APPLIED OPTIMIZATION

Matti Pursula and Jarkko Niittymäki (Eds.)

MATHEMATICAL METHODS ON OPTIMIZATION IN TRANSPORTATION SYSTEMS

Springer-Science+Business Media, B.V.

Mathematical Methods on Optimization in Transportation Systems

Applied Optimization

Volume 48

Series Editors:

Panos M. Pardalos
University of Florida, U.S.A.

Donald Hearn
University of Florida, U.S.A.

Mathematical Methods on Optimization in Transportation Systems

Edited by

Matti Pursula

and

Jarkko Niittymäki

Helsinki University of Technology, Finland

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

Library of Congress Cataloging-in-Publication Data

Mathematical methods on optimization in transportation systems / edited by Matti Pursula, Jarkko Niittymäki.
p. cm. -- (Applied optimization ; vol. 48)
Selected papers from the 7th EURO Working Group Meeting on Transportation, August 2-4, 1999, Helsinki University of Technology.
ISBN 978-1-4419-4845-8 ISBN 978-1-4757-3357-0 (eBook)
DOI 10.1007/978-1-4757-3357-0
1. Transportation--Planning--Mathematical models--Congresses. 2. Transportation--Management--Mathematical models--Congresses. I. Pursula, Matti. II. Niittymäki, Jarkko. III. EURO Working Group on Transportation. Meeting (7th : 1999 : Helsinki, Finland) IV. Series.

HE147.7 .M38 2001
388--dc21

00-066764

Printed on acid-free paper

Originally published by Kluwer Academic Publishers in 2001
Softcover reprint of the hardcover 1st edition 2001

Contents

Part II GENERAL TRANSPORT MODELS

Preface

This book contains selected papers from the presentations given at the 7^{th} EURO-Working Group Meeting on Transportation, which took place at the Helsinki University of Technology (HUT), Finland, during August 2-4, 1999. Altogether 31 presentations were given and 14 full papers have been selected in this publication through a peer review process coordinated by the editors.

The papers in this book cover a wide range of transportation problems from the simulation of railway traffic to optimum congestion tolling and mode choice modeling with stated preference data. In general, the variety of papers clearly demonstrates the wide areas of interest of people who are involved in the research of transportation systems and their operation. They as well demonstrate the importance and possibilities of modeling and theoretical approaches in the analysis of transportation systems and problem solving.

Most of the papers are purely theoretical in nature, that is, they present a theoretical model with only a hypothetical example of application. There are, however, some papers, which are closer to the practice or describe applications of and give interesting results of studies made by known methodologies. It is especially noteworthy, that half of the accepted papers deal with planning and operation of public transport.

The editors would like to thank the publisher and all the authors and reviewers of the papers, as well as other persons involved in the process of editing the papers and preparing the final publication. It is our hope, that this collection of interesting and timely papers will give inspiration and new ideas to researchers and practitioners in all fields of transportation research and planning.

Espoo, Finland, June 22, 2000

MATTI PURSULA AND JARKKO NIITTYMÄKI

I
PUBLIC TRANSPORT MODELS

Chapter 1

MANAGING AND PREVENTING DELAYS IN RAILWAY TRAFFIC BY SIMULATION AND OPTIMIZATION

Leena Suhl, Taïeb Mellouli, Claus Biederbick, and Johannes Goecke
Decision Support & OR Laboratory, Department of Business Computing, University of Paderborn, Warburger Str. 100, 33098 Paderborn, Germany

Abstract When a disturbance occurs within a railway network, a dispatcher has to decide 'online' about changes in the schedule in order to reduce induced delays and disadvantages for passengers. Computer assistance for dispatchers is needed. In an earlier work, a system architecture for a decision support system for operations control is proposed. This paper concerns the simulation part of this system, needed to manage and prevent delays. Besides operations control, we stress the usefulness of simulation already in the planning phase to prevent delays at operations. Analyzing (types of) disturbances, suitable distributions of delays are generated, and simulation is used to test the robustness of the timetable against disturbances. As a test vehicle, a computer-based environment configured for German Rail's network has been developed. Robustness depends on dimensions of conflicts and of passengers involved. Conflicts and their causes directly depend on the 'waiting time rules' in use. By a simulation study, the quality of these rules can be evaluated and corrected. Preventing delays may be achieved by a better planning, too. Special optimization models can be used to increase buffer times without need of extra resources.

Keywords: operations control, railways, preventing delays, simulation, optimization

1. INTRODUCTION

Providers of public transportation systems, such as railways, have to pay special attention to their acceptance by passengers as they are the main source of revenue. This acceptance is positively influenced by providing high quality trains/trips and passenger connections, but negatively influenced by

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 3–16.

delays and missed connections at operations. Because scheduled traffic is always subject to external factors, many types of disturbances cannot be avoided which make short-term changes to a given schedule necessary. When a disturbance such as a technical defect, accumulated vehicle lateness, missing crew members, or congestion, occurs, a dispatcher has to react within a few minutes, and decide about changes in the schedule, such as delaying connecting trains and reallocation of resources (vehicles and crew). In rail and air traffic, it often takes several days to reconstruct the planned schedule or a consistent one after ad hoc changes.

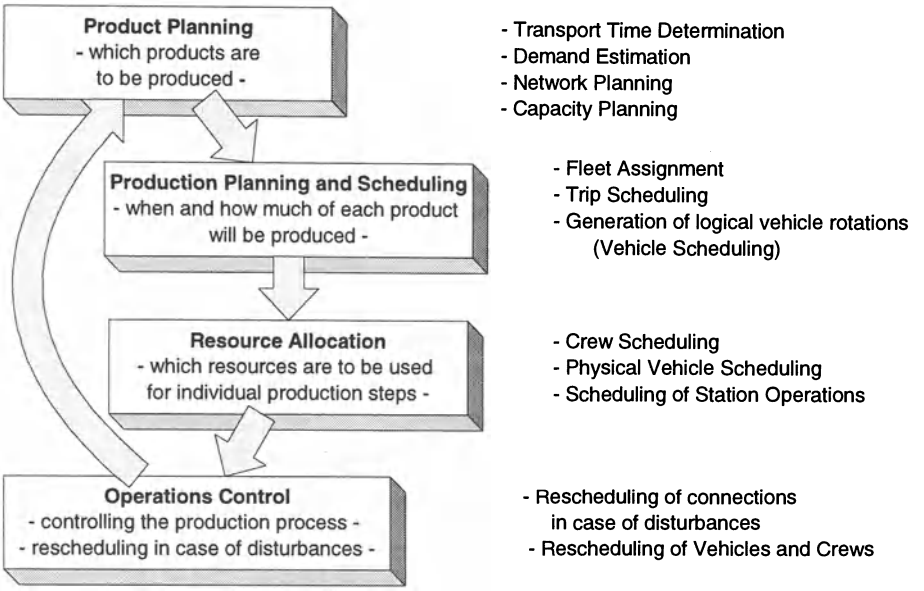

Fig. **1** Phases in Production Planning, Scheduling, and Control

We elaborate on methods to manage and prevent delays in public transportation to ensure passenger traffic with the best possible quality. At first glance, this task appears to be only related to operations control. This corresponds to the last phase of a cycle of product planning, scheduling, and control for public transport companies (see Fig. 1). Thus, the main goal is to provide decision support tools for dispatchers to manage the huge amount of data related to the actual state of the schedule and to carry out what-if analyses for their decisions. In Suhl and Mellouli (1999) we studied the requirements for an operations control system for railways based on a case study of German Rail and designed a knowledge-based object-oriented system architecture that includes problem solving components for automatically propagating delays, detecting conflicts, and simulating net-wide effects of delays for connecting trains enforced by dispatchers.

In this paper, however, we wish to stress the importance of planning quality to *prevent* delays. That is, we elaborate on techniques that allow us to draw conclusions along the return arc in Fig. 1 from operations control phase (or by utilizing an operations control system) towards the planning phases.

2. THE IDEA

First, we observe that rescheduling experience and what-if analyses of dispatchers help in detecting sensitive connections in the planned timetable/ schedule. This information is very valuable in order to improve the planning of passenger connections in the future. After having developed a computer-based prototype for operations control in our institute, this observation led us to the following idea:

Using computer simulation, one may get much of this valuable information from dispatchers to improve passenger connections already in the planning phase, before starting operations. We analyze disturbances and their types and generate delays of different extent within the network using well-known and/or self-developed distributions, and then use the simulation component of the operations control system to test the robustness of the timetable/schedule against disturbances. We discuss the utility of such an approach based on simulations with different distributions of delays using a computer-based test environment configured for a part of the German Rail network which incorporates 30,000 trips daily.

Type		**Remarks**
ICE	Inter City Express	Long distance train between major cities
IC/EC	Inter City/Euro City	Long/middle distance between bigger cities
IR	Inter Regio	Regional long distance train
SE	Stadt (city) Express	Short distance city connector
SB	Stadtbahn	City rail
Other (including D, RB, RE)		Other middle and short distance trains

Tab. **1** Train Types of German Rail

Here, an important criterion of robustness is the number of conflicts induced and number of passengers involved. A conflict is defined as a missed passenger connection, assuming that waiting times of connecting trains are regulated by the *regular waiting time rules* of German Rail. These rules specify a maximum time a connecting train has to wait depending on the types of connecting trains without need of dispatcher's analysis. At the time of operation a conflict is handled by the dispatcher who decides whether a train waits violating the regular waiting time rules.

These rules consider *minimum transit times* (*MTT*) needed by passengers to proceed from one railway platform to another within a station, as well as

regular waiting times (*RWT*) depending on the types of connecting trains (cf. Table 1). Global minimum transit times are set at 2 (or 3) minutes, if the connecting train leaves on the same platform (or that on the opposite side), and to at least 5 minutes otherwise (since passengers have to change platform by using a tunnel). Examples of global regular waiting times are: SE-trains for local traffic wait for IC/EC-trains up to 10 minutes but IC/EC-trains do not wait for SE-trains. The ICE-, IC/EC-, and IR-trains wait for each other up to 5 minutes. SE-trains and other local trains also wait for each other up to 5 minutes. Local exceptions of these rules are given in a book.

The main aim of our simulation study is to evaluate the quality of regular waiting time rules, to suggest global or local corrections of these rules, and to develop simulation-based methods for these corrections to reach a better robustness against delays.

3. SIMULATION AS A SOLUTION APPROACH

Computer-based simulation has gained more and more importance for many planning problems arising in the logistics and transportation area, especially for scheduling and dispatching problems (cf., e.g., Manivannan (1998)). Its ability to handle even very complex systems makes it well-suited for this special purpose network-based applications. Thus, we propose the use of simulation not only for visualization and real-time monitoring of transportation systems, but also for analyzing and–with a growing planning-knowledge–*preventing* delays before execution of a given timetable.

If it is known, which connections or parts in a timetable are sensitive against disturbances or unstable, either the timetable or rules influencing its execution could be improved in future rearrangements. Experienced dispatchers usually have an excellent, but often implicit, knowledge about such instabilities. Unfortunately, this knowledge is difficult to extract and could be revealed only through extensive examination of dispatchers' decisions, if at all. The intelligent employment of simulation methodology and technologies offers many possibilities to gather much of the desirable information needed for quality planning. Some examples of questions to be answered by our simulation study are:

- Are there typical delay profiles for some trains which could be changed by, e.g., varying buffer times or waiting time rules?
- Are there significant correlations between delay times of two trains which are directly or not directly connected? How can these correlations be formalized and represented?
- What are appropriate statistical distributions for delay times of trains in special stations and what are good estimates for their parameters?
- What are the net-wide effects of a generated disturbance if all global (and local) waiting time rules for trains are applied?

- How can these net-wide effects be measured in delay/waiting times for passengers?
- Could the total delay time of all passengers be decreased by using a different set of waiting time rules (what-if-analysis)? How do delay profiles of trains vary then?

Studying the effects of a generated disturbance, we took the total delay time for passengers as a central measure as it directly influences customers' behavior. It can be roughly divided into two parts: the total of delay times affected at the moment of analysis and the total of delay times induced by delaying connecting trains either using waiting time rules or by dispatcher's decisions in case of induced conflicts. Therefore, when we consider induced conflicts in other stations and try to minimize their number, this will indirectly decrease passengers' total delay.

To study all the problems outlined above, it is necessary to represent the network itself and the events occurring correctly. Therefore, the simulation model had to be validated precisely using a–in a statistical sense–sufficient amount of data from the real system, e.g., a log-file from the running system containing operating data. To perform this task, we used a data recording of a dense part of German Rail's network.

4. SIMULATION SYSTEM – PRINCIPLES AND CHARACTERISTICS

The simulation system itself is event-oriented and distributed, i.e., one computer processes the timetable, one handles simulation time, etc. This functional fragmentation was necessary because of the enormous complexity of the system and eases the use of low-cost hardware (2-3 Intel Pentium/AMD Athlon processors at approx. 500 MHz). The basic system consists of two components. The first concerns analysis of delays and pre-processing for simulation. By analyzing the data set from German Rail, plausible assumptions about the system's behavior are determined, for instance, possible consequences of expert decisions, propagation of delays, and numbers of passengers. Moreover, the preliminary simulation results were extensively examined in order to generate consistent delay profiles of trains, possible correlations, and distributions of delays.

Pure simulations are carried out by the second component of the system. For implementation of this part, the kernel of a prototype developed in Goecke (1996) and Stelbrink (1998) is used. This prototype, implemented in Visual C++ and JAVA, already includes simplified mechanisms for propagating delays, for recognizing conflict situations, and for performing what-if analyses to support estimating net-wide effects of a dispatcher's decision. The underlying object-oriented data management component contains a complete timetable of the German Rail. It is able to manage the

scheduled, actual, and expected state for all trips at all their intermediate stations, as well as minimum delays enforced by expert decisions. In addition to this, special simulation tools, e.g., for generating distributions or managing simulation time, were realized. Currently, the simulation prototype is developed further towards incorporating more realistic assumptions.

The described components are part of the decision support tool architecture for operations control presented by Suhl and Mellouli (1999). The system developed in this study combines the simulation and the statistics module in the problem solving components area: The simulation module is able to calculate the propagation of a given delay and, therefore, to determine its net-wide effects. Propagation algorithms make use of the waiting time rules and assumptions about delay distributions as well as the resulting delay profiles of trains. The statistics module collects and analyses information about disturbances and dispatcher's decisions at operations. Based on this, it generates multiple types of disturbances in the net and passes them to the simulation module. Simulation results are then passed again to the statistics module. Both modules generate large amounts of data to be evaluated after a simulation in a what-if-analysis. To ease data handling we developed a first version of an underlying data warehouse component which should perform most of the regular calculations which are necessary to analyze–or just to manage–the current and historical data, as well as to automatically check the validity of statistical parameters in the future.

5. A NOTE ON OPTIMIZATION

Before giving details of the simulation study, we stress the impact of appropriate optimization tools to prevent delays. By collecting suggestions of timetable corrections from the simulation study, optimization models can be developed for timetable redesign. The goal is to minimize total delay/waiting time for passengers while maintaining the quality of a schedule (frequency of service, good connections).

Further, using appropriate optimization models in planning phases may prevent delays at the time of operation. A main task is to increase buffer times between trips in vehicle rotations without need of additional trains. This is important, since delays may enforce rescheduling of vehicle rotations when buffer times are small. Here, first-in first-out (or FIFO) strategy at a station level delivers a good distribution of buffer times. Using optimization models based on time-space networks with connection lines for stations, it is possible to make use of this strategy for hard extensions, as well. In Mellouli (1998), a state-expanded aggregated time-space network is presented to handle maintenance requirements for IC/EC-trains of German Rail. The resulting mixed-integer models were solved for large-scale real-life problem instances. A computed optimal flow on the time-space network is viewed as a class of aggregated schedules. It is shown, that a FIFO strategy can be

applied here within a flow decomposition algorithm in a second step to extract an optimal schedule out of this flow. Here, the dispatching by the FIFO strategy is performed on connection lines at a *station-state* (maintenance state) level. This provides as much buffer time as possible within train rotations while sustaining optimality and maintenance requirements.

6. UNDERLYING ASSUMPTIONS IN THE SIMULATION

Definitions. Let W_P be the accumulated waiting time of all passengers induced by delays, $B := \{b_1,...,b_m\}$ the set of all stations, and $F := \{f_1,...,f_n\}$ the set of all trips to be served. Both sets are determined by a given schedule. In order to simplify the simulation, both the set of stations and the set of trips are divided into subsets of (1) large, (2) middle, and (3) small stations, and (1) ICE train, (2) EC/IC/IR train, (3) D/SE/RB/RE train, (4) SB train, and (5) other trains, respectively. In the study, all types of stations and the first three major classes of trains are regarded. (Let K be a function which assigns the indices in brackets to the different types of trains and stations.)

A single trip f_i is an ordered set of arrival and departure events:

$$f_i := [(b_{i1}, t_{i1}^{arr}, t_{i1}^{dep}, N_{i1}, v_{i1}^{arr}, v_{i1}^{dep}); (b_{i2}, t_{i2}^{arr}, t_{i2}^{dep}, N_{i2}, v_{i2}^{arr}, v_{i2}^{dep}); ...;$$
$$(b_{ik}, t_{ik}^{arr}, t_{ik}^{dep}, N_{ik-1}, v_{ik}^{arr}, v_{ik}^{dep}); ...; (b_{ik_i}, t_{ir_i}^{arr}, t_{ir_i}^{dep}, N_{ik_i-1}, v_{ir_i}^{arr}, v_{ir_i}^{dep})],$$

where $b_{il} \in B$ denotes the l-th station of trip i, $t_{il}^{arr/dep}$ the scheduled time of arrival/departure in station b_{il}, N_{il} the number of passengers in the train at the departure in station l and at the arrival in l+1, and $v_{il}^{arr/dep}$ the delay at arrival/departure in b_{il}.

Passenger Propagation. One of the main difficulties of the study–caused by a lack of passenger data from the real system–was determining consistent assumptions for the passenger propagation.

For simplicity, we initialize each trip-type with a constant number of passengers N_{i1}, e.g., an IC-train is initialized with 150 + z passengers, where z is a normally distributed random variable. After that, some passengers a_{il} will *leave* the train i at each station b_{il}, some (u_{il}^r) will *change* to connecting trains r, and some will *get into* train i. All the necessary (deterministic) calculations are based on simple functions which take size and reach of trains and the size of stations they meet at into account.

In case of passengers leaving the system, their delay has to be added to the accumulated waiting time, i.e., $W_P = W_P + a_{il} v_{il}^{arr}$. It has to be

increased, too, if u_{il}^{r} passengers cannot reach their connection f_r caused by a dispatchers' decision. We assume that passengers will wait for the next trip f_r on this line rather then search for an alternative connection, if it is near enough in time (e.g., less than M minutes in the future). Nevertheless, if the time to wait is too long, they will take other trains on different routes (we suppose that there *is* one) and arrive at their destinations before f_r arrives there. We use a simple estimate to model this circumstance. At last, the number of passengers entering a train equals the sum of passengers leaving it and those changing to another train. This number is again varied by a normally distributed random variable.

Delay Propagation. Every train can lose or regain time during a trip. To model this, we define normally distributed random variables for each type of train reflecting the relative change of delay on a trip between two stations. Extensive analysis of data from the real system led to this assumption.

Disturbance Generation. In reality, a "disturbance" of a train could have many causes and result in various actions for a dispatcher (cf. Introduction). For instance, bad weather conditions imply little delays for all trains in a whole region, while a closed track just influences some trains which must be rerouted, afterwards having a "big" delay.

In the current preliminary study, the type (the reason) or topology of disturbances is not considered, although this implies slightly unrealistic behavior of the system. All possible disturbances are assumed to result in delays of the concerned trains. Because these delays will propagate from one train to some of its connectors, a single delay could have net-wide effects. Therefore, ceteris-paribus-studies can be conducted to compare different dispatching scenarios which are described below.

Artificial disturbances in a simulation run can be generated automatically or manually. We used two exponentially distributed random variables, one for the inter-arrival times for disturbances and one for their delay length. We also carried out experiments with uniform and triangular distributions for the length of delays. The train to be delayed was chosen randomly from all trains or only from the driving/standing ones, i.e., dense parts (tracks and stations) of the network of course get more delays than sparse parts. Additionally, train-independent delays are introduced at each station to model congestion on outgoing tracks.

In future studies, we plan to use historical data of German Rail to determine a more realistic distribution. Furthermore, a typology (and topology) of delays will be integrated in the simulation model to facilitate and conduct more accurate statistical examinations of disturbances.

7. WAITING TIME RULES AND DISPATCHING

At operations of German Rail, one of the most difficult tasks is to decide, whether a train should wait for late connectors or not. To simplify this, waiting time rules are introduced at German Rail based on practical experience (cf. second section for some examples). Whenever these rules are violated, a conflict occurs and the dispatcher has to decide. In the system, both waiting time rules and dispatchers' decisions are modelled.

Regular Waiting Times, *RWT*, and **Minimum Transit Times, *MTT*.** *RWT* parameters are given by matrix $(RWT_{K(f_i),K(f_j)}) =: RWT \in R^{3x3}$ for each combination of trains. *MTT* is defined globally for the entire network. *RWT* and *MTT* have local exceptions at some stations and for some connections. *RWT** and *MTT** denote *RWT* and *MTT*, respectively, taking into account these local exceptions.

Global Waiting Time Rules and Conflict Determination. Let f_i be the incoming train, f_j the connection, and b_k the meeting-station. If t_i is the scheduled arrival of f_i in b_k and t_j the scheduled departure of f_j in b_k, then the following waiting time rules are used to determine conflicts which have to be handled by dispatchers:

1. If $t_i + v_i + MTT^* \le t_j + v_j$, everything is ok.
2. If $t_i + v_i + MTT^* \le t_j + RWT^*_{K(f_i),K(f_j)}$, the connecting train has to wait up to $RWT^*_{K(f_j),K(f_i)}$ minutes
3. If $t_i + v_i + MTT^* > t_j + RWT^*_{K(f_i),K(f_j)}$, a conflict message is generated.

Strategies for Dispatchers' Decisions. In the simulation system the computer instead of the dispatcher has to decide, whether a connecting train will wait or not. Since there is no fixed set of rules a dispatcher follows, we are not able to model dispatcher behavior correctly in all cases. Dispatchers consider many fuzzy parameters in parallel, e.g. passenger requests submitted through the conductor, service rate on this trip, number of connecting trains in the following stations, track occupancy, etc.

For the simulation we had to simplify this complex task. We tested different dispatching strategies by using fast heuristics.

The first idea led us to the exploitation of the estimated number of directly induced conflicts at the following stations in case of waiting and the number of passengers trying to reach the connection (e.g., calculated as described above). Because this estimation implies some additional assumptions which are

difficult to determine, we used a second possibility for testing the behavior, the distribution, and the consequences of artificial disturbances in the network: To get a simple classification of conflicts, we slightly modified the conflict determination rule (3) by a parameter λ :

If $t_j + RWT^*_{K(f_i),K(f_j)} < t_i + v_i + MTT^* \leq t_j + RWT^*_{K(f_i),K(f_j)} + \lambda$, the system encounters a *minor conflict*, otherwise, i.e.,

$t_i + v_i + MTT^* > t_j + RWT^*_{K(f_i),K(f_j)} + \lambda$, a *major conflict* occurs.

With this definitions, we considered four different scenarios:

A) A train will not wait for incoming trains at all.
B) A connecting train will wait until all its regular feeder trains arrives.
C) A connecting train will wait for its regular feeder trains, but only until regular waiting times are reached.
D) A connecting train will wait for its regular feeder trains within the regular waiting times and in case of minor conflicts.

Scenarios A and B reflect extreme dispatching strategies (no waiting at all, always wait), scenarios C and D are intermediate strategies.

Two classes of trains are distinguished in the simulation: Trains of the first class comprising types (1) and (2), mentioned above, wait for each other for up to 5 min and do not wait for trains of the second class (containing all other types), which wait for each other for up to 5 min. Thus, *RWT* reduces to 2x2-matrix, which is set to (5, 0; 10, 5). This corresponds to the global regular waiting times currently in use at German Rail. In the study, scenario D helps in drawing some conclusions about the effect of small changes of these regular waiting times. Logging all conflict situations occurring after propagating artificial disturbances helps in conducting various statements to local and regional instabilities in the schedule. In this paper, we concentrate on conflict behavior and propagation in the railway network. Preliminary results from the above scenarios are stated in the next section.

8. PRELIMINARY RESULTS

Tables 2 and 3 show the results of a series of simulation runs under the above assumptions. We conducted 30 runs for each simulation experiment. For each run over approximately one and a half days, 30-40 disturbances are distributed in the average with an expected delay between 3 and 7 minutes. For each of the scenarios, the same disturbances are distributed for λ =2;3;5.

For the extreme scenarios A and B, Table 2 shows the effect of propagating the disturbances generated by the distribution. Since connecting trains do not wait for incoming trains at all for scenario A, the number of (minor and major) conflicts simply reflects the number of delays induced at subsequent stations of those lines where the disturbances are generated (one

level). This is the first extreme, the second (strategy B) shows the dimension of propagation of delays and thus of conflicts throughout the network over several levels. This is because trains always wait in scenario B.

The number of conflicts per generated disturbance (shown in the second column of each conflict type) can be taken as a unifying normative indicator. Thus, for $\lambda = 2$ a disturbance induces in the average 1.395+0.067 delays on the first level (concerning connections on the same line) and 106.09+0.3958 delays over all levels within the considered trains of German Rail's network. The first number in the summations indicates the number of major delays.

Scenario	λ	MAJOR CONFLICTS			MINOR CONFLICTS		
		Total #	Confidence Interval: α=0.05	Average per Disturbance	Total #	Confidence Interval: α=0.05	Average per Disturbance
A	2	50.03	[48.05; 52.02]	1.3950	2.40	[1.85; 2.95]	0.0669
	3	47.97	[46.05; 49.88]	1.3374	4.47	[3.79; 5.14]	0.1245
	5	44.57	[43.08; 46.05]	1.2426	7.87	[6.90; 8.83]	0.2193
B	2	3681	[3474; 3889]	106.09	13.73	[7.87; 19.60]	0.3958
	3	3675	[3468; 3881]	105.90	20.47	[13.08; 27.85]	0.5898
	5	3655	[3449; 3860]	105.32	40.27	[29.74; 50.79]	1.1604

Tab. **2** Results for extreme dispatching strategies A and B

An interesting conclusion from Table 2 (scenario B) is that, when trains always wait, both the *number* and the *length* of delays increase drastically. This is indicated by the number of major delays which increases by a factor of 105-106 (for λ-values of 2, 3, and 5). This factor varies only slightly in the simulation, for instance, between 100 and 112 for a confidence of $\alpha = 0{,}05$ in the case of $\lambda = 2$. This shows the necessity of the waiting time rules and of dispatching decisions which have to reduce the amount of overall delays and at the same time minimize missed connections for passengers.

Scenario	λ	MAJOR CONFLICTS			MINOR CONFLICTS		
		Total #	Confidence Interval: α=0.05	Average per Disturbance	Total #	Confidence Interval: α=0.05	Average per Disturbance
C	2	52.50	[49.16; 55.84]	1.4100	3.70	[2.75; 4.65]	0.0994
	3	50.07	[47.05; 53.09]	1.3447	6.13	[4.59; 7.67]	0.1647
	5	46.33	[43.44; 49.22]	1.2444	9.87	[7.94; 11.79]	0.2650
D	2	141.4	[56.88; 225.85]	3.9488	126.6	[0.00; 280.3]	3.5363
	3	173.1	[17.88; 328.3]	4.8388	242.2	[0.00; 484.5]	6.7717
	5	1309	[783.2; 6374]	34.904	3266.	[1818; 4713]	87.083

Tab. **3** Results for intermediate dispatching strategies C and D

Now, it is interesting to see the impact of waiting time rules for reducing the propagation of delays within the net. Table 3 shows results for scenarios C and D for different values of λ. Recall, that scenario C corresponds to dispatching decisions solely by the waiting time rules, that is, trains do not wait in case of conflicts. For scenario D, trains also wait in case of minor conflicts, this is equivalent to relaxing the waiting time rules by adding λ minutes to the maximum waiting times given by the *RWT* matrix.

The results for scenario C in Table 3 show that the waiting time rules currently in use by German Rail only slightly increase the numbers of major and minor conflicts relatively to the scenario A where trains do not wait all. Because of a dense network, a slight increase of waiting times by 2 or 3 minutes (equivalent to the scenario D with λ = 2, 3) approximately triples numbers of major conflicts. (Minor conflicts are then not considered as conflicts since trains wait in case of minor conflicts in scenario D.)

The "astronomical" increase in number of major conflicts for scenario D, λ = 5, shows the dimension of induced delays, if trains wait 5 minutes more than the regular waiting times: From λ = 3 to 5 (*only* 2 min), the number of major conflicts increases by more than *seven* times. This already corresponds to more than one *third* of the conflict dimension of the extreme strategy B. This is because, a single major delay at the morning may propagate for the whole day and cause many major delays at other stations. Thus, an increase of regular waiting times may only be acceptable at the end of the day for latest connections, or only locally if the local network is not dense.

Therefore, it appears convenient to rather decrease the regular waiting times globally by a few minutes and perhaps increase these waiting times only locally in order to reduce missed connection at certain stations. To found conclusions in this direction, we have to extend simulation results by local considerations and to integrate passenger waiting times.

9. FUTURE WORK

After generating first preliminary results, future studies have to be conducted in order to localize–both spatially and temporally–instabilities of the timetable. This information should be taken into account in forthcoming planning periods. This means that, as an example, departure times and buffer times of localized trips have to be adjusted. Besides timetable redesign, simulation results may suggest local corrections of regular waiting times:

In our preliminary study, we observed that in a frequently served rail line, like Düsseldorf → Essen → Dortmund, accumulated short delays may rapidly cause conflict situations. Since the frequency of service is high, it is convenient to *decrease* regular waiting times *locally* for trains on such a line. This lets some passengers wait at that moment for the next trains, but prevents the occurrence of conflicts that would cause more waiting time for other passengers. For connecting trains into a rural district, however, the situation is opposite. The frequency of service is low, such as every 60 or 120 minutes, and passengers wishing to reach the connecting train may often arrive late owing to delays of trains of the frequently served line. In this case, either the departure time of such connecting trains may be corrected or the regular waiting times may be *increased locally*.

In conclusion our results imply that a station-dependent definition of waiting time rules may significantly improve the railway system performance from the passenger point of view. Further, bottlenecks shown by the simulation should be considered in the planning and scheduling phase of the next period, in order to indicate an improved level of passenger service.

It is worth mentioning that some researchers make use of a (max,+)-algebra to model the relationship between a connecting train and its feeder trains, cf. Bacceli et al. (1992). "+" is used for calculating the arrival time at a station out of the starting time at its preceding station of the same line and "max" refers to the fact that the earliest departure time of a connecting train is equal to the maximum of the arrival times of its feeder trains (actually including minimum transit times). Without delays, the (max,+)-algebra is used to model synchronization within timetable design. Egmond (1998) used this algebra to propagate delays. Here, the scenario B is assumed as a dispatching strategy, that is, connecting trains always wait for feeder trains. In our opinion, this is an alternative way to accomplish simulation under simplified requirements. A nice feature is to compute the maximum delay at a station j that does not reach station i for all pairs of stations (for scenario B) in a pre-processing step. This information may be useful for dispatchers.

Our present work, however, makes use of different dispatching strategies using waiting time rules and, to an increased extent, considering passenger flows and waiting times according to reliable data in our future works. Thus, simulation seems to be adequate for handling increasing complexity of real life constraints and various dispatching strategies within dense, large and

heterogeneous transportation networks, such as that of the German Rail. Optimization can be very helpful here in modeling sub-problems related to finding better dispatching strategies.

REFERENCES

Bacceli, F., Cohen, G., Olsder, G.J., and Quadrat, J.P. (1992). Synchronization and linearity – An algebra for discrete event systems. John Wiley & Sons, Chichester.

Egmond, R.-J. van. Propagation of delays in public transport. Presented at the 6th Meeting of the EURO Working Group on Transportation. September 9-11, 1998. Gothenburg.

Goecke, J. (1996). Entwicklung eines graphisch-interaktiven Systems zur Unterstützung der netzweiten Konfliktlösung bei Zugverspätungen der Deutschen Bahn AG. Diploma thesis. Decision Support & OR Lab. University of Paderborn.

Manivannan, M.S. (1998). Simulation of Logistics and Transportation Systems. In Handbook of Simulation – Principles, Methodology, Advances, Applications, and Practice (Banks J.), pp. 571-604. Wiley. New York.

Mellouli, T. (1998). Periodic Maintenance Routing of German Rail's IC/EC Trains by a Flow Model based on a State-Expanded Time-Space Network. Presented at the 6th Meeting of the EURO Working Group on Transportation. September 9-11, 1998. Gothenburg.

Stelbrink, M. (1998). Konzeption und prototypische Implementierung eines verteilten, echtzeit-basierten Kundeninformationssystems bei der Deutschen Bahn AG unter Verwendung von Intranet-Technologie. Diploma thesis. Decision Support & OR Lab. University of Paderborn.

Suhl, L. and Mellouli, T. (1999). Requirements for, and Design of, an Operations Control System for Railways. In Computer-Aided Transit Scheduling (Wilson N.H.M.), LNEMS, Vol. 471, pp. 371-390. Springer. Berlin – Heidelberg.

Chapter 2

HEURISTICS FOR SCHEDULING BUSES AND DRIVERS FOR AN EX-URBAN PUBLIC TRANSPORT COMPANY WITH BUS-DRIVER DEPENDENCIES

Taïeb Mellouli and Leena Suhl
Decision Support & OR Laboratory, Department of Business Computing, University of Paderborn, Warburger Str. 100, 33098 Paderborn, Germany
mellouli|suhl@uni-paderborn.de

Abstract This paper deals with the scheduling of buses and drivers for *ex-urban* service with *bus-driver dependencies*—a corporate rule influencing solution structure and methodology. An integration of crew rostering into a combined vehicle and crew scheduling process is proposed. A scheme letting five drivers *share* two buses is developed: The trips served by two buses during the week are partitioned into five rosters for drivers, who *rotate* by interchanging their rosters weekly. Heuristic components are presented which construct sets of rotations in a multiple-depot problem setting, where each rotation consists of one or two feasible driver shifts. These components are then combined to solve the problem. A *two-phased* strategy computes rotations for workdays and weekend, then groups them into weekly schedules for drivers using the two-bus-five-driver scheme. A *best-group-first* strategy rather merges both phases: Each iteration builds 'principle rotation parts' of the 'best' weekly schedule of a driver group with the same home depot. The rotations are then completed in several passes. Computational results are presented and discussed.

