles passing during a certain time (intensity) is gathered by induction loops. This is done at two sites and consequently two stochastic signals are obtained, which are compared.

The time shift necessary for matching the two signals (the maximum correlation) equals the mean travel time in the case of uncongested traffic. In the case of congested traffic
this correlation disappears, as no characteristic fluctuations can be measured at the second measuring site ([13], [14], [15]).

In its present form the method is not suitable for determining current travel times in the case of congested traffic.

5. Cumulative distributions
In this method individually registered vehicle passings are added as result of which a cumulative time-dependent distribution accumulates [16]. The cumulative distributions of successive measuring sites are compared. The time shift needed to match both distributions, gives the travel time between the measuring sites. If desired this can be done separately for different categories of vehicles if they can be detected as such. Especially if vehicle categories drive at a different speed this may considerably improve the accuracy of the method.

A possible disadvantage of this method is that when a (small) percentage of vehicles is not or wrongly detected this could lead to an increasing error in the calculated travel time. This critically depends on the way of comparing both signals.

Evaluation
The first two methods using recognition of individual vehicles and groups of vehicles are not applicable because these methods require very powerful hard- and software and the results with distances between the detectors of more than 1 or 2 kilometres are not accurate any more. Research showed that no improvement could be expected. The third method using speeds and intensities gives travel times which are not very reliable and can only be applied to very short distances (approx. 500 m). The fourth method using correlation between intensity-flows gives accurate and reliable travel times, but only in free flow situations. Method five using cumulative distributions presents no more than an outline and will have to be defined more precisely by further research.

Departing from the descriptions above we choose a combination of the fourth and fifth alternative. By means of alternative four reliable and accurate travel times can be determined, while alternative five is to be used to guarantee the reliability and accuracy in case of congested traffic.

A method based on combining these two alternatives will be elaborated below.

5. METHOD
The method developed is based on simple counting in combination with pattern recognition.

The determination of the travel times is done separately for different stretches of motorway. We take into consideration a stretch of motorway bounded by A and B, where A and B are measuring sites at which induction loop detectors are placed. These induction loop detectors measure a vehicle passing a detector and record the point in time and the speed of the passing.

Basically, for the developed method we only use the point in time at which the vehicle passed the measuring site. The spreading of this information is much smaller than the spreading created by means of the delay-function to relate the intensity and capacity [16].
Therefor it should be possible to get more accurate results with the developed method than is possible with the traditional methods using the delay-function.

The vehicle passings, measured by the induction loop detectors, are added up cumulatively for each measuring site. This results in two cumulative time-dependent distributions (see figure 2). The determination of the travel times takes place by comparing these distributions. The time shift needed to match both distributions gives the travel time between the measuring sites.

**The basic principle**

Let us assume that $t_B$ is a point in time at which the travel time (of vehicles passing B at $t_B$) is known. In other words the former point in time $t_A$ at which the same vehicles passed in A is known. In formula: $t_B - t_A = \tau(t_B)$. These "starting points in time" $t_B$ and $t_A$ are being determined over and over in time ("re-calibrated"). How exactly this is done will be described later.

The travel time $\tau(t)$ at a later point in time $t$ can now easily be determined as follows. In the cumulative distribution for counting point B (see figure 1) the number of vehicles - say $N$ - that has passed between $t$ and $t_B$ in B can directly be found. Then in curve A the point in time $t'$ at which (since $t_B$) also $N$ vehicles have passed can be found. The travel time (for vehicles passing B at $t$) then is $\tau(t) = t - t'$.

When $\delta t$ is the aggregation period (e.g. 1 minute) $t$ can essentially walk through the whole registration-interval with steps of $\delta t$. For each aggregation period $\delta t$ a corresponding "characteristic" travel time $\tau$ will be found.
The inaccuracy

The effect of the induction loops frequently "missing" vehicles (1-2%, [18]) can be analyzed by assuming that the chance of missing a vehicle in the case of individual detections is equal to $p$. We also assume that these missings are mutually independent and that $p$ is independent of the intensity, road type, weather, and suchlike. Further we restrict ourselves to total intensity (so no division into vehicle categories, which would essentially be possible).

After $N$ vehicle passings the mean error is $pN$ (missings). With an intensity $I$ it also takes $pN / I$ of time longer before the correct number ($= N$) of vehicles has passed B, assuming that A counts correctly and B does not (worst case !). The error in the travel time increases also with $N$.

As $N = T I$, the error in the travel time is $pT$ (where $T$ is the period in which the cumulative curves are built without "re-calibration"; see The Re-calibration process).

When counting further using $p=0,02 = 2\%$ it follows that the error in travel time after $T=5$ minutes is $pT=0.1$ minutes, after $T=30$ minutes is $pT=0.6$ minutes, etc.

If we assume that the intensity $I$ is constant (also if $I$ varies this valuation is in this same order), then at a travel time $\tau$, which is for example 3 minutes, the relative inaccuracy in $\tau$ after $T = 30$ minutes has increased up to $0.6 / 3 = 20\%$ ! This means that after half an hour (or at the most after one hour) a "re-calibration" is necessary. This re-calibration time slightly depends on the distance between the measuring sites, the mean travel speed and of course on $p$.

In case of congestion "re-calibrations" are difficult or impossible (see The Re-calibration process), but then $\tau$ increases, so the relative error ($= p * T / \tau$) decreases.

The Re-calibration process

We showed that, especially when the algorithm is working for some time, the error due to counting errors at both measuring sites is too large to determine a sufficiently accurate travel time. To correct this error as well as to increase the reliability of the determination of the travel times the "starting points in time" $t^A$ and $t^B$ have to be fixed regularly so that $t_B - t_A$ equals the travel time (at $t_B$).

The re-calibration process is based on fluctuations in the intensity and on the assumption that characteristic fluctuations that are signalled at measuring site B can be recognized at measuring site A [15].

A time interval around the point in time $t$, at which the determination of the travel time takes place, is singled out of the B curve (B-block). Next, in the A curve around the point in time $t'$ which has already been fixed using the basic principle, a time interval is singled out (A-block) (see figure 3).

Subsequently we try to fit the parts of the curves within respectively the A- and the B-block to each other. Since a characteristic change in the intensity flow is represented by distinguishable increases in the cumulative curves, the characteristic fluctuations in the intensity at the measuring sites A and B are thus linked in the best possible way.
Corrections
Although wrong re-calibrations as a consequence of coincidental resemblance are rare, they are not excluded. A small error is of little importance because this will be corrected by the next re-calibration. A big error would upset the process. To avoid this some check-criteria have been built in. When these criteria detect a re-calibration to be questionable it is simply left undone. This will not cause any problems, because (in the case of uncongested traffic) re-calibrations can be made very often but are not needed so frequently.

In one respect a re-calibration implies that the variables $t_A$ and $t_B$ are assigned the values of the determined points in time $t'$ and $t$ at respectively the A- and the B-curve. In another respect this implies that the total number of measured vehicle passings in both measuring sites are equalized by increasing the number of measured vehicle passings in the measuring site that has measured the fewest passings.

To conclude, the whole procedure (basic principle and re-calibration process) can be applied to the traffic as a whole as well as to separate vehicle categories. The latter case will result in several (different) travel times for cars, (heavy) trucks, etc. This also gives good possibilities for re-calibrations, even in the case of congested traffic.

6. MODEL
The algorithm that is described in the former paragraph is implemented on a computer. Below this computer model and the first test results will be discussed.

The input
For the input of the computer model we used real traffic data, collected by induction loop detectors. The basic data collected by the induction loops concerns information about individual vehicles. This data is converted into half minute-intensities, so that measurement time is divided into periods of 30 seconds. Furthermore, only trucks are taken into account.

Figure 4 shows a sketch of the motorway where the measurements were carried out.
The output

Figure 5 shows the cumulative curves of vehicles passing the (arbitrary) sites A and B.

The horizontal axis represents time and the vertical axis the number of measured car passings. The vertical lines illustrate moments in time where "re-calibrations" took place.

The starting point of the determination of travel times always is at the beginning of curve B. During the complete measurement process the step size equals the aggregation-period (in this case 30 seconds). In the curves in figure 4 two corresponding points in time (one in curve A and one in curve B) are marked by an arrow.

Analogous to figure 4 the determination of the travel time takes place by shifting the arrow in curve B one position to the right and subsequently finding a matching position for the arrow in curve A.

Validation

In figure 6 the travel times are denoted as a function of time. For this, we used two measuring sites 1 and 2 from figure 4 (a mutual distance of 8.4 kilometres and one on and off ramp in between). Furthermore only trucks are taken into account. On the right vertical axis of this figure the mean speed (in km/h) on the given road segment is also specified. This mean speed is calculated by dividing the distance between the measuring sites A and B by the computed travel time. The speeds calculated in this way vary between approx. 30 km/hour and approx. 90 km/hour. These results show that the determined travel times are realistic.

Almost simultaneously with the present study, another study was commissioned by the Transportation and Traffic Research Division of Rijkswaterstaat, dealing with the same subject but using the signal-analysis package PRIMAL [15]. As these two studies made use of the same traffic data and the objectives overlapped it is interesting to compare the results.
During the period from 17:15 to 18:05 the results of the PRIMAL-study showed deviant responses (which is an indication of congested traffic).

The road-monitoring system (MCSS, Motorway Control and Signalling System [19]) detected congested traffic during the period from 17:20 to 18:15.

From the results of the computer model it can be derived that there was congested traffic during the period starting at 17:17 and ending at 18:15. As a criterium for this, the lasting high travel time from 17:15 to 18:14 is used. The similarity indicates that the travel times determined by using the computer model are probably accurate, though a direct check (for instance using video observations) has not been made yet.

When testing the model with these data it appeared that the longest period of time during which no re-calibrations took place started at 17:14 and ended at 18:13. This means that, in the worst case, after approximately 60 minutes a re-calibration occurred.

The travel time \( t = 18:13 \) is about 7\( \frac{1}{2} \) minutes. The absolute error in the travel time then amounts to \( (p \times T = 2\% \times 60 = ) \) 1.2 minute and the relative error then amounts to \( (p \times T/t = 2\% \times 60 / 7\frac{1}{2} = ) 16\% \). These are maximal errors after a congestion lasting for almost an hour.

The influence of ramps and the distance between the measuring sites
To study the influence of the distance between the measuring sites and the presence of on and off ramps, we tested the computer model using different combinations of measuring sites. The results are shown in figures 7 to 10.

The study showed that the number of possible re-calibrations (and so the accuracy) gradually decreases as the distance between the measuring sites increases. We recommend to place the loop detectors at a mutual distance of no more than about 10 kilometres. If much larger distances are used the results will not be reliable any more and the travel times will lose their actuality (with respect to the longer travel time between the sites). Further study is needed to determine the optimal distance between two measuring sites, weighting out costs against benefits.

When using combinations of measuring sites with no ramps in between the number of re-calibrations is very high, because there are no ramps to influence the traffic flow and consequently the fluctuations in the curves are better preserved.

On parts of the motorways where congestion frequently occurs and where it is therefore of great importance to have reliable and rather accurate travel times available, we recommended to place the detection loops in such a way that they do not have any ramps in between (or else to place loops on the ramps themselves).
7. CONCLUSIONS

The study showed that the computer model based on the developed algorithm can produce reliable (mean) travel times on road segments with lengths of 5 to 10 kilometres.

Theoretical analysis confirms that after further refinement and optimisation of (the criteria in) the computer model the accuracy can be increased and travel times can be calculated for different vehicle categories. Only in the case of a traffic jam lasting longer than one hour or a highly occupied on/off ramp the accuracy of the travel times decreases.

In general, the presence of a traffic-jam or an on/off ramp between the measuring sites presents no problem for the reliability and the accuracy of the travel times. So it is not necessary to place induction loops at all ramps on the complete motorway. However, in the case of a highly occupied ramp the results become doubtful, so it is advisable to install an induction loop on this ramp.
8. REFERENCES

[12] Travel times on highways, Groninij NV, Department of Infrastructure, February 1990.
Effectiveness of information systems in networks with and without congestion.

R. Hamerslag and E.C. van Berkum

In: Transportation Research Record 1306, In-Vehicle Information Systems: Modelling Traffic Networks and Behavioral Considerations 1991. Transportation Research Board, Washington, DC
Effectiveness of Information Systems in Networks With and Without Congestion

Rudi Hamerslag and Eric C. van Berkum

The use of road transport informatics (RTI) is a recent development that optimizes the use of existing facilities in the transportation system and serves three main goals: alleviation of congestion, diminution of air pollution, and reduction of incidents. RTI instruments deal with traffic information. Examples of RTI systems are pretrip planning, roadside displays, radio data system—traffic message channel, and in-car navigation. To model the effects of providing the road user with information a method is used in which stochastic and deterministic assignments were compared for both networks with and without congestion. To let information also affect destination choice and the spatial distribution of activities, the assignment models were combined with different distribution models. The amount of information that travelers have was translated to a "level of uncertainty" measure. The more informed a traveler is, the lower the level of uncertainty. Since the effects appeared to be network dependent, a number of different networks were examined. Simulations show that the amount of kilometers driven decreases when travelers are provided with better and more information.

The use of road transport informatics (RTI) is a recent development that optimizes the use of existing facilities in the transportation system and serves three main goals: alleviation of congestion, diminution of air pollution, and reduction of incidents. RTI instruments deal with traffic information. Systems such as pretrip planning, roadside displays, radio data system—traffic message channel, and in-car navigation are all part of RTI. From a planners' viewpoint, it is essential to know the possible impact of RTI on the traffic system. One way to predict effects of RTI is to model individual travel behavior and to incorporate information explicitly as a model component. In this way, the effect of information on travel behavior can be simulated. Before this can be done, however, it is necessary to model the current situation, in which the traveler is not perfectly informed and therefore makes non-optimal choices.

In many existing models it is assumed that people have perfect knowledge of all travel alternatives. This assumption means that the usefulness of providing information to travelers cannot be determined. In the approach presented in this paper, the classic four-stage model is central. The key issue is, however, that the perceived travel times instead of the (perceived) travel times and the outcomes for different levels of uncertainty are compared, it is possible to get an insight in the effects of information.