Keywords: bus transit, ex-urban service, integrated vehicle and crew scheduling/rostering

1. Introduction

The construction of schedules for buses (vehicles) and drivers (crews) constitutes a main planning task for bus transit. Approaches to solve this task are primarily influenced by the type of transport, and further by the corporate rules of certain companies. *Urban service* providing public bus transit within

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 17–35.

a city is one of the most studied types in literature. It is characterized by a single depot and short distances between stations. This offers several degrees of freedom for linking timetable trips with deadhead trips and for driver transfer from bus to bus. The situation differs for long-distance bus transit as it is the case for *ex-urban* (or *extra-urban*) *service* which connects villages and minor towns with each other and with city centers as well as with some train stations. Ex-urban service is generally carried out from several depots and differs from *sub-urban service* which links a center of a city to its suburbs. Having trips of longer distance in the ex-urban case, deadhead trips are not freely allowed, and drivers are dependent on the bus schedules (rotations) to reach the stations and depots of possible driver relieving or those where they started their daily work (and left their cars). Further, corporate rules or the wishes of certain public transport companies may impose special properties to vehicle and crew schedules, which necessitate dramatic changes of existing approaches in order to solve the scheduling problems automatically. In this paper, a requirement of this type is examined—a bus-driver dependency appearing very useful in the context of ex-urban public transport.

The paper is organized as follows. An overview of the transit network of our case study, emphasizing its size and ex-urban characteristics, is given, followed by a second overview of terminology and approaches in the field of vehicle and crew scheduling as well as crew rostering. After stressing the differences between the urban and the ex-urban case for which an integration of daily vehicle and crew scheduling is recommended in the literature, the bus-driver dependency is motivated and studied. We propose the integration of crew rostering as well, develop and evaluate a rostering scheme together with heuristic procedures, able to schedule buses and drivers over several months for the case of ex-urban service under these requirements.

2. CASE STUDY: OVERVIEW OF TRANSIT NETWORK

Our methods have been demonstrated on a case study that stems from the BVO (BusVerkehr Ostwestfalen GmbH), an ex-urban public bus transit company of Eastern Westphalia, one of the nearly 20 such companies in Germany. In Eastern Westphalia, there are two kinds of public transport by bus: Urban service carried out by local carriers and ex-urban service carried out by BVO. The BVO-network is managed by different offices: Bielefeld, Paderborn, Lübbecke, and Detmold. Each office has its own drivers and buses and covers a multiple-depot region. The region managed by the Paderborn office comprises the districts of Paderborn and Höxter with 6 bus depots currently. An additional BVO office in Brakel, in the Höxter district, offers only customers' service.

For bus transit a timetable is generally given by a set of lines and line schedules. The term "line" identified by a number, e.g. 360, and a short course description, e.g., Paderborn → Wewelsburg → Flughafen (Airport) → Büren, actually refers to a pair of lines, the second line being the corresponding return line. For instance, line number 360 comprises a line from Paderborn to Büren serving 36 stations and the return line from Büren to Paderborn. Each line's schedule gives departure and arrival times for its trips which are (one-way) traversals of this line. The terminal stations of trips may differ for the same line, e.g., for line 360 some trips ends in "Büren, Markt" and others in "Büren, Barkhäuser Straße." A line's schedule comprises three sectors—for Monday-Friday, for Saturday, and for Sunday/Holidays. Except for Sunday/Holidays, the timetables differ for schooldays and school vacations. For the Paderborn-Höxter districts with a mostly rural nature, transit of school students constitutes a main part of service.

The Paderborn office manages nearly 50 pairs of lines, which imply 829 trips daily for schooldays (554 for vacations, 324 for Saturday, and 66 for Sunday). At the moment, 38 buses (4 of which are reserve) and 75 drivers, operating from several depots spread over a large geographic area, are scheduled by the Paderborn office to cover a considerable part of these trips. The remaining trips are outsourced to private transport companies.

The Paderborn BVO office co-operates with local transport companies which offer urban service or have the "right" of some lines in order to optimize the transit network in the districts of Paderborn and Höxter. Lines connecting city centers with suburbs, which sometimes constitute a third type of service (sub-urban service), are managed either by a local transport company for urban service or by BVO for ex-urban service.

3. VEHICLE SCHEDULING – CREW SCHEDULING – CREW ROSTERING

After designing a timetable, resources for serving its trips have to be assigned. *Vehicle scheduling* is the task of assigning these trips to vehicles. For bus transit, a vehicle schedule is a set of feasible daily rotations for buses starting and ending at depots. Each *rotation* consists of a sequence of scheduled trips which can be served consecutively by the same bus (Fig. 1). Every two consecutive trips i and j of a rotation must be *compatible*, that is, $\textit{end-time}(i) + \textit{deadhead-trip-duration}(\textit{destination}(i), \textit{origin}(j)) \leq \textit{start-time}(j)$, where $\textit{end-time}(i) := \textit{arrival-time}(i) + \textit{turn-time}(\textit{destination}(i))$.

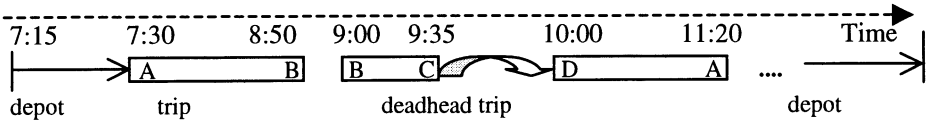

Fig. 1 Representing a bus rotation

The basic vehicle scheduling problem for bus transit with a single depot (minimizing the number of needed vehicles and costs of used deadhead trips) can be solved efficiently as it can be modeled as a minimum-cost network flow problem. The more difficult multiple-depot problem version requires that each vehicle must return to the starting depot and can be modeled as a multi-commodity flow problem. Among others, Löbel (1997) investigated this type of problem and solved large-scale real-life problems. A review on vehicle scheduling for bus transit is given in Daduna and Paixão (1995).

In the general case a driver may work with several buses during the day and change from one bus to another one at certain stations or depots called *relief points*. Parts of trips between two relief points are called driver trips or *d-trips*. (Cf. Fig. 2.) A *piece* (of work) for a driver is a sequence of d-trips from the same rotation. A *run* is a daily work schedule for a driver and consists of a piece or of a sequence of pieces on several buses, fulfilling all working time regulations. *Crew scheduling* for bus transit constructs runs for drivers, covering trips of the timetable. Clearly, crew scheduling is dependent on vehicle scheduling. The inserted deadhead trips for buses between scheduled trips or from and to a depot must be accomplished by drivers. The *changeovers* (changing buses during a driver run) are to be minimized and scheduled carefully; preferably with buffer times for drivers to avoid rescheduling at operations.

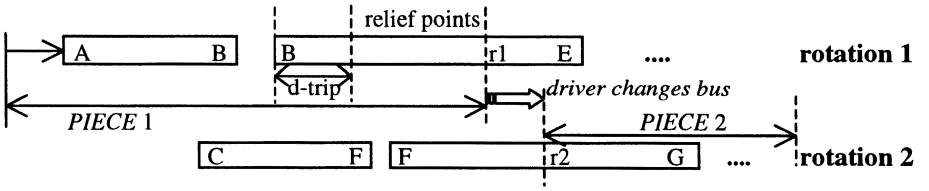

Fig. 2 Building runs for drivers (out of pieces from different rotations)

There are several ways to construct bus rotations, to decompose bus rotations into *pieces* of work for drivers, and to build daily driver schedules (*runs*) out of these pieces. Owing to the combinatorial complexity of the joint vehicle and crew scheduling problem, most solution approaches in practical use are sequential—scheduling vehicles prior to drivers.

Crew scheduling constitutes in and of itself a difficult problem. An overview on methods and systems for driver scheduling for urban bus transit can be found in Wren and Rousseau (1995). Only a few attempts are known to solve the joint problem. Freling et al. (1999) provided an overview of models and techniques for integrating vehicle and crew scheduling, mainly for the single-depot case (cf. next section for the multiple-depot case).

In addition to the construction of daily schedules for buses and drivers, a public transport company has to assign a working plan to each driver over

several weeks of the planning horizon (*crew rostering*). Such a plan consists of a sequence of weekly *rosters*. Each roster is an assignment of runs to days of the week such that weekly driving regulations are not violated. For drivers of the BVO, for instance, a daily rest of 11 hours (and 9 hours in exceptions) between two runs and a weekly rest of 45 consecutive hours are required. The average total work time of each employee is 38.5 hours per week. Again most solution approaches in practical use are sequential—vehicle and crew scheduling prior to crew rostering.

4. URBAN VERSUS EX-URBAN CASE

A transit network for urban service generally has one depot, so that drivers can easily change buses within the day—walk between nearby stations in the city center or travel in a short time between stations within the same city. Ex-urban public transport networks often have several depots for buses which are geographically spread over a large area. Due to trips of longer distance, the transfer of drivers between different locations without driving buses is severely limited. This makes drivers tight to their vehicle in order to reach relief points and depots. Moreover as for the BVO case study, each driver has to come back at the end of his/her daily working schedule (run) to the depot of residence, or more generally, to the depot where he/she began his/her daily working schedule and left his/her own car.

The trips of longer distance cannot be a-priori partitioned optimally among depots in order to solve single-depot problems. A difficult multiple-depot vehicle scheduling problem arises for ex-urban transit. In contrast, crew scheduling can significantly be simplified for this case without heavily affecting optimality. Gaffi and Nonato (1999) investigated methods to solve the joint vehicle and crew scheduling problem for ex-urban service under the following simplification: Drivers can relieve each other only at depots, being declared as the only relief points. This is achieved by imposing that each vehicle rotation consists of a sequence of *blocks* (starting and ending at any pairs of depots), which should collapse with pieces of works for drivers.

5. EX-URBAN CASE WITH BUS-DRIVER DEPENDENCIES – CURRENT STATE AND SOLUTION APPROACH

The BVO bus-driver dependency is now motivated for ex-urban service. This requirement turns out to be a main factor, influencing the scheduling of buses and drivers within the company. Recognizing drawbacks of the current BVO policy to deal with this dependency, a strategy is presented which avoids these drawbacks while retaining the advantages of the requirement.

Finally, our solution approach for scheduling buses and drivers for ex-urban service with the dependency requirement is discussed.

The current practice at BVO is that *two drivers share one bus* over the planning period. The daily bus rotation may contain two runs (or shifts), a morning/noon and an afternoon/evening run, each for one of the two drivers. Because of driver working regulations, this cannot be done over the whole week. Thus many daily bus rotations must consist of only one working shift for one driver, letting a free day for the other driver. Taking into account that each driver has an average work time of 38.5 hours per week including 30 to 60 min of break time per working shift, a bus is only used effectively for an average of 8 to 10 driving hours per day including short breaks.

This relatively short total driving time of a bus is a clear drawback of the one-bus-two-driver policy. Using this strategy, rotations for 34 buses and 68 drivers are built almost manually (of the 75 drivers in daily work, steadily approximately 7 have a vacation). The Paderborn BVO office temporarily employs additional drivers for replacement in case of drivers' illness. Currently, 40-50% of the trips are outsourced to private bus companies.

Upon being asked about the causes of this seemingly inefficient one-bus-two-driver strategy and of the dependency structure, the BVO officers argue as follows: "It is very important that a driver feels that he/she is driving his/her *own* bus. The two drivers take care that their bus is always clean and that it is checked whenever necessary. So there is no need to manage cleaning and maintenance of the buses by the office." Although there is a dramatic improvement of the usage (workload) of buses by fully abandoning this bus-driver dependency, the BVO officers rejected this strategy based on their own experience.

Searching for better solution strategies which take into account the advantages of a bus-driver dependency, we decided to study the impact of

partitioning drivers and buses into groups, such that
drivers of a group are assigned to at most *two* (or *three*) *buses*.

The objective is to maximize the workload of buses of the Paderborn BVO office, thus minimizing the number of trips and trip kilometers outsourced to external companies. The office agreed with such a strategy and agreed to employ new drivers in order to cover more trips of the timetable.

As discussed earlier, many arguments speak in favor of solving vehicle and crew scheduling jointly for the ex-urban case. With a bus-driver dependency such as that required by BVO, we propose to consider

vehicle/crew scheduling AND crew rostering as a joint problem.

The point here is that the bus-driver dependency makes working schedules of drivers dependent on those of buses. Especially, the weekly rosters of each driver group consist of the same subset of trips, which is partitioned into the rotations of two or three fixed buses over the week.

Considering possible ways of constructing groups of two buses and five drivers, we established that allowing drivers to change from one bus to another during a working day hardly improves the generated schedules. This led us to simplify vehicle and crew scheduling by enforcing that

> each bus rotation consists *of two* driver shifts (*straight runs*) *or of one* driver working schedule, possibly split by a long break (*split run*).

In order to solve the problem, a feasible–with respect to working time rules–and flexible scheme is designed, letting five drivers share two buses over the week (see next section, Fig. 3). Heuristics are then developed to "fill in" this scheme by *appropriate* rotations constructed from trips of the timetable. Thus, the rotations and the included runs must fulfill some conditions in order to fit into the scheme—these are conditions from the crew rostering step imposed on vehicle and crew scheduling.

To our knowledge, known approaches to scheduling buses and drivers do not handle a bus-driver dependency, and our approach to integrate crew rostering considerations within the vehicle and crew scheduling process is new. Therefore, our way of handling and solving the problem of scheduling buses and drivers for ex-urban service differs from, e.g., that of Gaffi and Nonato (1999) considerably. Note that the aim of developing heuristics in this paper is twofold, first, at improving upon the current state for the BVO case study, and second, at stimulating further research on exact and heuristic algorithms for the joint vehicle/crew scheduling and rostering problem for (the ex-urban case of) bus transit, cf. concluding remarks.

6. A TWO-BUS-FIVE-DRIVER ROSTERING SCHEME

The two-bus-five-driver scheme that we developed is depicted in Fig. 3. The rostering scheme is based on collecting two daily rotations R1 and R2 (the same except for the weekend) for two buses in such a way that:

(a) Each of R1 and R2 (Monday to Friday) consist of two runs for drivers (cf. Fig. 3), both starting and ending at the same depot.

(b) Only R1 can include (late-)night work.

(c) Between the finish time of R2 and the start time of R1, next morning, there is a minimum time span of 11 hours (or 9 hours in exceptions), in order to allow one switch from an afternoon to a morning shift.

(d) One or two rotations (each consisting of one or two runs) are added on Saturday and on Sunday in such a way that an average of 38.5 hours (per week and driver) is met as well as possible for the five drivers.

Five drivers are associated with each group of two buses serving R1 and R2 on Monday-Friday, RSat1 and RSat2 on Saturday, as well as RSun1 and RSun2 on Sunday, getting a *two-bus-five-driver group*.

In this scheme, weekly rosters numbered 1 to 5 for the five drivers cover the trips of a two-bus-rotation group. (Follow, for instance, roster 1 in Fig. 3.) Each roster contains four runs within Monday to Friday, two afternoon/evening runs (with a rest of 20-22 hours in-between) and then after a rest of 10-12 hours (switch from R2 to R1), cf. (c), two morning/noon runs (with a rest of 20-22 hours in-between). In this way, each roster contains a weekly rest of at least 45 consecutive hours already within Monday-Friday, since it includes a free day (e.g., Tuesday for roster 4) preceded by an afternoon and followed by a morning without working.

To build working plans for drivers over several weeks and months of the planning horizon, note that drivers shall rotate by interchanging their rosters weekly. That is, for instance, the first driver takes roster 1 in the first week, roster 2 in the second week, and so on, then roster 1 in the sixth week again (1 → 2→ 3 → 4 → 5 → 1). The second driver begins with roster 2 in the first week, roster 3 in the second week, ... Thus, the same cycle containing all rosters is served by each of the drivers repeatedly every 5 weeks.

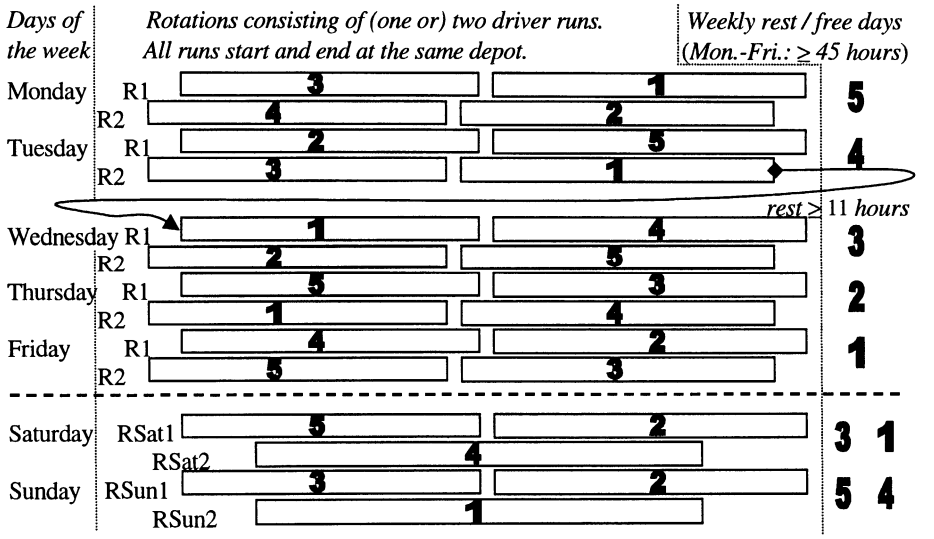

Fig. 3 Weekly rosters for drivers according to a two-bus-five-driver scheme

Having guaranteed a weekly rest within Monday-Friday, the advantage of our scheme is that all drivers or four of the five drivers get an additional free day in the weekend. This day is compound with the weekly rest getting a free time span of 2.5 days, two times every five weeks, namely, when serving roster 1 and when finishing from roster 4 and starting roster 5.

By our scheme, the total driving time of a bus can be increased by up to an average factor of 1/4 over the one-bus-two-driver strategy currently in use

at BVO. This is because two buses should now operate 5 times the weekly work time of drivers against only 4 times for the latter BVO strategy.

Finally, a three-bus-eight-driver scheme turned out to be not flexible in order to be realized effectively: 8 drivers should work for 38.5*8 = 308 h weekly, that is, each of the 3 rotations should log an average of (308/7)/3 = 14,67 h/day. This would often necessitate that each rotation of Saturday and Sunday should also log at least 10-12 h (two runs), however, there is a lack of trips in the weekend as the timetable is *uneven* with respect to *workdays* and *weekend.* Thus, many three-bus-seven-driver groups would occur inducing a worse workload than that achieved by the two-bus-five-driver scheme. Further, groups of two buses and five drivers are simpler to manage, and advantages of the bus-driver dependency are more likely to be saved.

7. HEURISTIC COMPONENTS

In order to realize this two-bus-five-driver scheme, heuristic components are developed, which construct sets of rotations in a multiple-depot problem setting, where each rotation consists of one or two feasible driver shifts. These components are now presented and will then combined to solve our scheduling problem according to various strategies in the next two sections.

We are given a set of scheduled trips T (for schooldays or vacations) partitioned into T_{MonFri} and T_{Sat} and T_{Sun} for Monday-Friday, Saturday, and Sunday/Holidays, respectively. Further, a set of possible depots D is given or computed heuristically in advance, e.g., requiring a minimum frequency of trip departures and arrivals at stations close to a depot as a criterion.

Component 1 (build trip pairs: *bTP*). *bTP(T_X, d, ts)* builds all possible pairs of trips (t1, t2), often belonging to the same line, from T_X (for X = MonFri, Sat, or Sun), where t1: A→ B, t2: B'→ A' with stations A and A' close to the depot d. Here, B and B' are close to each other and *deadhead-time(B,B')* ≤ *start-time*(t2) – *end-time*(t1) ≤ *ts* (*ts* is a given time span set to, e.g., 10 or 20 minutes). Whenever several trips t2 fit into pairs (t1,t2) prefer a t2 departing earlier and/or causing less deadhead time (heuristic decision).

The procedure *bTP(T_X, ts)* works as *bTP(T_X, d, ts)* and decides for a trip t1, whether to build trip pair (t1, t2) for depot d or, for instance, (t0, t1) for d' where trip t0: B''→ A'', d is close to A'' and d' is close to B, B', and B''. The criterion is to prefer earlier tip pairs and those with less deadhead times and time spans between end time of the first and start time of the second trip.

The result is a set of trip pairs $TP_{X,d}$ assigned to a given depot d for the first procedure and sets $TP_{X,d}$ to each depot $d \in D$ for the second procedure.

Component 2 (build rotations with gaps: *bRwG*). *bRwG($TP_{X,d}$, K)* builds K rotations out of the trip pairs $TP_{X,d}$. As not all the trips of T_X can be included

into trip pairs in the general case, most rotations will contain gaps. These gaps are filled in using remaining trips by component 3 below.

Recall that rotations should contain feasible driver runs (two or one), starting and ending at the same depot, in order to fit into our rostering scheme. Thus, working time regulations for runs are to be integrated here. For BVO, either one break of 30 minutes, two breaks of 20 minutes each, or three breaks of 15 minutes each must be integrated within or directly after each block of 4.5 hours driving time. An alternative so-called *sixth-rule* requires that the sum of all break minutes, considering only those breaks of a minimum of 10 minutes each, is greater than or equal to the sixth of the sum of driving times, happening no later than at the end of a 4.5 hour block.

The *bRwG*($TP_{X,d}$, *K*) algorithm linking trip pairs of $TP_{X,d}$ into rotations works as follows (cf. Fig. 4): First sort all trip pairs of $TP_{X,d}$ chronologically with respect to departure times. Then, maintain a list *UV* of partly used vehicles, that are represented by the corresponding partly constructed vehicle rotations $VR_1, VR_2, ..., VR_k$, each as a list of trip pairs to be served by a vehicle. If the next processed trip pair of $TP_{X,d}$ fits into some VR_j without violating the above working time rules, then choose one of them (to insert the new trip pair), namely that whose last trip pair arrives at latest (LIFO: last-in first-out strategy in dispatching vehicles at *d*). Otherwise, if $k+1 \leq K$ a new vehicle rotation VR_{k+1} is created containing this new trip pair as first element.

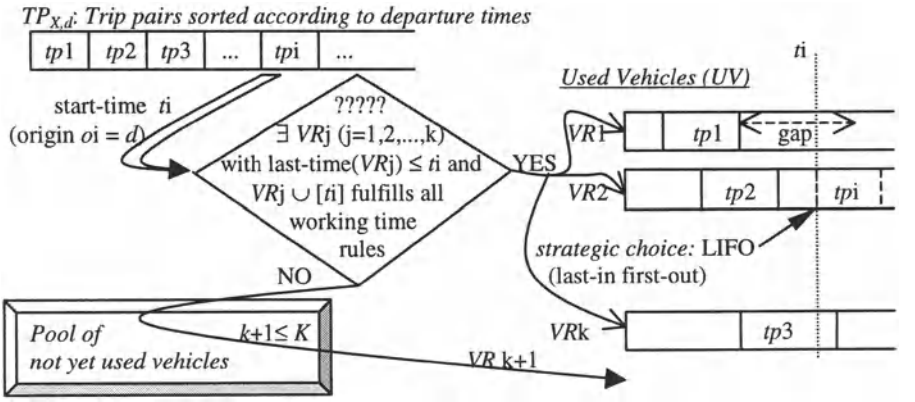

Fig. **4** An algorithm to construct rotations (with gaps), consisting of feasible runs

This algorithm is a modified version of a simple "greedy" algorithm used to construct aircraft routes (cf. Suhl (1995)), where an additional test that the origin of the next flight *oi* = last-destination(*VR*j) is omitted since both are equal to *d* here. Note that last-time(*VR*j) (and last-destination(*VR*j)) refer to the end time (respectively, to the destination) of the last trip pair assigned to *VR*j. The modified steps of the algorithm disallow linking further trip pairs into a rotation, if the latter becomes infeasible with respect to the working

time rules. A central idea is to use a LIFO strategy (last-in first-out policy for dispatching buses at each station). This strategy tries to build good runs for some rotations (cf. *VR*2) letting large gaps in the other rotations (cf. *VR*1), in order to insert remaining trips in a further step. Finally, strategies such as LIFO and FIFO (first-in first-out) can generally be realized more efficiently by keeping the list of used vehicles *UV* sorted with respect to last-time of the *VR*j's. This is possible when processing not only starting events but also ending events of trip(pair)s in the algorithm (cf. Mellouli (1997)).

The result of $bRwG(TP_{X,d}, K)$ is a set of K rotations $RwG_{X,d,K}$. If not all of the K rotations could be built, generate rotations containing a dummy trip during one minute at early morning, starting and ending at the same depot *d*.

Component 3 (fill in inner/outer gaps of rotations: *fiIG and fiOG*). $fiIG(rT_X, RwG_{X,d,K})$ and $fiOG(rT_X, RwG_{X,d,K})$: The rotations $RwG_{X,d,K}$ built by component 2 may contain inner gaps (caused by LIFO, cf. *VR*1 in Fig. 4) and "outer gaps" in the case the rotations do not start early in the morning or do not end at night. Remaining trips (rT_X) that are not included into trip pairs by *bTP* together with those of (disconnected) trip pairs, not used by *bRwG*, are now considered for the sake of filling the gaps of the rotations of $RwG_{X,d,K}$. This is done by *fiIG* for the inner and by *fiOG* for the outer gaps. In order to realize *fiIG* (analogously *fiOG*, see below), a trip-as-node network with nodes corresponding to trips from rT_X and arcs to compatibility relation between these trips is constructed (cf. Fig. 5).

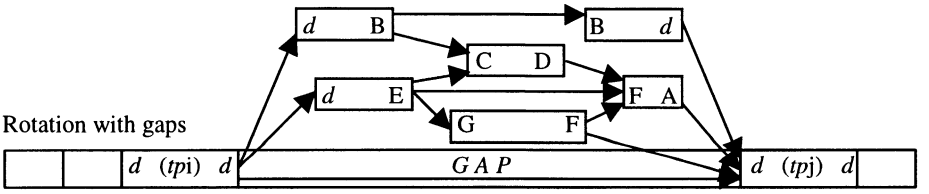

Fig. **5** Filling gaps by shortest-path computations based on a trip-as-node network

For each gap, a shortest-path algorithm is performed where the inserted source corresponds to the last trip before the gap (*tpi* in Fig. 5) and the target to the first trip after the gap (*tpj*). The costs on arcs are chosen principally in such a way that not only deadhead effort but also idle times are minimized. (An arc from the source *tpi* to the target *tpj* can be integrated with costs corresponding to an upper bound of the total effort allowed to fill in the gap.) For instance, the path [*tpi*, *d*-B, B-*d*, *tpj*] without deadhead trips but with a long break in B may be longer in the constructed network model than the path [*tpi*, *d*-E, G-F, F-A, *tpj*] with two deadhead trips and only short breaks.

If the trips of a computed shortest path–when being integrated into the gap–induce a violation of working time rules for a run of the corresponding rotation, the arc of this path that corresponds to the first trips-link causing a

violation is deleted from the trip-as-node network and a shortest path computation is again performed on the modified network.

The procedure *fiOG* applied to rotations assigned to a depot d works as *filG* after defining the gap 'size' by integrating a dummy trip starting and ending at d, as a source early in the morning or as a sink at night, for an outer gap at the left/right extremity of a rotation. Note that for very short rotations, e.g., starting and ending early in the morning, a sink dummy trip is integrated at say 14:00 in order to enforce the first run of a driver to end at the home depot. A second pass using the same procedure *fiOG* will then construct a second run for the other driver in the case that appropriate trips remain. The idea of applying the procedures *filG* and *fiOG* in several passes turned out to be very effective (cf. computational results below).

8. A TWO-PHASED STRATEGY

A solution approach used in early experiments consists of two phases. A first phase builds bus rotations, separately for Monday-Friday, Saturday, and Sunday/Holidays. For this, *bTP*(T_X, *ts*) with *ts* set to, e.g., 15 min firstly generates TP_d for each depot d. Applying the procedure *bRwG*($TP_{X,d}$, K) to each depot d and X = MonFri, Sat, Sun with a sufficiently large K, the best rotations are chosen heuristically (the same $2*G_d$ rotation number for X = Sun, Sat, as for X = MonFri, for each depot d). The trip pairs of rejected rotations are added to the sets of remaining trips rT_X. The procedures *filG* and *fiOG* are then applied to fill in the gaps of all generated rotations.

A second phase groups these rotations into weekly schedules for drivers (rosters) according to the rostering scheme: For each group two rotations for Monday-Friday are chosen which belong to the same depot and fulfill conditions (b) and (c) of the section before last. To achieve an average of 38.5 work hours per week for each driver, the goal of number of hours T_{we} to be accomplished in the weekend by the five drivers is computed by: 38.5 hours multiplied by 5 (for the five drivers) minus the total work time of the constructed two rotations (R1) and (R2) multiplied by 5 (number of workdays). Up to two rotations, each with one or two runs, are then chosen from the Saturday then from the Sunday rotations having approximately T_{we} work hours. In this step, guiding parameters are computed such as the percentage of remaining driving time for Saturday/Sunday and the numbers of the rotations remaining for Saturday/Sunday to determine the most likely best number of rotations to be added for Saturday/Sunday for each group. This favors a good distribution of work times within a week (workdays and weekend) over all groups, thus over all drivers.

9. A BEST-GROUP-FIRST STRATEGY

Experimenting with the above strategy, a strong relationship between its two phases is established in a sense that the quality of the groups built in the second phase depends on the quality of the rotations and their distribution among depots in the first phase. The drawbacks of the two-phased strategy can be summarized as follows: First, *bTP(T_X, ts)* may build trip pairs at a certain depot d which are disconnected later when choosing the best rotations. The trips of these pairs may, however, be convenient for another depot d'. Second, the chosen best $2*G_d$ rotations for a depot d may turn out to be inconvenient for obtaining a good partition into G_d two-rotation groups for Monday-Friday, mainly because of the condition (b) of the two-bus-five driver scheme, requiring a time span of a 11 hours between rotations.

To remedy these drawbacks, a 'best-group-first' strategy is developed which tries to merge both phases. The known number of a company's buses divided by 2 is designed by G, the number of groups to be constructed (note that numbers G_d of groups to be assigned to depots $d \in D$, respectively, are not set in advance). G is a parameter which can be set at the beginning of each run of the best-group first strategy. The algorithm's steps for the 'best-group-first' strategy are now presented and then explained within the text.

Algorithm. *Best-Group-First*

for each X = MonFri, Sat, Sun **do** set $rT_X := T_X$; // Initialization
for each group number $g = 1, ..., G$ **do**
  set $ts := f(g)$; *max-score* := 0; // $f(g)$ is a certain arithmetic function of g
(1) **for each** depot $d \in D$ **do**
    for each X = MonFri, Sat, Sun **do**
      bTP(rT_X, d, ts) ; ⇒ $TP_{X,d}$
      bRwG-modified($TP_{X,d}$, 2) ; ⇒ $RwG_{X,d,2}$
    end{for each X**}**
    Compute $score_d$ = score that evaluate a group consisting of the rotations of $RwG_{X,d,2}$ (X = MonFri, Sat, Sun), assigned to depot d
    if $score_d > max\text{-}score$ **then** *best-depot* := d; *max-score* := *score*; **end{if}**
    // Note that rT_X is not updated within this for-each-depot loop
  end{for each depot}
(2) // Reconstruct the rotations of the chosen g-th best group assigned to *best-depot*
  for each X = MonFri, Sat, Sun **do**
    bTP(rT_X, best-depot, ts) ; ⇒ $TP_{X,best\text{-}depot}$
    bRwG-modified($TP_{X,best\text{-}depot}$, 2) ; ⇒ $RwG_{X,best\text{-}depot,2}$
    Update rT_X ;
  end{for each X**}**
  Define the g-th two-bus-five-driver group assigned to the depot *best-depot* by the rotations of $RwG_{X,best\text{-}depot,2}$ (X = MonFri, Sat, Sun), containing gaps;
end{for each group number g**}**
for each *pass* = 1, ..., *max-pass* **do**

(3) **for each** group $g = 1, ..., G$ **do**
Let d be the depot of the g-th two-bus-five-driver group and
$RwG_{X,d,2}$ be the sets of its rotations for X=MonFri, Sat, Sun
for each X = MonFri, Sat, Sun **do**
fiIG(rT_X, $RwG_{X,d,2}$); ⇒ update of rT_X and $RwG_{X,d,2}$
fiOG-modified(rT_X, $RwG_{X,d,2}$); ⇒ update of rT_X and $RwG_{X,d,2}$
end{for each X**}**
// The gaps of the g-th group's rotations in $RwG_{X,d,2}$ (X=MonFri, Sat, Sun)
// are filled in (*pass* times) by trips of rT_X (which are deleted from rT_X).
end{for each group g**}**
end{for each *pass***}**

In order to determine the g-th best two-bus-five-driver group (g = 1,...,G), step (1) of the best-group-first algorithm computes–using the same sets of trips rT_X (X = MonFri, Sat, Sun)–a group to each depot $d \in D$. The constructed rotations (with gaps) of each group are evaluated and a score is given to each group/depot. For the 'winning' depot, *best-depot*, the rotations of the corresponding group, designated as the *g-th best*, are (re-)constructed in step (2). After building the rotations (with gaps) of all the G groups, *several passes* of *fiIG* and *fiOG* are performed in order to fill in the gaps of the rotations *gradually* with *less deadhead times* as possible (cf. refinements and remarks in the next section).