RELATED STUDIES

In recent years, many approaches have been presented to provide insight into the possible benefits of information systems in transport.

The feasibility of the Comprehensive Automobile Traffic Control project (1) was studied by using a simulation model in which the noninformed users choose their route on the basis of various factors, such as travel time, length of the route, number of lanes, number of turns, and so on, and the informed users choose their route solely on the basis of travel time. It was found that in Tokyo travel time could be reduced by 6 percent and fuel consumption by 5 percent. Tsuji et al. (2) investigated the effectiveness of a route guidance system by using a mathematical model. Among other factors, they used travel time reduction as a measure of effectiveness. The outcomes, however, must be related to the heavy assumptions under which the model is valid. The reduction in travel time was found to be 11 percent. van Vuren (3) tried to model the effectiveness of route guidance by using a multiuser class equilibrium and stating that the noninformed users behave greedily, as in a deterministic user equilibrium, whereas the informed users behave according to the principle of a system optimum. The results were found to be unrealistic because the uninformed users were better off.

Koutsopoulos and Lotan (4) modeled the impact of information on travelers by using a stochastic user equilibrium and stating that information systems reduce the variance in travel time. They found a reduction in travel time of about 5 percent, dependent on the assumed reduction in variance.

Mahmassani and Jayakrishnan (5) modeled the effectiveness of a real-time information system on a small test network with three parallel highways and a number of switching possibilities. The researchers chose one information supply strategy and focused on the users' reaction by defining them as bounded rational individuals. An important result was that the system performance might actually worsen by myopic local actions of the drivers. Van Berkum and van der Mede (6) presented a dynamic approach that simulates rational, uncertain, persistent individuals who base their decisions on experience and have a limited knowledge of alternatives.

R. Hamerslag, Departments of Civil Engineering and of Mathematics and Informatics, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands. E. C. van Berkum, Bureau Goudappel Coffeng, P.O. Box 161, 7400 AD Deventer, The Netherlands.
The approach that is followed in this paper is an extension of the work of Koutsopoulos and Lotan (4). The situation of recurrent congestion was also studied in this research. But whereas Koutsopoulos and Lotan restricted the effects of better information to route choice, the impact on destination choice and the location of activities has also been studied here. Another difference is that the amount of uncertainty in their approach was initially too small. Further, they examined one network, whereas different networks are studied here. Because the results are network dependent, it is difficult to compare results, but the results they found on route choice are on the same order of magnitude as the results presented in this paper. The results gained from the present research are not comparable with the results found by Mahmassani and Jayakrishan. They studied the reaction of people on dynamic traffic information that reports the actual traffic conditions. When drivers react myopically, this information becomes invalid. An adjustment process will occur, which in the end will lead to an equilibrium. This equilibrium is focused on in this paper. This further implies that the information given to the drivers is in some sense not real-time information but rather future-time information.

MODELING APPROACH

General

The main hypothesis of this study is that the fact that people are uncertain about travel times on links has more effects than only on route choice. There will also be effects on destination choice as well as on the spatial distribution of activities. People make trips because they want to perform activities that are spatially separated. In the traditional four-stage models, the spatial distribution of activities is fixed. In this study models are used that include the spatial distribution of activities as endogenous. Users choose a route by minimizing some measure of cost. In this study travel time will only be used as cost. Travelers do not possess perfect information about the network they travel on. This means that people do not minimize the objective time but rather the perceived travel time.

Destination choice can also be modeled by using a cost minimization procedure (8). Because of the observation made previously, this means that in determining the origin-destination (O-D) flows, the perceived cost or travel time must also be used. A basic assumption here is that route choice is made on the basis of the same perceived travel times as destination choice and the location of activities are made. Traffic information affects the perception of travel times in the network. The perceived travel times will be modeled as stochastic variables whose distribution is influenced by the amount of available information.

The approach that has been followed uses the traditional four-stage model as a basis, although an adjusted form has been developed. The following assumptions are therefore needed:

- All people base their decisions on what they know; and
- People base their route and destination decisions on the same perceived travel time.

In order to make the approach not too complex the following limitations have been adopted:

- The total number of trips is constant under all levels of uncertainty;
- All people have access to the same level of information;
- Information is assumed to be good and true; and
- No distinction has been made between different modes and purposes.

Route Choice

Link travel times on the network are defined as stochastic variables. The variance in travel times—that travel times are unpredictable to a certain extent—may be understood as uncertainty of travelers. Consequently, users will have different perceptions of travel times on the links.

A deterministic user equilibrium (DUE) can be defined as the situation in which no traveler can improve his or her travel time by unilaterally changing route (9). This definition assumes that every traveler has an exact knowledge of travel times and flows on all links in the network. A stochastic user equilibrium (SUE) can be defined as the situation in which every traveler thinks that he or she cannot improve the travel time by unilaterally changing routes (9,10). This definition assumes that travelers have different perceptions of travel times. Comparing a SUE with a DUE enables estimating the effect of providing information to travelers (or reducing their uncertainty) on the traffic system (4,11). Because this comparison can be translated as comparing travelers with exact knowledge of all travel times in the network with travelers with different perceptions of travel times in the network.

In networks without congestion the DUE assignment becomes a simple all-or-nothing assignment, where the SUE assignment becomes a classic stochastic assignment (12,13). The impedance \( Z_{sp} \) of a link \( a \) in a network for person \( p \) is a function of a number of variables \( X_{spa} \) such as time, cost, and distance and their relative importance \( \beta_i \) plus some measure of uncertainty. We define

\[
Z_{sp} = \sum \beta_i \cdot X_{spa} + \epsilon_{sp}
\]  

where \( \epsilon_{sp} \) is a noise term. The resulting route choice model depends on the distribution of \( \epsilon_{sp} \). It is supposed that \( \epsilon_{sp} \) is normally distributed with mean 0 (13), which yields a probit model for route choice. The introduction of the noise term \( \epsilon_{sp} \) can be explained by stating that (a) behavior cannot completely be explained by all \( X_{spa} \) ’s, (b) individuals have different perceptions of the \( X_{spa} \) ’s and their relative importance therefore may differ, and (c) individuals are uncertain about the exact value of the \( X_{spa} \) ’s, especially because these values differ in time. Instead of impedance, generalized cost, or generalized time only travel time will be considered as a measure for deterrence in this study.

The travel time on a link \( a \) in a network without congestion is

\[
Z_{a} = Z_{a} + \alpha \cdot R \cdot \sqrt{Z_{a}}
\]
where
\[ Z^* = \text{mean travel time of link } a, \]
\[ R = \text{draw from a normal } [N(0,1)] \text{ distribution, and} \]
\[ \alpha = \text{factor determining the variance (from now } \alpha \text{ will be}
\]
\[ \text{called level of uncertainty).} \]

The value of } \alpha \text{ is dependent on the chosen dimension (14). Given an O-D matrix, } \alpha \text{ can be determined by comparing true with model flows. When the dimension is minutes, it has been estimated that } 0.5 < \alpha < 1 \text{ for a regional network with relatively few alternative routes (15). Furthermore, Bovy (14) developed an efficient methodology for estimating } \alpha \text{ from observed flows.}

In reality, the uncertainty will, among other things, be a function of the frequency with which a person travels between a certain O-D pair. The lower the frequency, the higher the uncertainty. In this study, the uncertainty is assumed to be equal for all travelers.

The travel time of a link in a network with congestion is
\[ Z^* = Z^*_a + \alpha \cdot R \cdot \sqrt{Z^*_a} \quad (3) \]
with
\[ Z^*_a = Z^*_a \left[ 1 + \tau \left( \frac{q_a}{c_a} \right)^2 \right] \quad (4) \]
where
\[ Z^*_a = \text{the mean travel time of link } a, \]
\[ q_a = \text{the flow on link } a, \]
\[ c_a = \text{the capacity of link } a, \]
\[ R = \text{a draw from a normal } [N(0,1)] \text{ distribution,} \]
\[ \alpha = \text{the level of uncertainty, and} \]
\[ \tau = \text{a parameter dependent on the definition of capacity.} \]

**Destination Choice and the Location of Activities**

Because the distribution process is a utility maximization process (or disutility minimization), information will also have impact on destination choice resulting in a distribution of flows and the location of activities. In this study, the following interaction model with elastic constraints is used (16):
\[
\min \sum_i \left( \sum_j T_{ij} - m_i \cdot A_j \right)^2 + \sum_j \left( \sum_i T_{ij} - l_i \cdot D_j \right)^2 \quad (5)
\]
Subject to
\[
T_{ij} = 6l_i m_i Q_{ij} \exp(-0.4 \ln^2(Z_{ij} - d_{ij} + 1)) \quad (6)
\]
where
\[ T_{ij} = \text{number of trips between } i \text{ and } j, \]
\[ l_i, m_i = \text{equilibrium factors,} \]
\[ Q_{ij}, X_i = \text{polarities,} \]
\[ Z_{ij} = \text{objective travel time between } i \text{ and } j, \]
\[ A_j, d_{ij} = \text{arrivals and departures, and} \]
\[ d_{ij} = \text{difference between objective and perceived travel time between zones } i \text{ and } j. \]

In solving the model, the terms } 1/(1 + g) \text{ and } 1/(1 + h) \text{ become important. These terms will be called elasticities. Thus when } g \text{ and } h \text{ are both 0 the elasticities become 1 and the model turns into the classic gravity model with fixed constraints. To coordinate spatial planning, transportation development, and spatial development, the model with elastic constraints was developed. The value of the equilibrium factors in Equation 6 is a function of the extra effort needed to comply with the constraints. In poorly accessible areas, the value is high and, inversely, in easily accessible areas the value is low. When the number of arrivals and departures is seen as dependent, though not exclusively, on the accessibility, the objectives in Equation 5 must become elastic.

**Combining the Assignment and the Distribution Model**

To determine the effects of information on route choice, route and/or destination choice, and/or the location of activities, the following models must be compared:

- In the case of no congestion, a distribution model with and without elastic constraints will be compared with the same model but combined with a stochastic Burrel assignment.
- In the case of congestion, first the DUE assignment will be combined with the distribution model without (10,17) and with elastic constraints (18). Second, the same combination will be made, but with the SUE assignment.

To combine a SUE assignment with a distribution model, including the assumption that both models deal with the same perceived travel times, it is necessary to determine how the perceived travel times must be used in the distribution stage. In the proposed distribution model there is one value for travel time between each O-D pair. In reality this travel time is different for every individual (perceived travel time). Starting with } Z_{\text{upp}}, \text{ the perceived travel time between } i \text{ and } j \text{ along route } r \text{ of person } p, \text{ person } p \text{ chooses that route with the smallest perceived travel time. Therefore, it holds that}
\[
Z_{ij,p} = \min \ Z_{upp} \quad (7)
\]
Suppose the population } B \text{ is divided in two groups, } B_1 \text{ and } B_2. \text{ Persons belonging to } B_1 \text{ find route } i \text{ the best, and persons belonging to } B_2 \text{ do not, so}
\[
Z_{\text{upp}} = \min \ Z_{upp} \forall p \in B_1 \text{ and } r \neq 1 \quad (8)
\]
For persons belonging to } B_2 \text{ it holds that
\[
Z_{\text{upp}} = \min \ Z_{upp} \forall p \in B_2 \quad (9)
\]
So
\[
Z_{\text{upp}} = \min \ Z_{upp} \forall p \in (B_1 \cup B_2) \quad (10)
\]
Suppose there are } N \text{ persons in } B, \text{ then}
\[
Z_{ij} = \frac{1}{N} \sum_p Z_{ij,p} \text{ and } Z_{\text{upp}} = \frac{1}{N} \sum_p Z_{\text{upp}} \quad (10.1)
\]
Using Equation 10 it holds that

\[ Z_{yu} \leq Z_{yl} \quad \text{(10.2)} \]

The same result can be derived for every route \( r \), so

\[ Z_{yu} \leq Z_{yr} \quad \forall r \quad \text{(11)} \]

Thus the perceived travel time between any O-D pair used in the distribution stage is always less than or equal to the perceived travel time of any of the chosen routes between the OD pair.

The difference between the best route and the travel time between an O-D pair is dependent on the network. When, for instance, one route is by far the best so that every traveler between that OD pair will choose that route, the equal sign in Equation 11 holds for this particular route. When there is a spreading over the routes for all \( r \) the less than sign will hold. When the level of uncertainty \( \alpha \) becomes larger, the spreading in routes becomes larger and \( Z_{yu} \) will decrease, or in other words the difference between model travel time and the mean perceived travel time of the objectively seen best route (which is by definition the objective travel time of the best route) becomes larger. So in the distribution stage the following travel time is used:

\[ Z_{yu} - d_{yu} \]

Where \( Z_{yu} \) is the mean perceived travel time of the objectively seen best route between zone \( i \) and \( j \); \( d_{yu} \) is an increasing function of \( \alpha \) (obviously when \( \alpha = 0 \), also \( d_{yu} = 0 \)).

Models and Algorithms

To study the effects of more or better information on route choice the "A model" is used, which is a stochastic equilibrium assignment with a given, fixed O-D matrix. To study the effects on destination choice and on the resulting O-D flows too, the "A + D model" is used. In this model, a stochastic equilibrium assignment and distribution with fixed constraints are combined. In the O-D matrix, the numbers of departures and arrivals are fixed for each zone. The cell volumes solve Equation 5 subject to Equation 6 with \( g = h = 0 \).

To study the effects of activities on the locations, the "A + D + L model" is used. In this model, a stochastic equilibrium assignment and distribution with elastic constraints are combined. In the O-D matrix, the numbers of departures and arrivals are variable for each zone, but the total number of trips is fixed. The cell volumes solve Equation 5 subject to Equation 6 with \( g \) and \( h \) not necessarily equal 0.

In Figures 1 through 4, the separate algorithms for the congestion situation are depicted. Basically, the methodology as proposed by Evans (17) is followed. The steps that have to be executed more than once because the draw must take place \( m \) times have been depicted with a thick line. In the case of no congestion, the step where new travel times are computed becomes trivial.