Note that the *bRwG* procedure, described in the section before last, is called in step (1) with K=2 (number of rotations) and modified in order to generate two rotations fulfilling condition (c) of the rostering scheme: Whenever the second rotation is generated by the LIFO-algorithm (cf. Fig. 4) with start time *tstart*2, the latest end time *tend*1 of the first rotation (corresponding to (R2) in the rostering scheme) is computed by *tend*1 := 24h + *tstart*2 – 11h (or 9h). A trip pair whose end time exceeds *tend*1 is then not considered for extending the first rotation. The dependency of *tend*1 and *tstart*2 is also taken care of by a modification of *fiOG*, namely, when fixing start/end times of dummy trips influencing the 'size' of an outer gap to be filled in by trips of rT_X. These precautions ensure that rotations of a group are not too long, and thus additional tests, determining whether the 5*38.5 work hours (for 5 drivers) are exceeded by the two rotations of a group, become almost secondary. For a depot with several groups (as Paderborn for our case study), a better balance of work/free hours among groups can be achieved by exchanging rotations in a post-processing step.

10. REFINEMENTS AND COMPUTATIONAL RESULTS

Our algorithms are embedded into a planner's decision support system, where the above heuristics based on the two-bus-five-driver scheme are realized. The implemented algorithms for scheduling buses and drivers are

tested using data from the BVO company. Tables 1 and 2 below show computational results with the best-group strategy which turned out to be superior to the two-phased strategy. The test timetable for schooldays comprises 829 trips for Monday-Friday, 324 trips for Saturday and 66 trips for Sunday, that is, 4,535 trips with approximately 122,010 planned trip kilometers corresponding to 2,790 driving hours per week.

Knowing that the BVO actually uses 34 buses + 4 reserve buses and outsources approximately 50 % of its trips/trip kilometers to external companies, our first aim for the case study is at improving upon this current state while providing computer-based scheduling support. The results of Table 1 (and 2) show that 63.1% (68.7%) of the number of trips, 68% (74.3%) of the planned trip kilometers, and 69,5% (75.5%) of driving hours can be covered by 17 groups, 34 buses, and by 19 groups, 38 buses, respectively.

To illustrate the functionality of the best-group-first algorithm, statistics of the results are given after each of its main steps: after building rotations with gaps (first line) and after each pass of filling in inner and outer gaps (remaining lines). The statistics indicate the number of served trips, the amount of served trip kilometers and driving hours, as well as, the amount of deadhead kilometers and driving hours by inserted deadhead trips. Here, requirements of BVO are integrated: Deadhead trips during more than 40 minutes are not allowed and for computing deadhead distances and times, a detour factor of 1.8 and an average of 1.3 min/km are assumed.

	Number of Trips	%	Trip-KM	%	Trip-Hours	%	Deadhead-KM	%	Deadhead-Hours	%
BTP, bRwG	1,910	42.1	59,823	49.0	1,342	48.2	517	0.86	11.2	0.84
Pass 1: *fiIG*	2,151	47.4	64,188	52.6	1,471	52.8	2,062	3.21	44.7	3.04
fiOG	2,693	59.3	80,253	65.8	1,869	67.1	4,488	5.59	97.3	5.20
Pass 2: *fiIG*	2,759	60.8	80,880	66.3	1,884	67.6	4,997	6.18	108.3	5.75
fiOG	2,823	62.2	82,282	67.4	1,920	68.8	5,833	7.09	126.4	6.58
Pass 3: *fiIG*	2,833	62.4	82,332	67.5	1,921	68.9	5,961	7.24	129.2	6.72
fiOG	2,854	62.9	82,875	67.9	1,934	69.4	6,550	7.90	141.9	7.34
Pass 4: *fiIG*	2,859	63.0	82,908	68.0	1,935	69.4	6,640	8.01	143.9	7.44
fiOG	2,864	63.1	83,003	68.0	1,937	69.5	6,735	8.11	145.9	7.53

Tab. **1** Computational results, best-group-first strategy — 17 groups, 34 buses

The percentages indicate *efficiency values* by fractions of served trips, trip kilometers, and driving hours relative to the total of 4,535 timetable trips, 122,010 planned trip kilometers, and 2,790 driving hours, respectively, and *deadhead factors* by fractions of deadhead kilometers and driving hours relative to served trip kilometers and served trip hours, respectively.

	Number of Trips	%	Trip-KM	%	Trip-Hours	%	Deadhead-KM	%	Deadhead-Hours	%
BTP, bRwG	1,962	43.2	61,714	50.6	1,392	49.9	542	0.88	11.7	0.84
Pass 1: *filG*	2,230	49.1	66,604	54.6	1,533	55.0	2,165	3.25	46.9	3.06
fiOG	2,881	63.5	85,957	70.5	1,985	71.2	4,692	5.46	101.7	5.12
Pass 2: *filG*	2,948	65.0	86,730	71.1	2,004	71.9	5,500	6.34	119.2	5.95
fiOG	3,053	67.3	89,195	73.1	2,069	74.2	6,717	7.53	145.5	7.03
Pass 3: *filG*	3,078	67.8	89,744	73.6	2,082	74.7	7,103	7.91	153.9	7.39
fiOG	3,110	68.5	90,463	74.1	2,099	75.3	7,864	8.69	170.4	8.12
Pass 4: *filG*	3,115	68,6	90,482	74.2	2,100	75.3	7,880	8.71	170.7	8.13
fiOG	3,120	68.7	90,610	74.3	2,103	75.5	8,008	8.84	173.5	8.25

Tab. **2** Computational results, best-group-first strategy —19 groups, 38 buses

Several refinements of the best-group-first strategy are performed in order to increase efficiency values and decrease deadhead factors. The main ideas of the refinements are summarized in the following:

1. *Do not put minor depots at a disadvantage:* For the BVO case, the major Paderborn depot evaluates the best several times and may use trip pairs to minor depots (there and back), letting early starting trips from minor depots to Paderborn without connections. To avoid this, d = Paderborn is not considered for the first 6 groups ($g \leq 6$) and an extra bonus is assigned to early starting rotations by the groups' evaluation (besides the number of served trip pairs and the amount of served trip kilometers).
2. *Try to distribute trip pairs evenly among all constructed two-rotation groups:* For this, the value *ts* for *bTP* (building trip pairs) is set to a function of the group number g, in the experimentation, $ts = 10 + 2*g$, for $g < 10$, and $ts = 10 + 1.5*(g-10)$, for $g \geq 10$. Such settings of *ts* seem to enhance the quality of the last generated rotations, which avoids 1 to 2% of deadhead times, added in step (3) of the algorithm.
3. *Try to distribute remaining trips evenly among depots in order to fill in the gaps with less deadhead effort as possible:* A main difficulty is to decide which of the remaining trips are to be integrated in which gap in order to reduce total deadhead effort. For instance, trips used to fill in the gaps of rotations assigned to depot d may cause less deadheading when they are integrated in gaps of rotations assigned to another depot d'. Trying to avoid this, a complex cost function is used in the shortest-path approach of *filG* and *fiOG*: Costs on arcs between compatible trips are set to a linear function of standing time and deadhead driving time. Penalty costs are added to long deadhead trips, to deadhead trips to other depots, and to long deadhead trips before serving short trips. Setting a relatively low upper bound on shortest-path-lengths to fill in gaps (costs on the source-sink arc) and decreasing the costs for source-arcs for early outer

gaps and for sink-arcs for late outer gaps ensure that deadhead driving times are *only moderately* added by *fiIG* and *fiOG*, in several passes.

Results in Table 1 and 2 show that the first pass of *fiIG* and *fiOG* does not fill in the gaps completely which would induce a higher overall deadhead factor. Rather, the subsequent passes fill in the remaining gaps (including outer gaps for second driver runs) and those newly generated by *fiOG*.

11. CONCLUSION

Project. A main goal expressed by BVO at the start of this project was to minimize the amount of outsourced trips. This means to maximize the workload of their buses, as well. Note that a bus costs 600,000 German marks (≡ 300,000 EURO) and cannot be used for more than 10 years owing to restrictions of government subvention. Towards achieving this goal while retaining a bus-driver dependency required by BVO, the results with the best-group-first strategy show that approximately 70% of trip driving hours of the timetable can be covered with 7.5% of deadhead times by 34 buses.

Analyzing driving times of the 17 groups generated by the algorithm, 6 groups can be operated by 5 drivers, 6 groups by '4.5 drivers,' and 5 groups by 4 drivers. '4.5 drivers' means that 3 drivers operate halftime in one and halftime in another group. This is allowed by BVO for new drivers.

Therefore, 77 drivers, instead of 68 drivers by the current two-bus-four-driver strategy of BVO, can work on 34 buses of the Paderborn office (without reserve buses). Note that the required 38.5 work hours per week for each driver includes, besides driving times, approximately 5-6 hours regular breaks, 1-2 hours cumulative short breaks and 2-3 hours for accomplishing other activities, such as refueling, cashing up, and administrative needs. The results of the heuristic indicate that each of the 77 drivers covers approximately 27 driving (trip and deadhead) hours per week. It is still possible to insert additional trips, especially, those starting early in the morning from non-depots to depots. This can be achieved by additional passes of *fiOG* (and *fiIG*) with higher upper bounds on the shortest-path lengths.

Further improvements. The best-group-first algorithm is currently being improved in several ways. The step of building trip pairs (used in a subsequent step to construct rotations with gaps) should support their even distribution among depots and *external-company depots*, more carefully. It seems to be convenient for the BVO to outsource trips to those external companies which reside at non-depot locations far from depots, from which trips start early in the morning. An analysis of remaining trips indicates several trips of this kind. Thus, this improvement of the algorithm aims at minimizing the costs of operating remaining trips by external companies, as well. Here, a *make-or-buy* analysis is useful for public transport companies, especially for ex-urban bus

transit: Private bus companies in the tourist business, offering long journey trips mostly on weekends and vacancies, may serve the above mentioned early starting trips as well as school service and peak hours' trips, at lower costs than the public transport company itself.

Ongoing research. From a research point of view, our approach of integrating crew rostering considerations into the vehicle and crew scheduling process seems to be very useful for bus transit, especially for the ex-urban case. This is based on firstly designing general and flexible rostering schemes imposing special *conditions* on working schedules for buses and drivers, *which* are then integrated into a combined vehicle and crew scheduling process.

The heuristic components *bTP* and *bRwG* turn out to be very effective in building principal parts of rotations (consisting of feasible driver runs) in the multiple-depot problem setting for ex-urban service. A considerable part of trip hours can be covered in this way with a low deadhead factor (less than 1%, cf. results in Table 1 and 2). The best-group-first strategy can be developed further towards an optimization-based heuristic. A potential way is to replace step (3) of the algorithm by an optimization model.

Recall that step (3) fills in the gaps in several passes by applying shortest-path computations. This *sequential* proceeding (gap by gap) is merely a heuristic way to solve the step-(3)-problem (all gaps at once) optimally as a *multi-commodity flow* problem with *resource variables* and *constraints*, modeling driving times and break requirements. Principally, we developed such a mixed-integer model for the entire multiple-depot problem with a subset of working time rules. Model instances with up to 300 trips (for Saturday) could be solved. The point is that an improved version of *bTP* (with/without *bRwG*) reduces the size of the mathematical models by linking trips. Thus, the step-(3)-problem is expected to be optimally solvable for problems of larger scale, as for the Monday-Friday case of the BVO timetable.

The best-group-first algorithm can also be developed further towards a *column generator* to solve a *set-partitioning* problem where each column is the subset of trips of the rotations of a group and each row covers a trip.

REFERENCES

Daduna, J.R. and Paixão, J.M.P. (1995). Vehicle Scheduling for Public Mass Transit – An Overview. In Computer-Aided Transit Scheduling (Daduna et al.), LNEMS, Vol. 430, 76-90. Springer. Berlin – Heidelberg.

Gaffi, A. and Nonato, M. (1999). An Integrated Approach to Ex-Urban Crew and Vehicle Scheduling. In Computer-Aided Transit Scheduling (Wilson N.H.M.), LNEMS, Vol. 471, 103-128. Springer. Berlin – Heidelberg.

Löbel, A. (1998). Optimal Vehicle Scheduling in Public Transit (172 pages). Shaker. Aachen.

Mellouli, T. (1997). Improving Vehicle Scheduling Support by Efficient Algorithms. In Operations Research Proceedings 1996 (Zimmermann et al.) 307-312. Springer.

Suhl, L. (1995). Computer-Aided Scheduling: An Airline Perspective (248 pages). Gabler–DUV. Wiesbaden.

Freling R., Wagelmans, A.P.M., and Paixão J.M.P. (1999). An Overview of Models and Techniques for Integrating Vehicle and Crew Scheduling. In Computer-Aided Transit Scheduling (Wilson N.H.M.), LNEMS, Vol. 471, 103-128. Springer.

Wren, A. and Rousseau, J.-M. (1995). Bus Driver Scheduling – An Overview. In Computer-Aided Transit Scheduling (Daduna et al.), LNEMS, Vol. 430, 76-90. Springer.

Chapter 3

COMPUTER AIDED PLANNING OF RAILROAD OPERATION

Prof. Dr.-Ing. Thomas Siefer, IVE, University of Hanover, Germany
Dipl.-Math. Dirk Hauptmann MSc, IVE, University of Hanover, Germany

Abstract The construction and optimization of a stable timetable has a large number of boundary conditions. The simulation program Simu++ and the vehicle rostering program Dispo++ enable the railway-companies to make transportation better and cheaper. The article gives some information about the functionality of the program, developed at the Institute of Transport, Railway Construction and Operation at the University of Hanover, Germany.

Keywords: Simulation, Timetable-Construction, Railway-Operation, Railway-Infrastructure, Rolling-Stock

1 INTRODUCTION

All railway companies have the complex task of optimizing the planning of their resources in order to serve the arising transportation demand at a minimum of resulting costs. This demand is becoming more important in the last years. Decisions of European parliament gave the railway-companies on one side more responsibility for their cost-structure. On the other side - in most of the countries - the railway-companies now have real competition in the

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 37–47.

field of transportation. If they want to be busy in transportation in future, they must be better and cheaper than other companies.

For this task, the Institute of Transport, Railway Construction and Operation (IVE) at the University of Hanover, Germany, has developed an integrated software tool to generate a timetable, check the stability of that timetable and to create an optimal scheduling of existing rolling stock. We started with a program just called SIMU. For more than 20 years member of the Institute develop and improve the model for railway-simulation.

For details in the timetable-construction we combined Simu++ with the running-time calculation program DYNAMIS. It allows the user to design a complete concept for the railway operational process. Iteratively applied, the programs support the user throughout all steps to design a timetable with respect to the needed amount of rolling stock such as locomotives or railcars.

For the construction and optimization of a stable timetable a large number of boundary conditions such as the amount of rolling stock and running times have to be complied. The result itself has a feedback to the boundary conditions and in some cases the timetable is no more valid with respect to the altered boundary conditions. In these cases it has to be adopted again. This leads to an iterative planning process. An experienced planer or a planning team can hardly overview all restrictions during the construction phase. A computer aided method can help the planer to improve this process. The time for timetable-construction is getting shorter. It is not longer possible to ask the different transport companies to give more than one year before the timetable will start a final order for all trains. The market-needs make it necessary to react to different transport behavior shortly. That's another reason for a computer aided method of planning. Only the help of computerized planning allows you to test in a short time period some different ideas for future timetable constructions. So you can construct timetables with a high standard and at the same time you can for example compare different strategies in solving problems. In Germany a typical problem is to find the right solution for timetable construction in a case of a bottleneck in the infrastructure.

For the iterative process of creating a time-table you need

- Initialization and data collection
- Timetable construction
- Allocation of rolling stock
- Stability check of timetable

In order to answer the complex questions arising within the whole planning process an integrated software tool based on the systems described above was necessary. For each step of the planning process the Institute for Transportation, Railway Construction and Operating (IVE) at the University of Hanover, Germany, has developed computer programs. We started with batch programs for the first generation of computer, created programs for workstations and now it is possible to work on a normal personal computer.

2 OVERVIEW OF THE INTEGRATED PROGRAM MODELES

2.1 Track Editor

First we need the infrastructure. With the tool "Track Editor" you can create a new infrastructure or build up an existing network. The infrastructure is the basis for the timetable. With an graphical editor all modeled elements of the infrastructure include switches, signals, stations, stopping points, speed indicators, platforms and tracks are edited. They have various attributes such as length, maximum velocity, aspect ratio, geometrical position etc. The level of details takes into account all relevant aspects of the infrastructure, the signaling system and the railway operation process. For a station as well as for a line we need all tracks with these important information. For the signaling system we have all German types included. For studies, we already did, we have signaling-systems from Belgian, Italy, Denmark and Australia available. Modern signaling systems such as ETCS (European Train Control System), radio controlled operation or ATC/ATO (Automatic Train Control/Operation) can be simulated too.

With the signaling system we build up all the relevant times, which influence the operational scheme. For the case of research we can although create new operational forms to find out their influence for example to the quality of a railway system. It is possible to work out the improvements, new radio-controlled system could have. For trams, tubes, light-rails and commuter-systems we have some different signal-systems, even a system without signals, when trams drive in the street together with the cars.

For each signal we build up the ways, they are possible in the real infrastructure. The model gives us a real impression from the reality. Behind every signal there has some space to be free for misbraking, this length is although build up. The planer can get all information in the window-system, so he will have an impression, whether all parameters are build up correctly.

2.2 DYNAMIS: Detailed computation of running times

When we finished to create the infrastructure, the iterative time-tabling process can start. We have to know the acceleration and brake-force of each train. We can create several model-trains, therefor we need information about locomotives and wagons.

It is necessary to calculate an exact running-time, which can be used for a timetable-construction. Therefore, all important vehicle and infrastructure parameters (traction effort, air-resistance, braking-power, speed restrictions, gradient, radius or tunnel-resistance, etc.) have to be considered. The IVE developed program DYNAMIS can be used for:

- the calculation of running times,
- the determination and the check of the signaling infrastructure,
- the planning of speed restrictions,
- and the optimization of energy consumption to economize a train run.

For the running time account we have to know the traction unit requirements. We have to recognize the train protection equipment, because this can influence the allowed maximum speed. In some cases it would be interesting to know something about energy saving driving modes. We have to check the alignment. For the use in Austria we build up the engine temperature, it could be necessary to reduce the speed in the mountains because the engine is getting too hot. All the knowledge we get with the DYNAMIS system, we gave with an interface to the timetable interface.

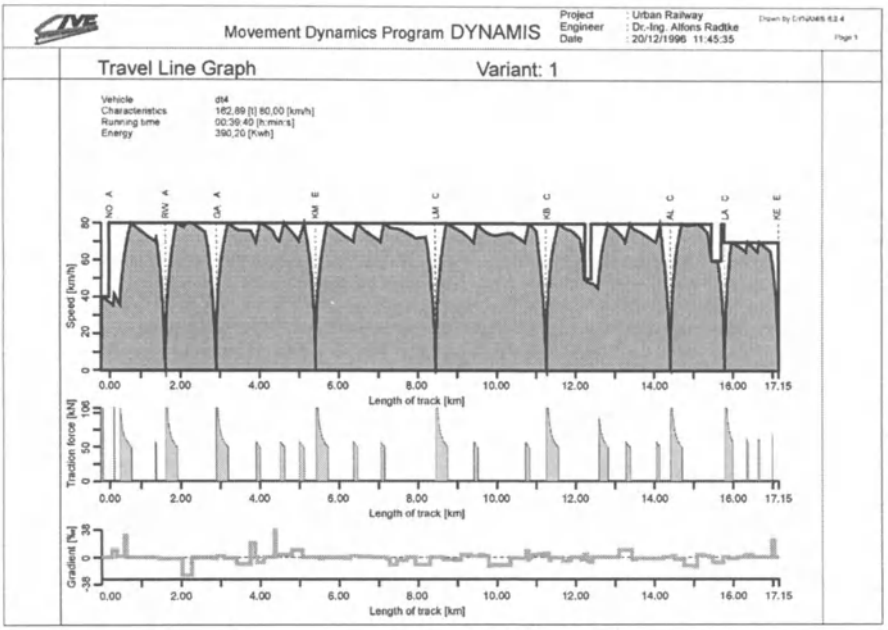

Figure 1: DYNAMIS - Running times

An example of the computed results is shown in Figure 1 in a speed-profile graph of a train run. The second diagram in the figure illustrates the necessary traction force for that train run. The gradient at the bottom completes the picture.

2.3 Timetable construction and simulation

The program system Simu++ has the software features to simulate accurately and easily the railway operation on any given network. The level of details takes into account all relevant aspects of the infrastructure, the signaling system and the railway operation process such that the resulting system has quick turnaround times on the one hand and allows the exact modeling of the problem on the other hand.

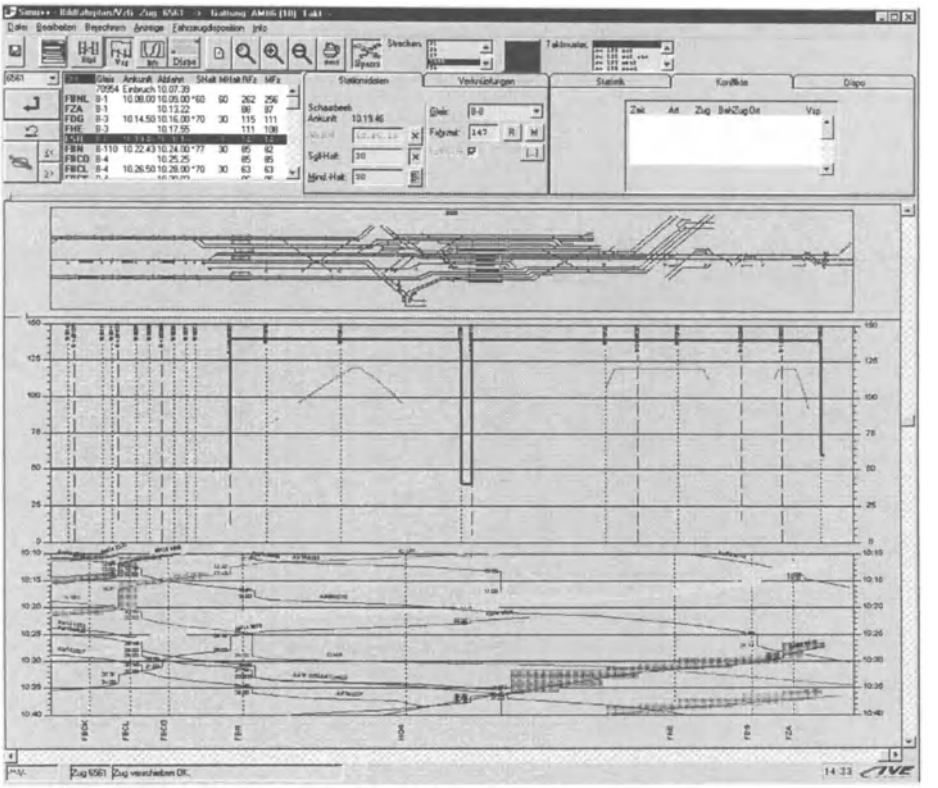

Figure 2: Timetable construction

The timetable is constructed with Simu++ interactively at the timetable graph. All train runs are computed with an integrated running time calculation which was verified using the program DYNAMIS. During this process, additionally, the occupation times of every block section of each train run are computed. In combination with the respective train locations this data is used to detect conflicts with other train runs such as occupation conflicts in stations or train headway conflicts. The running times and the conflicts are computed after every modification of the timetable. By according actions (moving a train run, changing of tracks, changing of running times etc.) a conflict free timetable can be produced by the planer.

Figure 2 shows the main screen of the timetable editor. The lower part of the screen-shot shows the timetable with a selected train. In the dialog box at the upper part of the window the data of the train can be viewed and edited. The folder named 'Konflikte' shows detected conflicts with other trains. The conflicts are also shown in the timetable by concrete black boxes.

Creating a new train is as follows. After choosing a train type the train route is determined automatically by a user-defined starting and arrival point and the stations or junctions in between. Then, the tracks in the stations, the departure times, stopping times and the running times can be chosen. During all steps the above described conflict-detection algorithm helps the user to handle complex timetables for single lines as well as for large railway networks.

To check the stability of the operational program (i. e. the constructed timetable) a given number of timetables (between 50 and 100) are automatically generated and overlaid with disturbances. These timetables in reality represent a series of operational days. The simulation is run with these timetables. Integrated dispatching algorithms simulate the actions of the human dispatchers to control the disturbed traffic flow. Finally, the result of the simulations is statistically evaluated. The stability of the timetable is not satisfying if, for example, delays of selected train types exceed predefined limits or if disturbed trains can not reduce delays within their run such that the delays are propagated through the whole simulated network.

Another typical applications of the simulation is to minimize the amount of infrastructure being necessary to operate a certain timetable. For example, the number of sidings needed along the track can be determined. On the other hand for a fixed track layout the simulation provides information for the optimization of the timetable. Or you can test, with train-mix allows you the highest number of train running during a special period of time on your infrastructure. With the simulation you can see how you can improve the efficiency of your railroad-system in different ways.

2.4 DISPO++: Rostering of rolling stock

The program Dispo++ has been developed since 1990 to optimize the amount of rolling stock for a given timetable. Due to the extent of the boundary conditions which have to be considered a computer aided method is necessary for the optimization of vehicle use of traffic systems in extensive networks. The computer model Dispo++ carries out an optimization with algorithms giving a mathematically optimal solution. Hence, the computed number of necessary vehicles represent the minimum amount which is necessary according to the timetable and the infrastructure. The solution contains the minimization of empty runs and empty runing kilometers of the used locomotives. Further features of Dispo++ are the planning of maintenance stops and the planning of several vehicle or locomotive classes. These features are based on heuristic algorithms.

Dispo++ takes into account all relevant boundary conditions of the real world problem such as reversing times, the location of train depots or the average velocity of the empty train runs. For the calculation of the number of necessary locomotives a graph is constructed. The edges of the graph represent the train runs and possible connections between two train runs. During the construction of the graph (especially the connections) the boundary conditions are tested. If one of the boundary conditions is violated the edge will not be introduced in the graph model.

After the graph construction the MINIMUM-COST-MAXIMUM-FLOW-Problem is solved. The resulting solution gives the optimal number of necessary locomotives and the allocation of the locomotives to the train runs. It also includes the required empty train runs. These empty train runs are constructed automatically using a predefined average speed.

For the planning of maintenance stops an adopted Treshold Accepting Algorithm is implemented. With this algorithm a feasible solution according to the boundary conditions (e. g. distance between two stops, duration of a maintenance stop) is computed. For the initial solution, the above described solution of the optimization problem is used. The procedure now is to generate feasible solutions that satisfy the boundary conditions. The empty running kilometers are the objective function for the optimization program.

In figure 3 a part of a rostering plan is shown. The rows represent one vehicle each.. The concrete bars are the resulting empty train runs.

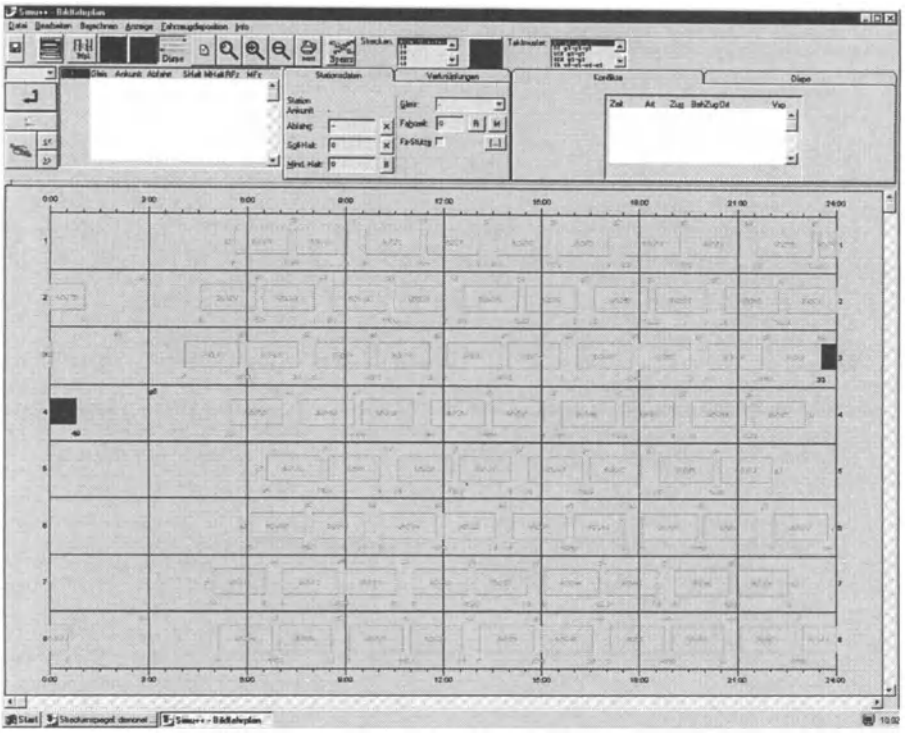

Figure 3: Timetable construction - Rostering plan (black bars: resulting empty train runs)

2.5 Performance Evaluation (Statistics)

When the timetable-construction is finished, the simulation is done, you wish to know the statistical evaluation. For this reason we developed the performance evaluator. It allows you to show the simulation results on different ways. You can get information separated for train-types, special lines or some stations. You can see the average
delay per train, per train type, train direction. You can see how many trains are in time and you can choose some class of delays. It is although possible to get information about delays, shown as a part of the infrastructure. For a lot of planning teams this is the best way to get information about the situation in the network. Depending on the form of the line they can recognize parts with problems in the system.

When you do case studies like a break of the signal-system for a certain time or a breakdown of a train on an important line, it is important to see what will

happen, when the problem is solved. So we can visualize in what time the system will be in time and stable again. For this studies it is possible to show the average delay in steps of 15 minutes.

The simulation system Simu++ now is a very comfortable system. It can be used for timetable-construction and simulation as well as for the determination of the infrastructure.

3 APPLICATIONS

The system enables the planer iteratively to set up a complete operational scheme having regard to complex boundary conditions such as energy consumption and vehicle allocation. The software system can be applied by all railway companies operating on sufficiently large track networks. The complete system has been tested at the IVE on the railway network of the eastern part of the Belgium railways (SNCB/NMBS). The considered network consists of 191 stations (including the large nodes Liege and Brüssel), 1866 kilometers of track and 2769 signals. The simulated and planned timetable included 1800 train runs during 24 hours. For an example, the low range traffic of the AM140-railcars have been assigned (more than 500 train runs).

To optimize the integrated timetable in a eastern part of Germany we work together with consultants. For this item we have now simulated a network with more than 4000 kilometers of track. As a result we got information about the infrastructure, which is necessary to run the new timetable. We although got information about the estimated stability of the new system. With these information it is possible to make a forecast about new passengers, who we choose the train. The new quality of public transport is the best way to win new passengers.

There will be further improvements in the simulation-system Simu++. A system develop at a university is always an object of research. According to new questions in the field of railroad-operation we are looking for answers. The use of the program-family by railway-companies and consulting-groups guaranties always the state of the art.

References

Radtke A. (1995), Dispositionsmodell für den optimierten Betriebsmitteleinsatz der Eisenbahn. Dissertation, Institute for Transportation, Railway Construction and Operating ,Hanover

Radtke A., Hörstel J, Müller L., Schumacher A. (1996), Entwicklung einer optimierten Instandhaltungsplanung für spurgeführte Betriebsmittel in großen Verkehrsnetzen Institute for Transportation, Railway Construction and Operating ,Hanover

Kaminsky R., Hauptmann D., Radtke A.(1996), Integrated planning-system for railways - A tool to improve the planning process. World Congress on Railway Research (WCRR), Colorado, Colorado, USA, Volume 1, pp 81-86

Burmeister P, Siefer Th.. (1999), Fahrdynamischer Workshop bei der Hamburger Hochbahn AG. Verkehr und Technik , Vol. 3, pp 99-105

Chapter 4

URBAN MULTIMODAL INTERCHANGE DESIGN METHODOLOGY

Ricardo García
Departamento de Matemáticas. E.U.P. Almadén. Universidad de Castilla-La Mancha. Plaza Manuel Meca, 13.400-Almadén, Ciudad Real, Spain.
rgarcia@pol-al.uclm.es

Angel Marín
Departamento de Matemática Aplicada y Estadística. E. T. S. I. Aeronáuticos. Universidad Politécnica de Madrid. Plaza Cardenal Cisneros, 3. 28.040, Madrid, Spain.
amarin@dmae.upm.es

Abstract In this paper an Urban Multimodal Interchange Design model is proposed, which considers simultaneously the interchange location problem in a main transit network and the design problem of a secondary transit network which feeds the interchanges, at strategical level. The problem of the design of these interchange facilities, such as the capacity and fares of parking lots is also considered at tactical level. The problem has been formulated by means of a bi-level model. At upper level the design decisions are considered and at lower level, the combined multimodal demand share. To solve this some heuristic algorithms based on the simulating annealing and greedy techniques have been proposed. Computational results in some test networks are presented.

Keywords: Interchange network design, Bi-level programming, Location, Combined multimodal demand share, Greedy heuristic, Simulated Annealing

1. INTRODUCTION

At the present, mobility continues growing, and people travel greater distances causing increase in travelling times and growing congestion. To reduce the amount of personal vehicle traffic, the use of inter-connecting

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 49-79.

high-quality public transport systems such as the metro and local rail networks must be encouraged. This can be achieved through the introduction of urban multimodal interchanges where it is possible to transfer to a different mode of transport. Secondary forms of public transport lines (e.g. tram and bus) as well as other methods of transport (car, bicycle, taxi, walking, etc.) act as feeders for the main lines. These types of trips using different modes are called combined mode trips.

In most large European urban areas, present parking policies sustain the use of public transport and try to reduce private traffic by promotion of the park'n ride trips and the development of superior transit modes. The European Union promoted this research in its Research Frameworks. In particular the Euritrans group was formed to present a project to the IV European Research Program Framework: Task 5.3: "Transition in multi-modal transport". In the Euritrans research project, [16], described the "macro methodology" to study the interchange design. These researches are also of preferential interest in Spain, whose Science Ministry and the Madrid Community are giving support to cover this research.

The proposed approach at opposed to with the classical ones consider simultaneously the traffic assignment and the interchange design. Some of the approaches to the problem have been done from the parking location and design points of views.