A generalized description of the used algorithm is

1. Read network;
2. Draw link travel times for every link;
3. Determine travel times from shortest routes between every O-D pair;
4. Repeat No. 2 and No. 3 \( m \) times;
5. IF model = A THEN read O-D matrix
   ELSE determine mean travel times with the travel times per draw determined in No. 3. Determine O-D matrix with elastic constraints (A + D + L) or with fixed constraints (A + D) using a Gauss-Seidel iteration procedure to solve Equation 5 subject to Equation 6.

![FIGURE 1 A model, congestion.](image-url)
FIGURE 2 A + D model, congestion.

FIGURE 3 A + D + L model, congestion.

FIGURE 4 Explanation for Figures 1 to 3.
6. Subdivide the O-D matrix in \( m \) equal parts and load them to the routes determined in No. 3, yielding loads \( q^*_i \) for link \( a; \)
7. load link \( a \) in iteration \( i \) the network with \( q^*_i = [q^{i-1}_a (1 + q^*_i)]/i; \)
8. In case of congested networks: determine new travel times; and
9. Go to No. 2 until stop criterion is reached.

EXPERIMENTS

The experiments were performed using the research facilities of the Teacher Friendly Transportation Programs V90.2 (19).
In the stochastic assignments \( m, \) the number of draws was 4 and the number of iterations was 8. Because for every tree of shortest paths, new travel times were drawn, the total number of draws is 32 times the number of zones. Convergence was no problem in all test networks. The number of iterations was far less than expected in a combined distribution-assignment procedure (10,20).

Networks

Earlier it was observed that the spreading of chosen routes determines to some extent the value of \( d^a \). The amount of spread is not only dependent on the size of the variance as used in the stochastic assignments, but also on the presence of (relevant) alternative routes. Obviously in a situation in which there are hardly any alternative routes, the spread will be small. Therefore it is important to investigate different networks. In this study four regional networks with a diameter of about 40 km (called REGIO, RING, SLOW, and CBD) and one urban network with a diameter of about 15 km (TOWN) were examined. For the regional networks only, the situation without congestion is considered. For the urban network both the situation with and without congestion are considered.

In Figures 5 to 7 some of the networks with their spreading of activities and flows are shown. In Figure 8 the notation of the activities is shown. The networks RING, SLOW, and FAST are the same size (number of links, number of nodes, distances) as CBD.

![FIGURE 5 Network CBD.](image)

![FIGURE 6 Network REGIO.](image)

![FIGURE 7 Network TOWN.](image)

![FIGURE 8 Notation of activities.](image)

In Figure 8 the radius of the outer circle is proportional to

\[
\text{max} \left( \sum_i T_u, \sum_i T_s \right)^{\frac{1}{2}}
\]  

(12)

The radius of the inner circle is proportional to

\[
\text{abs} \left( \sum_i T_u - \sum_i T_s \right)^{\frac{1}{2}}
\]  

(13)
In a smaller level of uncertainty, it can be stated that providing road users with information reduces the amount of carkilometers. The results show that the gains differ per network. A network means not only the set of links and nodes, but also the initial trips ends. This observation implies that it is hard to compare the results of other studies with one another and with these results, because different networks are used in all studies.

The results for the networks as listed in Tables 3, 4, and 5 are more or less comparable. These results were calculated with models that did not deal with congestion. The results for the TOWN network show a larger increase in carkilometers when \( a \) increases (See Table 2). This can be explained by the fact that the TOWN network obviously contains more alternative routes than all the other networks. The spread in route choice will be bigger for this network since there simply exist more alternatives. Because the network outcomes in Tables 3 to 5 reflect few alternative routes, the effect on route choice is small compared with the effect on destination choice (compare the outcomes in Tables 3 and 4). The extra effect on the location of activities is also small compared with the effect on destination choice (compare the outcomes in Tables 4 and 5).

When looking at the network TOWN, the effects on route choice are the largest. Change in destination choice and in the location of activities are marginal compared with this effect. Because this network is more realistic than the other ones, this observation may be generally true. By comparing Tables 1 and 2, it follows that the effect of the provision of information is larger in the network with congestion than without congestion.

### Tables

**Table 1** Carkilometers for TOWN Network, Without Congestion, Under Different Levels of Uncertainty (KM for \( a = 0 \) Are 100)

<table>
<thead>
<tr>
<th>( a )</th>
<th>A</th>
<th>A+D</th>
<th>A+D+L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>117</td>
<td>128</td>
<td>130</td>
</tr>
<tr>
<td>0.7</td>
<td>112</td>
<td>119</td>
<td>121</td>
</tr>
<tr>
<td>0.5</td>
<td>104</td>
<td>107</td>
<td>108</td>
</tr>
<tr>
<td>0.3</td>
<td>101</td>
<td>102</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 2** Carkilometers for TOWN Network, With Congestion, Under Different Levels of Uncertainty (KM for \( a = 0 \) Are 100)

<table>
<thead>
<tr>
<th>( a )</th>
<th>A</th>
<th>A+D</th>
<th>A+D+L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>124</td>
<td>129</td>
<td>136</td>
</tr>
<tr>
<td>0.7</td>
<td>118</td>
<td>121</td>
<td>126</td>
</tr>
<tr>
<td>0.5</td>
<td>106</td>
<td>109</td>
<td>110</td>
</tr>
<tr>
<td>0.3</td>
<td>102</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 3** Carkilometers Driven for Different Networks, Under Different Levels of Uncertainty with the A-Model (KM for \( a = 0 \) Are 100)

<table>
<thead>
<tr>
<th>( a )</th>
<th>CBD</th>
<th>RING</th>
<th>SLOW</th>
<th>FAST</th>
<th>REGIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>101</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>0.5</td>
<td>101</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
</tr>
</tbody>
</table>

**Table 4** Carkilometers Driven for Different Networks, Under Different Levels of Uncertainty with the A+D-Model (KM for \( a = 0 \) Are 100)

<table>
<thead>
<tr>
<th>( a )</th>
<th>CBD</th>
<th>RING</th>
<th>SLOW</th>
<th>FAST</th>
<th>REGIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>103</td>
<td>103</td>
<td>112</td>
<td>120</td>
<td>111</td>
</tr>
<tr>
<td>0.5</td>
<td>102</td>
<td>103</td>
<td>106</td>
<td>107</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 5** Carkilometers Driven for Different Networks, Under Different Levels of Uncertainty with the A+D+L-Model (KM for \( a = 0 \) Are 100)

<table>
<thead>
<tr>
<th>( a )</th>
<th>CBD</th>
<th>RING</th>
<th>SLOW</th>
<th>FAST</th>
<th>REGIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>123</td>
<td>113</td>
</tr>
<tr>
<td>0.5</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
CONCLUSION

Trips and activities are a result of decisions people make. These decisions concern route and destination choice as well as activity choice. The actual choices depend on the perceived travel times, rather than on the objective travel times. As a result, travelers think they choose the best route, but this route is not necessarily the best from an objective point of view. Also destinations are chosen because they appear to be close. This causes extra, unnecessary carkilometers.

The approach presented in this paper has a number of assumptions and limitations about information:

- Information is seen as an abstract entity; it is not possible to evaluate a specific information system or different types of information.
- Because of the equilibrium approach the presented method is able to predict the long-term effects of the provision of information in a situation of recurrent congestion.

The results of this study should be looked at in light of these assumptions as well as in light of the limitations this approach has.

It was proven that the perception of two or more independent routes is always less than or equal to the perception of each of two or more routes together. The travel time of the chosen route is systemically being underestimated. Providing information reduces the difference between perceived travel time and objective travel time. This has an impact on the choice of route, destination, and activity. As a result, the amount of carkilometers decreases. The different test cases show that the form of the network, with respect to the presence of alternative routes, is of importance. Further, the simulations show that in a situation with congestion, the decrease of carkilometers is larger than in the situation with no congestion. Currently it is not possible to quantify the effects of information precisely because the present and future values of $\alpha$ are not exactly known, uncertainty will only partially be influenced by information, and only a part of the travelers will use the information. On the other hand, through route guidance, delays on intersections may be minimized (27) and the influence that information about incidents could have is neglected. With the above considerations in mind it seems valid to state that information systems may decrease the amount of carkilometers in urban networks by 15 to 20 percent and in regional networks by 5 to 10 percent.

REFERENCES


A method to determine the influence of information for cardrivers on the performance of different networks.

R. Hamerslag and E.C. van Berkum

Abstract

Through the introduction of Road Transport Informatics the information that cardrivers base their decisions on will improve. It may be expected that the decisions of the individual will become better. This contribution deals with improvements of uncertainty on the whole network.

Cardrivers are uncertain about the travel times in the network. The better the information about the network is, the lower the uncertainty. Uncertainty is modelled by defining the travel times on links as stochastic variables with different variances. By comparing model outcomes for different variances the influence of better information can be determined. A mathematical model has been specified. Simulations were carried out with it for different networks to determine the effects of better information on route choice, destination choice and the location of activities.

The results show that the influence of better information on the total amount of kilometers driven can be quite substantial.

1. Introduction

The use of road transport informatics (RTI) is a more recent development aiming to optimize the utilization of existing facilities in the transportation system, serving three main goals: alleviation of congestion, diminution of air pollution and reduction of incidents. RTI instruments deal with traffic information. Systems such as pre-trip-planning, roadside displays, RDS-TMC and in-car navigation are all part of RTI.

In recent years a number of approaches have been developed to get an insight in the possible benefits of information systems in transport [1], [4], [13], [16], [17], [19]. The approach that is followed in this paper is an extension of the work of Koutsopoulos and Lotan [16]. They modelled the impact of information on travellers by using a stochastic user equilibrium and stating that information systems reduce the variance in travel time. They restricted the effects one route choice solely, in one network with congestion. In this paper also the impact on destination choice and the location of activities has been studied. The calculations were carried out for more networks with and without congestion.

2. Basic assumptions

The model that has been used is an extension of the so-called aggregate interaction model, that has been used for years for traffic and transport planning. In this traditional approach the spatial distribution of activities is fixed. In this study also some experiments were carried out where an interaction model, where the spatial distribution of activities is endogenous, is used.
In the traditional models it is assumed that people have perfect knowledge of all travel alternatives. In reality this knowledge is limited, but will be improved by RTI-measurements. Improving knowledge means lessening uncertainty. Therefore a measure of uncertainty is introduced. So by using models for different levels of uncertainty it is possible to get an insight in the effects of better information.

So the main assumption in this study is that people minimize the perceived travel time and not the objective travel time.

The perceived travel time will be modelled as stochastic variables whose distribution is influenced by the amount of available information.

The following assumptions are made:
- all people base their decisions on what they know
- the total number of trips is constant under all levels of uncertainty, instead of the total number of trips per zone in the traditional models.
- all people have access to the same level of information.
- information is assumed to be good and true
- route choice is based on the same perceived travel times as destination choice and the location of activities.

In order to make the approach not too complex the next limitations have been adapted:
- no distinction has been made between different modes and purposes
- the uncertainty is assumed to be equal for all cardrivers
- the influence of congestion on departure times has been neglected.

Due to the equilibrium approach the presented method is able to predict the long-term effects of the provision of information. Because information is seen as an abstract entity it is not possible to evaluate a specific information system, or different types of information.

3. Modelling approach

3.1 Stochastic equilibrium

Link travel times on the network are defined as stochastic variables. The variance in travel times may be understood as uncertainty of travelers. A stochastic user equilibrium (SUE) is defined as the situation in which every traveler thinks that he or she cannot improve the travel time by unilaterally changing routes [13,18]. This definition assumes that travelers have different perceptions of travel times.

Comparing a stochastic with a deterministic user equilibrium [18] provides a possibility to estimate the effect of providing information to travellers (or reducing their uncertainty) on the traffic system [16,19], since this comparison can be translated as comparing travellers with exact knowledge of travel times in the network with travellers with different perceptions of travel times in the network.

3.2 Route choice

The impedance (generalized cost or generalized time) $Z_{ap}$ of a link $a$ in a network for person $p$ is a function of a number of variables $X_{ap}$ like time, costs, distance and their relative importance $\beta_a$ plus some measure of uncertainty. Further in this paper only travel time will be used as impedance.

We define:

$$Z_{ap} = \sum \beta_a X_{ap} + e_{ap}$$
where \( e_{it} \) is a noise-term.

The resulting route choice model depends on the distribution of \( e_{it} \). We suppose \( e_{it} \) to be normally distributed with mean 0 [8], which yields a probit model for route choice.

The introduction of the noise-term \( e_{it} \) can be explained by stating that

(a) behavior can not completely be explained by all \( x_{it} \)'s,
(b) individuals have different perceptions of the \( x_{it} \)'s, and their relative importance therefore may differ,
(c) individuals are uncertain about the exact value of the \( x_{it} \)'s, especially since these values differ in time.

The travel time on a link \( a \) in a network without congestion is

\[
Z^*_a = Z_s + \alpha R \sqrt{Z_s}
\]

where \( Z_s \) = the mean travel time of link \( a \)
\( R \) = a draw from a normal \( N(0,1) \)-distribution
\( \alpha \) = a factor determining the variance. From now \( \alpha \) will be called level of uncertainty.

The value of \( \alpha \) first is dependent on the chosen dimension [3]. Given an OD-matrix \( \alpha \) can be determined by comparing true and model-flows. When the dimension is minutes, it has been estimated that \( 0.5 < \alpha < 1 \) for a regional network with relatively few alternative routes [10]. Further more Bovy [3] developed an efficient methodology for estimating \( \alpha \) from observed flows.

In reality the uncertainty will among other things be a function of the frequency in which a person makes a trip between a certain OD-pair. The lower the frequency, the higher the uncertainty. In this study the uncertainty is assumed to be equal for all travelers.

The travel time of a link in a network with congestion is

\[
Z^*_a = Z'_s + \alpha R \sqrt{Z'_s}
\]

with

\[
Z'_s = Z_s (1 + \tau (Q_s/C_s)^4)
\]

where \( Z_s \) = the mean travel time of link \( a \)
\( Q_s \) = the flow on link \( a \)
\( C_s \) = the capacity of link \( a \)
\( R \) = a draw from a normal \( N(0,1) \)-distribution
\( \alpha \) = the level of uncertainty
\( \tau \) = a parameter dependent on the definition of capacity

3.3 Destination choice and the location of activities

Since the distribution-process is a utility-maximization process (or disutility-minimization) information will also have impact on distribution and the location of activities. In this study the following interaction-model with elastic constraints is used [9].

\[
\min_{i,a} \{ \sum_j (E_i T_{ij} - m_j h_j A_j)^2 + \sum_i (E_i T_{ij} - 1 h_i D_i)^2 \} 
\]
where

\[(3.3.2) \quad T_{ij} = \Omega \cdot l_{ij} \cdot Q_{ij} \cdot m_{ij} \cdot X_{ij} \cdot \exp\{-0.4 \cdot \ln(z_{ij} + 1)^2\}\]

with
- \(T_{ij}\) the number of trips between \(i\) and \(j\)
- \(l_{ij}, m_{ij}\) equilibrium factors
- \(Q_{ij}, X_{ij}\) polarities
- \(z_{ij}\) travel time between \(i\) and \(j\)
- \(A_{ij}, D_{ij}\) the arrivals and departures
- \(\Omega\) a constant

When \(g\) and \(h\) are both 0, this model turns into the classic interaction model with fixed constraints.

In order to coordinate spatial planning, transportation development and spatial development the model with elastic constraints was developed. The value of the equilibrium-factors in (3.3.1) is a function of the extra effort needed to comply with the constraints. In poorly accessible areas the value is high and, inversely, in easy accessible areas the value is low. When the number of arrivals and departures is seen as dependent, though not exclusively, on the accessibility, the objectives in (3.3.1) must become elastic.

3.4 Combining the assignment and the distribution model

In order to be able to determine the effects of information not only on route choice, but also on destination choice a SUE-assignment and a distribution with and without elastic constraints has been combined. So it is necessary to determine how the perceived travel times must be used in the distribution stage.

Let us start with \(z_{ijr,p}\), the perceived travel time between \(i\) and \(j\) along route \(r\) of person \(p\). Person \(p\) chooses that route with the smallest perceived travel time, so it holds that

\[(3.4.1) \quad z_{ijr,p} = \min, z_{ijr}\]

Suppose the population \(B\) is divided in two groups \(B_1\) and \(B_2\). Persons belonging to \(B_1\) find route 1 the best, and persons belonging to \(B_2\) don't, so

\[(3.4.2) \quad z_{ijr,p} = z_{ijr} \text{ for } p \in B_1 \text{ and } z_{ijr,p} < z_{ijr} \text{ for } p \in B_2\]

So it follows that

\[(3.4.3) \quad z_{ijr,p} \leq z_{ijr} \text{ for } p \in (B_1 \text{ or } B_2)\]

Suppose there are \(N\) persons in \(B\), then by definition

\[(3.4.4) \quad z_{ij1} = (\sum_p z_{ijr}) / N \text{ and } z_{ij} = (\sum_p z_{ijr}) / N\]

Using 3.4.3 it holds that

\[(3.4.6) \quad z_{ij} \leq z_{ij1}\]

The same result can be derived for every route \(r\), so

\[(3.4.7) \quad z_{ij} \leq z_{ijr}\]

Thus the perceived travel time between any OD-pair used in the distribution-stage is always less or equal than the shortest travel time between the OD-pair.
In the distribution-stage the following travel time is used:

\[ Z_{ij} - d_{ij} \]

Where \( Z_{ij} \) is the shortest travel time of the objectively seen best route between zone i and j; (travel time by perfect knowledge)

\( d_{ij} \) is an increasing function of \( \alpha \) (obviously when \( \alpha = 0 \), also \( d_{ij} = 0 \)).

3.5 Models

To study the effects of more, or better information on route choice the so-called A-model is used, which is a stochastic equilibrium Assignment with a given, fixed OD-matrix.

To study the effects on destination choice and on the resulting OD-flows too, the so-called A+D-model is used. In this model a stochastic equilibrium Assignment and Distribution with fixed constraints are combined. In the OD-matrix the number of departures and arrivals are fixed for each zone. The cell-volumes solve (3.3.1) s.t. (3.3.2) with \( g = h = 0 \).

To study the effects on the locations of activities too, the so-called A+D+L-model is used. In this model a stochastic equilibrium assignment and distribution with elastic constraints are combined. In the OD-matrix the number of departures and arrivals are variable for each zone, although the total number of trips is fixed. The cell-volumes solve (3.3.1) s.t. (3.3.2) with \( g \) and \( h \) not necessarily equal 0.

In combining an interaction-model and an assignment model basically the technique as presented by Evans [7] is followed.

4. Experiments

The experiments were performed using the research facilities of the TFTP-program, V90.2 [11]. In the stochastic assignments the number of draws was 4 and the number of iterations was 8. Convergence was no problem in all test-networks.

4.1 Networks

In 3.4 it was observed that the level of uncertainty \( \alpha \) determines to some extent the value of \( d_{ij} \). The value \( d_{ij} \) is not only dependent on \( \alpha \), but also on the presence of (relevant) alternative routes. Obviously in a situation were there hardly any alternative routes the spread will be little.

Therefore it is important to investigate different networks. In this study four regional networks with a diameter of about 40 kilometers (called 'REGIO', 'RING', 'SLOW' and 'CBD') were examined and one urban network with a diameter of about 15 kilometers called ('TOWN'). The networks 'RING', 'SLOW' and 'FAST' are of the same size (number of links, number of nodes, distances) as 'CBD'.

CBD network with speedways (100 km/h) pointed to a central zone; other roads are 40 km/h.
FAST network with only speedways (100 km/h)
SLOW network with only secondary roads of 40 km/h
RING network with speedways in a ring around a central area
REGIO network like 'CBD', but more dense near the center
TOWN urban network; there are no trips to or from the surrounding areas

For the regional networks only the situation without congestion is considered. For the urban network also the situation with congestion is considered.
4.3 Level of uncertainty $\alpha$

The simulations have been performed for all networks with level of uncertainty $\alpha = 0; \alpha = 0.5$ and $\alpha = 1$. For the network "TOWN" also $\alpha = 0.3$ and $\alpha = 0.8$ were used.

4.4 Results

The results of the simulation with the regional networks are given in Table 1 and in the urban network in Table 2. For all networks we see that the amount of carkilometers increases when the level of uncertainty increases. Since provision of information can be translated in a smaller level of uncertainty it can be stated that providing the road users with information reduces the amount of carkilometers.

With a network is meant not only the set of links and nodes, but also the initial tripends. This observation implies that it is hard to compare this study with the results of other studies, since in all studies different networks are being used.

The results in table 1 show that the gains differ per network. The results for the networks as listed in table 1 are more or less comparable. They were calculated with models that did not deal with congestion.

The results for the "TOWN"-network show a larger increase of carkilometers when $\alpha$ increases (See Table 2). This can be explained by the fact that the "TOWN"-network obviously contains more alternative routes than all the other networks. The spread in route-choice will be bigger for this network since there simply exist more alternatives.

Due to the fact that the networks in the tables 1 contain that little alternative routes the effect on route-choice is little compared to the effect on destination choice. The extra effect on the location of activities is also small compared with the effect on destination choice.

When looking at the network "TOWN" (tables 2) the effects on route-choice are the largest. A change in destination choice and in the location of activities are marginal compared to this effect. Since this network is more realistic than the other ones this observation may be generally true. The effect of the provision of information is larger in the network with congestion than without congestion. (table 2)

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>SLOW</th>
<th>FAST</th>
<th>REGIO</th>
<th>FAST</th>
<th>REGIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>102</td>
<td>103</td>
<td>101</td>
<td>105</td>
<td>103</td>
</tr>
<tr>
<td>0.5</td>
<td>100</td>
<td>103</td>
<td>101</td>
<td>101</td>
<td>101</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>A+D-model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>108</td>
<td>109</td>
<td>112</td>
<td>120</td>
<td>111</td>
</tr>
<tr>
<td>0.5</td>
<td>102</td>
<td>103</td>
<td>105</td>
<td>107</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>A+D+L-model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.0</td>
<td>113</td>
<td>113</td>
<td>113</td>
<td>123</td>
<td>115</td>
</tr>
<tr>
<td>0.5</td>
<td>103</td>
<td>103</td>
<td>105</td>
<td>107</td>
<td>104</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
Table 2 Carkilometers for the "TOWN"-network, without congestion under different levels of uncertainty using the A-, A+D- and A+D+L-model. (kilometers for α=0 are 100)

<table>
<thead>
<tr>
<th>α</th>
<th>A</th>
<th>A+D</th>
<th>A+D+L</th>
<th>A</th>
<th>A+D</th>
<th>A+D+L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>117</td>
<td>128</td>
<td>130</td>
<td>124</td>
<td>129</td>
<td>136</td>
</tr>
<tr>
<td>0.7</td>
<td>112</td>
<td>119</td>
<td>121</td>
<td>118</td>
<td>121</td>
<td>126</td>
</tr>
<tr>
<td>0.5</td>
<td>104</td>
<td>107</td>
<td>108</td>
<td>106</td>
<td>109</td>
<td>110</td>
</tr>
<tr>
<td>0.3</td>
<td>101</td>
<td>102</td>
<td>103</td>
<td>102</td>
<td>103</td>
<td>103</td>
</tr>
<tr>
<td>0.0</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

4. Conclusion

Trips and activities occur from decisions people make. These decisions concern route and destination choice as well as the choice of activities. The actual choices are dependent on the perceived travel times, rather than on the objective travel times.

As a result a traveler thinks he chooses the best route, but this route is not necessarily the best from an objective point of view.

It was proven that the perception of two or more independent routes is always less or equal than the perception of each of the two routes together. The travel time of the chosen route is systematically being underestimated. This means that destinations are chosen since they appear to be closer than they really are. This causes extra, unnecessary carkilometers.

Evidently the results that were gained should be looked at in the light of the assumptions made, as well as in the light of the limitations this approach has.

Providing information means that the difference between perceived travel time and objective travel time become smaller. This has impact on the choice of route and destination as well as on the location of activities. As a result the amount of carkilometers decreases.

The different testcases show that the form of the network, with respect to the presence of alternative routes is of importance. Further the simulations show that in a situation with congestion, the decrease of carkilometers is larger than in the situation when there was no congestion. The knowledge about congestion also influences the location pattern of the activities. (see the figure of the "TOWN"-network)
At the present it is not possible to quantify the effects of information precisely since the present and future value of $\alpha$ are not exactly known, uncertainty will only partially be influenced by information and only a part of the travelers will use the information.

It seems valid to state that information systems may decrease the amount of car kilometers in urban networks by 15-20% and in regional networks by 5-10%. There are 5,000,000 cars in the Netherlands. Each car does about 16,000 kilometer per year. The variable cost are 0.44 guilder per kilometer, including taxes, which are used for traffic, transportation and environment expenses. So RTI-measurements will reduce the total amount of kilometers by approximately 7.5 billion a year, which resembles social benefits of about 3 billion Dutch guilders or 1.5 billion ECU's a year.

5. REFERENCES


[12] HAMERSLAG, R. (1990): Controlled traffic with information systems and model applications (in Dutch), in F. Middelham (ed.): Symposium behoers(t) verkeer, Delft University, Delft


Analyzing the Netherlands Travel Survey. Methods for analyzing and some findings.

R. Hamerslag
ANALYSING
THE
NETHERLANDS' NATIONAL TRAVEL SURVEY
Methods of Analysing and some Findings

Paper prepared for the "5th DVWG seminar Verkehrstatistische Informatiesystemen für Wissenschaft, Management und Verwaltung"
September 13, 14, Heilbronn, BRD

Prof Dr Ir R. Hamerslag
Delft University of Technology 31 August 1990
SUMMARY

The Netherlands' National Transportation Survey is a home survey that has been held every week since 1978. A very large database is available to analyse the travel behaviour. The purpose of this contribution is to give an overview of the methods used to analyse these data, and to present some relevant findings.

The first part deals with making homogeneous groups of people with respect to travel behaviour. The principles of the used likelihood-based grouping methods are dealt with. The vital characteristics for the car-availability group are car ownership and a personal net income. For these groups the ratio between car kilometres during peakhours and weekdays has been determined.

The second part of the paper deals with getting a better insight into the factors that can explain the differences between the homogeneous groups. The weighted poisson estimator has been used to analyse the car-availability and the car kilometres for each of the six years between 1979 and 1984.

The most important influence factors are personal net income, age, gender, and the number of inhabitants of the dwelling place.

The results of these studies are used to forecast the car-availability in the next 15 years.

The data of the national travel survey are also used to estimate deterrence functions, to be used for the determination of Origin and Destination matrices. The method to estimate these functions is described in the paper. The methods used to fit the estimated OD matrix to traffic counts are briefly mentioned.
1 The Netherlands' National Travel Survey

The Netherlands' National Travel Survey (OVG) is a home interview that is held in the Netherlands. Specific for this home interview is that since 1978 from 800 up to 1000 households have been surveyed every month. In 1981 e.g. 24490 persons who made 131,029 trips and 1,442,720 trip kilometres. So in 10 years' time approximately 1/4 million persons have been interviewed who made more than 1 million trips and more than 10 million trip kilometres.

The sampled data are about the same as in the traditional home interview. They refer to the characteristics of the trips made (e.g. mode of transport, trip purpose), the characteristics of the persons (e.g. age, gender, car ownership, personal income, occupation, etc.) and to the characteristics of the household (household size, the number and age of children, the number of cars in the household, household income, etc.).

The origin and destination are coded on a municipal level. For 1981 however, origins and destinations on the more detailed post codes' level are available as well.

So a very large data set is available. These data are distributed on tape by the Netherlands' Central Bureau for Statistics.

Computer programs have been made to access these data and to use the database for analysis. The purpose of this contribution is to give an overview of the methods used to analyse these data and to give some findings.

Three major issues will be dealt with:

- methods to obtain homogeneous groups of data in terms of travel demand by the "likelihood-based grouping method";
- analysis of the differences among these groups in relation to their characteristics by means of the "weighted poisson method";
- the estimation of the deterrence functions by mode of
transport.