One of the first papers in parking design was proposed by Florian and Los ([9]). More recently, Carrese et al. ([5]) consider the problem with inclusion of two decision levels: parking location and demand allocation. A verification of different scenarios for parking demand satisfaction was used to integrate both levels.

García and Marín ([13], [14]) have developed an equilibrium model with combined modes in order to evaluate the demand at the interchange in function of macro characteristics of the system of transport such as capacity of the interchange, mean distances at the interchanges, the level of service in the public transport system and the level of congestion of the road network.

Coppola in [6] also considers a model of mode/parking choice with elastic parking demand. In relation with the parking location two recient references are: Nickel et al. ([22]) present approaches based on network design models, and Mesa and Ortega ([21]), which consider the location of park and ride station areas comparing private travel time with that using a combination of modes.

Oppenheim ([23]) describes the network design process as a bi-level programming problem, the upper level problem being the supply problem (in our case a transport supply problem) and the lower level problem

being the demand problem of these services. The system designer (leader) designs the system of transport taking into account how the users (followers) employ it. The leader is assumed to have knowledge of the responses of the followers. This situation is known in game theory as a Stackelberg game.

A classical taxonomy of the network design problems (NDP) is the continuous network design (CNDP) and the discrete network design problem (DNDP). The continuous network design problem takes the network topology as given and is concerned with the parametrization of the network. The discrete network design problem is concerned with the topology of the network such as the selection of optimal facility locations (interchanges in our case) .

Leblanc ([18]) made a first approach to the DNDP. Poorzahedy and Turnquist ([25]) presented a typical heuristic algorithm for solving the integer programming DNDP. Various variations of the DNDP were also formulated and solved by Boyce and Janson ([4]) and Chen and Alfa ([7]). Yang and Bell ([26]) provide a comprehensive review of the models and algorithms for NDP.

CNDP selects the network capacity assignment of the the transport network in order for the system optimum to minimize the total travel time in the network. Abdulaal and LeBlanc ([1]) formulated the CNDP under user equilibrium as a bi-level programme and Hook-Jeeves heuristic is used to solve it.

Under the name of network design (so they are not studied in this paper) other urban transportation planning applications may be considered. For instance, the problem of allocating a given bus fleet between the transit lines of a given urban transportation system. These problems, as with the interchange network design here studied, are formulated using bi-level programming. Han and Wilson ([15]) and Leblanc ([19]).

Generally, bi-level programming is very difficult to solve in practice due to its inherent nonconvexity and nonsmoothness. For these models the use of probabilistic search methods, such as simulated annealing is recommended. These methods are appropriate when a global optimum is sought.

Anandalingam et al. ([2]) illustrate the use of simulated annealing in solving bi-level linear programming problems. Friesz et al. ([10],[11]) applied simulated annealing to solve the bi-level programming formulation of the continuous network design problem. Meng et al. ([20]) present a bi-level programming model CNDP, where that is transformed into a single level optimization problem by using a marginal function tool.

In this paper, first the interchange design problem is presented. After an equilibrium model with combined modes and the lower level model of a

bi-level model are introduced. The equivalence between the equilibrium conditions and the lower level problem is shown in an appendix. The upper level problem is the interchange design model, that joined to the previous lower level model define a bi-level model. This is followed by the description of the Greedy and Simulated Annealing heuristics adaptation to this model. Finally some test results and conclusions are presented.

2. INTRODUCTION TO THE URBAN INTERCHANGE DESIGN PROBLEM

In this paper it is assumed that two transit systems are operated (see Figure 4.1). The main transit network (transit mode) offers fast line travel for longer trips. The secondary transit network (access mode), on the other hand, permits access trips to the transit network. The transit mode, for example, may be regional rail or metro, and the access mode is defined by a local bus.

We dealt with the problem of stations location on the physical transit network. These stations with special facilities, such as parking lots, are called urban multimodal interchanges. If we open a new station then several rail lines could stop at it. We assume that a model to design lines of transit and their frequencies is available, and it produces travelling time on the new infrastructure of the main transit network.

In this work the following three aspects have been considered:

- The location of the interchanges in the main transit network.
- The design of the access mode to the interchange. This problem has not dealt with the way of designing a service network where routes, frequencies, etc. are defined. It has dealt with the evaluation of specific designs of the secondary transit network in the context of the general transport network.
- The design of facilities at these interchanges. In this model the capacity and the fares of the parking lots of the interchanges are considered.

The general problem of designing urban interchanges can be divided into two sub-problems:

- The network equilibrium -subproblem: that is to say how the users choose the mode of transport, the interchange and the route on the system of transport. To provide a modelling framework for the consideration of combined mode trips, it is important to decide which choices are modelled by the mode choice model (demand)

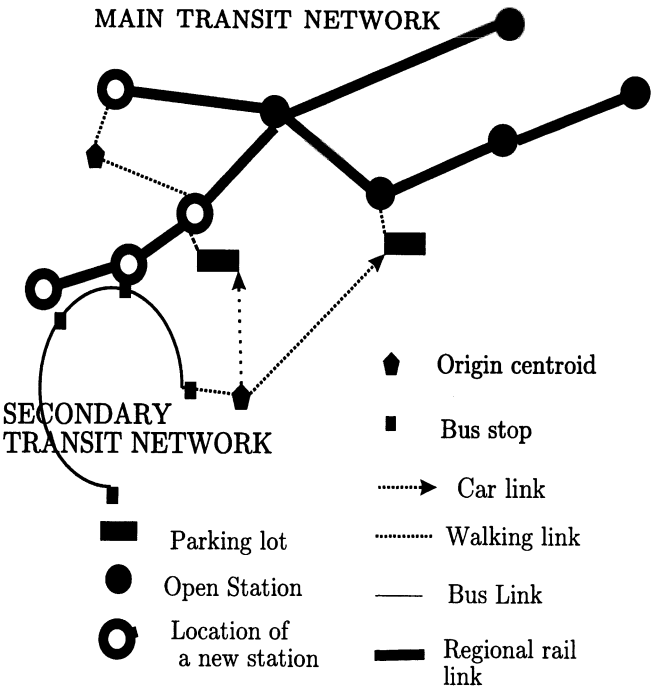

Figure 4.1 Hierarchic transit system.

and which by the route choice on the transport network (supply). In this stage an equilibrium between supply and demand which leads to a modal split and traffic assignment is achieved.

- The design-subproblem: that is the choice of the location where the interchange will be established, the choice of the dimension of its parking lot and fares, and the type of service of the secondary transit to the interchange.

Figure 4.2 shows the interrelation between both problems. In the network equilibrium phase, the users (demand) evaluate the attractiveness of the design system of transport, and choose the mode of transport and interchange. The decision-maker, usually an authority, evaluates the cost-benefit of the effective state of the system, which depends on the demand at the interchanges. The decision-maker generates a new plan on intervention in the system. This consists of the making of two types of decisions.

The first ones are strategical decisions about the topology of the main transit network and the choice of the type of feeders. The second ones

are tactical about the dimensioning and the fare policy of the parking lots.

The demand allocation problem is formulated by means of an equilibrium model with combined modes, where the demand decisions are modelled by a nested logit model and the transportation network has been simplified to a transportation model. The model describes the user behaviour as a function of the generalized transportation costs.

The design network problem is formulated by means of an integer mixed non-linear optimization model. Both models have been integrated using bi-level mathematical programming.

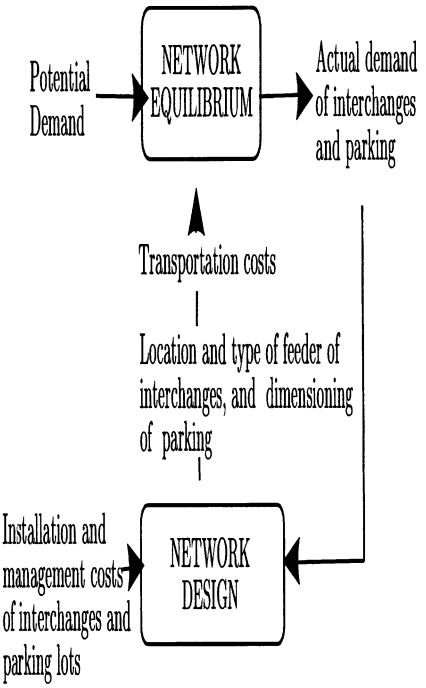

Figure 4.2 Urban Interchange Design Problem

3. NETWORK EQUILIBRIUM MODEL WITH COMBINED MODES.

In order to describe the equilibrium model, it has been divided into two parts. The first one is demand modelling and the second one is supply modelling.

3.1 DEMAND MODELLING

We assume that a potential demand for our planning period of trips between origin i and destination j is known. We denote as $\{\bar{g}_\omega\}_{\omega \in W}$ the origin-destination matrix, where $\omega = (i, j)$ is a pair origin-destination and W is the set of pairs O/D. We have used a nested-logit model to desegregate the demand. Ben-Akiva and Lerman (1985).

We have considered the following hierarchical structure:

1. Mode choice. The first choice is the mode of transport. We assume that the users choose between the following four alternatives of transport; where (s), s∈ {a, b, c, d}, signifies the mode of transport:

   (a) Park'n ride where the first part of the trip is taken by private car; then the car is parked, and the trip is completed by taking a transit mode and by walking to the final destination.

   (b) Transit with access by bus. In this alternative the trips are taken by transit and the access to main transit network is taken by local bus (access mode).

   (c) Transit with access by walking or cycling. In this alternative the trips are taken by transit mode and the types of access to the transit network are walking or cycling.

   (d) Others. This alternative clusters the other available forms of transport which do not use the interchanges, such as: car, motorcycle, bus, etc.

   A logit function G^k gives the proportions of trips taken on each mode according to the formula

$$G^k_\omega(\mathbf{U}_\omega) = \frac{\exp\{-(\alpha^k + \beta_1 U^k_\omega)\}}{\sum_{k' \in \{a,b,c,d\}} \exp\{-(\alpha^{k'} + \beta_1 U^{k'}_\omega)\}}, \qquad \omega \in W, \quad k \in \{a, b, c, d\} \tag{4.1}$$

   where U^k_ω is the user's perception of the generalized cost of travelling between the O-D pair ω by mode k, that corresponds to a user optimal route choice of the transport network, $\{\mathbf{U}_\omega\}$ is the vector of generalized costs for all the modes present, and α^k, β_1 are parameters.

2. Interchange choice. In order to determine the proportion of park'n ride trips travelling between O-D pair ω through the interchange t

an additional logit type model is introduced. (Fernández et al. in [8])

$$G^a_{\omega,t}(\mathbf{U}^a_\omega) = \frac{\exp\{-(\alpha_t + \beta_2 U^a_{\omega,t})\}}{\sum_{t' \in I_\omega} \exp -\{(\alpha_{t'} + \beta_2 U^a_{\omega,t'})\}}, \quad \omega \in W, \quad t \in I_\omega \tag{4.2}$$

The utility for the alternative (a), $\mathbf{U}^a_\omega$, is computed as a "log-sum" of the utilities through each used interchange, $\mathbf{U}^a_{\omega,t}$,

$$U^a_\omega = \frac{-1}{\beta_2} \log \left(\sum_{t' \in I_\omega} \exp -\{(\alpha_{t'} + \beta_2 U^a_{\omega,t'})\} \right), \tag{4.3}$$

where I_ω is the set of available interchanges for the demand $\omega \in W$. The parameters α_t represent the relative attractiveness of transfer node t, due to factors not included in the user's generalized cost perception $\mathbf{U}^a_{\omega,t}$ such as: security, protection, comfort, the means of buying tickets, etc, and β_2 ponders the importance of that cost perception in the transfer node choice decision process. For the alternatives (b) and (c) it is assumed that the users choose the interchange that minimizes the total time (or generalized cost) of the journey.

3. parking type choice. A conclusion of the work of Hunt and Teply (1993) is that when modelling parking choice using a logit model, it is appropriate to use the nested form with a hierarchical structure that acknowledges the effects of the greater similarities among off-street parking locations and on-street facilities. We will refer to each form as 1 and 2 respectively. Another logit function has been introduced in order to model this choice.

$$G^{a_s}_{\omega,t}(\mathbf{U}^a_{\omega,t}) = \frac{\exp\{-\left(\alpha^s{}_t + \beta_3 U^{a_s}_{\omega,t}\right)\}}{\sum_{s' \in \{1,2\}} \exp\{-(\alpha^{s'}_t + \beta_3 U^{a_{s'}}_{\omega,t})\}}, \tag{4.4}$$
$$s \in \{1,2\}, \ t \in I_\omega, \ \omega \in W$$

The modal split is given by the expression $g^k_\omega = G^k_\omega(\mathbf{U}_\omega)\bar{g}_\omega, \quad k \in \{a,b,c,d\}$, the number of users of the park'n ride mode through the interchange t is given by $g^a_{\omega,t} = G^a_{\omega,t}(\mathbf{U}^a_\omega)g^a_\omega$, and the number of travellers that use the park'n ride with parking type s at the interchange t is $g^{a_s}_{\omega,t} = G^{a_s}_{\omega,t}(\mathbf{U}^a_{\omega,t})g^a_{\omega,t}$. The above expressions can be written in a shortened form:

$$\mathbf{g} = \Phi(\mathbf{U}, \mathbf{g}) \tag{4.5}$$

The demand share is given by a function, Φ, of the generalized cost which is estimated by traffic assignment, which is also a function of the demand share.

3.2 SUPPLY MODELLING

The transport supply is considered by a multimodal transportation network which is defined by a classical road network, a transit network and transfer nodes between networks, called urban multimodal interchanges. We assume that the users choose the route into the road network according to the first principle of Wardrop. For more details see Patriksson ([24]).

We assume that the demand level does not significantly affect the generalized costs in the modes of transport $\{b, c, d\}$, because the potential zones to locate the interchanges are placed out of congestion zones (in suburban zones) and the first part of the combined trip is not affected by congestion.

These considerations lead us to consider as a good approximation that the generalized costs U_ω^b, U_ω^c, and U_ω^d are independent of the demand. We assume that these values are known in function of the network topology. For example they can be computed using a traffic assignment model in a given multimodal network. U_ω^a will be considered as a function of the demand, and therefore we have introduced a transfer cost, which depends on the parking lot capacity and the fares to use the parking lot. Other factors like parking and walking times, etc. are taken into account in the transfer cost parameters.

The transfer cost function is defined for the two types of parking: off-street and on-street, and it may be of the follow formulation, for each interchange t:

$$c_t^s(f, u, v) = v_t^s + B_t^s \left(\frac{f}{u_t^s}\right)^{n_s} \quad s \in \{1, 2\} \tag{4.6}$$

where c_t^s is the cost of parking with a parking flow, f. n_s is a positive parameter, u_t^1 is the capacity of the parking lot t, u_t^2 is the capacity of the "neighbouring" parking zone at the interchange t. The parameter u_t^2 is considered fixed and u_t^1 is a design variable. The parameter v_t^1 represents the monetary cost of the parking lot t by means of the formula $v_t^1 = \tilde{v}_t^1 + \mu d_t$ where $\tilde{v}_t^1$ is the parking cost and d_t is the walking time from the parking lot to interchange, and μ is a parameter which transforms time cost into monetary cost. The parameters v_t^2, d_t and B_t^s may be calibrated by using survey data.

The number of users in the parking lots is computed by

$$g_t^{a_s} = \sum_{\omega \in W_t} g_{\omega,t}^{a_s}, \quad t \in I, \ s \in \{1, 2\}, \tag{4.7}$$

where I is the set of the interchanges, $W_t = \{\omega \in W : t \in I_\omega\}$, and the vehicular flows in the parking lots are computed by

$$g_t^s = \frac{g_t^{a_s}}{\tau}, \tag{4.8}$$

where τ is the car occupancy rate mean; g_t^s represents the number of vehicles in the parking lot of the interchange t ($s = 1$), and the number of vehicles parked on the street ($s = 2$).

The capacity constraints of the type $g_t^s \leq u_t^s$ for all t and $s \in \{1, 2\}$ have not been considered explicitly, but they are taken into account by means of the generalized cost in each parking zone. A value of the demand close to the capacity makes up a large value of the generalized cost and the demand will be forced to satisfy the capacity constraint.

The generalized cost for the combined mode a_s, for the O-D pair ω via parking in the interchange t will be computed as

$$U_{\omega,t}^{a_s}(g_t^s) = c_t^s(g_t^s) + \bar{U}_{\omega,t}^{a_s} \tag{4.9}$$

where $\bar{U}_{\omega,t}^{a_s}$ is the generalized cost of transport at equilibrium in the road network union transit network.

3.3 LOWER LEVEL MODEL

The expression (4.9) shows that the generalized cost for the alternative park'n ride depends on the demand and the design variables: capacity and fares of the parking lots. We have assumed that $\mathbf{U}^a, \mathbf{U}^b, \mathbf{U}^c$ do not depend on the demand but depend on the design variables: location of the interchanges and the type of feeders used.

In Figure 4.3 five available locations for the interchanges in relation with a main public transportation line, joined with some demand and some feeder secondary lines are represented. The objective of the figure is to show graphically the relations between the design variables and the generalized costs.

The black points represent interchanges already located. The pentagon, the origin of a demand. The arrows from this represent the modes car, bicycle and walking used by the demand to arrive at interchanges. The continuous closed curves represent the secondary bus lines which feed the interchanges. In the figure the bus lines associated to origin of a generic demand are represented.

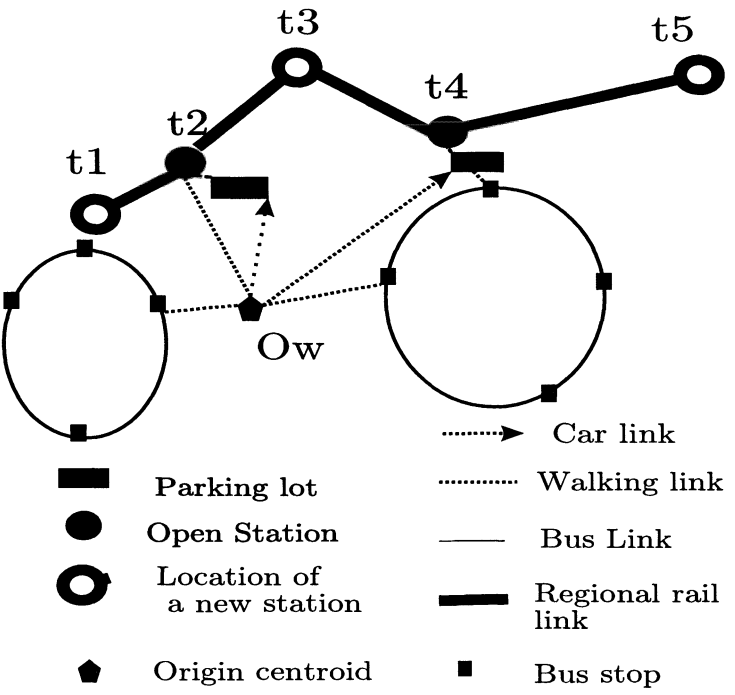

Figure 4.3 Generalized costs versus design variables.

For the given demand ω with origin in O_ω and destination in D_ω, we go on to consider the dependence of the generalized cost in relation with the design variables.

The $\mathbf{U}^a_\omega$ is the cost of a user of the park'n ride mode, which may use the interchange t_2 or t_4, and his decision depends on the interchange location variable, $\mathbf{y}$, and the interchange parking capacity and fare; $\mathbf{u}, \mathbf{v}$: $\mathbf{U}^a_\omega(\mathbf{y}, \mathbf{u}, \mathbf{v})$. In this case, $I_\omega = \{t_2, t_4\}$.

The $\mathbf{U}^b_\omega$ is the cost of the generic user of an interchange feeder line, which depends on the decision to install the bus line, $\mathbf{z}$. This dependence is a function of the line frequency, etc., and of course, of the decision to locate the interchange t_4: $\mathbf{U}^b_\omega(\mathbf{y}, \mathbf{z})$

The $\mathbf{U}^c_\omega$ is the cost assuming that the user is walking or travelling by bicycle to the station. He only has access to the interchange t_2: $\mathbf{U}^b_\omega(\mathbf{y})$. And finally, $\mathbf{U}^d_\omega$ is the cost of the user that uses a car to arrive at his/her destination. In this case the generalized cost is not directly dependent on the design decision variable.

If we denote as $\mathbf{x}$ the design variables then we can express the generalized cost as a function of the demand and the design variables, i.e $\mathbf{U}(\mathbf{g}, \mathbf{x})$. The explicit expression of $\mathbf{U}(\mathbf{g}, \mathbf{x})$ is defined in (4.6)-(4.9). This

relationship and (4.5) lead us to the first formulation of the equilibrium conditions by means of the following fixed point formulation:

$$\mathbf{g} = \Phi(\mathbf{U}(\mathbf{g}, \mathbf{x}), \mathbf{g}) = \Gamma(\mathbf{g}, \mathbf{x}) \tag{4.10}$$

Under the constraints for the parameters $\beta_j > 0$, and $\beta_1 < \beta_2 < \beta_3$ we have proved (see appendix) that the equilibrium condition can be stated as an optimization model (lower level problem) as follows

$$\mathbf{LLP}(\mathbf{x}): \quad \min_{\mathbf{g}} T(\mathbf{g}, \mathbf{x}) = \sum_{\omega \in W} \left[\sum_{s \in \{1,2\}} \sum_{t \in I_\omega} \left(\int_0^{g_t^s} c_t^s(s, \mathbf{x}) ds + C_{\omega,t}^{a_s} g_{\omega,t}^{a_s} \right) + \sum_{k \in \{b,c,d\}} U_\omega^k g_\omega^k \right] + G(\mathbf{g}) \tag{4.11}$$

Subject to:

$$\sum_{k \in \{a,b,c,d\}} g_\omega^k = \bar{g}_\omega, \qquad \omega \in W \tag{4.12}$$

$$\sum_{t \in I_\omega(x)} g_{\omega,t}^a = g_\omega^a, \qquad \omega \in W \tag{4.13}$$

$$\sum_{\omega \in W_t} g_{\omega,t}^{a_s} = g_t^{a_s} \qquad t \in I,\ s \in \{1,2\} \tag{4.14}$$

$$g_{\omega,t}^{a_1} + g_{\omega,t}^{a_2} = g_{\omega,t}^a \qquad \omega \in W_t,\ t \in I \tag{4.15}$$

where $G(\mathbf{g})$ is stated as follows:

$$\begin{aligned} G(\mathbf{g}) &= (1/\beta_1) \sum_{k \in \{a,b,c\}} \sum_{\omega \in W} g_\omega^k (\ln g_\omega^k - 1 + \alpha^k) - (1/\beta_2) \sum_{\omega \in W} g_\omega^a (\ln g_\omega^a - 1) \\ &+ (1/\beta_2) \sum_{\omega \in W} \sum_{t \in I_\omega} g_{\omega,t}^a (\ln g_{\omega,t}^a - 1 + \alpha_t) - (1/\beta_3) \sum_{\omega \in W} \sum_{t \in I_\omega} g_{\omega,t}^a (\ln g_{\omega,t}^a - 1) \\ &+ (1/\beta_3) \sum_{\omega \in W} \sum_{t \in I_\omega} \sum_{s \in \{1,2\}} g_{\omega,t}^{a_s} (\ln g_{\omega,s}^{a_s} - 1 + \alpha_t^s) \end{aligned}$$

$G(\mathbf{g})$ is the term of the objective function that corresponds to the demand share obtained by a nested logit model. The logit parameters can be calibrated with the methodology developed in García and Marín (1998).

From the above results, we arrive at the conclusion of the equivalence between the equilibrium conditions (1.7) with the LLP(x), that it may be expressed in a shortened form by:

$$\mathbf{LLP}(\mathbf{x}): \qquad \min_{\mathbf{g} \in \Omega(x)} T(\mathbf{g}, \mathbf{x}),$$

where $\Omega(\mathbf{x})$ denotes the feasible set of LLP(x).

4. URBAN MULTIMODAL INTERCHANGE DESIGN PROBLEM

The design problem must take strategical and tactical design decisions. At strategic level, the location of the interchanges, $\mathbf{y}$, and design of the secondary bus feeder lines, $\mathbf{z}$ are studied. At tactical level, the interchange parking capacity, $\mathbf{u}$, and the parking fare, $\mathbf{v}$ must be decided. We denote as $\mathbf{x}$ the set of design variables associated with the decisions of the upper level. $\mathbf{x} := (\mathbf{y}, \mathbf{z}, \mathbf{u}, \mathbf{v})$.

We have used an objective function which is defined as the difference between the cost and the benefit in the transport system. To define the objective function all economic and non-economic factors are changed into monetary worth. This difficulty leads us to take great care with the calibration of the parameters of the model.

As benefit, we consider a single decision-maker, normally an authority, which controls the potential sites to locate the interchanges, denoted $t = 1, \ldots, m$. We assume that the decision-maker has a criterion for the evaluation of the level of service. We consider that this consists of two components, first a measure of the effective state of the transport system, which depends on the demand variable $\mathbf{g}$. Second, it also depends on the parking fare, $\mathbf{v}$. This relation will be denoted as $B(\mathbf{g}, \mathbf{v})$.

The cost has three components: the location cost, $L(\mathbf{y})$, which is a fixed cost of opening an interchange at the site t; the cost of the parking installation and management, $P(\mathbf{u})$, which depends on its capacity; and the third one is the design cost of the secondary transit networks, $R(\mathbf{y}, \mathbf{z})$, which is a function of the strategic design variables.

The mathematical formulation of the design problem considers that each decision has its own constraint set and the strategic variables are binary, $\mathbf{y}$, and integer, $\mathbf{z}$, meanwhile, the tactical decisions, $\mathbf{u}, \mathbf{v}$ are continuous and non negative.

The interchange design model is the following:

$$\mathbf{ULP}(\mathbf{g}): \quad \min_{\mathbf{x}} \Psi(\mathbf{x}, \mathbf{g}) = L(\mathbf{y}) + R(\mathbf{y}, \mathbf{z}) + P(\mathbf{u}) - B(\mathbf{g}, \mathbf{v})$$

$$\text{Subject to:}$$

$$\mathbf{y} \in Y \subset \{0,1\}^m$$

$$\mathbf{z} \in Z \subset \mathcal{Z}^m$$

$$(\mathbf{y}, \mathbf{z}) \in S \subset \{0,1\}^m \times \subset \mathcal{Z}^m$$

$$\mathbf{u} \in U \subset \mathcal{R}^m$$

$$\mathbf{v} \in V \subset \mathcal{R}^m$$

where $\mathbf{y}$ is a binary vector which components are equal to 1 if an interchange is located at site t, and $\mathbf{z}$ is an integer vector, which components represent a selected type of feeder bus line. In this formulation, the sets Z and Y represent other investment constraints such as budget constraints. The set S takes into account the interactions between the problems of location of interchanges and design of the feeder lines. This means that if we open an interchange, the transit line must stop at it. Also, the sets U and V represent resource assignment constraints.

We can integrate the network equilibrium model and the design network model by means of a bi-level model. In the upper level the decision-maker locates and designs the interchange facilities. In the lower-level, the users choose the mode of transport, the interchange and the type of parking in function of the designed facilities in the upper level.

The bi-level model (BLM) is formulated as follows:

$$\begin{aligned} \mathbf{ULP(g)}: \quad & \min_{\mathbf{x}\in X} \Psi(\mathbf{x},\mathbf{g}) \\ \mathbf{LLP(x)}: \quad & \mathbf{g} = \text{Arg} \ \min_{\mathbf{q}\in\Omega(x)} T(\mathbf{q};\mathbf{x}) \end{aligned}$$

where the upper feasible set has been expressed in a shortened form as $\mathbf{x} \in X$.

5. HEURISTIC APPROACHES FOR NETWORK DESIGN PROBLEM

The strategical variables $(\mathbf{y},\mathbf{z})$ are associated with a problem of expansion of transit networks. An aspect of this problem is to determine the characteristics of the facilities of the interchanges. We deal with the capacity of the parking lots and their fares.

BLM solves both problems simultaneously, but it could also be used to solve them separately. If we fix a type of variables, $(\mathbf{y},\mathbf{z})$ or $(\mathbf{u},\mathbf{v})$, then a new bi-level model appears and this is used to determine the other variables.

In this section we have assumed that the tactical variables $(\mathbf{u},\mathbf{v})$ are fixed, and we deal with heuristic algorithms in order to solve the new bi-level model defined only with the strategical variables. This is stated as:

$$\begin{aligned} \min \Psi(\mathbf{y},\mathbf{z},\mathbf{g}) \\ \mathbf{y} \in Y \subset \{0,1\}^m \\ \mathbf{z} \in Z \subset \mathcal{Z}^m \\ (\mathbf{y},\mathbf{z}) \in S \subset \{0,1\}^m \times \subset \mathcal{Z}^m \\ \mathbf{g} = \arg \min_{\mathbf{q}\in\Omega(\mathbf{y},\mathbf{z})} T(\mathbf{q};\mathbf{y},\mathbf{z}) \end{aligned}$$

The implicit relationship between the demand at equilibrium and the variables $(\mathbf{y}, \mathbf{z})$ which define the topology of the transport network, defines an explicit function which is expressed by the function $\mathbf{g} = \Theta(\mathbf{y}, \mathbf{z})$. This is due to the fact that the lower lever problem is a strictly convex program where the feasible region is a compact set, and for any feasible value of the variables $(\mathbf{y}, \mathbf{z})$ there exists only one demand at equilibrium.

The resulting combined location-equilibrium problem can be compactly expressed as: $(\widehat{\text{BLM}})$

$$\begin{aligned} \min \bar{\Psi}(\mathbf{y}, \mathbf{z}) = \Psi\left(\mathbf{y}, \mathbf{z}, \Theta(\mathbf{y}, \mathbf{z},)\right) \\ \mathbf{y} \in Y \subset \{0,1\}^m \\ \mathbf{z} \in Z \subset \mathcal{Z}^m \\ (\mathbf{y}, \mathbf{z}) \in S \subset \{0,1\}^m \times \subset \mathcal{Z}^m \end{aligned}$$

5.1 ALGORITHMS FOR LLP

One major difficulty with this formulation is that the function $\Theta(\mathbf{y}, \mathbf{z})$ is not explicitly known, but it is implicitly defined by means of the equilibrium model (section 3.) There are two ways to compute the demand at equilibrium: one is by solving the optimization model LLP. To this end, we can adapt a type of Evans' algorithm, García and Marín ([13]), or embeding this algorithm in a simplicial decomposition scheme such as García and Marín ([12]). The other way is by using the fixed point formulation (4.10). The natural method to solve this system of equations is by using the functional iteration scheme

$$\mathbf{g}^{i+1} = \Gamma(\mathbf{g}^i, \mathbf{x})$$

If the sequence $\{\mathbf{g}^i\}$ converges towards a point $\hat{\mathbf{g}}$, then this point is the solution of the system of equations because Γ is a continuous function. An improvement of the functional iteration method is derived from the structure of the system of equations (4.10), that is:

$$\begin{aligned} &\Gamma_\omega(\mathbf{g}_\omega, \mathbf{g}_t, \mathbf{x}) = \mathbf{g}_\omega, \quad \forall \omega \text{ demand}, \\ &\sum_{\omega \in W_t} g_{\omega,t}^{a_s} = \gamma g_t^s, \quad \forall t \text{ interchange}, \end{aligned} \tag{4.16}$$

where Γ_ω is defined by the relationships (4.1)-(4.4).

If we know the level of service of the parking zones, which are the variables g_t^s at the interchanges t, we can break down the system into a $|W|$ small system of equations, one system of equations for each demand $\omega \in W$. Moreover, each small system can be solved using a functional

iteration scheme. This leads to an algorithm of the Gauss-Seidel type in order to solve the system (4.16). This method fixes all the variables in each iteration, with the exception of $\mathbf{g}_\omega$ and $\mathbf{g}_t$, to their current value, so the corresponding system is solved using the functional iteration method. If we choose cyclically the pair ω that is not fixed and the sequence generated by this method is convergent, then the limit point is the solution of the system of equations. Summarising the Gauss-Seidel approach:

Initialization. Let two integer numbers be n_W and n, and an initial point $\mathbf{g}^0$.

For $i \in \{1, \ldots, n\}$
{
  For $\omega \in \{1, \ldots, |W|\}$
  {
    Let $\mathbf{q}_\omega^0 = \mathbf{g}_\omega^i$ and $\mathbf{q}_t^0 = \mathbf{g}_t^i$
    For $j \in \{0, \ldots, n_W - 1\}$
    {
      $\mathbf{q}_\omega^{j+1} = \Gamma_\omega(\mathbf{q}_\omega^j, \mathbf{q}_t^j, \mathbf{x})$
      Compute $(q_t^s)^{j+1} = \frac{1}{\gamma}\left\{(q_{\omega,t}^{a_s})^j + \sum_{\omega' \in W_t, \omega' \neq \omega} (g_{\omega',t}^{a_s})^i\right\}$
    }
    Let $\mathbf{g}_\omega^{i+1} = \mathbf{q}_\omega^{n_W}$,
    Let $(g_t^s)^i = (q_t^s)^{n_W}$.
  }
  Let $(g_t^s)^{i+1} = (g_t^s)^i$.
}

5.2 GREEDY HEURISTICS FOR $\widehat{\text{BLM}}$

The idea of the forward greedy algorithm (FGA) is simple. Given a set of locations defined by means of the variable $\mathbf{y}$, the algorithm explores all possibilities of opening k new facilities, and the next location is one that gives the greatest decrease in $\min \bar{\Psi}(\mathbf{y}, \mathbf{z})$ value over them. Moreover, once a station is located it is kept throughout the algorithm. To state FGA, we first define the forward k-neighbourhood:

$$\mathcal{N}_k^+(\mathbf{y}) = \left\{ \mathbf{y}' \in Y : \sum_{j=1}^{m} \left| y_j' - y_j \right| \leq k \text{ and } \mathbf{y} \leq \mathbf{y}' \right\} \text{ for } \mathbf{y} \in Y$$

The set $\mathcal{N}_k^+(\mathbf{y})$ gives all locations for k new stations at the current location $\mathbf{y}$. Moreover, we must fix the type of secondary transit network to feed these new stations. This leads us to consider an extended k-

neighbourhood:

$$\mathcal{S}_k^+(\mathbf{y}) = \left\{(\mathbf{y}', \mathbf{z}') \in \mathcal{N}_k^+(\mathbf{y}) \times Z : (\mathbf{y}', \mathbf{z}') \in S\right\} \text{ for } \mathbf{y} \in Y$$

The FGA is the following:

Step 1 Initialization. Find an initial solution $(\mathbf{y}^0, \mathbf{z}^0)$. Initialize $i = 1$.