The first part of this contribution (par 2.1) deals with the determination of people having a homogeneous travel behaviour. The principles are outlined by the used likelihood based grouping method. It is found that the total population can be divided into groups with a different travel behaviour. For these groups the ratio between car kilometres during peakhours and weekdays has been determined (par 2.2)

The second part of the paper deals with an estimator for analysing the mobility of these groups. To get a better insight into the factors that can explain the differences between the homogeneous groups of data, higher dimensional matrices are made. The cells of these matrices are defined by the characteristics and classes. Because traditional regression technics cannot be used, the weighted poisson estimator has been developed. The paper will deal with the principles of this method (3.1). This method has been used to analyse the car-availability and the car kilometres (3.2). It has been done for each of the six years between 1979 and 1984. The results of these studies are used to forecast the car-availability in the next 15 years (3.3). The final part of the paper deals with the fact that the data of the national travel survey are also used to estimate deterrence functions. These functions are used for the determination of Origin and Destination matrices (4.1). The methods to estimate these functions and some findings are described. The deterrence functions are estimated for car trips and simultaneously for car drivers, car passengers and public transport users. The methods used to fit the estimated OD matrix to traffic counts are briefly mentioned (4.2).
Homogeneous population groups

2.1 The likelihood based grouping method

Tables are made by using certain attributes. The cells of these tables are more or less homogeneous, depending on the attributes used.

Aggregation of data causes loss of information. This is illustrated in figure 1. The small points show the trip/person per day made by individuals. The bold points show the result after aggregation into three groups. It is clear that this may

Figure 1. Illustration of the loss of information by the aggregation of data
lead to a wrong conclusion about the influence of the income on travel behaviour in this illustrative example. At the bottom of the figure aggregation into six groups is shown. The loss of information is less. The groups are more homogeneous. The influence of income on the number of trips is much better shown in this way. The aggregation of the upper part of the figure is less suitable for forecasting purposes.

Cluster and segmentation methods (see e.g. (1)(2)) can be used to make homogeneous groups of data. The "distance" between two groups of data is measured by a dissimilarity measure. Different measures are used. The resulting "homogeneous groups" depend on the dissimilarity measure used. A measure being consistent with the likelihood estimation theory has been developed. The principle of this method is the following:

The dissimilarity $D_{lk}$ between two groups $(l,k)$ of observations is defined as the difference of log-likelihood after segmentation {$\ln(L_l) + \ln(L_k)$} and before segmentation {$\ln(L_{lk})$}

\[
D_{lk} = \ln(L_l) + \ln(L_k) - \ln(L_{lk})
\]

(2.1) $D_{lk} = \ln(L_l) + \ln(L_k) - \ln(L_{lk})$

The larger the value of $D_{lk}$, the more homogeneous the groups $l$ and $k$ are.

The value of $D_{lk}$ depends on the probability density function of the observation(4). The poisson distribution is often used, because it is characterized by exclusively non negative integer observations. In that case the dissimilarity is:

\[
D_{lk} = \sum \{ x_{al} \ln(x_{al}/N_l) + x_{ak} \ln(x_{ak}/N_k) -
\]

\[
(x_{al}+x_{ak}) \ln((x_{al}+x_{ak})/(N_l+N_k))\}
\]

(2.2) $D_{lk} = \sum \{ x_{al} \ln(x_{al}/N_l) + x_{ak} \ln(x_{ak}/N_k) - (x_{al}+x_{ak}) \ln((x_{al}+x_{ak})/(N_l+N_k))\}$

Other dissimilarity measures are derived for normal and binominal probability density functions.
The method has been used to analyse car ownership (5),
kilometres by mode of transport(6), and the mobility of people
older than 55 (7).

2.2 Travel performance per mode of transport

Here are determined population groups with respect to travel
performance (kilometres by travel mode).
The likelihood has been determined for age, car-availability,
children in the household, city size, education level,
employment status, gender, household income, personal income,
income/adult in the household, marital status, position in the
household, and for the variable "railway station nearby".

The car-availability gives
the highest value
of the distance
measure
factor(table 1). A
segmentation of the
data on the
characteristic car-
availability will
result in the most
homogeneous groups
of people. Within
the car-
availability group
personal income
gives the most
homogeneous
subgroups.

The results are given in figure 2. The vertical length of a
block indicates the group size. The horizontal width
indicates the mean travel distance by mode. So the surface
represents the trip kilometres made by the group by travel

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>No of classes</th>
<th>Dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All Classes</td>
<td>Final Step</td>
</tr>
<tr>
<td>Age</td>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>Car-availability</td>
<td>3</td>
<td>201</td>
</tr>
<tr>
<td>Children in the household</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>City size</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>Education level</td>
<td>7</td>
<td>51</td>
</tr>
<tr>
<td>Employment status</td>
<td>5</td>
<td>49</td>
</tr>
<tr>
<td>Gender</td>
<td>2</td>
<td>74</td>
</tr>
<tr>
<td>Household income</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td>Personal income</td>
<td>6</td>
<td>105</td>
</tr>
<tr>
<td>Income/adult in the household</td>
<td>6</td>
<td>13</td>
</tr>
<tr>
<td>Marital status</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>Position in the household</td>
<td>5</td>
<td>101</td>
</tr>
<tr>
<td>A railway station nearby</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>
Figure 2 Example of homogeneous population groups, with respect to travel performance.
mode.

After segmentation a great difference is find in kilometres made in the homogeneous groups of people. There is also a great difference in the used travel modes. So different groups of people have different travel needs.

3 Estimation of n-dimensional matrices

3.1 Weighted Poisson Estimator

To get a better insight into the factors that can explain the differences between the homogeneous groups of data, higher dimensional matrices are made. The cells of these matrices are defined by the characteristics and classes. In general the number of cells in the matrices become very large, so that most of the cells have a zero value. If a model is specified in a product form, then linearisation by a log transformation is not possible. So traditional regression technics cannot be used.

Therefore, the weighted poisson estimator has been developed. This method is based on the likelihood estimation theory. The probability function of the surveyed data is poisson distributed. The model has a product form.

\[(3.1) L^* = \ln(L) = \sum_a \ln \{\text{Prob}(T_a | T'_{a})\}\]

with the poisson distribution

\[(3.2) \text{Prob}(T_a \mid T'_{a}) = \exp(- T_a) (T'_{a})^{(T_a)} / (T_a !)\]

and a model in a product form

To analyse multi-dimensional trip tables the following equation is used:

\[(3.3) T'_{a} = P_a \cdot \pi_k \cdot b_{k1}\]

With:
The observed number of trips in cell $n$ of the matrix; $T_n$

The estimated number of trips in cell $n$ of the matrix; $T'_n$

$P_n$ is the weighting factor: in this case the number of persons in the group;

$\beta_{ki}$ the coefficients $l$ of the influence factor $k$.

The maximum value of the log likelihood is found by setting the first partial derivatives of $\beta_{ki}$ to zero

$\frac{\delta L^*}{\delta \beta_{kl}} = 0$ for all $k$ and $l$

The result is a set of equations. Solving these equations results in the coefficients of best fit.

### 3.2 Estimation of car ownership and car kilometres

This model has first been used to analyse accidents on roads(6) and later to analyse car-ownership(5) as well. Here we present some recent research results of the use of this method in order to analyse The Netherlands' National Travel Survey(OGV) data. The theme was the number of kilometres made by car drivers from 1979 to 1984.

$T'_{cilha} = P_{cilha} \cdot \beta_o \cdot \beta_{Cc} \cdot \beta_{Hi} \cdot \beta_{Hh} \cdot \beta_{Aa}$

with

$T'_{cilha}$ the number of trip kilometres made by people who have a car available in cell cihl of the matrix;

$P_{cilha}$ the number of people in cell $n$ of the matrix;

$\beta_o$ a constant;

$\beta_{Cc}$ factor for the number of inhabitants (class c) in the municipality;

$\beta_{Hi}$ factor for personal income class $i$;

$\beta_{Hh}$ factor for the difference between household and personal income in class $h$;

$\beta_{Aa}$ factor for age and sex group $a$. 
### Table 2: Estimation of car availability (Cars/person) and performance (Car-kilometers/person) with the weighted poisson estimator

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Cars in 1984</th>
<th>Car Availability Mean</th>
<th>Standard Deviation</th>
<th>Car-km in 1984</th>
<th>Performance Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>12369</td>
<td>.413</td>
<td>0.005</td>
<td>455523</td>
<td>14.142</td>
<td>0.311</td>
</tr>
<tr>
<td>Inhabitants in 10^3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 no station</td>
<td>5372</td>
<td>1.067</td>
<td>0.007</td>
<td>211140</td>
<td>1.130</td>
<td>0.035</td>
</tr>
<tr>
<td>30 station</td>
<td>2797</td>
<td>1.022</td>
<td>0.009</td>
<td>105068</td>
<td>1.045</td>
<td>0.035</td>
</tr>
<tr>
<td>30 - 100</td>
<td>1861</td>
<td>0.953</td>
<td>0.005</td>
<td>64965</td>
<td>0.905</td>
<td>0.027</td>
</tr>
<tr>
<td>100 -</td>
<td>2340</td>
<td>0.897</td>
<td>0.021</td>
<td>74349</td>
<td>0.803</td>
<td>0.029</td>
</tr>
<tr>
<td>Personal income in FL.10^6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no</td>
<td>1022</td>
<td>0.390</td>
<td>0.024</td>
<td>45312</td>
<td>0.430</td>
<td>0.012</td>
</tr>
<tr>
<td>- 8</td>
<td>299</td>
<td>0.558</td>
<td>0.046</td>
<td>15209</td>
<td>0.569</td>
<td>0.029</td>
</tr>
<tr>
<td>8 - 17</td>
<td>1562</td>
<td>0.953</td>
<td>0.032</td>
<td>47171</td>
<td>0.754</td>
<td>0.036</td>
</tr>
<tr>
<td>17 - 24</td>
<td>3381</td>
<td>1.407</td>
<td>0.026</td>
<td>104410</td>
<td>1.175</td>
<td>0.017</td>
</tr>
<tr>
<td>14 - 38</td>
<td>3141</td>
<td>1.668</td>
<td>0.025</td>
<td>145633</td>
<td>1.749</td>
<td>0.068</td>
</tr>
<tr>
<td>38 -</td>
<td>2184</td>
<td>1.847</td>
<td>0.018</td>
<td>97787</td>
<td>2.420</td>
<td>0.094</td>
</tr>
<tr>
<td>Difference Household &amp; Personal income</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Small</td>
<td>10234</td>
<td>0.992</td>
<td>0.004</td>
<td>363915</td>
<td>0.978</td>
<td>0.013</td>
</tr>
<tr>
<td>Large</td>
<td>2135</td>
<td>1.043</td>
<td>0.016</td>
<td>91608</td>
<td>1.101</td>
<td>0.067</td>
</tr>
<tr>
<td>Age &amp; Sex Man</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 - 24</td>
<td>905</td>
<td>0.193</td>
<td>0.056</td>
<td>40960</td>
<td>1.623</td>
<td>0.122</td>
</tr>
<tr>
<td>25 - 44</td>
<td>5096</td>
<td>1.305</td>
<td>0.017</td>
<td>190181</td>
<td>1.325</td>
<td>0.038</td>
</tr>
<tr>
<td>45 - 64</td>
<td>2833</td>
<td>1.190</td>
<td>0.029</td>
<td>96905</td>
<td>1.064</td>
<td>0.060</td>
</tr>
<tr>
<td>65 -</td>
<td>1007</td>
<td>0.792</td>
<td>0.070</td>
<td>27630</td>
<td>0.574</td>
<td>0.089</td>
</tr>
<tr>
<td>Woman</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 - 24</td>
<td>327</td>
<td>0.485</td>
<td>0.033</td>
<td>17082</td>
<td>0.660</td>
<td>0.042</td>
</tr>
<tr>
<td>25 - 44</td>
<td>1308</td>
<td>0.725</td>
<td>0.021</td>
<td>56116</td>
<td>0.795</td>
<td>0.052</td>
</tr>
<tr>
<td>45 - 64</td>
<td>686</td>
<td>0.568</td>
<td>0.049</td>
<td>21791</td>
<td>0.471</td>
<td>0.051</td>
</tr>
<tr>
<td>65 -</td>
<td>207</td>
<td>0.172</td>
<td>0.031</td>
<td>4859</td>
<td>0.135</td>
<td>0.031</td>
</tr>
</tbody>
</table>
In table 2 the mean of the coefficients that are estimated for each of the six years with the standard deviation. The most important findings are:

- in municipalities with less than 30,000 inhabitants car ownership and car use are larger than in municipalities with more inhabitants;
- car-availability and car use are strongly influenced by personal income;
- car-availability and car use of men is greater than of women;
- most estimated coefficients in the period from 1979 until 1984 are stable over time. There is but little increase in the car-availability and car use in the higher age groups during that period.
3.3 The relationship between peak hours and weekdays

The ratio is determined between car kilometres made during weekdays and during peak hours (16.30 - 17.30). The result of the analyses is given in table 3.

The ratio between peak hours and non-peak hours is different for several groups.

As may be expected participation in the labour market is an important influence factor. Young people are less free to avoid peak hours than older people. Of the trip kilometres 83% are made by the working population, although the trip purpose work-home is much lower. The population in the higher income groups (24% of the total population) made 58% of the trip kilometres in peak hours and 55% during daytime.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Persons Day</th>
<th>Peak</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inhabitants municipalities</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in 1000 persons</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>- 30 no station</td>
<td>39</td>
<td>45</td>
<td>0.143</td>
</tr>
<tr>
<td>- 30 with station</td>
<td>22</td>
<td>23</td>
<td>0.131</td>
</tr>
<tr>
<td>30 - 100</td>
<td>17</td>
<td>15</td>
<td>0.136</td>
</tr>
<tr>
<td>100 -</td>
<td>22</td>
<td>17</td>
<td>0.128</td>
</tr>
<tr>
<td>... +</td>
<td>100</td>
<td>100</td>
<td>0.136</td>
</tr>
<tr>
<td>Personal income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in F 1000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>no and &lt; 8</td>
<td>37</td>
<td>13</td>
<td>0.085</td>
</tr>
<tr>
<td>8 - 17</td>
<td>19</td>
<td>11</td>
<td>0.126</td>
</tr>
<tr>
<td>17 - 24</td>
<td>20</td>
<td>24</td>
<td>0.140</td>
</tr>
<tr>
<td>24 - 38</td>
<td>16</td>
<td>30</td>
<td>0.160</td>
</tr>
<tr>
<td>38 -</td>
<td>8</td>
<td>23</td>
<td>0.136</td>
</tr>
<tr>
<td>... +</td>
<td>100</td>
<td>100</td>
<td>0.136</td>
</tr>
<tr>
<td>Age and gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 - 24 y</td>
<td>7</td>
<td>7</td>
<td>0.153</td>
</tr>
<tr>
<td>25 - 44 y</td>
<td>21</td>
<td>44</td>
<td>0.153</td>
</tr>
<tr>
<td>45 - 64 y</td>
<td>13</td>
<td>21</td>
<td>0.128</td>
</tr>
<tr>
<td>65 -</td>
<td>7</td>
<td>3</td>
<td>0.093</td>
</tr>
<tr>
<td>... +</td>
<td>48</td>
<td>77</td>
<td>0.142</td>
</tr>
<tr>
<td>Women</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 - 24 y</td>
<td>8</td>
<td>4</td>
<td>0.162</td>
</tr>
<tr>
<td>25 - 44 y</td>
<td>22</td>
<td>13</td>
<td>0.107</td>
</tr>
<tr>
<td>45 - 64 y</td>
<td>14</td>
<td>5</td>
<td>0.105</td>
</tr>
<tr>
<td>65 -</td>
<td>8</td>
<td>1</td>
<td>0.076</td>
</tr>
<tr>
<td>... +</td>
<td>52</td>
<td>23</td>
<td>0.116</td>
</tr>
<tr>
<td>Men and women100</td>
<td>100</td>
<td>100</td>
<td>0.136</td>
</tr>
<tr>
<td>Participation in the labor market</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>35</td>
<td>25</td>
<td>0.094</td>
</tr>
<tr>
<td>No</td>
<td>45</td>
<td>75</td>
<td>0.150</td>
</tr>
<tr>
<td>... +</td>
<td>100</td>
<td>100</td>
<td>0.136</td>
</tr>
</tbody>
</table>
3.4 Forecast of the mobility 1985-2000

The analyses are used to forecast the mobility in the next 15 years.
Increase in car mobility is expected because of:
- the increase in adult population;
- the increase in personal income;

Figure 3 Relation between personal income, age and car availability
the penetration of car ownership in higher age groups by cohort effects.
The mean of the car-availability is about 40% in the period 1989-1984. However, there are great differences between different income and age groups.
In fig 3 the relationship is shown between income, age and car-availability for various income groups. As may be expected, car ownership is considerably influenced by income. The car-availability is greater in the higher income groups than in the lower income groups. In high income groups and age groups of 25-44, the car-availability is near its absolute limit (95%). Notice the difference between the 25-44 age groups and between the 45 and 65 age groups. It is expected that the difference between these age groups will disappear in the next 20 years (Cohort effect).
So it is expected that the number of cars will increase because of an increase in adult population, growth in income, and by the cohort effect.

The personal income also influences the number of car kilometres. Figure 4 gives the relationship between income and kilometres/person/day made by the car-availability group. The growth in car kilometres depends on the number of cars, the growth in income and the spread of the income.
A backwards forecast has been made for the years 1970-1985 (figure 5). The forecast follows the development very well in the past.
Figure 4: Relation between personal income and kilometers/person/day in the car availability group.

Figure 5: Backwards forecast (1970-1985) of the car kilometers.
The increase in car kilometres for the period 1985 to 2000 can be summarized as follows:

- Increase in the adult population: 15%
- Cohort effect: 10%
- Increase in personal income: 10 - 20%

Total: 35 - 45%

The car use will increase the next 15 years by 35-45% and in the next 25 years by 65-75%, caused by a growth in population, cohort effects and higher personal income. Notice that the growth of population and cohort effects are pretty certain. Less certain is this of the personal income. The increase will be realized by women and lower income groups. This growth can hardly be influenced by the given income and woman emancipation policies.

About 20% of the forecast increase has already been realized in 1990.

4 Estimation of deterrence functions

4.1 Deterrence functions with OD data and Weighted Poisson Estimator

Origin Destination (OD) matrices are widely used in transportation planning. Origin Destination matrices are observed by home interviews in the Netherlands' National Travel survey on munucipality level. These matrices can be used to estimate deterrence functions. These functions are used for forecasts. However, most cells of the matrix as surveyed contain zeros, so that a logarithmic transformation cannot be used. Therefore, the weighted poisson estimator is used to estimate deterrence functions which may be used in forecasts. They are estimated for:

- car traffic, and
- simultaneously for car drivers, car passengers and public transport users.
If OD matrices are partly observed (10), e.g. if screenline interviews are available, this method can also be used (11). Then the method is called "partial matrix technique".

To estimate multimode Origin-Destination matrices, the following model is used (9). The interaction model is:

\[(4.1) \quad T'_{ijm} = \mu \cdot (l_i \cdot Q_i) \cdot (m_j \cdot X_j) \cdot F(Z_{ijm}),\]

in which

- \(T'_{ijm}\) estimate of the number of trips between zone \(i\) and zone \(j\) by mode \(m\);
- \(Q_i\) and \(X_j\) are the polarities of \(i\) and \(j\);
- \(l_i\) and \(m_j\) are the equilibrium factors of \(i\) and \(j\);
- \(Z_{ijm}\) are the (generalized) travel times between \(i\) and \(j\) by mode \(m\);
- \(F(Z_{ijm})\) is the deterrence function.

Let \(q_i \stackrel{\text{def}}{=} (l_i \cdot Q_i) ; x_j \stackrel{\text{def}}{=} (m_j \cdot X_j)\) and the deterrence function with discrete values \(F_{km} \stackrel{\text{def}}{=} F(Z_{ijm})\).

If only a part of the matrix is observed, as in cordon interviews, then \(d_{ijm} = 1\) if \(ijm\) has been observed, and \(d_{ijm} = 0\) if \(ijm\) has not been observed. So (4.1) may be written as

\[(4.2) \quad T'_{ijm} = \mu \cdot q_i \cdot x_j \cdot F_{km} \cdot d_{ijm}\]

Because most cells of the observed matrix \(T_{ijm}\) are zero, it is not possible to make it linear by a log transformation. Therefore, the use of the weighted poisson estimator is necessary.

\[(4.3) \quad \delta L^*/\delta q_i = 0 \quad \text{for all} \ i\]

\[\delta L^*/\delta x_j = 0 \quad \text{for all} \ j\]
Figure 6 Estimated deterrence function for the regions of 't Gooi, Haarlem, Leiden, Utrecht, Breda, Tilburg, Den Bosch, Einhoven, Maastricht, Mijnstreek, Nijmegen, Arnhem, Twente and Groningen.
19

\[ \frac{\delta L'}{\delta F_{km}} = 0 \text{ for all } k \text{ and } m \]

\[ \frac{\delta L'}{\delta \mu} = 0 \]

Solving this set of equations results in the coefficients \( \mu \), \( q_i \), \( x_j \), and the discrete values of the deterrence function \( F_{ij} \) in (4.2)

The estimates of the OD matrix \( T'_{ij,m} \) fit optimal to the observations \( T_{ij,m} \).

This method has been widely used in the Netherlands with available OD data. Recent estimations with NNTS data are made for intra regional trips (11) (figure 6) and for national trips (12) (figure 7).

Deterrence functions have recently been estimated for car drivers, public transport users, car passengers and in the agglomeration also for bike and feet. The travel time is used as an influence factor in the national study. The functions are estimated for Car Available (CA) and Non Car Available (NCA) people for different trip purpose groups.

The log-normal function that fits the discrete function values \( F_{km} \) best, is specified as:

\[ (4.4) F(Z_{ij,m}) = \exp(-0.4 \ln^2 (Z_{ij,m} + 1)) \]

The exponent (-0.4) is found in several studies
Specific for these NNTS (OVG) studies is that the same function can be used for car drivers and public transport if the time difference between these modes is taken into account.

Another interesting outcome of this study (12) is that the exponent (-0.4) is influenced by the personal income. This value for car drivers (work-home) is in low income groups - 0.46, for middle income groups - 0.41, and for high income groups - 0.39.
Figure 7 Estimated deterrence function for national travel in the Netherlands (12)
The method has been used for traffic in telephone networks. The resulting deterrence functions are estimated. The difference is that the distance between origin and destination is divided by 10. Good results are also achieved by analysing OD matrices for postal services.(14)

4.2 Estimation of OD matrices using traffic counts

In general a great deal of information is available in traffic counts. A large number of methods are used to estimate OD matrices with the help of traffic counts.

We will describe three methods that are used in the Netherlands, i.e.

+ entropy estimators with hard constraints;
+ entropy estimators with elastic constraints;
+ likelihood estimators.

Well-known is the entropy estimator and the information minimization of Van Zuylen.(15) These methods have the disadvantage that every constraint (count) needs its coefficient. The original model changes, which can become a problem when making forecasts. This problem can be overcome by using elastic instead of hard constraints(19)

The methods that are based on the likelihood estimation theory do not have this disadvantage too. A chi-square method known as FASCAL(17) and a likelihood estimator known as BIKAL(18) are used by different consulting firms in the Netherlands. In order to compare the methods, a pilot study has been carried out.

The different estimation methods lead to different OD matrices. It is advisable to use the well accepted likelihood estimation theory.
5 Concluding remarks

The Netherlands' National Transportation Survey is a home survey that has been held every week since 1978. A very large database is available to analyse travel behaviour.

It is found that the total population can be divided into groups with a different travel behaviour. The vital characteristics for the car-availability group are car ownership and personal net income. There are great differences in the total number of kilometres and the used travel modes.

For these groups the ratio between car kilometres during peak hours and weekdays has been determined. As may be expected, participation in the labour market is an important influence factor. Young people travel more during peak hours than older people. Of the population 24% with higher incomes make more than 55% of the car kilometres during peak hours.

The traffic to central areas of communities with more than 100,000 inhabitants is 5%-7%. Because the public transport system to these areas is good and worse to other areas, the possibilities to influence growth in car use by an improvement of the public transport system will be limited.

To get a better insight into the factors that can explain the differences between the homogeneous groups of data, higher dimensional matrices are made. The cells of these matrices are defined by the characteristics and classes.

The weighted poisson estimator has been used to analyse the car-availability and the car kilometres for each of the six years between 1979 and 1984. The most important influence factors are personal net income, age, gender, and the number of inhabitants of the dwelling community.

The results of these studies are used to forecast the car availability in the next 15 years. An increase in the number
of cars is expected caused by an increase in the adult population, a penetration in higher age groups by cohort effects and an increase in personal income. The number of car kilometres of the car-availability group will also be increased by a growth in personal income. The car-use will increase the next 15 years by 35-45% and in the next 25 years by 65-75% The increase will be caused by women, and by increasing the income of the lower income groups. This growth can hardly be influenced by the prevailing income and woman emancipation policies.

The data of the national travel survey are also used to estimate deterrence functions, to be used for the determination of Origin and Destination matrices. The method to estimate these functions is described in the paper. The deterrence functions are estimated for car trips and simultaneously for car drivers, car passengers and public transport users. The same functions for car drivers and public transport are found if excess times are taken into account. The functions may also be employed to estimate telephone traffic.

**Literature**


6. Hooning J.J.(1985) et al "Een verkenning naar mobiliteitspatroon van ouderen"(Mobility study people elder then 55 year. DHV Consultants, Amersfoort

7. R. Hamerslag (1982), J.P. Roos en M. Kwakernaak


15. Zuylen H.J. van(1979) and L.G.Willumsen "The most Likely Trip Matrix Estimated from Traffic Counts" Transportation Research (14b) pp 281-293


Quantitative relationship between traffic, infrastructure and environment.

R. Hamerslag and M. Westerman

Proceedings 25th ISATA, Road Transport Informatics/Intelligent Vehicle Highway Systems, Florence, Italy, pp 221-228
ABSTRACT
At an expected increase of traffic in the coming 25 years of 65-75%, emissions can be reduced to half of the present level, provided that all proposed measures are implemented at the same time.

Improved environmental effects are shown by pricing measures, more economical and cleaner cars, regulating the speed limit on freeways down to 90 km/h and the application of route-guidance and trip-planning systems for car drivers.

Though public transportation plays an important part in maintaining accessibility, quantitative modelling studies show that the effectiveness of measures improving the public transportation system in favour of environmental goals may be overestimated.

In order to prevent the further spreading of traffic and activities, it is in some cases advisable to increase the capacity of some freeways while taking into account the scope of the environmental goals.

1. INTRODUCTION
Within the scope of education and planning of future research at Delft University of Technology, Department of Transportation Planning and Highway Engineering there is a need to anticipate current traffic problems. This need was the occasion of this study.

We paid attention to the possible growth of car traffic, environmental effects of this growth, problems caused by a continuously worsening accessibility and had a closer look at instruments and measures by means of which it might be possible to limit the negative consequences.

In this paper we present relevant results of this study. First, in paragraph 2, we will pay attention to the growth of car traffic in the coming 25 years and in paragraph 3 to objectives and interests of individuals, companies and society. Next, in paragraph 4, we will examine the extend to which price measures, a stimulation of alternative means of transport, the application of trip planning and route guidance, push measures, an expansion of the capacity and a decrease of the harmful emission per car kilometre, contribute to accessibility and the environmental objective.

2. INCREASE IN CAR TRAFFIC
To be able to forecast the growth of car traffic we classified the population into homogeneous population groups with respect to travel behaviour (5). Using such a
classification we forecast a 65-75% growth of car traffic in the west of the Netherlands on the highway network in the coming 25 years, assuming a growth of the adult population on the one hand and a growth of car ownership and a growth of the number of car kilometres in the car ownership group brought about by a growth of income on the other hand. By now, this growth on the highway network has already been 35% (16), so we fear that our forecast is still too low.

Up to about the 1950s only locations within the range of public transport were able to develop. Residential areas in commuter cities and major concentrations of employment developed in the surroundings of public transport. The increase of car ownership fundamentally changed this situation (10). Employment areas and residential areas developed independently of the public transport system and are are now mainly concentrated around the highway network.