Step 2 Let $(y^i, \mathbf{z}^i) = \arg \min \left\{\bar{\Psi}(\mathbf{y}, \mathbf{z}) : (\mathbf{y}, \mathbf{z}) \in \mathcal{S}_k^+(\mathbf{y}^{i-1})\right\}$.

Step 3 If $\bar{\Psi}(\mathbf{y}^i, \mathbf{z}^i) \geq \bar{\Psi}(\mathbf{y}^{i-1}, \mathbf{z}^{i-1})$, then stop. $(\mathbf{y}^{i-1}, \mathbf{z}^{i-1})$ is an FGA solution.

Step 4 If $\mathcal{S}_k^+(\mathbf{y}^i) = \{\emptyset\}$ then $(\mathbf{y}^i, \mathbf{z}^i)$ is an (FGA) solution. Otherwise, let $i = i + 1$ and return to Step 2.

This type of greedy algorithm has a second version. Assuming that at the beginning all the facilities are located, one can drop the facilities so that their level of service will be deficient. We have called this a backward greedy algorithm (BGA). The algorithm is obtained by interchanging the set $\mathcal{S}_k^+(\mathbf{y})$ in FGA for

$$\mathcal{S}_k^-(\mathbf{y}) = \left\{(\mathbf{y}', \mathbf{z}') \in \mathcal{N}_k^-(\mathbf{y}) \times Z : (\mathbf{y}', \mathbf{z}') \in S\right\} \text{ for } \mathbf{y} \in Y$$

where

$$\mathcal{N}_k^-(\mathbf{y}) = \left\{\mathbf{y}' \in Y : \sum_{j=1}^{m} \left|y_j' - y_j\right| \leq k \text{ and } \mathbf{y} \geq \mathbf{y}'\right\} \text{ for } \mathbf{y} \in Y$$

5.3 A K-INTERCHANGE ALGORITHM FOR BLM

Two main disadvantages of both BGA and FGA are that the facilities are selected only once and the initial benefit could change when the algorithm progresses, and the computational time spent to find the optimal solution on $\mathcal{S}_k^{\pm}(\mathbf{y})$ is very intensive. These algorithms explore all the elements of the k-neighbourhood. We have relaxed this step to find one solution which reduces the value of the objective function. The first problem has been avoided by means of considering a new k-neighbourhood

$$\mathcal{S}_k(\mathbf{y}) = \mathcal{S}_k^+(\mathbf{y}) \cup \mathcal{S}_k^-(\mathbf{y}) \text{ for } \mathbf{y} \in Y$$

This modification allows the opening of a new station or the dropping of an existing one in each iteration.

The k-Interchange Algorithm (IA) is the following:

Step 1 Initialization. Find an initial solution $(\mathbf{y}^0, \mathbf{z}^0)$. Initialize $i = 1$.

Step 2 Given a feasible point $(y^{i-1}, \mathbf{z}^{i-1})$, if there is a point $(\mathbf{y}', \mathbf{z}') \in \mathcal{S}_k(\mathbf{y}^{i-1})$ with $\bar{\Psi}(y', \mathbf{z}') < \bar{\Psi}(y^{i-1}, \mathbf{z}^{i-1})$, then let $(y^i, \mathbf{z}^i) = (y', \mathbf{z}')$, and let $i = i + 1$ and repeat the iteration. Otherwise stop, $(y^{i-1}, \mathbf{z}^{i-1})$ is a $k-$ interchange solution.

5.4 A SIMULATED ANNEALING ALGORITHM FOR $\widehat{\text{BLM}}$.

The FGA, BGA, and IA stop when they find a locally optimal solution relative to the chosen neighbourhood. A possible solution is to run these heuristics many times with randomly chosen starting points. A different way to try to obtain a global optimum is using a simulated annealing algorithm (SAA). This method occasionally allows obtain better local solutions. In particular, given a solution $(\mathbf{y}, \mathbf{z})$, we select a candidate solution $(\mathbf{y}', \mathbf{z}')$ from the neighbourhood $\mathcal{S}_k(\mathbf{y})$. The solution is accepted if it is cost improving. Otherwise, $(\mathbf{y}', \mathbf{x}')$ may be accepted with the probability

$$\exp - \left(\frac{\bar{\Psi}(\mathbf{y}', \mathbf{z}') - \bar{\Psi}(\mathbf{y}, \mathbf{z})}{K\mathcal{T}} \right) \tag{4.17}$$

where $\mathcal{T}$ is a positive constant, K is a constant which depends on the system cost; and it is rejected with the complementary probability. $\mathcal{T}$ is called the temperature of the process, and it plays the role of defining large or small moves for the optimization variables.

The empirical efficiency of SAA depends on the neighbourhood structure and the selection of parameters for SAA, but there is no general rule for selecting the parameters that characterize them. We present the sets of rules that have given us the best set of solutions in most of the computational experiments.

(a) In practice, the temperature parameter is decreased. We denote the initial and final values of the temperature as $\mathcal{T}_0$ and $\mathcal{T}_f$ respectively. The rule for decreasing this parameter is $\mathcal{T} = \alpha\mathcal{T}$ where α is an annealing parameter such that $0 < \alpha < 1$. When the temperature is low enough (less than $\mathcal{T}_f$) the search finishes. We have used the values of $\mathcal{T}_0 = 1.$, $\mathcal{T}_f = 0.1$ and $\alpha = 0.95$

(b) When the system is at equilibrium the temperature of the system must be changed. Usually, this equilibrium is characterized by two parameters: NAC= maximum number of accepted configurations.

$NRC=$ is the maximum number of consecutive rejected configurations. To define the parameters NAC and NRC the size of the problem is usually taken into account. We have used the following values in computational experiments: $NAC = [0.85m]$, where m is the dimension of the vector $\mathbf{y}$ (potential sites to locate the interchanges), and $[\cdot]$ means 'integer part of'; and $NRC = [2m]$.

(c) SAA, in its theoretical formulation, converges with probability one to a global minimum under the assumption that the each set in a neighbourhood $\mathcal{S}_k$ is chosen with equal probability. This means of selecting a new element from the neighbourhood is of great theoretical interest , however, the experimental results have not been satisfactory. In the experiments we have systematically selected the components of vector $\mathbf{y}$. This procedure consists of two phases. In the first, the component of $\mathbf{y}$ whose value in the current solution are 0, are selected in a consecutive manner, so that the component selected is fixed at 1. When the last of these components is reached, the second phase begins. This consists of consecutively selecting components whose existent value is 1, and changing this value to 0. When all the components have been analized the phase changes.

If the component is fixed at a value of 1 this implies that the interchange is open, so it is necessary to decide upon the type of feeder, that is to say, to decide upon the value of the variable $\mathbf{z}$. The selection of the type of feeder is randomly effected in accordance with a certain distribution of probability.

(d) The probability that a worse solution than the current is accepted is defined as (4.17). Taking for $\mathcal{T}$ initially $\mathcal{T}_0 = 1$, the probability of accepting a new configuration in this initial stage is

$$\Pr = \exp - \left(\frac{\Delta c}{K}\right)$$

where Δc is the increment of the system cost: $c' - c^0$ where c' and c^0 are the costs of the new and current configuration respectively. If we consider that this probability must be a certain value, p, then K must be

$$K = \frac{-\Delta c}{\log(p)}$$

This relation is used at the initiation stage to calculate the value of the constant K. We have considered a probability of $p = 0.1$

which assumes a solution which is 5% worse than the cost of the initial solution c^0. In this case $\Delta c = 0.05\,|c^0|$ and we obtain:

$$K = \frac{-0.05\,|c^0|}{\log(0.1)} = 0.0217\,|c^0|$$

The adaptation of the SA method to BLM takes into account the above rules. The SAA is the following:

To summarize the SAA

Step 1 Initialization. Find an initial solution (y^0, z^0). Let c^0 be the cost. Define $\mathcal{T}_0$, $\mathcal{T}_f$, and α. Define K according to (d). Select the parameters NAC and NRC as in (b). Define the optimal solution $\mathbf{y}^{op} = \mathbf{y}^0$ and $\mathbf{z}^{op} = \mathbf{z}^0$. Actualize the counters: cNAC=0 and cNRC=0. Let $i = 0$

Step 2 Given $(\mathbf{y}^i, \mathbf{z}^i)$, select $(\mathbf{y}', \mathbf{z}') \in \mathcal{S}_k(\mathbf{y}^i)$ using (c). Let c' be the cost of the solution $(\mathbf{y}', \mathbf{z}')$.

Step 3 If $c' < c^i$, then $(\mathbf{y}^{i+1}, \mathbf{z}^{i+1}) = (\mathbf{y}', \mathbf{z}')$, and $c^{i+1} = c'$. If $c' < c^{op}$ then $(\mathbf{y}^{op}, \mathbf{z}^{op}) = (\mathbf{y}', \mathbf{z}')$ and $c^{op} = c'$.

Step 4 Otherwise, $(\mathbf{y}^{i+1}, \mathbf{z}^{i+1}) = (\mathbf{y}', \mathbf{z}')$ with probability $p = \exp -[(c' - c^i)/K\mathcal{T}]$, and $(\mathbf{y}^{i+1}, \mathbf{z}^{i+1}) = (\mathbf{y}^i, \mathbf{z}^i)$ with probability $1 - p$.

Step 5 If the solution $(\mathbf{y}', \mathbf{z}')$ has been accepted then $cNAC = cNAC + 1$, and $cNRC = 0$. Otherwise, $cNRC = cNRC + 1$. If $cNAC = NAC$ or $cNRC = NRC$ then decrease the temperature $\mathcal{T} = \alpha\mathcal{T}$, let $cNAC = 0$, and $cNRC = 0$.

Step 6 If $\mathcal{T} < \mathcal{T}_f$ stop. Otherwise $i = i + 1$, and return to Step 2.

6. COMPUTATIONAL EXPERIMENT

The objective of the computational experiment has been to compare the efficiency of the different heuristic methods.

We have randomly generated the test problems for this computational experiment. Figure 4.4 illustrates the procedure for this generation. Primarily, the procedure generates three circles of different radius. In circle A, the destination centroid is uniformly distributed. The circular sector obtained on removing circle A in circle B, randomly generates the different possibilities for location of the interchanges. In the circular sector which is defined upon removing circle B in circle C, the origin centroids are randomly located. Follow, the demand pairs are generated. Given

the pair $\omega = (O, D)$ (see Figure 4.4) we calculated the k minor distances of going from origin O to destination D via an interchange. These will be the only possible interchanges which can use demand ω. After the costs of the journey by each mode of transport are calculated, this is $C_\omega^a, C_\omega^b, C_\omega^c$ and C_ω^d. We have assumed that these costs are proportional to the euclidean distance (d_i) in each mode of transport. The proportionality constant plays the role of mean velocity in each mode of transport. For example, the combined trip in the drawing via interchange t, will be the sum of the time takes to travel the distance d_1, plus the time taken to travel distance d_2. The time taken in travelling the first part depends on whether the journey has been made by car, bus or on foot, while the time taken in distance d_2 will be the same for all combined trips, and depends on the mean velocity of the transit line. The cost of undertaking the journey by car will be proportional to the car distance d_3.

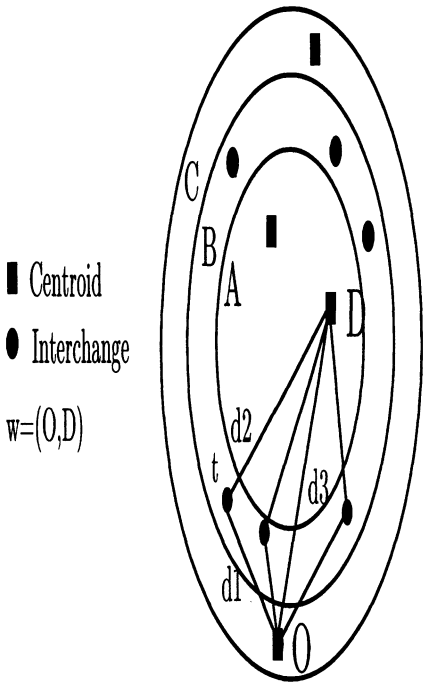

Figure 4.4 Illustration of a test network

We have considered three different ways of feeding each interchange. Each one of these means is moved using 3 different velocities of access to the interchange by bus mode. We have also assumed three different costs for implanting and making arrangements for this type of bus line.

In order to generate the parameters of the logit function, we have assumed two interchanges produce two different combined trips for the same demand, and if they have the same cost then they are equally attractive. This implies that all the logit parameters are the same as each other. We have taken $\alpha_t = 0$. On the other hand, we have assumed that the parking on-street and off-street only depends on the cost in each alternative. This leads to $\alpha_t^1 = \alpha_t^2$, and we have fixed these parameters at a value of 0. We have first set up a modal split of the demand before hand, independent of the transport costs, which has allowed us to estimate the parameters α^k. After, we have taken into account the costs and the weight least square estimation has been used in order to estimates β_1 and β_2, in such a way that the modal split taken as reference is as similar as possible to the prediction one effected by the logit model.

We have considered that the measure of the effective state of the transport system $B(\mathbf{g})$ depends on the demand variable $\mathbf{g}$. We have used the function:

$$B(\mathbf{g}) = \sum_{t=1}^{m} B_t(\mathbf{g}) + \varsigma \left(\bar{U}^d - \sum_{\omega \in W} g_\omega^d U_\omega^d \right), \tag{4.18}$$

where for each interchange t, B_t is defined by:

$$B_t(\mathbf{g}) = \varsigma \sum_{\omega \in W_t} \left(g_{\omega,t}^a + g_\omega^b \delta_{\omega,t}^b + g_\omega^c \delta_{\omega,t}^c \right),$$

where $\delta_{\omega,t}^k = 1$ if the journey ω by mode k uses the interchange t, and $\delta_{\omega,t}^k = 0$, otherwise.

This benefit function takes into account the economic benefits produced for the public transport system by users (that is to say the number of tickets sold), plus the social benefit measured by the overall reduction in emissions and energy consumption in private traffic for the same total demand. The parameter ς represents a pseudo-cost of the ticket. The parameter μ transforms the number of hours spent travelling by car to monetary units. The value $\bar{U}^d$ is the mean number of hours spent by the private car mode without intervention in the system of transport.

We have considered the costs of opening and management of the interchanges as

$$L(\mathbf{y}) = \sum_{t=1}^{m} f_t y_t$$

We have assumed that all parameters f_t have the same value.

The mathematical formulation of the design of feeders considers that the variable z_t belongs to a subset of integer values and each value is

Table 4.1 Size of the test problems

Test Network	$\|W\|^a$	m^b	CPU per LLP^c
NET1	300	25	0.29
NET2	300	50	0.26
NET3	300	100	0.24
NET4	500	100	0.38
NET5	1000	50	0.47
NET6	3000	25	3.12

[a] Number of demand pairs. [b] Number of available sites to locate the interchanges. [c] Mean CPU time spent in solving LLP

associated with a type of design to feed the interchange t. We have associated these variables with the function $R(\mathbf{y}, \mathbf{z})$ which gives the design cost of secondary network.

We have only considered that only three different types of feeders exist. This means that the variables z_t, type of feeder of interchange t, could only have the values $1, 2, 3$. We have assumed a function $C_R(\cdot)$ defined on $1, 2, 3$ and the cost of the design of the secondary transit network is given as

$$R(\mathbf{y}, \mathbf{z}) = \sum_{t=1}^{m} C_R(z_t) y_t$$

We have omitted the values, and the methodology to estimate the parameters in order to reduce the extent of this paper.

Clearly the amount of work per iteration in this algorithm depends on crucially on k (size of the neighbourhood), and for the heuristics to be fast we limit k to value 1. The hardware and software used have been a Pentium PC, speed 400 MHz, with FORTRAN 77 code.

The size of the test problems is shown in Table 4.1. The fourth column shows the CPU time spent in the solution of one LLP. This time depends on the number of pairs of demand and the number of iterations done by the Gauss-Seidel algorithm. We have observed that using $n_W = 3$ and $n = 4$ for the values of $\mathbf{u}$ given, the sequence generated $\{\mathbf{g}^i\}$ is convergent and the relative error $\frac{\|\mathbf{g}^{i+1} - \mathbf{g}^i\|}{\|\mathbf{g}^i\|}$ is of approximately 1.%. This algorithm is a good procedure for solving the LLPs. The main algorithm disadvantage is its not convergence when the capacity of the parking lots ($\mathbf{u}$) is very small with respect to the potential demand of parking. In this case the algorithm generates an oscillate sequence.

Table 4.2 Computational results

Test Network	FGA	BGA	IA	SAA
NET1	-473240.10[a]	-488721.5	-483845.2	-481622.3
	660[b]	882	136	4153
NET2	-466172.6	-485436.0	-469551.8	-472346.2
	1602	3624	288	10013
NET3	-635517.9	-630940.6	-608787.5	-627120.1
	4692	15648	1094	34438
NET4	-512012.4	-506577.9	-490027.4	-494691.6
	5430	15300	1547	46533
NET5	-412825.4	-437841.5	-395456.9	-404498.1
	2241	3564	513	33805
NET6	-466086.2	-471755.7	-469353.4	-474846.6
	741	696	104	9587

[a] Value of the objective function. [b] Number of evaluations of the objective function.

The experimental results are shown in Table 4.2. The conclusion is that the choice of an algorithm could depend on the size of the problem. For very intensive computational requirements we could use IA or SAA, and for medium or small requirements, it could be recommendable to use FGA or BGA. This affirmation is illustrated Figure 4.5. It shows that IA and SAA are better at the begining, with more iterations this trend is reversed.

Note that SAA achieves the best cost in few evaluations, but we have done a lot of iterations to check the possibilities of the algorithm.

7. CONCLUSIONS

A bi-level model has been defined as an approach to the Urban Multimodal Interchange Design Problem. At an upper level the transport system regulator's decisions are considered, and at a lower level the demand share decisions are considered.

At upper level two binary design variables are the interchange location and the feeder network decisions, both at strategic level. Meanwhile, at tactical level are considered the capacity and the fare of the interchange parking decisions.

The upper level must be informed about the demand share reactions in relation to its design decisions (in the context of a Stackelberg equilibrium). The lower level is a demand share equilibrium depending on

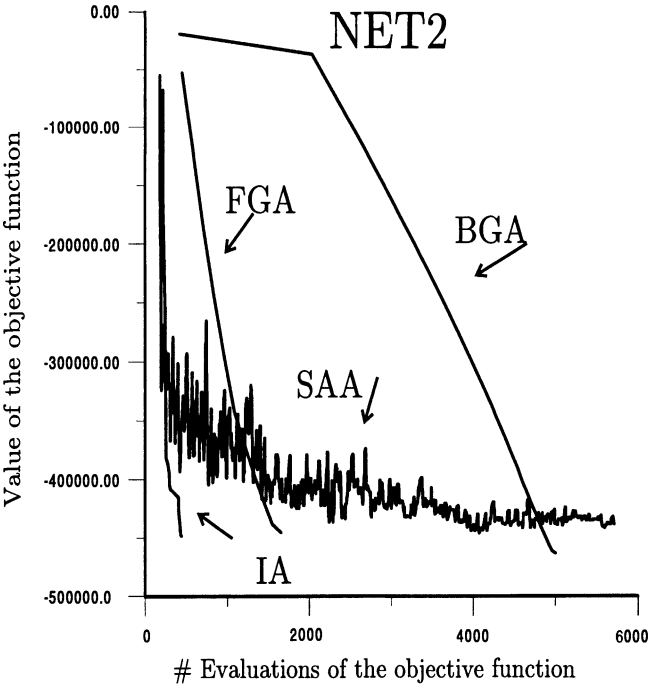

Figure 4.5 Progress of the algorithms.

design variables, where the traffic decisions (with inclusion of combined modes) have been characterized by Nested Logit Models.

The equivalence between the equilibrium conditions for the lower level and an optimization model has been proved.

We have proposed several heuristic algorithms based on greedy techniques and a simulated annealing method. These procedures require the solution of LLP($\mathbf{x}$), which is obtained using a Gauss-Seidel approach.

The proposed methodology is computationally expensive, but it yields solutions for realistics medium-sized problems.

Acknowledgment

This research is based on work supported by Ministerio de Educación y Cultura of the Spanish Government under the grant TRA99-1156-C02-01. The authors wish to thank the anonymous referees for their suggestions and helpful comments on this paper.

References

[1] Abdulaal, M. and LeBlanc, L.J. (1979), Continuous Equilibrium Network Design Models. Transportation Research, 13B, pp. 19–32.

[2] Anandalingam G., Mathieu P., Pittard C.L., Sinha, N. and Vernekar A. (1988), Nontraditional Search Technique for solving Bi-Level Linear Programming Problems. University of Pennsylvania., Philadelphia.

[3] Ben-Akiva, M. and Lerman, S.R.(1985), Discrete Choice Analysis, MIT Press Serie in Transportation Study.

[4] Boyce, D.E. and Janson, B.N. (1980), A Discrete Transportation Network Design Problem with Combined Trip Distribution and Assignment. Transportation Research, 14B, pp. 147–154.

[5] Carrese, S., Gori, S. and Picano, T. (1996), Relationship between Parking Location and Traffic Flows in Urban Areas. Advanced Methods in Transportation Analysis Transportation Science, Lucio Bianco and Paolo Toth, Eds. Springer Verlag, pp. 183-214.

[6] Coppola, P. (1998), A Joint Model of Mode/Parking Choice with Elastic Parking Demand. 6th EURO Working Group Meeting on Transportation, Göteborg.

[7] Chen, M. and Alfa, A.S. (1991), A Network Design Algorithm using a Stochastic Incremental Traffic Assignment Approach. Transportation Science, 25, pp. 215–224.

[8] Fernández E., De Cea J., Florian M. and Cabrera E. (1994), Network Equilibrium Models with Combined Modes. Transportation Science, 28, pp. 182–192.

[9] Florian M. and Los, M.(1979), Determining Intermediate Origin-Destination Matrices for the Analysis of Composite Mode Trips. Transportation Research, 13B, pp. 91–103.

[10] Friesz T., Cho H-J., Mehta N., Tobin R. and Anandalingam G. (1992) A Simulated Annealing Approach to the Network Design Problem with Variational Inequality Constraints. Transportation Science, 26, pp. 18–26.

[11] Friesz T., Anandalingam G., Mehta N., Nam K., Shah, S.J. and Tobin R. (1993). The Multiobjective equilibrium network design problem revisited: A Simulated Annealing Approach. European Journal of Operational Research, 65, pp. 44–57.

[12] García, R. and A. Marín (1997a). Using RSD within Partial Linearization Methods. 16th International Symposium on Mathematical Programming, Lausanne .

[13] García, R. and A. Marín (1997b). Urban Multimodal Interchanges (Macro Vision). EURO XV–INFORMS XXXIV Joint International Meeting, Barcelona.

[14] García, R. and Marín, A. (1998). A Bi-level Programming Approach for Estimation of Origin-Destination Matrix and Calibration of Parameters of a Network Equilibrium model with Combined Modes. 6th EURO Working Group Meeting on Transportation, Göteborg.

[15] Han, A.F. and Wilson, H.M. (1982). The allocation of buses in heavily utilised networks with overlapping routes. Transportation Research-B, 16B, 3, pp.221–232.

[16] Hansen, I., Marín, A. and Rystam, A.(1996). Euritrans: "macro" methodology. Euritrans Group Meeting, Rotterdam.

[17] Hunt, J. D. and Teply, S. (1993) A Nested Logit Model of Parking Location Choice. Transportation Research-B, 4, pp. 253–265.

[18] Leblanc, L.J. (1975) An Algorithm for the discrete network design problem. Transportation Science, 9, pp. 183-199.

[19] Leblanc, L.J. (1988) Transit System Network Design. Transportation Research-B, 22B, 5, pp. 383-390.

[20] Meng, Q., Yang, H. and Bell, M.G.H. (1999) An Equivalent Continuously Differentiable Model and a Locally Convergent Algorithm for the Continuous Network Design Problem. Transportation Research-B, on Bi-level Traffic Modeling and Optimization.

[21] Mesa, J.A. and Ortega, F.A. (1999) Park-and-Ride Station Catchment Areas in Metropolitan Transit Systems. 7th EURO-Working Group Meeting on Transportation, Helsinki.

[22] Nickel, S., Chöbel, A. and Sonneborn, T. (1999) Hub Location Problems in Urban Traffic Network. 7th EURO-Working Group Meeting on Transportation, Helsinki.

[23] Oppenheim, N. (1994) Urban Travel Demand Modelling, Wiley-Interscience, N.Y

[24] Patriksson, M. (1994) The Traffic Assignment Problem. Models and Methods., VSP, Utrecht, The Netherlands.

[25] Poorzahedy, P. and Turnquist, M.A. (1992), Approximate Algorithms for the Discrete Network Design Problem, Transportation Research-B, 16, pp. 45–56.

[26] Yang, H. and Bell, M.G.H. (1998) Models and algorithms for road network design and how to avoid it. Transportation Research, 18B, pp. 257-278.

Appendix: Formulation of the equilibrium conditions by means of mathematical programming.

In the following we derive the equilibrium conditions (4.5) by applying the necessary conditions for optimality to the mimimization problem LLP.

LLP is of the form: (Min T(g), s.t: $a_i\mathbf{g} = b_i, \mathbf{g} \geq 0$); therefore , we consider the problem Min $\mathcal{L} = T + \sum_i \lambda_i(b_i - a_i\mathbf{g}), \mathbf{g} \geq 0$, where $\mathcal{L}$ represents the lagrangian function and λ_i denotes the lagrangian multiplier of the constraint i. The Kuhn-Tucker conditions are necessary and sufficient for the optimal solution provided that the objective function (4.11) is convex. By inspecting (4.11) one notes that only the term involving variables $g^{a_s}_{\omega,t}$ may be nonconvex, if $\beta_3 > 0$.

In particular, for any ω the term

$$(1/\beta_2)g^a_{\omega,t}(\ln g^a_{\omega,t} - 1 + \alpha_t) - (1/\beta_3)g^a_{\omega,t}(\ln g^a_{\omega,t} - 1) \tag{A.1}$$

must be convex, since this part of the objective function is separable by O-D pair ω and interchange t. (A.1) may be rewritten as

$$[(1/\beta_2) - (1/\beta_3)]g^a_{\omega,t}(\ln g^a_{\omega,t} - 1) + (1/\beta_2)g^a_{\omega,t}\alpha_t \tag{A.2}$$

Since the entropy function $g^a_{\omega,t} \ln g^a_{\omega,t}$ is convex and $g^a_{\omega,t}\alpha_t$ is linear and increasing ($\alpha_t \geq 0$), if

$$\beta_2 > 0,\ \beta_3 > 0 \text{ and } \beta_3 > \beta_2 \tag{A.3}$$

then

$$(1/\beta_2) \sum_{\omega \in W} \sum_{t \in I_\omega} g^a_{\omega,t}(\ln g^a_{\omega,t} - 1 + \alpha_t) - (1/\beta_3) \sum_{\omega \in W} \sum_{t \in I_\omega} g^a_{\omega,t}(\ln g^a_{\omega,t} - 1)$$

is indeed a convex function. Using a similar argument we obtain that if

$$\beta_1 > 0,\ \beta_2 > 0 \text{ and } \beta_2 > \beta_1 \tag{A.4}$$

then the addend

$$(1/\beta_1) \sum_{k \in \{a,b,c\}} \sum_{\omega \in W} g^k_\omega(\ln g^k_\omega - 1 + \alpha^k) - (1/\beta_2) \sum_{\omega \in W} g^a_\omega(\ln g^a_\omega - 1)$$

is a convex function, so we may conclude the convexity of the function if the assumptions are verified.

The minimization problem is stated as

$$\min_{\mathbf{g}} \mathcal{L} = T + \sum_{\omega \in W} \lambda_\omega \left(\sum_{k \in \{a,b,c,d\}} g_\omega^k - \bar{g}_\omega \right) + \sum_{\omega \in W} \lambda_\omega^a \left(\sum_{t \in I_\omega} g_{\omega,t}^a - g_\omega^a \right) +$$

$$\sum_{t \in I, s \in \{1,2\}} \lambda_t^s \left(\sum_{\omega \in W_t} g_{\omega,t}^{a_s} - g_t^{a_s} \right) + \sum_{t \in I} \sum_{\omega \in W_t} \lambda_{\omega,t}^a \left(g_{\omega,t}^{a_1} + g_{\omega,t}^{a_2} - g_{\omega,t}^a \right)$$

The primal variables, as stated, are the following:

$$\left\{ g_\omega^a,\ g_\omega^b,\ g_\omega^c,\ g_\omega^d,\ g_{\omega,t}^a,\ g_{\omega,t}^{a_s},\ g_t^{a_s} \right\}$$

Multiplier λ can be interpreted as the shadow price for this alternative which has a set of several sub-alternatives and different costs.

We compute first the partial derivates of Lagrangian with respect to all the variables of the problem:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial g_\omega^k} &= (1/\beta_1) \left(\log g_\omega^k + \alpha^k \right) \\
&+ U_\omega^k + \lambda_\omega, \quad k \in \{b,\ c,\ d\},\ \omega \in W \qquad \text{(A.5)} \\
\frac{\partial \mathcal{L}}{\partial g_\omega^a} &= (1/\beta_1) \left(\log g_\omega^a + \alpha^a \right) - (1/\beta_2) \log g_\omega^a \\
&+ \lambda_\omega - \lambda_\omega^a,\ \omega \in W \qquad \text{(A.6)} \\
\frac{\partial \mathcal{L}}{\partial g_{\omega,t}^a} &= (1/\beta_2) \left(\log g_{\omega,t}^a + \alpha_t \right) - (1/\beta_3) \log g_{\omega,t}^a \\
&- \lambda_{\omega,t}^a + \lambda_\omega^a,\ t \in I, \omega \in W_t \qquad \text{(A.7)} \\
\frac{\partial \mathcal{L}}{\partial g_{\omega,t}^{a_s}} &= (1/\beta_3) \left(\log g_{\omega,t}^{a_s} + \alpha_t^s \right) + C_{\omega,t}^{a_s} \\
&+ \lambda_t^s + \lambda_{\omega,t}^a,\ t \in I, \omega \in W_t, s \in \{1,\ 2\} \qquad \text{(A.8)} \\
\frac{\partial \mathcal{L}}{\partial g_t^{a_s}} &= c_t^s(g_t^s) - \lambda_t^s \qquad t \in I, s \in \{1,\ 2\} \qquad \text{(A.9)}
\end{aligned}$$

According to Kuhn-Tucker conditions, and by using (A.9), one obtains

$$\text{if } g_t^{a_s} > 0, \text{then: } \lambda_t^s = c_t^s(g_t^s) \qquad \text{(A.10)}$$

Using (A.10) to replace λ_t^s in (A.8) and (4.9), we obtain that when $g_{\omega,t}^{a_s} > 0$,

$$\log g_{\omega,t}^{a_s} = -\beta_3 U_{\omega,t}^{a_s} - \alpha_t^s - \beta_3 \lambda_{\omega,t}^a$$

and one obtains

$$g_{\omega,t}^{a_s} = \exp\left(-\{\alpha_t^s + \beta_3 U_{\omega,t}^{a_s}\} \right) \exp(-\beta_3 \lambda_{\omega,t}^a) \qquad \text{(A.11)}$$

where $\alpha_t^s + \beta_3 U_{\omega,t}^{a_s}$ is the subjective evaluation given by travellers to the park'n ride path joining pair ω and using interchange t, and parking type s in the choice process represented by the demand model (4.4). Then using (A.11) it is easy to see that the proportions of trips made by travelling between ω that use interchange t, and type of parking s is given by

$$\frac{g_{\omega,t}^{a_s}}{\sum_{s' \in \{1,2\}} g_{\omega,t}^{a_s'}} = \frac{\exp\{-\left(\alpha^s{}_t + \beta_3 U_{\omega,t}^{a_s}\right)\}}{\sum_{s' \in \{1,2\}} \exp\{-(\alpha_t^{s'} + \beta_3 U_{\omega,t}^{a_{s'}})\}},$$

Using the relationship $g_{\omega,t}^{a_1} + g_{\omega,t}^{a_2} = g_{\omega,t}^{a}$ and (A.11)

$$g_{\omega,t}^{a} = \sum_{s \in \{1,2\}} \exp\left(-\{\alpha_t^s + \beta_3 U_{\omega,t}^{a_s}\}\right) \exp(-\beta_3 \lambda_{\omega,t}^{a}) \tag{A.12}$$

By taking log on both sides of (A.12) we get

$$\lambda_{\omega,t}^{a} = \frac{-1}{\beta_3} \log g_{\omega,t}^{a} - L_{\omega,t}^{a} \tag{A.13}$$

where $L_{\omega,t}^{a}$ is computed as the "log-sum" of the utilities of the subalternatives of $g_{\omega,t}^{a}$, i,e

$$L_{\omega,t}^{a} = \frac{-1}{\beta_3} \log \left(\sum_{s \in \{1,2\}} \exp\left(-\{\alpha_t^s + \beta_3 U_{\omega,t}^{a_s}\}\right) \right)$$

Substituting (A.13) in (A.7), we obtain

$$\frac{1}{\beta_2}(\log g_{\omega,t}^{a} + \alpha_t) + L_{\omega,t}^{a} + \lambda_{\omega}^{a} = 0$$

Removing $g_{\omega,t}^{a}$ from the previous relationship

$$g_{\omega,t}^{a} = \exp\left(-\{\alpha_t + \beta_2 L_{\omega,t}^{a}\}\right) \exp(-\beta_2 \lambda_{\omega}^{a}) \tag{A.14}$$

The proportion of park'n ride trips travelling between ω that use the interchange t is given by

$$\frac{g_{\omega,t}^{a}}{\sum_{t' \in I_\omega} g_{\omega,t'}^{a}} = \frac{\exp\{-\left(\alpha_t + \beta_2 L_{\omega,t}^{a}\right)\}}{\sum_{t' \in I_\omega} \exp\{-(\alpha_t + \beta_2 L_{\omega,t'}^{a})\}},$$

Using (4.13) and (A.14), one obtains

$$g_{\omega}^{a} = \sum_{t \in I_\omega} \exp\left(-\{\alpha_t + \beta_2 L_{\omega,t}^{a}\}\right) \exp(-\beta_2 \lambda_{\omega}^{a}) \tag{A.15}$$

and removing λ_ω^a from the previous relationship

$$\lambda_\omega^a = \frac{-1}{\beta_2} \log g_\omega^a - L_\omega^a \tag{A.16}$$

where L_ω^a is given by (4.3). Substituting λ_ω^a in (A.6), and if $g_\omega^a > 0$ then,

$$\frac{1}{\beta_1} (\log g_\omega^a + \alpha^a) + \lambda_\omega + L_\omega^a = 0$$

Removing g_ω^a in the previous relationship

$$g_\omega^a = \exp(-\{\alpha^a + \beta_1 L_\omega^a\}) \exp(-\beta_1 \lambda_\omega) \tag{A.17}$$

Removing g_ω^k in (A.5), we get

$$g_\omega^k = \exp(-\{\alpha^k + \beta_1 U_\omega^k\}) \exp(-\beta_1 \lambda_\omega) \tag{A.18}$$

and using (A.17) and (A.18), we obtain that the modal split is given by

$$\frac{g_\omega^k}{\sum_{k' \in \{a,b,c,d\}} g_\omega^{k'}} = \frac{\exp - \left(\alpha^k + \beta_1 U_\omega^k\right)}{\sum_{k' \in \{a,b,c,d\}} \exp - \left(\alpha^{k'} + \beta_1 U_\omega^{k'}\right)},$$

where $U_\omega^a = L_\omega^a$ that it is the "log-sum" of the utilities $L_{\omega,t}^a$, where $t \in I_\omega$.

Using (4.12) and (A.7), we obtain the multiplier λ_ω

$$\lambda_\omega = \frac{-1}{\beta_1} \log \bar{g}_\omega + \frac{1}{\beta_1} \log \left(\sum_{k' \in \{a,b,c,d\}} \exp -(\alpha^{k'} + \beta_1 U_\omega^{k'}) \right) \tag{A.19}$$

and if we substitute the relationships (A.13), (A.16), and (A.19) that have been obtained for the multipliers in (A.18), (A.15), and (A.11) then we obtain the equilibrium conditions defined for the equation (4.5).

Chapter 5

PARK-AND-RIDE STATION CATCHMENT AREAS IN METROPOLITAN RAPID TRANSIT SYSTEMS

Juan A. Mesa
Departamento de Matemática Aplicada II. Universidad de Sevilla.
Paseo de los Descubrimientos s/n. 41092 Sevilla, Spain.
jmesa@cica.es

Francisco A. Ortega
Departamento de Matemática Aplicada I. Universidad de Sevilla.
Reina Mercedes s/n. 41012 Sevilla, Spain.
riejos@cica.es

Abstract Park-and-Ride facilities are common in commuter transit systems as well as in metro network stations situated in residential areas. Catchment areas are useful for different purposes; in particular, to evaluate the coverage of transit lines, assuming that the population covered by the line bears relation to the expected number of trips.

In this paper, catchment areas for riderships using park-and-ride facilities are obtained by comparing the total travelling time using just private mode with that using a combination of both modes. This methodology leads us to establish a region whose boundary is typically a branch of hyperbola with foci located in the nearest station S from user and the destination D.

Moreover, when the trips in central and suburban districts are assumed with different average speeds, the curve limiting catchment areas is obtained and belongs to the same kind of conic curves.

Finally, the boundary between (level-)catchment areas of adjacent stations for riderships using park-and-ride facilities is also characterised as a branch of hyperbola, and an optimisation problem is proposed in relation with the line coverage.

Keywords: Transit systems, Park-and-ride facilities, Coverage

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 81–93.

1. INTRODUCTION

One of the most widely-used decision criteria in location, routing and transportation network design is the maximisation of the population covered, i.e. the number of people or demand points which are within a prefixed threshold of distance from the service.

In network design problems, coverage can be (one of) the objective function(s) [1]. More specifically, coverage is used as the only objective function in the problem of locating a rapid transit line [2], and it is a crucial measure of effectiveness in planning projects for urban transit systems. In fact, coverage is one of the index used for assessing the offer of existing [3] or planned [4] transit systems.

Basically, there are four different modes of access to the stations: feeder bus (or tram), park-and-ride, kiss-and-ride, and walking, with taxi and cycling being other additional and, sometimes, relevant means of access. Catchment areas are useful for evaluating the coverage of stations which can be applied to forecasting the number of riderships. Geometrical shape of the catchment areas depends on the access mode. The determination of the catchment areas is the first step of the procedure proposed by Bolger, Colquhoun and Morrall ([5]), being applied to the Calgary LRT to estimate the park-and-ride demand. Planners can use this estimation to calculate the size of park-and-ride facilities, to choose the stations to which these facilities should be added and to decide the location of stations provided by park-and-ride facilities along the line, in accordance with the existing trip patterns and/or the new proposed ones regarding the land uses (see [6]).

The organisation of the paper is as follows: in Section 2, the park-and-ride coverage component is analysed by studying the catchment areas determined by the stations in different settings. Park-and-ride coverage provided by the transit line is defined in Section 3, and an optimisation problem is proposed in relation with the location of q stations along the line. The conclusion and further research follow in Section 4.

2. PARK-AND-RIDE STATIONS CATCHMENT AREAS

According to access guidelines for the suburban LRT stations of Calgary, described in the paper by Bolger, Colquhoun and Morrall ([5]), approximately one-third of the users arrive by private automobile, either as passenger drop-off (the so-called kiss-and-ride) or using a park-and-ride lot. In fact the modal share suggested for this system is as follows:

Bus, 60-65 %; park-and-ride, 15-20 %; kiss-and-ride, 15 % and walking, 5 %.

For each access mode a different catchment area (and the corresponding levels of attraction) can be defined and the provided coverage calculated. Thus, the total coverage provided by station S is:

$$R(S) = R_b(S) + R_p(S) + R_d(S) + R_w(S),$$

where $R_p(S)$ is the coverage provided by the park-and-ride facility to the station (our main aim to be analysed) and

- Walking-coverage (as well as bicycle-coverage), denoted by $R_w(S)$, can be calculated by using the τ_p norm, whose parameters, $\tau > 0$ and $p \in [1, 2]$, are a result of an estimation process [7] which consists of calibrating these parameters so that a global function of the deviations with respect to actual distances is minimized. The weighted l_p norm is the most accurate function to estimate walking distances in the street network, as some computational experiments have shown [8]. This notion of coverage has generated several optimisation problems which have received attention in the relevant literature; for instance, recently Laporte, Mesa and Ortega ([9]) have dealt with the problem of locating a prefixed number of stations so that walking-coverage is maximised.
- Bus(tram)-coverage $R_b(S)$ can be obtained as a line-cover whose calculation must take into account the times spent (from the origin to the boarding bus stop, waiting for the bus and that from the alighting bus stop to the station) and the competition with the private means of transport.
- As far as the authors are aware there are no theoretical studies on the behaviour of riders driving passengers for dropping-off at stations ($R_d(S)$), but it can be assumed that the majority of them corresponds to those trip patterns in which one resident drives to work and another takes the public transit mode to the final destination.

When the station has a park-and-ride facility, a case which is common in commuter or regional transit systems as well as in metro network stations situated in residential areas, the determination of the catchment area should take into account the private car travelling times to the station. For this kind of trip, the catchment areas are not circle-shaped but bell-shaped, because the users do not always go to the nearest station, but to the station for which the total travelling time of the combination of the two modes is shortest. Furthermore, in this case, the transit system is clearly in competition with the private mode.

2.1 TRANSIT MODE COMPETITION IN A HOMOGENEOUS AREA

Let us denote $t_c(X, S)$ and $t_c(X, D)$ as the expected travelling times when using a private car, from home X to suburban station S and destination station D (point located in the central business district), respectively. Let $t_t(S, D)$ be the expected travelling time from station S to destination D using the transit system. Analogously, let us denote $d_c(A, B)$ as the actual distance when a private car is used from point A to point B along a street network and $d_t(A, B)$ as the corresponding distance on the transit system.

In the simplest deterministic model, let us suppose that commuters choose the bi-modal trip if the corresponding time does not exceed the one using private mode plus a term δ (positive, negative or null, depending on other factors such as fare, comfort, walking time from the parking lot, waiting time, etc.):

$$t_c(X, S) + t_t(S, D) \leq t_c(X, D) + \delta$$

Initially the additive term δ will be assumed constant in order to obtain the geometrical shape of the corresponding catchment area border given by the pair (S, D), and subsequently the dispersion of users' choices will be considered by assuming the existence of random attributes in the definition of term δ.

Proposition 1
The locus of points of the plane satisfying $t_c(X, S) + t_t(S, D) = t_c(X, D) + \delta$ is a branch of hyperbola with foci S and D.

Proof

Let v_c and v_t be the average speed for private cars and the commercial speed for the transit mode, respectively. Then,

$$\frac{d_c(X, S)}{v_c} - \frac{d_c(X, D)}{v_c} = \delta - \frac{d_t(S, D)}{v_t}; \quad \text{or } d_c(X, S) - d_c(X, D) = 2a.$$

The locus of points of the plane satisfying this relation is a branch of hyperbola with foci S and D when $d_c(A, B)$ represents any weighted Euclidean distance ($\tau l_2(\vec{AB})$, $\tau > 0$). ⊗

It should be noted that empirical experiences show the prevalence of travel time over the cost associated with the distance factor; moreover, other typical components of the utility function can be considered constant (as modal penalty) or involved in the trip time (as waiting time). Therefore, the previous result remains true if the travel time is the only relevant variable when the generalised costs by car, $\text{cost}_c(A, B)$, and by

the transit system, $\text{cost}_t(A,B)$, are modelled proportionally to time; i.e. $\text{cost}_c(A,B) = \alpha_c t_c(A,B)$, $\text{cost}_t(A,B) = \alpha_t t_t(A,B)$, where α_c and α_t are assumed as non-negative constants and determined by estimation.

In the following figures, we have assumed these coordinates for points S and D: $S = (0,0)$ and $D = (d_1,0)$, where $d_t(S,D) = d_1 = 10$. Moreover, the additive term δ has not been taken into account ($\delta = 0$) when constant $2a = v_c\Big(\delta - \dfrac{d_t(S,D)}{v_t}\Big)$ is defined.

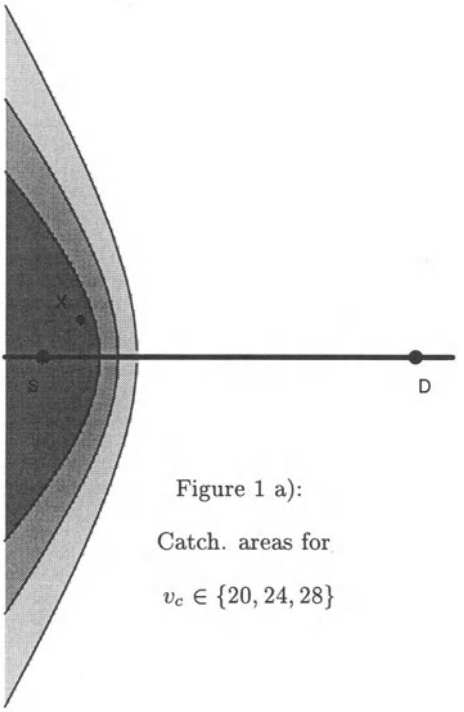

Figure 1 a):

Catch. areas for

$v_c \in \{20, 24, 28\}$

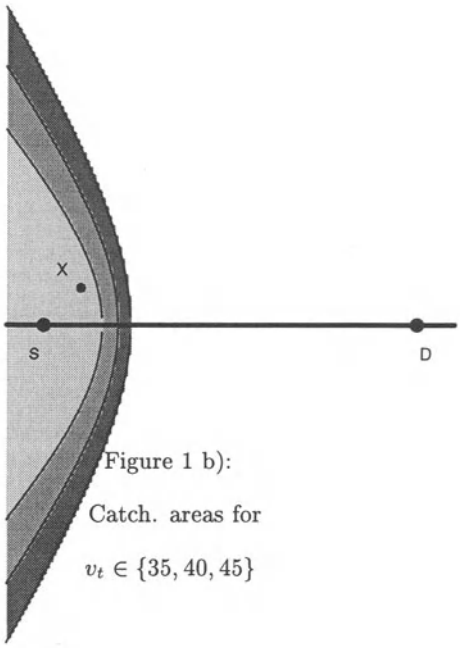

Figure 1 b):

Catch. areas for

$v_t \in \{35, 40, 45\}$

The Euclidean distance is considered as the actual distance for calculating distances from point $X = (x,y)$ by car. Hence,

$$d_c(X,S) = \sqrt{x^2+y^2};\ d_c(X,D) = \sqrt{(x-d_1)^2+y^2}$$

By imposing the condition $d_c(X,S) - d_c(X,D) = 2a$, in terms of coordinates, the following expression is obtained:

$$\sqrt{x^2+y^2} - \sqrt{(x-d_1)^2+y^2} = 2a$$

and, by algebraic manipulation, we deduce the explicit expression

$$y = \pm\sqrt{\Big(-\frac{d_1}{2a}x + \frac{d_1^2}{4a} - a\Big)^2 - x^2},\ \forall x \leq \frac{d_1^2 - 4a^2}{2d_1 + 4a}$$

which corresponds to the different catchment area boundaries.

In Figure 1*a*), car speed takes the values 20 (light grey), 24 and 28 (dark grey) (mantaining $v_t = 40$); on the other hand, in Figure 1*b*), $v_c = 24$ is fixed, while transit system speed takes the values 35 (light grey), 40 and 45 (dark grey).

The eccentricity $\dfrac{2d_t(S,D)}{v_c\left(\dfrac{d_t(S,D)}{v_t} - \delta\right)}$ of the hyperbolas obtained by varying D along the alignment yields a constant when parameter δ remains constant.

It should be noted that the non-spatial attributes, such as the idiosyncrasies of each individual, have not been taken into account in the previous development. In the random utility theory [10] the individual q selects the maximum-utility alternative by comparison of

$$U^q_{\mathrm{c\&t}}(X) = V_{\mathrm{c\&t}}(X) + V^q_{\mathrm{c\&t}} + \epsilon^q_{\mathrm{c\&t}} \quad \text{and} \quad U^q_{\mathrm{c}}(X) = V_{\mathrm{c}}(X) + V^q_{\mathrm{c}} + \epsilon^q_{\mathrm{c}},$$

where subindex c&t denotes the combination of means (car and transit system) and subindex c indicates exclusive car usage;

- $U^q_{(\cdot)}(X)$ is the global utility associated to the individual q located in X,
- $V_{(\cdot)}(X)$ is the gross utility of achieving the destination from point X,
- $V^q_{(\cdot)}$ is the component associated to the individual q and
- $\epsilon^q_{(\cdot)}$ is the random part reflecting the particular taste of individual q.

The individual choice depends on the comparison $U^q_{\mathrm{c\&t}}(X) \leq U^q_{\mathrm{c}}(X)$, i.e.,

$$V_{\mathrm{c\&t}}(X) \leq V_{\mathrm{c}}(X) + (V^q_{\mathrm{c}} - V^q_{\mathrm{c\&t}}) + (\epsilon^q_{\mathrm{c}} - \epsilon^q_{\mathrm{c\&t}})$$

If the gross utility is assumed depending on only the travel time (as before), the term corresponding to the constant δ in the deterministic spatial model must be considered analytically as the sum of the improvement $V^q_{\mathrm{c}} - V^q_{\mathrm{c\&t}}$, due to objective attributes perceived by individual q, and the residual term $\epsilon \equiv \epsilon^q_{\mathrm{c}} - \epsilon^q_{\mathrm{c\&t}}$.

Let us assume momentarily that there is no randomness in the system and that the difference $V^q_{\mathrm{c}} - V^q_{\mathrm{c\&t}}$ may take the values $\{\mathrm{v}_1, \ldots, \mathrm{v}_n\}$. Then for each v_i, $i = 1, \ldots, n$, we obtain one hyperbola and, thus, a family of n branches of hyperbola H_i, with the same foci S and D but

different eccentricities, is achieved. Therefore, if the values of the variables which represent the attributes of a particular individual are known, his choice between the combined and private modes can be deduced by means of the relative position of the individual with respect to the hyperbola determined by these values.

Moreover, when the individual uncertainty component, given by the ϵ-terms, is assumed to be a random variable with mean 0 and a certain probability distribution, then each previous hyperbola H_i, $i = 1, \ldots, n$, is the instance $\epsilon = 0$ of a parametrical family of branches of hyperbola, again sharing foci S and D, where each member has a probabilistic interpretation provided by the parameter $\epsilon \in \mathbb{R}$ and different eccentricity (because the value of $\epsilon \in \mathbb{R}$ would take part in the calculation).

2.2 TRANSIT MODE COMPETITION IN A NON-HOMOGENEOUS SETTING

Following on, we consider a new scenario where point E is the 'gate' to the central or congested area (see Figure 2), and speed in the suburban district is considered much higher than in central areas.

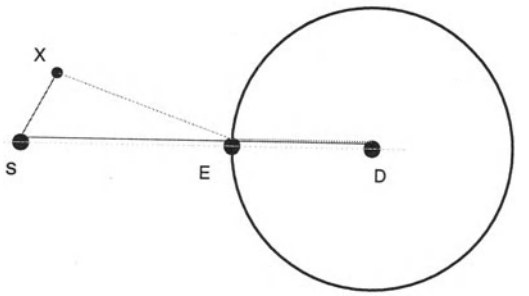

Figure 2: Access to D via gate E

Although other gates can exist in the compact form of the central area, one of these (the closest gate) is considered exclusively by the user in the radial trip to the centre. Let us denote $\tau_c(A, B)$ as the expected travelling time when using a private car in a central district from point A to B. Therefore, the user behaviour depends on the inequality

$$t_c(X, S) + t_t(S, D) \leq t_c(X, E) + \tau_c(E, D) + \delta,$$

Proposition 2
The locus of points of the plane satisfying

$$t_c(X, S) + t_t(S, D) = t_c(X, E) + \tau_c(E, D) + \delta,$$

is a branch of hyperbola with foci S and E.

Proof

Let v_c and v_c' be the average speeds in residential and central areas, respectively. Then $\frac{d_c(X,S)}{v_c} + t_t(S,D) = \frac{d_c(X,E)}{v_c} + \frac{d_c(E,D)}{v_c'} + \delta$; hence,

$$d_c(X,S) - d_c(X,E) = \frac{v_c}{v_c'} d_c(E,D) + v_c (\delta - t_t(S,D)) = 2a'$$

⊗

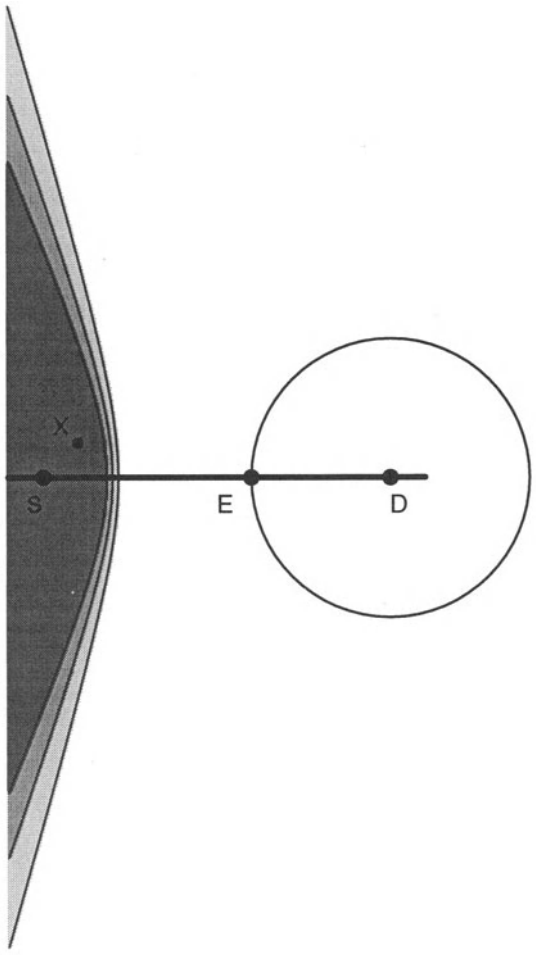

Figure 3: Catchment areas for $v_c \in \{20, 24, 28\}$

In a similar manner to the homogeneous case, the result remains true when substituting Euclidean distances by the estimated inflated

Euclidean distances ($\alpha_c d_2(A,B)$ and $\alpha'_c d_2(A,B)$) corresponding to the actual car distances in suburban and central zones, respectively.

In Figure 3, the values $v'_c = 12$ and $d_c(E,D) = 4$ have been added to the other parameters which appeared in the previous case ($v_c \in \{20,24,28\}$ and $v_t = 40$).

While varying the attraction levels, i.e. allocating different values to the difference of distances, the curves limiting the level-catchment areas are determined.

2.3 CATCHMENT AREAS OF ADJACENT PARK-AND-RIDE STATIONS

A decomposition of region A under consideration in catchment areas $A(c)$, $A(S_1)$, $A(S_2),\ldots,A(S_q)$ is obtained when considering all stations of the line provided by park-and-ride facilities. By fixing the values of the parameters, the curves limiting the catchment areas determine three regions $A(S)$, $A(S')$ and $A(c)$, which are patronized by the riders of adjacent park-and-ride stations S and S' and the private mode, respectively.

Proposition 3

The boundary between (level-)catchment areas of adjacent stations for riderships using park-and-ride facilities, is also a branch of hyperbola

$$t_c(X,S) + t_t(S,D) = t_c(X,S') + t_t(S',D)$$

Proof

The above expression leads us to

$$d_c(X,S') - d_c(X,S) = v_c\left(\frac{d_t(S,D)}{v_t} - \frac{d_t(S',D)}{v_t}\right) = 2a''$$

⊗

In Figure 4 these regions are represented when $a'' = 0$, $v_c = 20$, $v_t = 40$, $d_t(S,D) = 10$ and $d_t(S',D) = 6$, S'' being a station without a park-and-ride facility.

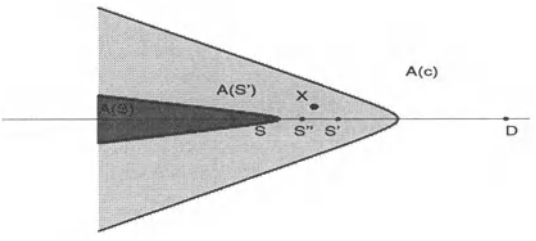

Figure 4: Catchment regions corresponding to stations S and S'.

3. CATCHMENT LEVELS. PARK-AND-RIDE COVERAGE OF THE LINE

Let us suppose that the considered region A is divided into census tracts A_j, $A = \bigcup_{j=1}^{J} A_j$, where each zone A_j is assumed to be a polygonal region appropriately weighted by $\rho_j > 0$, $\forall j = 1, \ldots, J$, which can represent car availability densities as well as other socio-economic features.

To estimate travel times inside region A by using the private car, the weighted Euclidean norm, $t_c(X, Y) = \tau_c \parallel X - Y \parallel_2$ can be used. In terms of distances, the relevant function is the difference $r(X) = d_c(X, S) - d_c(X, D)$, for all points X inside region $A(S)$.

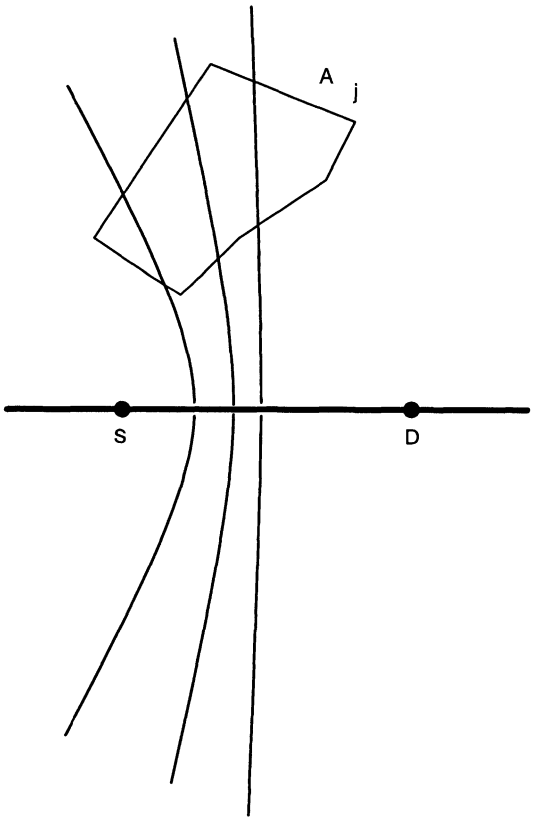

Figure 5: Attraction levels from station S.

By using a continuous gravitational model, the park-and-ride coverage can be expressed as follows:

$$R_p(S) := \sum_{j=1}^{J} \rho_j \int_{A(S)\cap A_j} \frac{dX}{r(X)^2}$$

For purposes of numerical calculation the function which defines the park-and-ride coverage of station S will be discretised; in fact, K different levels of attraction ($r_k > 0$) will be considered. Since the range of the function $d_c(X,S) - d_c(X,D)$, $\forall X \in A(S)$, is $[-d_c(S,D), 2a]$, the attraction level r_k is

$$r_k = -d_c(S,D) + k\,\frac{2a + d_c(S,D)}{K},\ k = 0,1,\ldots,K.$$

Therefore, level r_k determines the region $B_D(S,r_k)$ whose boundary is a branch of the hyperbola with foci in points S and D (see Figure 5):

$$X \in B_D(S,r_k) \Longleftrightarrow d_c(X,S) - d_c(X,D) \leq r_k$$

In order to calculate the park-and-ride coverage $R_p(S)$ of each station, regions B'_k are determined by the respective intersections with the catchment area $B_D(S,r_k) \equiv B_k$ (subindex D is not present because it is assumed by default): $B'_k = B_k \cap \left(\bigcup_{j=1}^{J} A_j\right) = \bigcup_{j=1}^{J} \left(B_k \cap A_j\right)$. Assuming $B'_0 = \emptyset$ and modelling the attraction by the function derived from the gravitational model [10], the discretised park-and-ride coverage of station S follows:

$$R_p(S) := \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\rho_j}{(\frac{r_{k-1}+r_k}{2})^2}\, Area\Big((B'_k \setminus B'_{(k-1)}) \cap A_j\Big).$$

Let us note that the definition of the park-and-ride coverage can be adapted to the case in which the non-spatial components (deterministic and random) of the individual utility are considered. Then, the definition of function $r(X)$ and integral operator dX must be modified in order to involve the non-spatial individual attributes.

4. CONCLUSION AND FURTHER RESEARCH

When stations have park-and-ride facilities, the potential user does not only take into account the time to reach stations by private means but also the time spent in the transit system. This assumption leads us to establish the catchment regions whose boundaries are typically

hyperbolas. The same kind of conic curves appears in different settings, characterising the coverage by the stations. In particular, this also happens in the more general case in which the utility used for the choice model includes terms corresponding to the individual attributes which cannot be spatially explained. However, the decomposition of the utilities into spatial and individual components enables the behaviour of the potential users to be analysed with regards to their relative locations with respect to park-and-ride stations.

Finally, in order to get insight into the general problem of locating the stations of an alignment by maximising the overall coverage, a problem to be considered consists of locating q stations provided by park-and-ride facilities so that $\max_{S_i \in L \cap (CIS)} R_p(L) := \sum_{i=1}^{q} R_p(S_i)$, in which technical characteristics of the transit system force us to include the so-called constraint on the interstation spacing 'CIS' [9]. This optimisation problem deserves further research.

Some knowledge about the particular behaviour of the people living in the area under consideration should be required before applying the results given in this paper to a real case. Additionally, other factors such as land use plans and layout of both stations and park-and-ride facilities have an important influence in the actual ridership.

References

[1] Current, J., ReVelle, C. and Cohon, J. (1985). "The maximum covering/shortest path problem: a multiobjective network design and routing formulation". European Journal of Operational Research 21, pp 189–199.

[2] Dufourd, H., Gendreau, M. and Laporte, G. (1996). "Locating a transit line using tabu search". Location Science 4, pp 1–19.

[3] García, A. and Cristóbal, C. (1996). "Cobertura de las redes ferroviarias de transporte público (metro y cercanías) de la Comunidad Autónoma de Madrid, utilizando un sistema de información geográfico". II Simposium de Ingeniería de Transportes (Madrid, Spain), pp. 455–462 (in Spanish).

[4] Plan Director de Infraestructuras para el Área Metropolitana de Sevilla (1998). Plan Director de Infraestructuras de Andalucía 1997–2007. Consejería de Obras Públicas y Transportes de la Junta de Andalucía (Sevilla, Spain), pp. 216–219 (in Spanish).

[5] Bolger, D., Colquhoun, D. and Morrall, J. (1992). "Planning and design of Park-and-Ride facilities for the Calgary Light Rail Transit System". Transportation Research Record 1361, pp 141–148.

[6] Lutin, J.M. and Benz, G.P. (1992). "Key issues in light rail transit station planning and design". Transportation Research Record 1361, pp. 117–124.

[7] Love, R.F., and Morris, J.G. (1988). "On Estimating Road Distances by Mathematical Functions". European Journal of Operational Research 36, 251–253.

[8] Brimberg, J., Love, R.F. and Walker, J.H. (1995). "The Effect of Axis Rotation on Distance Estimation". European Journal of Operational Research 80, pp 357–364.

[9] Laporte, G., Mesa, J.A. and Ortega, F.A. (1998). "Location Stations on Rapid Transit Lines". Accepted in Computers and Operation Research.

[10] Ortúzar, J.D. and Willumsen, L.F. (1990). Modelling Transport. J. Wiley, New York.

Chapter 6

HUB LOCATION PROBLEMS IN URBAN TRAFFIC NETWORKS

Stefan Nickel

Anita Schöbel

Tim Sonneborn
Institut für Techno– und Wirtschaftsmathematik (ITWM) e.V.
Gottlieb–Daimler–Straÿe 49
D–67663 Kaiserslautern
Germany
sonnebor@itwm.uni-kl.de

Abstract In this paper we present new hub location models which are applicable for urban public transportation networks. In order to obtain such models we relax some of the general assumptions that are usually satisfied in hub location problems, but which are not useful for public transportation networks. For instance we do not require that the hub nodes have to be completely interconnected. These new models are based on network design formulations, in which the constraint that all flow has to be routed via some hub nodes is formulated by a flow conservation law. We present some solution approaches for these new models and illustrate the results on a numerical example.

Keywords: hub location, urban public transportation, network design

1. INTRODUCTION

In this paper we will look at an urban public transportation network, in which two different kinds of vehicles (e.g. buses and undergrounds) can be used. Passenger requests are given for every origin–destination

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 95–107.

pair of the network nodes. The task is to design a public transportation network such that the overall costs are minimized. The location of the transshipment points (so called hub nodes or hubs) and the allocation of the other nodes to these hubs are of special interest in this process.

During the last years, different kinds of hub location problems have been discussed in literature (for an overview of some basic problems see [3]). Main applications of hub location problems have been air passenger and cargo transportation, telecommunication and postal delivery services. The main types of problems which are dealt with are p–hub location, where the number of hubs to be located is fixed to p (see e.g. [14]), and fixed charge hub location problems, where this number is unlimited, but a certain fixed cost has to be paid for establishing a hub facility (see e.g. [5], [3]). Furthermore it can be distinguished between single allocation (see e.g. [6], [13]) and multiple allocation (see e.g. [7]) problems. In the single allocation case every non–hub node must be allocated to exactly one hub, while in the multiple allocation case a non–hub node can be allocated to several hubs.

In this paper we consider an application of hub location problems to public transportation, which to the best of our knowledge has not been studied so far. The problem which might be the most useful for applications in public transportation is the fixed charge multiple allocation hub location problem. In Subsection 2.1 we will consider a modified uncapacitated version of this problem, which will be the basis for this paper. In order to obtain new models that are more useful for urban public transportation networks in Subsection 2.2 we relax some of the general assumptions that are usual satisfied in hub location problems . These new models correspond to network design problems. We will give an overview of different possible solution ideas in Section 3. In Section 4 we present a numerical example. Finally we give some conclusions in Section 5.

2. MODELING HUB LOCATION PROBLEMS

Hubs are special kinds of facilities, which collect the flow from a set of other facilities arriving at the hub, regroup it and distribute it to other hubs or to their final destination. A hub network can be seen as a two-level network: The hub level network connects the hubs among each other, and the spoke level network connects the non–hub to the hub nodes (see Figure 6.1). As we allow multiple allocation, each non–hub node may be allocated to several hubs.

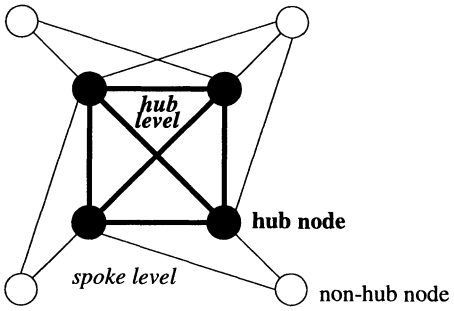

Figure 6.1 Example for a hub network (multiple allocation)

In public transportation the hub edges may represent fast underground lines while the spokes can be bus lines. Passengers can change at the hub nodes from bus to underground or vice versa.

The Network Hub Location Problem consists of two parts, which are dependent on each other: In the location part the nodes which should serve as hubs are selected, while in the allocation part the non–hub nodes have to be allocated to the hub nodes.

In common hub location models the following general assumptions are used:

(a) The hub level network is a complete graph.

(b) Using interhub connections has a lower price per unit than using spoke connections. So we have a discount factor $\alpha \in [0, 1]$ for using interhub connections.