As a result, only 17% of car traffic nowadays has its origin or destination in central cities areas. The share of city centres in this amounts to approximately 5%.

3. OBJECTIVES
The expected growth of car traffic will cause spatial and financial problems which will be hard to solve. We doubt that it will be possible to deal with this growth if we do not take drastic measures in time. When searching for these measures we should bear in mind that there exist different and often conflicting interests.

The interests of individuals
It is in the interest of traffic participants to reach their destinations in time, sufficiently fast, at an affordable price and without accidents, to carry out activities.

The interests of companies
It is in the interest of companies to deliver the goods at the desired point in time making costs that are as low as possible.

The interests of society
The interests of society with respect to traffic and transport concern facilities, accessibility, environment and traffic safety and can be summarized as follows:
- Supplying the different groups of traffic participants with adequate transport facilities including the group of people not owning a car.
- Guaranteeing or stimulating the accessibility of areas such as city centres, residential areas, office establishments, ports and airports.
- Decreasing air pollution, consumption of energy, noise pollution, destruction of the landscape caused by traffic, etc.
- Decreasing the number of accidents.

Accessibility and environment are interests of society which seem to force conflicting measures. We will show that this need not be the case if we take the correct measures in time. In the following we limit ourselves to the interests of society with respect to environment and accessibility.
4. CONTRIBUTIONS TO THE ACCESSIBILITY AND ENVIRONMENTAL OBJECTIVES

The objective with respect to accessibility can be measured in number of car kilometres during peak hours. The objective with respect to environment can be met by reducing the total number of car kilometres, the fuel consumption per car kilometre and the harmful emission per car kilometre.

4.1 Price measures

An important foundation on which the Dutch government is building her policy with respect to limiting the growth of car traffic is the use of price measures. An increase of the fuel price has a positive effect on the fuel consumption in the long term. In literature we find a price elasticity of -1 \((15)(2)\). This effect will probably be less (in the Netherlands), because people will substitute this increase in costs by using fuels which are less expensive (gas or diesel). As a result of the substitution-effect an increase in fuel prices will in the long term have very little effect on the total number of car kilometres. Therefore this measure is not very effective in dealing with the accessibility and environmental issues.

4.2 Stimulating alternative means of transport

4.2.1 Stimulating public transport

Environmental objective

Although there are a great many possibilities to improve the present situation, the Dutch public transport system is one of the best in the world. For the last 25 years the public transport network has been expanded at many places. A comparison of the public railway network in the rim-shaped agglomeration of cities in the western part of the Netherlands with railway networks in other world cities shows that in the Netherlands we are well behind as far as the density of the railway network is concerned. The possibilities for expansion are limited. The origin and destination table consists of a large number of very small relations. To keep the exploitation shortages within acceptable limits the size of the transportation flows must be sufficiently large. Only a few of the relations are big enough to justify a public transport relation. Because of this the expansion of the railway network only consists of some main relations. Even with large investments in mass public transport we still do not expect the improvements to be more than local.

The relation between car drivers on the one hand and public transport passengers who have a car at their disposal on the other hand during the peak hours amounts 20:1 (see figure 1). Even the very optimistic assumption that the number of public transport passengers in this car-available group would be doubled at the expense of car drivers in the other group would only change this relation into 19:1.

In conclusion we can say that we do not expect public transport to be able to contribute significantly to push back the total number of car kilometres and the emission of harmful pollution.

Figure 1 Trip kilometres in the car-availability group
**Accessibility objective**
The share of public transport in direct relations often amounts to more than half of the total number of movements. Because city areas often have many direct relations with their surroundings the role public transport plays with respect to accessibility is vital. This can be shown by means of a model simulation.

In the figures the equilibrium situation in spatial development is simulated with (figure 2) and without (figure 3) a public transport network. As we can see employment moved out of the city centre in the case when no public transport network was available. The function of public transport clearly is preserving the employment in the centre and stimulating developments elsewhere.

![Figure 2 Spatial development with public transport network](image)

**Figure 2 Spatial development with public transport network**

![Legend](image)

**Legend**

![Figure 3 Spatial development without public transport network](image)

**Figure 3 Spatial development without public transport network**

### 4.2.2 Stimulating car- and van-pool
Stimulating people to travel together in a car- or van-pool is a possibility to restrict the amount of traffic. A van-pool is understood to be a group of on the average 11.5 persons who use a VAN for commuter traffic (11). The van has been paid for by the participants and is driven by one of them. An advantage is that no extra driver and no empty return rides are needed.

The contribution to the environmental objective of car- and van-pooling will be of little importance. To the accessibility objective the contribution will be vital, mainly in relations with employment concentrated around the highway network.

### 4.2.3 Stimulating the use of bicycles
For use in relations with a short distance (city areas) the bicycle is an outstanding example of a suitable mean of transport, because of its very favourable spatial and environmental component. The transport performance of the bicycle equals the one of public transport. The use of bicycles should be stimulated as much as possible (7).

### 4.3 Trip planning and route guidance
By means of route guidance systems information about the route is given during the trip. By means of trip planning systems (6) a prediction about the travel time is made, which is given in advance of or during the trip.
Trip planning and route guidance for passenger car-traffic
To study the effect of route guidance and trip planning we carried out a traffic flow calculation using different networks (7). The results appear to be sensitive to the number of alternative routes and the existing uncertainty and should therefore be interpreted indicatively.

<table>
<thead>
<tr>
<th>Congestion</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Routeguidance</td>
<td>11%</td>
<td>16%</td>
</tr>
<tr>
<td>Trip planning</td>
<td>18%</td>
<td>23%</td>
</tr>
</tbody>
</table>

Table 1 Decrease of the number of car kilometres

Contrary to expectation the results show less traffic during the peak hour as well as outside this period. This implies less fuel consumption and less harmful pollution.
Route guidance as well as trip planning reduce congestion and contribute positively to the environmental objective.

Trip planning and route guidance for truck-traffic
A combination of monitoring systems for truck-traffic and trip planning systems will make it possible to produce trip schedules which avoid congestion. When such a system is used orders can be dispatched in another sequence, so passing bottlenecks through can be put off till after the peak hour. The disadvantages of congestion can be reduced and goods can be delivered within the correct period of time.

4.4 Push-measures
We consider calculated jamming of the traffic by means of not taking certain measures as a possible push-measure. It is assumed that the loss of time as a consequence of congestion on highways results in traffic diverting to relations with less or no congestion.

To verify this assumption we carried out a model study with which we computed the consequences of congestion on the number of car kilometres, taking in consideration feedback at:
- route choice,
- route choice and destination choice,
- route choice, destination choice and changes in the pattern of activities.

The results of this model study are shown in the following two figures and table.
When we compare the results of the computations with and without congestion, we hardly see a decrease in car kilometres. This is caused by traffic avoiding the congestion. In this way the size of the congestions and the duration of the delays during the peak hours keep within bounds. The activities spread over a larger area. By allowing the traffic to be jammed no decrease in car kilometres and no reduction in air pollution is accomplished, whereas the decentralisation of the activities will have a negative influence on public transport.

Because we have little congestion beyond the peak hours this kind of push-measures does not effect the total number of car kilometres either.

4.5 Expanding the capacity of the highway network
An expansion of the capacity of the highway network will lead to an increase in accessibility. The effects of this expansion do not necessarily conflict with environmental objectives.

4.5.1 Additional lanes
A physical increase of the capacity by judiciously adding some lanes is an obvious measure for maintaining the accessibility. Expanding the highway network by adding new highways should be avoided as much as possible, because of its negative impact on environment.

(14) Shows that an addition of new lanes may result not only in a reduction of congestion but also in a 'back to peak hours' effect and a traffic-attracting effect of the highways.

4.5.2 Vehicle navigation
We consider vehicle navigation to be keeping the right distance between cars using electronic means (1).

Besides enhancing traffic safety a major advantage is an increase in the capacity of existing highways. In non-automated systems the mean period of time between two vehicles is 1.8 to 2 seconds. The mutual distance is changes continuously so continual correction by the car driver is needed to keep the right distance. Therefore the distance between the cars is often larger than necessary. Automating by means of vehicle navigation means reducing the human reaction time interval, through which the mutual distance can become slightly shorter. A decrease of the mutual distance with 0.5 seconds results in an increase of the capacity of 25%.

4.5.3 Road metering
The flow on the highway network is maximal at speeds between 50 and 90 km/hour. By means of metering and with the aid of special purpose hard- and software (3) maintaining speeds of at least 70 km/hour can be aimed at.
In conclusion we can say that expanding the highway network increases the chance of expansion of activities within the range of the highway network and so contributes to accessibility. This will not lead to adverse effects of harmful pollution, because traffic is attracted from elsewhere. A combination of these measures with stimulation of car- and van-pooling could lead to better results.

In default of such a measure the activities will spread over a larger area, which is much more harmful with respect to the environmental objective.

4.6 Decrease of harmful emission per car kilometre

It is possible to reduce the emission by cars of nitrogen and hydrocarbon. (13) And (16) expect a 30-40% cleaner fleet of cars in 2010, accomplished by improved transmissions, application of catalyst and poor-mixed engines.

5. QUANTIFYING THE RESULTS

By means of quantitative model simulations we have tried to determine the effect of each above-mentioned measure on fuel consumption and congestion.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Energy</th>
<th>Congestion</th>
<th>City</th>
<th>Highway</th>
<th>City</th>
<th>Highway</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reduction of the fuel consumption per car kilometre amounts to 30% when the fuel price is increased by 30-40%.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reducing the speed on highways from 120 to 90 km/hour and regulation on the highway network is possible with more control and results in 25% less fuel consumption. If we likewise take into account the increase of travel times the fuel consumption even decreases by 33%.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A 30-40% reduction of the harmful emissions is possible by improving the vehicle.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The effects of trip planning systems for passenger cars with respect to the reduction of the total number of car kilometres amount to 18%, while route guidance systems in city area networks show 16% less congestion during peak hours and 11% outside the peak hours (provided that all cars are equipped with on board units).

The effect of all proposed measures as a whole will be a reduction of the harmful emission of 60-65% in city areas and of 70-75% on the highway network. When we likewise take into account a growth of 65-75% of the number of car kilometres the emission will reduce by 30-40% of the present level in city areas and by 50-60% on the highway network.
6. CONCLUSIONS

Public transport plays a vital role in the accessibility objective and in maintaining urban concentration. With respect to environment and compared to other measures public transport is of little importance.

Vehicle navigation, trip planning, route guidance show positive effects in dealing with traffic issues.

Contrary to expectation a judicious expansion of the capacity of existing highways with relation to the environmental objective is advisable, in order to prevent a further spreading of traffic and activities.

Our conclusion is that it should be possible to reduce the harmful emission by 50% and keep the congestion from exceeding the present level. The inevitable growth of traffic and the increase of road capacity are not necessarily conflicting environmental issues.

LITERATURE

8. Heuvel, M.G v/d. Dutch Railways in city regional transport: participation or integration (in Dutch), In: VPC90
11. Kumar A(1989) and M. Molov A study of Van-pools in Los Angeles TRB annual meeting, paper 890570
1987

Teacher-friendly transportation programs.

R. Hamerslag

A series of programs has been developed especially for educational purposes and to support the lectures on urban and interurban modelling. These programs are collectively called “Teacher-Friendly Transportation Program/Student Proof Version” or TFTP/SPV.

The programs consist of four levels. The first level illustrates the working of the calculation methods used in detail. The second level gives the interconnection between the different parts of the transportation forecasting model. The third level demonstrates the estimation of base year origin-destination matrices, estimation of deterrence functions, and the nonlinear weighted Poisson estimation method for multidimensional matrices.

The programs are used in lectures, but can also be used as prototypes in feasibility studies of new research developments (Level IV). The programs can be used on microcomputers working under MS-DOS.

INTRODUCTION

In the past, network-based highway and transit planning computer programs have been developed to forecast urban and interurban traffic and transportation. Well-known examples are, for instance, the Los Angeles Transportation Study, Chicago Area Transportation Study, London Traffic Survey, etc. The calculations for these studies have then been carried out on large mainframes. In the meantime the methods used in these studies have been improved scientifically and technically. However, the improved versions are not in general very user-friendly. It takes a lot of time to learn how to use this kind of program. To find a path through the labyrinth of the shelves with handbooks of the transportation programs takes weeks of study. It often turns out that the handbooks need some updating because the computer programs have changed. To arrange the necessary input data takes a lot of time which makes the programs less user-friendly.

The better of these programs have been modified in order to be used on microcomputers [21]. These “downloaded” programs have become a lot more user-friendly than before. Examples of these programs areTRANPLAN, TRANPRO, MICROTRIPS, MUNITP, TMODEL, EMME/2 [2, 20].

As I found out experimenting with programs in the Netherlands, some of the problems which exist in the older programs also exist in the “downloaded” versions. The same kind of problems are reported in [21] and [21] about some programs developed in the USA [2, 21].

To use these programs for educational reasons would not be advisable. The most important reason for not doing so is that the available “downloaded” programs have not been developed for educational purposes. They are less suitable for use during lectures, since only the results of the calculations are shown and not the succeeding steps in the calculation process. Another restriction concerns the limited availability of time during the lectures. The response time of the “downloaded” programs is too long.

Therefore, a series of programs has been developed especially for educational purposes and to support the lectures on urban and interurban modelling. The set of programs has been developed to give insight in the possibilities of the (large) professional models mentioned above. They can also be used in student workshops on this subject. The programs have been used in courses for professional transportation planners as well. Furthermore, they already have proved to be a good instrument for research purposes.

DESCRIPTION OF THE FUNCTION OF THE PROGRAMS

Since the computer programs were to be used during the lectures, they consist of independent programs for:
1. Forecasting traffic and transportation flows;
2. Estimating coefficients;

Much attention has been paid to the scientific mathematical principles of the programs. To obtain the calculation results within the limited available time of the lectures, the input for the network specification must be held as simple as possible.

In order to facilitate the demonstration of the calculation results, extra attention is given to the graphical reproduction of the results on the screen. Examples are the calculation of the trees of shortest routes, exogenously and endogenously determined tripends, and assignment results.