(c) Direct connections between non–hub nodes are not allowed.

(d) Costs are proportional to (e.g. Euclidean) distances and satisfy the triangle inequality.

From (a) and (d) it follows that transportation in the hub level is always done directly. Together with (c) this means that every flow is routed via 1 or 2 hubs.

2.1 THE UNCAPACITATED FIXED CHARGE MULTIPLE ALLOCATION HUB LOCATION PROBLEM

Now we will give a modified formulation of the uncapacitated fixed charge multiple allocation hub location problem as a mixed integer program. It differs from the one used so far [3] because we not only consider fixed costs for hub nodes, but for hub edges, too. Let $\mathcal{N}$ be the set of all

facilities and $W_{ij} \geq 0\,(i,j \in \mathcal{N})$ the given flow from facility i to facility j. In public transportation it is often assumed that $W_{ii} = 0$ for all $i \in \mathcal{N}$. Let $F_k \geq 0\,(k \in \mathcal{N})$ be the fixed cost for establishing a hub facility at node k, $I_{kl} \geq 0\,(k,l \in \mathcal{N} : k \leq l)$ the fixed cost for establishing an undirected hub edge between nodes k and l and $C_{ijkl} \geq 0\,(i,j,k,l \in \mathcal{N})$ the transportation cost per unit of flow that is routed from node i to node j via the (potential hub) nodes k and l (in this direction). If $d_{ij} \geq 0$ is the usual cost for shipping one unit of flow directly from i to j and $\alpha \in [0,1]$ is the discount factor for using the interhub connections, then $C_{ijkl} = d_{ik} + \alpha\, d_{kl} + d_{lj}$.

As variables to be determined we need

$H_k \in \{0,1\}$, which is 1, if facility k is a hub node, 0 otherwise $(k \in \mathcal{N})$;

$U_{kl} \geq 0$, which is 1, if facilities k and l are hubs, 0 otherwise $(k,l \in \mathcal{N} : k \leq l)$;

$X_{ijkl} \geq 0$, which determines the fraction of flow from i to j which is routed via nodes k and l $(i,j,k,l \in \mathcal{N})$ (in this direction).

We note by definition that $U_{kl} = H_k \cdot H_l$ for all $k,l \in \mathcal{N} : k \leq l$.

(UFC)

$$\min \quad \sum_{i\in\mathcal{N}} \sum_{k\in\mathcal{N}} \sum_{l\in\mathcal{N}} \sum_{j\in\mathcal{N}} W_{ij}\, C_{ijkl}\, X_{ijkl} + \sum_{k\in\mathcal{N}} F_k\, H_k + \sum_{k\in\mathcal{N}} \sum_{l\in\mathcal{N}:l\geq k} I_{kl}\, U_{kl} \tag{6.1}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{N}} \sum_{l\in\mathcal{N}} X_{ijkl} = 1 \qquad \forall\, i,j \in \mathcal{N}, \tag{6.2}$$

$$\sum_{l\in\mathcal{N}} X_{ijkl} \leq H_k \qquad \forall\, i,j,k \in \mathcal{N}, \tag{6.3}$$

$$\sum_{k\in\mathcal{N}} X_{ijkl} \leq H_l \qquad \forall\, i,j,l \in \mathcal{N}, \tag{6.4}$$

$$U_{kl} \geq H_k + H_l - 1 \qquad \forall\, k,l \in \mathcal{N} : k \leq l, \tag{6.5}$$

$$U_{kl} \geq 0 \qquad \forall\, k,l \in \mathcal{N} : k \leq l, \tag{6.6}$$

$$H_k \in \{0,1\} \qquad \forall\, k \in \mathcal{N}, \tag{6.7}$$

$$X_{ijkl} \geq 0 \qquad \forall\, i,j,k,l \in \mathcal{N}. \tag{6.8}$$

In the objective function (6.1) we minimize the total (variable plus fixed) costs. All flow between two nodes i and j has to be routed via one or two nodes k and l (6.2), but only if k and l are hub nodes ((6.3) and (6.4)). In (6.5) and (6.6) we ensure that in an optimal solution $U_{kl} = 1$ if and only if $H_k = H_l = 1$ (because (UFC) is a minimization problem).

2.2 MODELS FOR APPLICATIONS IN PUBLIC TRANSPORTATION

The current hub location models are not useful for applications in public transportation, because in this case the general assumptions (a), (c), and (d) are often not satisfied. We want to relax the requirement that routing in the hub level network must be done directly (see (a) and (d)), because it may often be the case that the distance graph is not complete, the triangle inequality is not valid in public transportation networks, or the fixed cost for establishing a direct hub edge is higher than the cost for non-direct transportation. At first we keep the general assumption (c), but show that it can easily be relaxed, too.

In the hub location models used so far this non–direct transportation cannot be modeled because the X_{ijkl} variables are not sufficient to describe the whole flow paths. Therefore, we will present new models based on network design formulations (following a hint given by [11], see also [9]; for an overview on network design problems see [12], [2]).

We will show two alternative models. The meaning of the first one is similar to the meaning of (UFC): Spoke connections are allowed only for the first and last edge of every flow path, and they must begin or end at a hub node. It can also be required that direct connections must be routed via at least one hub (corresponding to general assumption (c)).

With $\mathcal{E}$ we describe the set of edges which can be established in the whole hub–and–spoke network, e.g. $\mathcal{E} = \{\{i,j\} \in \mathcal{N}^2 : i \leq j\}$. In the remainder of this paper we always locate undirected edges, though a directed version of the models may be possible, too (for directed network design problems see e.g. [12]).

The variables get an edge–oriented meaning: Now X_{ijkl} defines the fraction of flow from commodity (i,j) which is routed via the hub edge $\{k,l\}$ (in this direction). In addition we define new variables $S_{ijkl} \geq 0\,(i,j,k,l \in \mathcal{N})$, which determine the fraction of flow from commodity (i,j) which is routed via the spoke edge $\{k,l\}$ (in this direction). As new binary variables we need $Y_{kl}\,(k \leq l)$, which is 1, if the edge $\{k,l\}$ is established as a hub edge, 0 otherwise.

Then constraint (6.2), which requires that all flow has to be routed, must be reformulated as a flow conservation law for the flow of commodity (i,j). Constraints (6.3) and (6.4) are rewritten by means of the new Y_{kl} variables.

The mixed integer formulation for the Public Transportation Hub Location Problem (PT) goes like this:

(PT)

$$\min \quad \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} W_{ij}\Big[\sum_{\{k,l\}\in\mathcal{E}} \alpha\, d_{kl}(X_{ijkl}+X_{ijlk}) + \sum_{k\in\mathcal{N}} d_{ik}\, S_{ijik} + \sum_{l\in\mathcal{N}:l\neq i} d_{lj} S_{ijlj}\Big] + \sum_{\{k,l\}\in\mathcal{E}} I_{kl}\, Y_{kl} + \sum_{k\in\mathcal{N}} F_k\, H_k \tag{6.9}$$

$$\text{s.t.} \quad \sum_{l\in\mathcal{N}} (X_{ijkl} + S_{ijkl} - X_{ijlk} - S_{ijlk}) = \begin{cases} +1, & \forall\, i,j,k \in \mathcal{N} : k = i, i \neq j, \\ -1, & \forall\, i,j,k \in \mathcal{N} : k = j, i \neq j, \\ 0, & \forall\, i,j,k \in \mathcal{N} : k \neq i, k \neq j, \end{cases} \tag{6.10}$$

$$\sum_{l\in N} (X_{iiil} + S_{iiil}) = 1 \quad \forall\, i \in \mathcal{N}, \tag{6.11}$$

$$\sum_{l\in N} (X_{iili} + S_{iili}) = 1 \quad \forall\, i \in \mathcal{N}, \tag{6.12}$$

$$X_{ijkl} \leq Y_{kl} \qquad \forall\, i,j \in \mathcal{N}, \{k,l\} \in \mathcal{E}, \tag{6.13}$$

$$X_{ijlk} \leq Y_{kl} \qquad \forall\, i,j \in \mathcal{N}, \{k,l\} \in \mathcal{E}, \tag{6.14}$$

$$S_{ijik} \leq H_k \qquad \forall\, i,j,k \in \mathcal{N} : k \neq j, \tag{6.15}$$

$$S_{ijkj} \leq H_k \qquad \forall\, i,j,k \in \mathcal{N} : k \neq i, \tag{6.16}$$

$$S_{ijij} \leq H_i + H_j \qquad \forall\, i,j \in \mathcal{N}, \tag{6.17}$$

$$S_{ijkl} = 0 \qquad \forall\, i,j,k,l \in \mathcal{N} : k \neq i, l \neq j, \tag{6.18}$$

$$Y_{kl} \leq H_k \qquad \forall\, \{k,l\} \in \mathcal{E}, \tag{6.19}$$

$$Y_{kl} \leq H_l \qquad \forall\, \{k,l\} \in \mathcal{E}, \tag{6.20}$$

$$S_{ijkl}, X_{ijkl} \geq 0 \qquad \forall\, i,j,k,l \in \mathcal{N}, \tag{6.21}$$

$$Y_{kl}, H_k \in \{0,1\} \qquad \forall\, k,l \in \mathcal{N}. \tag{6.22}$$

The first constraints are the flow conservation law for the flows of commodity $(i,j), i \neq j$ (6.10) and (i,i) ((6.11) and (6.12)). The latter is only needed if $W_{ii} \neq 0$ for some $i \in \mathcal{N}$. In (6.13) and (6.14) it is required that interhub connections must be routed via hub edges. Spoke connections must begin or end at a hub, which is required in (6.15) and (6.16). Direct connections must also be routed via at least one hub (6.17), which is exactly general assumption (c). These constraints may be dropped in some applications for public transportation. Spoke connections are only allowed for the first and last edge of every flow path (6.18). Hub edges must begin and end at a hub node ((6.19) and (6.20)).

As the edge $\{k,l\}$ will be used only in one direction on a path from i to j, (6.13) and (6.14) can be replaced by

$$X_{ijkl} + X_{ijlk} \leq Y_{kl} \quad \forall i,j \in \mathcal{N}, \{k,l\} \in \mathcal{E} \tag{6.23}$$

An aggregation over the constraints (6.23) is possible by summing up over all commodities (i,j). This will reduce the number of constraints, but unfortunately it also makes the LP relaxation weaker (see [2]).

We remark that there exists an optimal solution of (PT) in which all X_{ijkl} and S_{ijkl} variables are binary, as it is usual in uncapacitated multiple allocation hub location problems. This means that all flow between each pair of nodes is routed only via one path. We also note that $Y_{kl} = 1$ for some $\{k,l\} \in \mathcal{E}$ implies $S_{ijkl} = 0$ for all $i,j \in \mathcal{N}$ in an optimal solution of (PT).

Now we introduce a second model that allows a more generalized application: We introduce fixed costs for establishing spoke edges, but we allow these spoke edges to be used without any restriction.

The MIP formulation is again a network design problem with two different kinds of edges to be established.

In addition to the variable definitions in (PT) we introduce a new binary variable Z_{kl} for all $\{k,l\} \in \mathcal{E}$, which is 1, if edge $\{k,l\}$ is a spoke edge; 0 otherwise. As a new parameter we have $J_{kl} \geq 0$ to describe the fixed cost for establishing an undirected spoke edge between nodes k and l. We assume that $J_{kl} \leq I_{kl}$ for all $\{k,l\} \in \mathcal{E}$.

The Generalized Public Transportation Hub Location Problem (GPT) is formulated as follows:

(GPT)

$$\begin{aligned}\min \quad & \sum_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}} W_{ij}\Big[\sum_{\{k,l\}\in\mathcal{E}} d_{kl}\big(\alpha(X_{ijkl}+X_{ijlk})+S_{ijkl}+S_{ijlk}\big)\Big] \\ & + \sum_{\{k,l\}\in\mathcal{E}} I_{kl}\,Y_{kl} + \sum_{\{k,l\}\in\mathcal{E}} J_{kl}\,Z_{kl} + \sum_{k\in\mathcal{N}} F_k\,H_k \qquad (6.24)\end{aligned}$$

$$\begin{aligned}\text{s.t.} \quad & (6.10),(6.11),(6.12),(6.13), \\ & (6.14),(6.19),(6.20), \\ & S_{ijkl} \leq Z_{kl} \qquad \forall\, i,j \in \mathcal{N}, \{k,l\} \in \mathcal{E}, \qquad (6.25) \\ & S_{ijlk} \leq Z_{kl} \qquad \forall\, i,j \in \mathcal{N}, \{k,l\} \in \mathcal{E}, \qquad (6.26) \\ & S_{ijkl}, X_{ijkl} \geq 0 \qquad \forall\, i,j,k,l \in \mathcal{N}, \qquad (6.27) \\ & Y_{kl}, Z_{kl}, H_k \in \{0,1\} \qquad \forall\, \{k,l\} \in \mathcal{E}. \qquad (6.28)\end{aligned}$$

In (6.25) and (6.26) it is required that spoke connections must be routed via spoke edges. Certainly (6.25) and (6.26) may be summarized and aggregated as it was shown for (6.13) and (6.14). Note that there exists an optimal solution in which for every $\{k,l\} \in \mathcal{E}$ either Y_{kl} or Z_{kl} is zero.

A weakness in model (GPT) may be that there is no limitation for the number of changes in the type of edge (from spoke edge to hub edge and vice versa) on every flow path. In applications for public transportation this means some passengers may have to change their means of traffic very often, which will be inconvenient. Therefore in model (GPT) we can limit the number of such changes not be greater than q, where $q \in \mathbb{N} \cup \{0\}$. As a new binary variable we need C_{ijl}, which is 1, if the type of edge is changed on the trip from i to j at node l; 0 otherwise.

The additional constraints are

$$C_{ijl} \geq \sum_{k \in N} (X_{ijkl} - X_{ijlk}) \quad \forall i, j, l \in \mathcal{N} : l \neq i, j \tag{6.29}$$

$$C_{ijl} \geq \sum_{k \in N} (X_{ijlk} - X_{ijkl}) \quad \forall i, j, l \in \mathcal{N} : l \neq i, j \tag{6.30}$$

$$\sum_{l \in \mathcal{N}} C_{ijl} \leq q \quad \forall i, j \in \mathcal{N}, \tag{6.31}$$

$$C_{ijl} \in \{0, 1\} \quad \forall i, j, l \in \mathcal{N}. \tag{6.32}$$

In (6.29) and (6.30) the value of C_{ijl} is 1 if and only if the type of edge used for the flow from node i to node j is changed at node l, i.e., iff the amount of flow of commodity (i, j) arriving at node l via hub edges is different from the amount of this flow departing from node l via hub edges. The sum of changes in the type of edge on every flow path must be less or equal to q (6.31).

3. SOLUTION APPROACHES

The new models (PT) and (GPT) have approximately the same size as the usual ones, e.g. (UFC). There are $O(|\mathcal{N}|^4)$ variables and $O(|\mathcal{N}|^4)$ linear constraints required. However, for network design problems several exact and good heuristic solution algorithms are known, e.g. dual ascent methods [2] and branch–and–bound algorithms using Bender's Decomposition [12] so that those methods can be applied also for solving the new hub location problems (PT) and (GPT).

Here we will look at solution methods based on shortest path algorithms. As already mentioned in Section 1, usual hub location problems consist of two parts: a location and an allocation part. If the set of hubs $\mathcal{H}$ is already fixed, i.e. the location part is solved, the allocation part of an uncapacitated multiple allocation hub location problem can be solved by an all–pairs shortest–path algorithm, e.g. the Floyd–Warshall algorithm (see [1]), in time complexity $O(|\mathcal{H}| \cdot |\mathcal{N}|^2)$ (see [7]), where every shortest path must be routed only via hub nodes. For the location part one can apply an exact enumeration algorithm, branch–and–bound tech-

niques using clustering theory [8] and several heuristics, e.g. Greedy or interchange [7].

Now if we look at our new network design models (PT) and (GPT), the hub location problem consists of three parts: There are two location parts, one for the location of the hub nodes and one for the location of the hub edges, and the allocation part. The allocation part can be solved again by an all–pairs shortest path algorithm. In (PT) only hub edges must be used on every shortest path except for the first and last edge, while in (GPT) these shortest paths can be routed via hub and spoke edges as well. Now we have a closer look at the location part for the hub edges. If the set of hub nodes $\mathcal{H}$ is already fixed, then there are $2^{\binom{|\mathcal{H}|}{2}}$ possibilities to locate hub edges. An exact algorithm may only be useful for very small $|\mathcal{H}|$, so again a heuristic algorithm has to be developed. The idea of such a heuristic algorithm is to let the hub level be connected, because for transportation in the hub level it can be taken advantage of the discount factor α for many flow paths. A least expensive connected starting configuration would be a (fixed–cost) minimal spanning tree (MST), which can be computed efficiently in $O(|\mathcal{H}|^2)$ time by means of e.g. the shortest–path algorithm of Dijkstra (see [1]). We can apply a Greedy and/or interchange heuristic to this MST to improve the objective value of the hub location problem.

4. AN ILLUSTRATIVE EXAMPLE

We will now compare the three mixed integer programs on a numerical example. We use the 10–node problem of the Civil Aeronautics Board (CAB) data set described in [13]. This data contains the passenger flows and distances between ten major cities in the U.S. in 1970. As there are no fixed costs given in this data set, we introduce the following values:

$F_i := 50,000,000$ for all $i \in \mathcal{N}$ (fixed cost for establishing a hub node). $I_{kl} := 5,000 \cdot d_{kl}$ for all $k, l \in \mathcal{N}$ (fixed cost for establishing a hub edge). To include also (GPT) in the comparing process we set $J_{kl} := 0$ for all $k, l \in \mathcal{N}$ in (GPT) (fixed cost for establishing a spoke edge). We set q to different integer values from 0 to 2. We tested the models with AMPLPlus [10] and used XPRESS-MP [4] to solve the Mixed Integer Programs.

In Table 6.1 we compare the objective value (which in this case is the cost per unit of flow) and the location of hub nodes and hub edges for programs (UFC) and (PT). Remember that in (UFC) every edge between all pairs of hubs has to be located, so we can drop the column of hub edges for (UFC).

	(UFC)		(PT)		
α	opt.sol.	hubs	opt.sol.	hubs	hub edges
0.0	489.52	1,3,4,6,7,8	435.28	1,3,4,6,7,8	{1,6},{1,7},{3,6}, {4,6},{7,8}
0.2	601.59	1,3,4,6,7	560.33	1,3,4,6,7,8	{1,6},{1,7},{3,6}, {4,6},{4,7},{4,8}, {7,8}
0.4	692.08	3,4,6,7	668.25	3,4,6,7,8	{3,6},{4,6},{4,7}, {4,8},{7,8}
0.6	747.18	4,6,7	742.70	3,4,6,7	{3,6},{4,6},{4,7}
0.8	786.19	4,6,7	785.08	4,6,7	{4,6},{4,7}
1.0	815.06	4,6,7	804.50	4,6,7	—

Table 6.1 Comparison between programs (UFC) and (PT)

Another point of interest may be the gap between the optimal and the minimal spanning tree (MST) solution for program (PT). This is visualized for the 10 nodes problems of the CAB data set with different values of α (see Figure 6.2). The solid thick and thin lines represent the hub and spoke edges, respectively, of an MST solution. The edges described by dashed lines must be added to the set of hub edges of the MST solution to obtain the optimal solution. In case of $\alpha = 1$, the edges corresponding to dotted lines must be removed from the set of hub edges of the MST solution to obtain the optimal solution.

For $\alpha = 0$ the MST solution is already optimal. The reason for this is that transportation in the hub level is free, so no additional hub edges are needed. All non–hub nodes are allocated to the nearest hub. If α increases up to 0.2, more hub edges are built in order to take advantage of the discount factor. However, a further increase of α leads to a reduction of hub edges again because the discount is getting less. This results again in MST optimal solutions for $\alpha = 0.6$ and 0.8. For $\alpha = 1$ no discount is given any more, so hub edges are not built at all. Starting with the MST solution all hub edges must be removed to obtain the optimal solution. We note that the MST gap, i.e. the relative gap between MST and optimal solution, is always $\leq 1.5\%$ in this example.

In Table 6.2 we compute the optimal solution for program (GPT). First we solve (GPT) with the additional constraints (6.29) — (6.32) for $q \in$

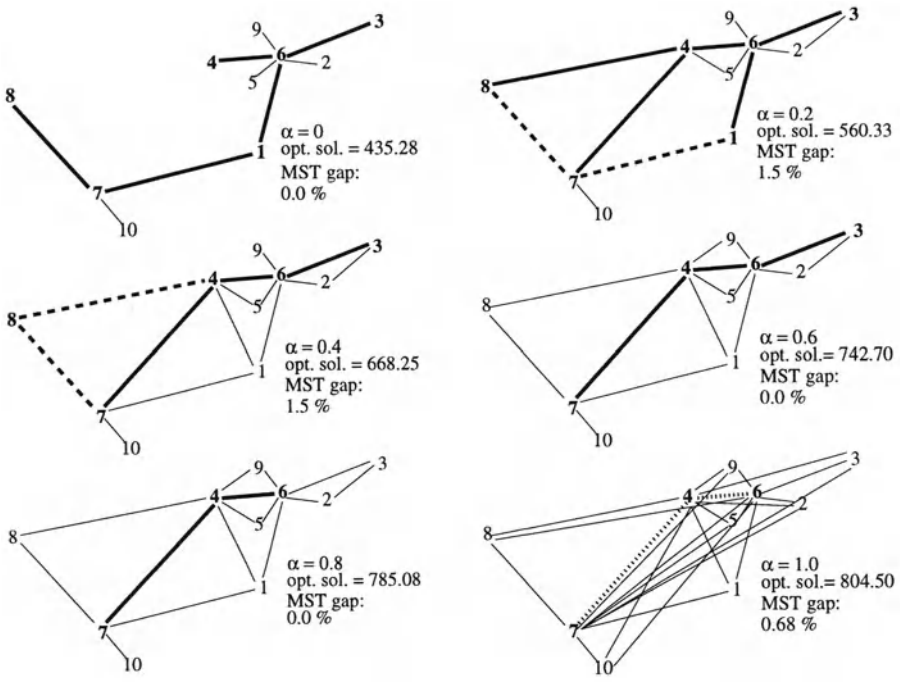

Figure 6.2 Comparison of MST and optimal solution for (PT)

	(GPT)			
α	$q = 0$	$q = 1$	$q = 2$	basic version
0	509.69	446.81	433.35	433.35
0.2	619.07	565.53	557.02	557.02
≥ 0.4	619.07	619.07	619.07	619.07

Table 6.2 Optimal solutions of (GPT) for different q

$\{0, 1, 2\}$, then we solve the basic version without additional constraints. We note that for all $\alpha \in [0, 1]$ the optimal solutions of $q = 2$ and the basic version are the same. This means that for model (GPT) without the additional constraints (6.29) – (6.32) we will get an optimal solution where every passenger has to change the means of traffic at most twice. For $\alpha \geq 0.4$ it is not worth to build any hub node or hub edges, so the optimal solution is to establish only spoke edges. This is the reason why the optimal solution is the same for all $\alpha \geq 0.4$ and all $q \geq 0$.

5. CONCLUSIONS

In this paper we presented new mixed integer formulations for hub location problems which are applicable for urban public transportation networks. Starting with the Uncapacitated Fixed Charge Multiple Allocation Hub Location Problem (with additional fixed costs for hub edges) we relaxed the requirement that the hub level has to be a complete graph. The new formulations are based on network design problems, in which every flow path can use more than one hub edge. While the first new model still requires that spoke edges are only allowed to be used as the first and last edge of every flow path, hub and spoke edges can be located in the second new model in an arbitrary way.

In future research, the solution approaches described in this paper will be implemented and tested on numerical examples. Bounds gained from heuristics can be used to construct branch–and–bound–algorithms. The feasibility polytopes of the LP relaxations may be examined to determine facets. From the modeling view even better applicable models can be found. As the models described here mainly consider the objective of cost minimizing, customer satisfaction (e.g. ticket prices, waiting times) can be included. Capacitated versions of the new models may be studied, i.e. some or all hub nodes and/or hub edges can only deal with a limited amount of passenger flow.

References

[1] Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Network flows: theory, algorithms and applications. Prentice Hall, Engelwood Cliffs, New Jersey.

[2] Balakrishnan, A., Magnanti, T. L., and Wong, R. (1989). A dual ascent procedure for large-scale uncapacitated network design. Operations Research, 37(5):716–740.

[3] Campbell, J. F. (1994). Integer programming formulations of discrete hub location problems. European Journal of Operational Research, 72:387–405.

[4] DASH (1999). XPRESS-MP. Internet: www.dash.co.uk.

[5] Ebery, J., Krishnamoorthy, M., Ernst, A., and Boland, N. (1998). The capacitated multiple allocation hub location problem: Formulations and algorithms. European Journal of Operational Research, forthcoming.

[6] Ernst, A. T. and Krishnamoorthy, M. (1996). Efficient algorithms for the uncapacitated single allocation p-hub median problem. Location Science, 4(3):139–154.

[7] Ernst, A. T. and Krishnamoorthy, M. (1998a). Exact and heuristic algorithms for the uncapacitated multiple allocation p-hub median problem. European Journal of Operational Research, 104.1:100–112.

[8] Ernst, A. T. and Krishnamoorthy, M. (1998b). Shortest–path based lower bounding methods for p–hub median problems. INFORMS Journal on Computing, 10(2):149–162.

[9] Fasbender, G. (1999). Problème de localisation de hubs sans capacité. Master's thesis, Université Libre de Bruxelles, Faculté de Sciences. (in French).

[10] Fourer, R., Gay, D. M., and Kerningham, B. W. (1993). AMPL: A Modeling Language for Mathematical Programming. The Scientific Press, San Francisco.

[11] Labbé, M. (1998). Private communication.

[12] Magnanti, T. L. and Wong, R. T. (1984). Network design and transportation planning: Models and algorithms. Transportation Science, 18:1–55.

[13] O'Kelly, M. E. (1987). A quadratic integer program for the location of interacting hub facilities. European Journal of Operational Research, 32:393–404.

[14] Skorin-Kapov, D., Skorin-Kapov, J., and O'Kelly, M. (1996). Tight linear programming relaxations of uncapacitated p-hub median problems. European Journal of Operational Research, 94:582–593.

Chapter 7

Stochastic assignment to high frequency transit networks:

models, algorithms, and applications with different perceived cost distributions

Giulio Erberto Cantarella
University of Salerno – Dept. Of Civil Engineering
via Ponte Don Melillo – 84084 Fisciano (SA) – Italy
giucanta@starnet.it tel +39-089-964118 fax +39-089-964045

Antonino Vitetta
University of Reggio Calabria– Dept. of Comp.Sci., Math., Electr., and Transportation
Feo di Vito – 89100 Reggio Calabria – Italy
vitetta@unirc.it tel +39-0965-875205 fax +39-0965-875227

Abstract In urban areas transportation demand is commonly served both by private cars and mass transit systems (bus, tram, metro, etc.), which are usually based on a network of partially competing and overlapping lines. Therefore, user pre-trip path choice behaviour in transit systemsrefers to overall strategies that specify which line will be boarded at each bus stop (more generally how the user will behave at diversion nodes). The topology of user strategies are effectively modelled by hyperpaths (introduced by Nguyen and Pallottino, 1988). The pre-trip choice among hyperpaths (say strategies) is currently simulated through a deterministic choice model. This assumption leads to deterministic network loading or user equilibrium assignment for uncongested or congested network respectively. In this paper the pre-trip choice among hyperpaths is simulated through probabilistic choice models derived from random utility theory. Resulting stochastic network loading and user equilibrium models are analysed as well as solution algorithms. Results of an application to a test system and a real one are also reported, using different perceived cost distribution.

Keywords: Transit Assignment, Stochastic assignment

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 109–129.

1. INTRODUCTION

The supply-demand interaction in a transportation system is commonly studied with equilibrium models, as introduced by Wardrop (1952), for deterministic choice behaviour. In early studies the stochastic assignment to uncongested networks, based on probabilistic choice behaviour, was proposed by Dial (1971) using a Logit choice model. Since then, this assignment has been used in several subsequent studies with the generalisation of Logit-based assignment to congested networks (Fisk, 1980) and with the use of Probit path choice model (Daganzo and Sheffi 1977). Reviews and references about equilibrium models are in Sheffi and Powell (1982), Patriksson (1994) and Cascetta (1998). More recently dynamic process models for transportation system have been proposed (a review in Cantarella e Cascetta, 1995).

Path choice behaviour embedded within transit network assignment is modelled through two different approaches:

- for high frequency systems user behaviour is generally *mixed pre-trip*, at the origin the user chooses a travel strategy possibly involving several lines, and *en-route*, at each bus stop the user chooses a transit line among those available; this is often the case of urban areas;
- for low frequency systems user behaviour is generally *fully pre-trip*, as usually occurs for inter-city travels

In this paper main emphasis is on the former case, usually approached assuming an average headway for each bus line. The latter case, not dealt with in this paper, would require a timetable based approach.

After some preliminary studies, Spiess (1984) introduced the concept of optimal strategy in a general form. Nguyen and Pallottino (1988) proposed the hyperpath concept to describe the users' strategy topology. Wu and Florian (1993) and Wu *et al.* (1994) have analysed deterministic assignment to congested transit network. These models are based on some simplifying assumptions about user behaviour at a bus stop. Recently Bouzaiene-Ayari *et al.* (1995, 1997) proposed an extended analysis of behaviour at a bus stop.

Whilst all the above models are based on deterministic choice behaviour, an urban transit system with high frequency lines can better be analysed through a stochastic approach, since the travel time associated to most links is better modelled as a random variable. For instance travel time on a line link shows dispersion since it depends on the traffic congestion on the transit mode and on the other interacting modes as well as waiting time at a bus stop due to service irregularity. Moreover, connector links between a centroid and the real network simulate different pedestrian routes within a traffic zone and the associated travel time should better be modelled as a random variable, for its high dispersion. Cantarella (1997) proposed a general fixed-point

approach for stochastic equilibrium assignment that can also be applied to congested transit networks, according to the above assumptions.

In this paper models and algorithms for stochastic assignment to high frequency transit networks are proposed and tested. New elements are introduced for modelling users path choice behaviour, testing several distributions for perceived cost as well as resulting network loading algorithms. In section 2 assignment models for a transit system are reported and in section 3 solution algorithms are discussed. Some numerical results for a test system and a real one are presented in section 4. Conclusions and some indications for further research development are reported in section 5.

2. ASSIGNMENT MODELS

Assignment models simulate how origin-destination demand flows affect link flows in a transportation network, and resulting performances. Traffic assignment models are generally made up by:

- a *supply model*, simulating how the network performances, such as travel times, are affected by user' choices;
- a *demand model*, simulating how user behaviour, such as route choice, is affected by network performances;
- a *supply/demand interaction model*, which simulates the interaction between users' behaviour and network performances.

These three models are briefly described below with reference to a transit system.

2.1 Supply model

Supply is generally simulated through a flow network model. In particular, the topology can be described through a graph, by concentrating origins and destinations of journeys into centroids, so that all users travelling between an origin-destination pair share a set of relevant paths. Throughout this paper, it is assumed that at least one path connects each O-D pair, and only elementary (say loop-free) paths are considered, thus only a finite number of paths exists for each user class. For each origin-destination pair OD, the relationship among the links and the (elementary) paths can be described by the link-path incidence matrix, Δ_{OD}, with entries $\delta_{OD,ij}$ equal to 1 if link i belongs to the path j and zero otherwise.

The link flow vector, $\mathbf{f} = \Sigma_{OD} \mathbf{f}_{OD}$, is obtained by summing up the vectors of the link flows $\mathbf{f}_{OD}$ coming from each OD pair. Each flow vector per OD pair $\mathbf{f}_{OD}$ is obtained from path flow vector $\mathbf{h}_{OD}$ through the link-path

incidence matrix Δ_{OD}, thus $\mathbf{f}_{OD} = \Delta_{OD}\, \mathbf{h}_{OD}$. Hence, the link flow vector is given by:

$$\mathbf{f} = \Sigma_{OD}\, \Delta_{OD}\, \mathbf{h}_{OD} \tag{1}$$

For congested networks, the link cost vector **c** depends on the flow vector:

$$\mathbf{c} = \boldsymbol{c}(\mathbf{f}) \tag{2}$$

Finally, the additive path cost vector, $\mathbf{g}_{OD}^{ADD}$, for each OD pair is given by:

$$\mathbf{g}_{OD}^{ADD} = \Delta_{OD}^{T}\, \mathbf{c} \tag{3}$$

The network simulating an urban bus (and/or tram, metro, etc.) transit system is made up starting from the pedestrian network, whose elements are the *pedestrian links* and *nodes*, including a node for each bus stop. *Centroids* as origins or destinations of journeys are suitably connected to the pedestrian network by *connector links*. Those elements allow simulating access and egress to/from the transit system. Then, each bus line is simulated through a linear graph, whose elements are the *line links* and *nodes*.

Finally, for each bus stop one *waiting node* is introduced, from which users may board different lines, through *boarding links*. Moreover, line nodes are connected through *alighting links* to the corresponding pedestrian stop node. Finally, this pedestrian node is connected to the waiting node through a *waiting link* (Fig. 1).

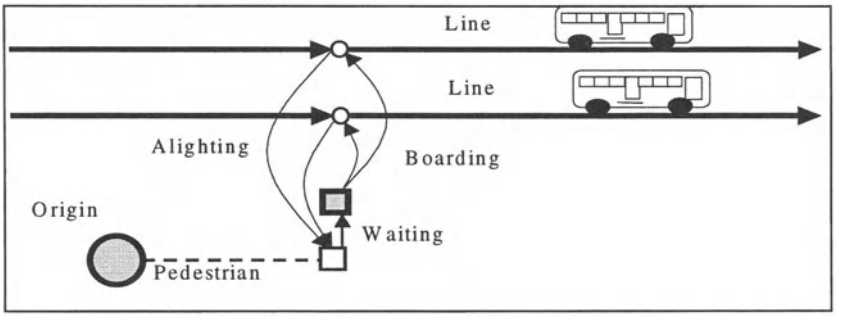

Fig. 1 - A bus stop representation

2.2 Demand model

In a high frequency transit system, the path actually followed by a user from an origin to a destination depends on a choice carried out before the trip starts and possibly on choices carried out during the trip. At the origin, before starting the journey, the user may consider as routing alternatives a single path or sets of partially overlapping paths, such that the path actually followed is defined by en-route choices at *diversion nodes*. In the latter case the user choice is a travel strategy expressing rules for en-route choices.