The pictures of networks, tripends, and traffic flows, stored on files, can be printed on screen. A hard copy of the screen can be made by normal MS-DOS routines.

There are four levels to demonstrate the modelling. The levels I and III contain partly independent programs (marked with [-]) which can be generated by menu manipulation. The programs of level II and some programs of the other levels are more comprehensive. In this respect they show some resemblance to the traditional programs for mainframes. They are marked with [+].

The first level (I) can be used to illustrate elements of the calculation process in detail:

- A tree-building algorithm [3, 13];
- The Gauss-Seidel iteration of the double constraint distribution process [17];
- A Monte-Carlo technique used for stochastic assignment [1];
- A nonlinear optimization used by the equilibrium assignment [11, 16].

The second level (II) pays attention to the interconnection between the different parts of the traffic forecasting model:

+ Network specification, tripends;
+ Calculation of shortest and minimum cost paths between origin and destination pairs;
+ Tripend by car, trip distribution, and spatial interaction;
+ All or nothing, stochastic and equilibrium assignment.

The third level (III) is meant for lectures on estimating coefficients for transportation models. Special attention is given to less traditional estimators, e.g.,

- The estimation of deterrence functions and origin-destination matrices with the weighted Poisson estimator [4, 5], and partial matrix techniques [12, 14];
- The estimation of origin-destination matrices of a base year matrix using traffic counts [11-14];
- A nonlinear, likelihood-based weighted Poisson estimator [7], for the estimation of coefficients in multiplicative models;
- Likelihood-based grouping methods [9] for making homogeneous groups of data.

The fourth level (IV) is the TFTP workbench and is used to try out new modelling techniques.

**FIGURE 1.** The main menu.

**Workfile: PHNX**

---

**Level I**

**Calculation algorithms**

- ROUTES
- DISTRIBUTION
- ALL-OR-NOTHING
- STOCHASTIC
- EQUILIBRIUM
- EXAMPLES

**Level II**

**Traffic flow forecasts**

- NETWORK
- ROUTES
- DISTRIBUTION
- ALL-OR-NOTHING
- STOCHASTIC
- EQUILIBRIUM
- USERSFILE

**Level III**

**Experimental research**

- ESTIMATORS, RESEARCH
- FAST PROGRAMS
- Other functions
- PICTURE
- WAIT
- FINISH
- CHANGE DISK

**Specification of new or changing of network & tripends**

< ENTER > to execute
DESCRIPTION OF THE PROGRAMS IN DETAIL

MENU

The main menu program (Figure 1) is used for selecting a program in different levels. The program prohibits the use of old files after changes in network, tripends, calculation of trips by car, and trip distribution. The menu program addresses the use of the right program after those changes. Therefore, it is impossible to use files from earlier alternatives if that is not explicitly wanted.

FILE MANAGEMENT

This program uses ASCII files. The files made in a certain computer run can be saved on disk as USERFILE for later use under a name chosen by the user. One of those USERFILES should be specified by the user as WORKFILE. The file management program is used to specify a WORKFILE from the available USERFILES or to save a WORKFILE on disk.

NETWORK AND TRIPENDS

This program has four functions:

1. Nodes. The user can determine the origin of the coordinate axis. The coordinates of locations of nodes are automatically determined. For demonstrations during lectures up to 7 nodes are mostly sufficient. For exercises during workshops 35 to 40 nodes are possible.

2. Roads. Roads are defined by a chain of two or more nodes. Road types are roads with the same speed and capacity. Road types can be selected by the user from a given set or can be specified by the user (Figure 2).

3. Tripends. Tripends (origins and destinations) can be addressed to nodes. Instead of tripends, population and employment per zone can also be used as input. The number of trips by car drivers can then be calculated as is described in a following section. Balancing procedures are used to equalize the total number of origins and destinations. The tripends can also be determined endogeneously in the distribution model under influence of the transportation network.

4. Save/change. Network and tripends can be saved after changes under a different name.

ROUTES AND TIMES BETWEEN ORIGIN-DESTINATION (OD) PAIRS

The program calculates routes and generalized cost between OD pairs. The calculated trees and calculated generalized cost are shown on the screen for every node in sequence (Figure 3). On level I the details of tree-builders algorithms are shown, e.g., the difference between the algorithms of Moore [13] and Dijkstra [3]. A fast version of this program working on level III is useful for calculations of larger networks during workshops.

TRIPS BY CAR, DISTRIBUTION, AND SPATIAL DEVELOPMENT

This program has three functions:

1. To calculate the number of trips by car drivers;
2. The distribution of trips;

Roads can be removed by using a zero value for the capacity. Coordinates are used to calculate distances. Distances and velocities are used to calculate travel times. Travel times and distances are used to calculate generalized cost in cost/min.

3. Tripends. Tripends (origins and destinations) can be addressed to nodes. Instead of tripends, population and employment per zone can also be used as input. The number of trips by car drivers can then be calculated as is described in a following section. Balancing procedures are used to equalize the total number of origins and destinations. The tripends can also be determined endogeneously in the distribution model under influence of the transportation network.

4. Save/change. Network and tripends can be saved after changes under a different name.

FIGURE 2. Legend.
3. To calculate the number of trips under influence of the level of service of the transportation system.

Trips of Car Drivers
If in the network program the population is used as input, this part of the program calculates the number of trips made by car drivers. The calculation uses the following equation for the destinations:

\[ A_{i}'n = A_{i}'o \times (\text{cars}'n/\text{cars}'o) \times (\text{trip}'n/\text{trip}'o) \times (\text{phf}'n/\text{phf}'o), \]

where \( A_{i}'n \) is the new and \( A_{i}'o \) is the old calculated number of destinations by trip drivers in zone \( i \). The trips are calculated under the influence of changes, e.g.,
1. in the car density (\( \text{cars}'n/\text{cars}'o \));
2. the trips made by car (\( \text{trip}'n/\text{trip}'o \)); and
3. ratio between the number of trips of a certain period (e.g., peak hour) and the number of trips during a day (\( \text{phf}'n/\text{phf}'o \)).

The calculation of the trips of car drivers from the total population is based on a similar equation.

Distribution
The model used for the distribution is [6]

\[ T(i, j) = \rho \times l(i) \times m(j) \times Q(i) \times X(j) \times F(i, j), \]

\[ \sum_i T(i, j) = O(i) \times l(i) \times (-g), \]

\[ \sum_j T(i, j) = D(j) \times m(j) \times (-h), \]

\[ F(i, j) = \exp(0.4 \times (\log(A \times Z(i, j) + B))^{2}), \]

\[ s = 1/(1 + g) \quad \text{and} \quad r = 1/(1 + g) \]

where

- \( T(i, j) \) is the number of trips between zones \( i \) and \( j \);
- \( \rho \) is the balancing factor for the total number of trips;
- \( l(i) \) and \( m(j) \) are balancing factors for rows and columns;
- \( Q(i) \) and \( X(j) \) are polarities;
- \( O(i) \) is the number of origins in \( i \);
- \( D(j) \) is the number of destinations in \( j \);
- \( Z(i, j) \) are the generalized times for OD pair \( i, j \);
- \( s, r \) are exogeneous coefficients (see below);
- \( g, h \) are endogeneous coefficients.

The deterrence function has a log-normal shape. The coefficients in the deterrence function \( A \) and \( B \) and the interzonal triplelength \( Z(i, j) \) have to be determined by the user.

The quantities \( s \) and \( r \) are elasticities. The values of these quantities determine whether the calculation is carried out with a single constraint model or a double constraint model [17]. If the values of \( s \) and \( r \) are unequal one or zero tripends can be calculated endogenously under the influence of the network (see Figure 4 for an example). This gives the opportunity to show the influence of the transportation network on the land-use development. The exogeneously specified and endogeneously calculated tripends are shown on screen. They can be saved on file for presentation later on. On level I some details of the Gauss–Seidel iteration process in
Exogeneous tripends

Deterrence function : \( F_{ij} = \exp[-0.4 \ln(z(i,j) \times 1.000 \times 1.0) \times 21] \)
elasticity ROW = -3 \& COLUMN = .5

FIGURE 4. Example of a level II forecasting model (see Figure 2 for Legend). The upper figure shows the input in this case in all tripends are equal. The lower figures gives the number of tripends, endogeneously calculated with the model.

the double constraint model are shown. On level III a fast version of the program is available for calculations of larger OD tables.

ASSIGNMENT

Tree assignment programs can be used for:
1. All-or-no assignment;
2. Stochastic assignment;
3. Equilibrium assignment.

On level II the total flow pattern is shown. The screen (flow pattern) can be saved on file for subsequent presentation.

All-or-No Assignment
On level I the traffic flows are shown simultaneously after assignment of each row of the OD matrix.

Stochastic Assignment
In the stochastic assignment [1], the origin–destination matrix will be assigned stepwise. For every step and for every tree calculation, the link times are computed randomly (Figure 5). The number of steps can be determined by the user.

The function to calculate the randomized cost/min on link is:

\[ Z_{\text{rand}}(i, k) = Z(i, k) + \text{Rand} \times A \times \text{SQR}(Z(i, k)) \]

where
- \( Z(i, k) \): the generalized items on link \( i, k \);
- \( \text{SQR}(Z(i, k)) \): the squareroot of \( Z(i, k) \);
- \( Z_{\text{rand}}(i, k) \): the randomized generalized time on link, \( i, k \);
- \( \text{Rand} \): a number randomized from a (pseudo) normal distribution;
A coefficient to influence the variance.

On level I the assignment is shown in detail and is illustrated for calculation of flows between two points in the network.

**Equilibrium Assignment [11, 16]**

In the equilibrium model, the link times are a function of the link loads. The delay function is used to determine the travel times (costs) in a loaded network. The shape of this function is:

\[
T_{\text{flow}}(l, k) = T_{\text{min}}(l, k) \times \left( 1 + A \times \frac{\text{Flow}(l, k)}{\text{Cap}(l, k)} \right)^B,
\]

where:

- \( T_{\text{flow}}(l, k) \) the time on the loaded link;
- \( T_{\text{min}}(l, k) \) the time on the unloaded link;
- \( \text{Flow}(l, k) \) the load on the link;
- \( \text{Cap}(l, k) \) the capacity of the link.

The user can determine \( A \) and \( B \) of this function. The equilibrium is determined by linear approximation [16]. The program gives some network performance measures. Figure 6 shows an example of this assignment method. On level III a fast version is available.

**ESTIMATORS (LEVEL III)**

To support the lectures on the use of nonlinear estimators, which are used in transportation modelling, the following computer programs are available:

1. **Estimating deterrence functions.** A simple example with 4 nodes shows the principle of the algorithm used for the estimation of complete [4, 5] and partial surveyed [12, 14] origin-destination matrices.

2. **Estimating OD matrices using traffic counts.** A simple example with 6 nodes shows the loss of the a priori information by using information minimization [8] or entropy maximization methods [19]. The model also demonstrates the advantage of using elastic instead of fixed constraints [10] to overcome this loss of information.

3. **Analysis of multidimensional tables.** This computer program demonstrates the use of the weighted Poisson estimator [7]. This is a multiplicative likelihood-based estimator. For demonstration, data of the Netherlands National Traffic Survey for the years 1979–1984 are used to estimate: i) car ownership (car/person), ii) car use (car km/person per day), and iii) train use (km/person per day).
Step 1

\[ \text{FLOW}(i) = \text{FLOW}(i') \times (0.181) \times \text{FLOW}(-1) \times (1 - 0.181) \]

Delay function: \[ T_{\text{flow}} = T_{\text{min}} \times (1 + 1.80 \times \text{FLOW} / \text{Cap})^{-4} \]

Step 6

**FIGURE 6.** Example of an equilibrium assignment. The upper figure shows the first step (all-or-nothing) and the lower figure shows the last step. The links with dark shades are more loaded than links with light shades. Notice the influence of the equilibrium assignment on the link loads. (See Figure 2 for legend.)

**RESEARCH (LEVEL IV)**

This computer program has a simple structure and is easy to handle. The computer model has proved to be a good instrument to develop prototypes of new transportation models and to study their feasibility (Figure 7). The [+ ] marked programs are or will be integrated with the level II programs:

- The development of likelihood-based grouping methods [9];
- The calculation of traffic flows and spatial development under influence of changes in the networks and spatial patterns [6];
- Research into the loss of a priori information by using information minimization [18] and entropy maximization methods [10] and how to deal with it;
+ Program for estimating the OD matrices using traffic counts;
+ A dynamic three-dimensional assignment model using the time space [8];
+ Program for optimization of road networks;
- Program for assignment in public transport networks is being developed and will be incorporated in level II;
- The development of a mathematical design method for public transport networks [15];
+ The assignment of different car types on a congested network.

The function of these prototypes is to confront researchers with possibilities and difficulties of different methods (models). Alternative approaches can easily be compared. Potential users learn the advantages of the pro-
posed new methods. Some of the prototypes are still to be developed for more professional systems. They are also partly used in the lectures.

**TUTORIAL EXPERIENCE**

The "Teacher-Friendly Transportation Program/Student Proof Version" (TFTP/SPV 87.5) has been used during a course on interactive graphical transportation information systems and computer programs. This course has been offered for transportation professionals from consulting firms and administrations. The TFTP/SPV version 87.5 has been used successfully during the 50 h lectures of the normal course by the author and in courses for transportation professionals from consulting firms and administrations.

Our experience with the use of the package shows that the understanding of transportation modelling improved. Also the time of learning decreases so there is more time to lecture new developments.

It took about 6 h to learn how to use the program to participants who have had at least a 4-h spreadsheet training. One hour was needed to go through the TFTP/SPV programs step by step guided by a lecturer. The next 4 h had been used for practicing the programs by the participants themselves. It turned out that in the beginning some help was needed in running the programs. However, after 3 h of training it took them about 25 min to calculate an alternative without help.

The program has been used for a better understanding of the professional programs. Most of the professional students also like to use the program for the following reasons:

1. To train their own staff;
2. To show policy makers the working of large programs;
3. To use the program for feasibility studies of setting up calculations on large computers;
4. To use the program for small communities.

**REFERENCES**