For simplicity's sake in the following only waiting nodes will be considered diversion nodes. Thus, en-route choices concern which line is boarded at a bus stop, whilst a strategy defines which lines are to be considered at each bust stop, called *attractive lines*, and how to choose among them. A commonly adopted rule is simply "take the first bus arriving at the stop", provided that it belongs to an attractive line. It should be stressed that a line may be attractive in a strategy but not in another one.

Once rules for behaviour at waiting (diversion) nodes are given a pre-trip strategy is completely defined by its topology, which in turn may be represented by an hyperpath (as introduced by Nguyen and Pallottino, 1988). Formally, an (elementary) hyperpath is an acyclic graph oriented from an origin node to a destination node so that each node is connected to the destination and the origin is connected to each node. In a path each node (a part from the origin and destination) has only one entering and one exiting link. Vice versa, in a hyperpath a waiting (diversion) node, where users carry out en-route choices, has as many exiting links as attractive lines, (but still only one entering link), whilst each of the other nodes has only one entering and one exiting link. As such a hyperpath is made up by some (elementary) paths overlapping at waiting nodes, or of course just one elementary path. It should be noted that a path may well belong to several hyperpaths.

2.2.1 En-route choice behaviour

En-route choice behaviour can be simulated through diversion probabilities defined for each waiting node within each strategy, say each hyperpath. In particular, in a hyperpath j, an *en-route diversion probability* η_{aj} is associated to each boarding link a (corresponding to an attractive line) exiting from a waiting node n:

$$\eta_{aj} \in [0,1] \quad a \in j \quad \text{boarding link}$$

with $\Sigma_{a \text{ exiting from the same waiting node n}} \eta_{aj} = 1$

For consistency, it is also assumed that

$$\eta_{aj} = 1 \qquad a \in j \qquad \text{non-boarding link}$$
$$\eta_{aj} = 0 \qquad a \notin j$$

It should be noted that a boarding link corresponding to a line non attractive does not belong to the hyperpath, thus it receives a zero diversion probability. If hyperpath j is just a single path the diversion probabilities are equal to one or zero.

Generally, and in this paper too, diversion probabilities are assumed independent of link flows. Recently Bouzaiene-Ayari *et al.* (1995, 1997) proposed an extended analysis of behaviour at a bus stop, including diversion probability depending on link flows.

The assumptions about en-route choice behaviour allow defining the waiting time for each waiting link. It should be noted that this time depends on the considered attractive lines. Thus, it depends on the hyperpath, and cannot be considered a network characteristic, such as the time associated to other types of links.

According to the commonly adopted rule "take the first bus arriving at the stop", en-route diversion probabilities at a waiting node n are proportional to the frequencies of attractive lines. In this case, the time associated to the waiting link entering node n is given by the inverse of sum of the frequencies of the attractive lines.

The *path-hyperpath probability* ω_{kj} of path k contained in hyperpath j is the product of the diversion probabilities of the links in the path:

$$\omega_{kj} = \Pi_{a \in k} \eta_{aj} \in [0,1]$$

thus

$$\omega_{kj} > 0 \Leftrightarrow k \subseteq j$$
$$\omega_{kj} = 0 \Leftrightarrow k \not\subset j$$

If hyperpath j is actually a path these probabilities are equal to one or zero.

Thus, en-route choice behaviour between pair OD can be modelled through the path-hyperpath probability matrix, Ω_{OD}, whose entries ω_{kj} represent the probability to use a path k within the hyperpath j (equal to zero if path k is not contained in hyperpath j). Clearly the sum of the entries in each column is equal to 1. In addition, since each path is also an hyperpath matrix Ω_{OD} contains an identity matrix.

2.2.2 Pre-trip choice behaviour

Pre-trip choice behaviour can be simulated assuming that users perception of hyperpaths connecting each OD pair can be described by the perceived utility vector $\mathbf{u}_{OD}$ modelled as a random variable, usually expressed as the sum of the expected value, or systematic utility, $\mathbf{v}_{OD} = E[\mathbf{u}_{OD}]$ and a random residual $\xi_{OD} = \mathbf{v}_{OD} - E[\mathbf{u}_{OD}]$.

The cost attributes associated to a hyperpath allow specifying the systematic utility. In particular, time values associated to connector, pedestrian, boarding, line and alighting links can be summed up to made up additive path costs as in equation (3). The corresponding additive cost attribute, x_j^{ADD} , for an hyperpath j is given by averaging additive costs, g_k^{ADD}, of paths contained in the hyperpath, say $k \in j$ or $\omega_{kj} > 0$, with respect to path hyperpath probabilities, ω_{kj}:

$$\mathbf{x}_{OD}^{ADD} = \Omega_{OD}^{T} \mathbf{g}_{OD}^{ADD}$$

On the other hand, the time values for waiting links depend on the hyperpath, thus cannot be defined by summing up network characteristics. Hence, the total waiting time associated to a hyperpath is not additive, x_j^{NA}, as the other costs previously discussed.

Thus, the systematic utility vector $\mathbf{v}_{OD}$ is given by the opposite of the sum of the additive cost vector $\mathbf{x}_{OD}^{ADD}$, obtained from link costs, and a non additive cost vector $\mathbf{x}_{OD}^{NA}$, representing waiting times (and possibly other non additive costs such as number of transfers, fare, and the like):

$$\mathbf{v}_{OD} = - \Omega_{OD}^{T} \mathbf{g}_{OD}^{ADD} - \mathbf{x}_{OD}^{NA} \qquad (4)$$

The probability of choosing hyperpath j, between the origin-destination pair OD, is given by the probability of hyperpath j being the maximum perceived utility one. Hence, the choice probability vector $\mathbf{p}_{OD}$ depends on the systematic utility (and the parameter of the random residual distribution):

$$\mathbf{p}_{OD} = \boldsymbol{p}_{OD}(\mathbf{v}_{OD}) \qquad (5)$$

In the following it will be assumed that a probabilistic choice model, derived from the random utility theory (see BenAkiva and Lerhman, 1987, for a review), specifies equation (5), thus relation in eqn (5) is a function[1] (Cantarella, 1997), whose expression depends on hyperpath perceived utility,

[1] Results in this paper can be extended to a deterministic choice model, even if in this case eqn (5) is not a function but a one-to-many map (also called a multi-valued relation).

or random residual, joint distribution. In particular, hyperpath perceived utility can be specified through link perceived utility modelled as a random variable (see section 4.1 for some examples). The hyperpath flow vector, $\mathbf{y}_{OD}$ is given by the choice probability vector $\mathbf{p}_{OD}$ multiplied by the demand flow d_{OD}:

$$\mathbf{y}_{OD} = d_{OD}\mathbf{p}_{OD} \tag{6}$$

Finally, the path flow vector, $\mathbf{h}_{OD}$, is obtained from the hyperpath flow vector, $\mathbf{y}_{OD}$, through the matrix Ω_{OD}:

$$\mathbf{h}_{OD} = \Omega_{OD}\,\mathbf{y}_{OD} \tag{7}$$

2.3 Demand-supply interaction model

The supply model can be specified by expressing path costs as a function of path flows, through a relation obtained by combing together the above equations (1), (2), (3):

$$\mathbf{g}_{OD}^{ADD} = \Delta_{OD}^{T}\, \boldsymbol{c}(\Sigma_{OD}\,\Delta_{OD}\,\mathbf{h}_{OD})$$

The demand model can be specified by expressing path flows as a function of path costs, through a relation obtained by combing together the above equations (4), (5), (6), (7):

$$\mathbf{h}_{OD} = d_{OD}\,\Omega_{OD}\,\boldsymbol{p}_{OD}(-\,\Omega_{OD}^{T}\,\mathbf{g}_{OD}^{ADD} - \mathbf{x}_{OD}^{NA})$$

Thus, demand-supply interaction models, according to an equilibrium approach, can be specified by a set of non-linear equations, given by the supply and the demand models. A brief analysis will be carried out below.

The link flow vector, $\mathbf{f}_{SNL}$, resulting from the assignment to an uncongested network, may be expressed as a function of the link cost vector obtained by combining equation (1) with equations (3-7), and named the Stochastic Network Loading (SNL) function:

$$\mathbf{f}_{SNL} = \boldsymbol{f}(\mathbf{c}) = \Sigma_{OD}\, d_{OD}\,\Delta_{OD}\,\Omega_{OD}\,\boldsymbol{p}_{OD}(-\Omega_{OD}^{T}\,\Delta_{OD}^{T}\,\mathbf{c} - \mathbf{x}_{OD}^{NA}) \tag{8}$$

For congested networks, where link costs depend on link flows, the stochastic user equilibrium (SUE) link flow vector, $\mathbf{f}_{SUE}$, can be defined as the solution of a fixed-point model, obtained by combing the SNL function (8) with the cost functions (2):

$$\mathbf{f}_{SUE} = f(c(\mathbf{f}_{SUE})) \tag{9}$$

In a connected network at least one solution exists for the fixed point model (9) if:

- cost functions $\mathbf{c} = c(\mathbf{f})$ are continuous;
- SNL function $\mathbf{f} = f(\mathbf{c})$ is continuous (or all choice functions $\mathbf{p}_{OD} = p_{OD}(\mathbf{V}_{OD})$ are continuous).

At most one solution exists if:

- SNL function $\mathbf{f} = f(\mathbf{c})$ is monotone non increasing with respect to link costs as:

$$[\mathbf{f}(\mathbf{c}_1) - \mathbf{f}(\mathbf{c}_2)]^T (\mathbf{c}_1 - \mathbf{c}_2) \leq 0 \qquad \forall\, \mathbf{c}_1, \mathbf{c}_2$$

  (this feature is assured for additive choice models where random residuals are distributed independently from systematic utility: $\phi(\xi/\mathbf{v}) = \phi(\xi)$)

- cost functions $\mathbf{c} = \mathbf{c}(\mathbf{f})$ are monotone strictly increasing with respect to link flows as:

$$[\mathbf{c}(\mathbf{f}_1) - \mathbf{c}(\mathbf{f}_2)]^T (\mathbf{f}_1 - \mathbf{f}_2) > 0 \qquad \forall\, \mathbf{f}_1 \neq \mathbf{f}_2$$

  (this condition can be relaxed to monotone non decreasing choice models which provide strictly positive probabilities $\mathbf{p}_{OD}$ for any systematic utility $\mathbf{V}_{OD}$ such as Logit, Probit, Gammit and the like (Cantarella, 1997))

Whilst in most road transportation systems link cost functions can be considered separable, this is likely not the case in a transit system, hence monotonicity can be more difficult to check. An example is reported in section 4.

3. ALGORITHMS

If hyperpath perceived utility values are specified by independently distributed link perceived utility values, the SNL function (8) can be computed through a MonteCarlo technique (introduced by Burrell, 1968; see also Sheffi, 1985). In this case the result is an unbiased estimate of $f(\mathbf{c})$, obtained by averaging several shortest hyperpath loading for different pseudo-realisations of link perceived utility values, with expected values given by the link costs. A procedure is reported below.

Procedure SNL for computing $f_{SNL} = f(\mathbf{c})$
it := 0; $\mathbf{f}_{SNL}^{0} := \mathbf{0}$
repeat it := it + 1
$\mathbf{c}^*$:= *pseudo-realisation of perceived costs* with mean $\mathbf{c}$
$\mathbf{f}^{it}_{AON}$:= *shortest hyperpath loading* with costs $\mathbf{c}^*$
$\mathbf{f}^{it}_{SNL}$:= [(it-1) $\mathbf{f}^{it-1}_{SNL}$ + $\mathbf{f}^{it}_{AON}$] / it
until ($\mathbf{f}^{it-1}_{SUE} \cong \mathbf{f}^{it}_{SNL}$)

The shortest hyperpath loading can be computed through a generalisation of shortest path loading algorithms (Nguyen and Pallottino,1988).

The fixed-point model (9) can be solved through an MSA algorithm. In particular with the Flow Averaging (MSA-FA) algorithm at each iteration link flows are updated (Daganzo and Sheffi, 1983). A procedure is reported below.

Procedure SUE for computing f_{SUE}
it := 0; $\mathbf{f}^{0}_{SUE}$:= *starting solution*
repeat it := it + 1
$\mathbf{c}^{it} := \boldsymbol{c}(\mathbf{f}^{it-1})$
$\mathbf{f}^{it}_{SNL} := \boldsymbol{f}(\mathbf{c}^{it})$
$\mathbf{f}^{it}_{SUE}$:= [(it-1) $\mathbf{f}^{it-1}_{SUE}$ + $\mathbf{f}^{it}_{SNL}$] / it
until ($\mathbf{f}^{it-1}_{SUE} \cong \mathbf{f}^{it}_{SNL}$)

With the Cost Averaging (MSA-CA) algorithm at each iteration link costs are updated (Cantarella, 1997). A procedure is reported below.

Procedure SUE for computing f_{SUE}
it := 0; $\mathbf{f}^{0}_{SUE}$:= *starting solution*
$\mathbf{c}^{0} := \boldsymbol{c}(\mathbf{f}^{0}_{SUE})$
repeat it := it + 1
$\mathbf{f}^{it}_{SNL} := \boldsymbol{f}(\mathbf{c}^{it-1})$
$\mathbf{c}^{it}_{SNL} := \boldsymbol{c}(\mathbf{f}^{it}_{SNL})$
$\mathbf{c}^{it}$:= [(it-1) $\mathbf{c}^{it-1}$ + $\mathbf{c}^{it}_{SNL}$] / it
until ($\boldsymbol{f}(\mathbf{c}^{it}_{SNL}) \cong \mathbf{f}^{it}_{SNL}$)

For both the algorithms the starting solution $\mathbf{f}^{0}_{SUE}$ can be computed through SNL with zero-flow link costs. (Anyhow it does not actually contribute to the solution at the first iteration, thus to the final solution).

If existence and uniqueness conditions hold the convergence of MSA-FA algorithm can be guaranteed if the Jacobian of link cost flow functions is symmetric. For link cost function with non-symmetric Jacobian the convergence of the MSA-FA algorithm is generally not guaranteed (a part

from more complicated conditions). For additive choice models, as defined in subsection 2.3, the Jacobian matrix of the SNL function is symmetric, in this case the convergence of the MSA-CA can be guaranteed, if existence and uniqueness conditions hold). The MSA-CA convergence is generally slower than the MSA-FA algorithm. Thus, two-stage algorithms should be preferred, where the starting solution is obtained through the MSA-FA scheme and final steps are performed through MSA-CA scheme. Multi-stage algorithms are also useful to prevent step becoming too small.

4. NUMERICAL RESULTS

Results of applications to a test and a medium-size real urban network are reported in this section. The applications are developed considering uncongested and congested networks using for boarding links non-separable cost functions. Deterministic and random link perceived costs are considered in order to compare the influence of randomness on the hyperpath choice model. The used link cost performance functions are reported in subsection 4.1. Results are compared and discussed in subsections 4.2 and 4.3, respectively.

4.1 Link performance models

Link cost functions per link type (Fig. 2) are described below. In uncongested conditions, cost corresponding to zero link flow is used.

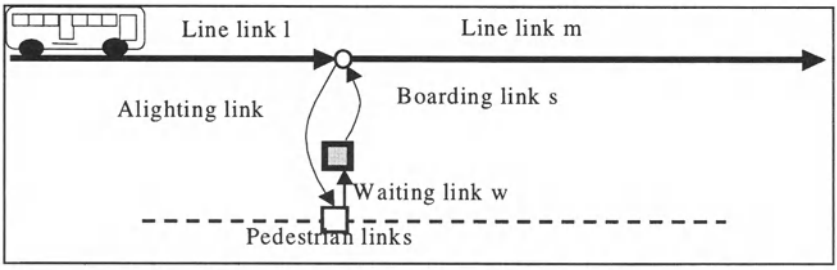

Fig. 2 - The links in a bus stop

Connector links. The connector link costs are not considered dependent on the link flows and the travel cost is equal to pedestrian travel time (speed equal to 1 m/sec).

Pedestrian links. The pedestrian link costs are not considered dependent on the link flows and the travel cost is equal to pedestrian travel time (speed equal to 1 m/sec).

Waiting links. The cost function c_w on the waiting link w, is:

$$c_w = \rho / \Sigma_{l \in j/w} \varphi_l \tag{12}$$

where
- φ_l is the line l frequency;
- ρ is a parameter belonging to [0.5, 1] interval in relation to the service irregularity (in this application $\rho = 0.5$ is used);
  j the considered hyperpath.

Alighting links. The alighting link costs are considered not dependent on the link flows and the alighting cost is considered equal to 20 seconds (it is the average observed alighting time).

Boarding links. The cost function c_s on a boarding link s, in a bus stop between line links l and m (fig. 2), is):

$$c_s = t_s [1 + \mu ((f_s + \nu (f_m - f_s))/ C_m)^{\gamma}] \tag{10}$$

where
- t_s is the free flow boarding time;
- f_s is the user flow on the link s;
- f_m is the user flow on the link m;
- C_m is the user capacity on the line containing the links m and l;
- μ, γ and ν are calibrated parameters geater than zero (in this application μ=0.1, ν=0.7 and γ=10, obtained calibrating the cost function from observed data).

Clearly, the cost increases with the flow on link s (f_s) and with the number of users on the bus ($f_m - f_s$).

Line links. The cost function c_m on the line link m, is:

$$c_m = t_m [1 + \tau (f_m / C_m)^{\gamma}] \tag{11}$$

where
- t_m is the running time on line link l for the bus when no users are on the bus;
- f_m is the user flow on the link m;
- C_m is the user capacity on the line containing the link m;
- τ and γ are calibrated parameters greater than zero (in this application τ=0.2 and γ=10, obtained calibrating the cost function from observed data).

The proposed link cost functions for boarding line links are monotone strictly increasing (hence the Jacobian is positive definite), if the following conditions hold:

$$0 < \nu < 1 \tag{12}$$

and

$$4 \tau (1 - \nu) > \mu \nu^2 \tag{13}$$

These conditions are satisfied by the used values.

Time on pedestrian links has been considered anyhow deterministic as well as frequency and waiting time on waiting links. For all the other links, time has been modelled both as a deterministic or random variable.

Three different (independent) distributions for link perceived costs have been considered: Normal, Log-Normal and Gamma. The route choice probability using a Normal (Probit Model), Log-Normal (Log-Probit model) or Gamma (Gammit) link perceived costs distributions cannot be expressed in a closed form; approximation functions (Abramowiz and Stegun, 1970) and MonteCarlo technique have been used for numeric route choice evaluation.

Probit model, based on Normal distribution, does not fit totally because it allows also negative perceived costs. Nielsen (1997) explores different ways of reducing this problem with symmetrical truncation of Normal or the use of Log-Normal distribution for perceived link costs (in this case independence from link segmentation is no longer assured, see also the end of this subs-section). Sheffi (1985) and Nielsen (1997) suggested the use of Gamma distribution for perceived costs, more recently Gammit choice models based on Gamma distribution have been deeply analysed by Cantarella and Binetti (2000) to overcoming all those drawbacks, but these models are more computing demanding.

Considering a link a with cost c_a, a reference link cost c_a^* not depending from link flow (the free flow link cost can be used) and a parameter θ greater than zero, the three distributions tested have the following parameters:

- *Normal*, with parameters $\mu_{N,a} = c_a$ (mean value) and $\sigma_{N,a}^2 = \theta\, c_a^*$ (variance);
- *Log-Normal*, with parameters $\mu_{L,a}$ and $\sigma_{L,a}^2$ obtained from $e^{(\mu_{L,a}+\sigma_{L,a}^2/2)} = c_a$ (mean value) and $e^{(2\mu_{L,a}+2\sigma_{L,a}^2)} - e^{(2\mu_{L,a}+2\sigma_{L,a}^2)} = \theta\, c_a^*$ (variance);
- *Gamma*, with parameters $\alpha_a = c_a/\theta$ and $\beta = 1/\theta$ (mean value $\alpha_a / \beta = c_a$ and variance $\alpha_a / \beta^2 = \theta\, c_a^*$).

With these assumptions, the three distributions considered for the link perceived cost have the same mean and variance values. Parameters could have been differentiated per link types, but for simplicity's sake this option has not been considered.

Considering a link a with perceived cost Normal distributed with parameters $\mu_{N,a}$ and $\sigma_{N,a}^2$, the additive cost of an elementary path k is also normal distributed with mean $\Sigma_{a\in k}\ \mu_{N,a} = \Sigma_{a\in k}\ c_a$ and variance $\Sigma_{a\in k}\ \sigma_{N,a}^2 = \theta\ \Sigma_{a\in k}\ c_a$ (for independently distributed link perceived costs). Also Gamma distribution, as long as link perceived costs are independently distributed with the same β value, has the additive property; in this case for path k, the path α parameter is given by $\Sigma_{a\in k}\ \alpha_a$ and the β value is equal to link parameter β. Generally this is not the case of Log-Normal distribution.

4.2 Test Network

The described models and solution procedures, with different distributions for link perceived costs, have been applied to a test system briefly described in the following and reported in Fig. 3. The network has three paths: the path 1 is composed only by one transit line; the paths 2 and 3 have an overlapping section. Besides these three (simple) hyperpaths, the network also contains other four (composed) hyperpaths. The total travel time on all the paths has been assumed equal for testing purpose.

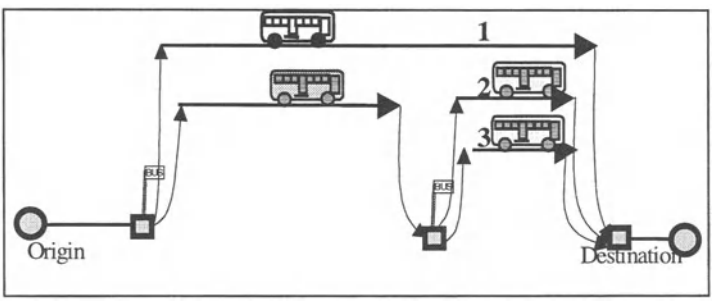

Fig. 3 - The transit test system

Tab. 1 – Elementary path choice probabilities for the test system (θ = 5 sec)

Link cost distribution	Normal	Log-Normal	Gamma
Path		Overlapping 0%	
1	33.33	33.33	33.33
2	33.33	33.33	33.33
3	33.33	33.33	33.33
		Overlapping 25%	
1	40.54	40.72	40.64
2	29.73	29.64	29.68

Link cost distribution	Normal	Log-Normal	Gamma
3	29.73	29.64	29.68
		Overlapping 50%	
1	43.18	43.54	43.34
2	28.41	28.23	28.33
3	28.41	28.23	28.33
		Overlapping 75%	
1	46.66	47.12	46.82
2	26.67	26.44	26.59
3	26.67	26.44	26.59

Tab. 1 reports the choice probabilities of each path with a SNL hyperpath assignment varying the overlapping degree between paths 2 and 3 and using a Normal, Log-Normal or Gamma link perceived costs. Parameter θ is considered identical and equal to 5 sec.

Four SNL assignments have been tested with overlapping degree between path 2 and 3 varying from 0% to 75%. The results obtained with Normal (and in some cases with Log-Normal) and Gamma distributions are comparable in term of path choice probabilities. Elaboration time is also comparable thus the use of Gamma or Log-Normal distributions, which avoid negative link costs perception, are preferable for practical applications. It should also be noted that the use of Log-Normal distribution does not guarantee independence from link segmentation, since it is not stable with respect to sum.

Some results for congested conditions are reported. The MSA-FA algorithm has been used, even if the Jacobian of link costs is asymmetric, to investigate it convergence under these conditions. It has been carried out in two macro-stages.

Stage 1 quickly provides a first guess of the equilibrium solution carrying out SNL procedures, within the SUE procedure, with 1 AON assignment. This stage id stopped after 50 iterations. Stage 2 allows to get closer to the equilibrium solution, carrying out SNL procedures, within the SUE procedure, with several AON assignments. This stage is stopped if the relative Euclidean distance between $\mathbf{f}^{it-1}_{SUE}$ and $\mathbf{f}^{it}_{SNL}$ is less than 0.01.

Fig. 4 reports the convergence values for 200 iterations of stage 2 (even if the convergence test is satisfied for fewer iterations) for the test network, using a Gamma distribution and parameter θ equal to 5 seconds, with different number of internal AON procedure in the stage 2.

In all applications the convergence depends on the number of internal AON procedure. Using 50 AON assignments in each SNL, the convergence index is between 0.5% and 6%; instead using 200 AON assignments, the convergence index is between 0.2% and 3%, as shown in Fig. 4. The number of internal AON assignments affects the limit value of the convergence

index, since this value is somehow lower bounded by the internal SNL variance, due to the MonteCarlo simulation.

4.3 Real Network

Results for a real system (transit system of Salerno, a medium size city in Italy) are also reported in uncongested and congested conditions. The real system is composed by 89 centroids, 3562 nodes and 5307 links. In the system 79 transit lines are present (Fig. 5). The number of peak hour users on the transit system is 4454.

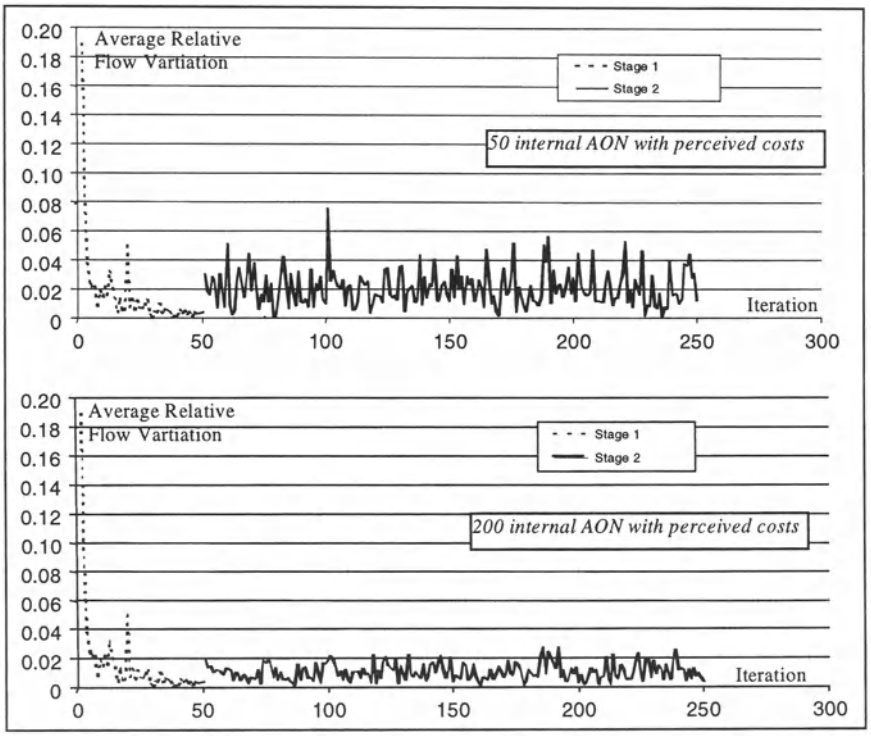

Fig. 4 - MSA algorithm convergence for the test network with different numbers of AON in SNL procedure

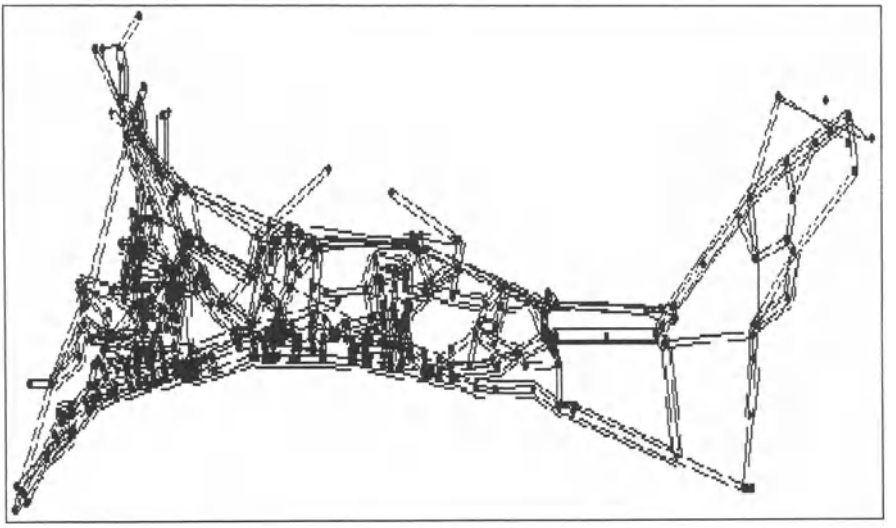

Fig. 5 – The graph of the real network

In Fig. 6 a comparison between AON and SNL assignment with Log-Normal or Gamma link costs perception distributions is reported for the transit line 12/A, which is a representative transit line in term of number of users and route. The results are obtained with two different variance levels (θ = 5 seconds and θ = 10 seconds). The flows from deterministic and stochastic assignment are rather different. In fact the system does not provide a perfectly regular service and deterministic transit assignment procedure is less realistic. The use of Log-Normal and Gamma distributions generate link flows not perfectly comparable.

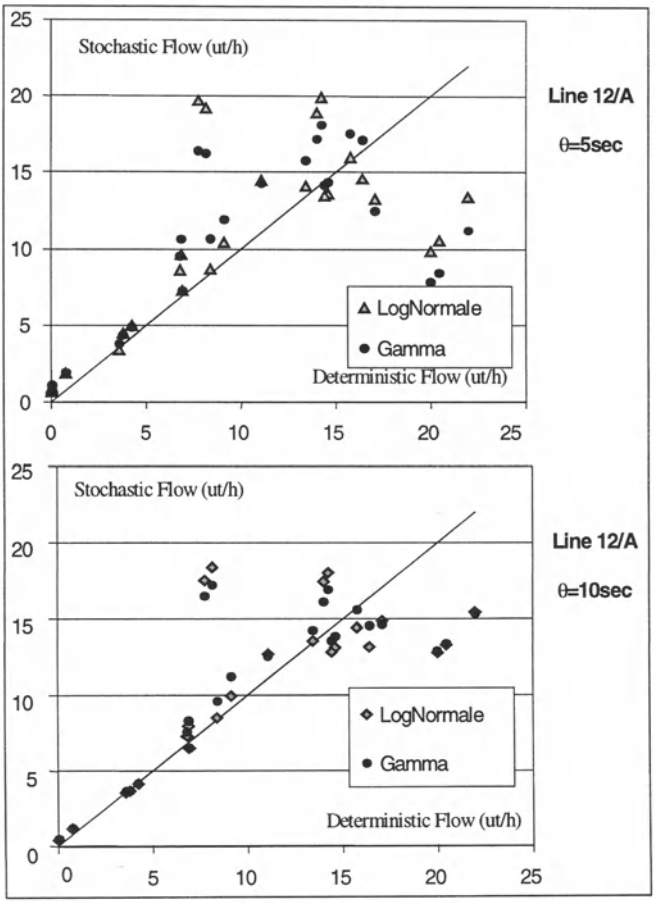

Fig. 6 - Comparison between deterministic (AON) and stochastic loading flows (SNL) on line 12/A of the real system

The cost variance plays a fundamental role. In Fig. 7 the average user journey time and the user expected maximum perceived time (disutility) are reported against the cost variance value (parameter θ), using Gamma distribution. These values depend on the variance and are identical in deterministic assignment, and diverge with the increase of the variance (indicator of service irregularity) as expected in theoretical formulation, due to randomness.

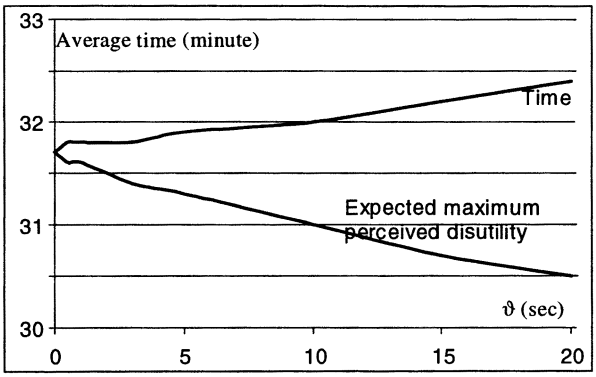

Fig. 7 - Average and expected maximum perceived disutility in the real system against parameter θ with Gamma distribution

In MSA procedure (Fig. 8) the convergence index with 50 AON within SNL, is between 7% and 8%. In this application, fixing the number of AON assignments inside SNL procedure, the internal SNL variance increases with respect the variance value in the test system and it is more stable.

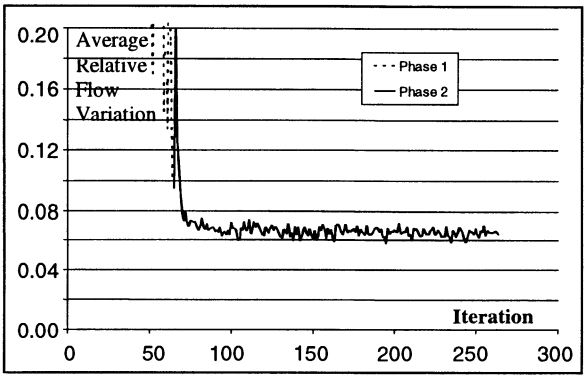

Fig. 8 - MSA algorithm convergence for the real network

5. CONCLUSIONS

In this paper models and algorithms for stochastic assignment to high frequency transit networks are proposed and tested, following a so-called average headway (within-day static) approach. The pre-trip hyperpath choice

behaviour is modelled through random utility models comparing Normal, Log-normal and Gamma distributions. The path choice probabilities are similar using the three different distributions. Gamma and Log-Normal distributions allow to avoid negative perceived costs. It should also be noted that the use of Log-Normal distribution does not guarantee independence from link segmentation, since it is not stable with respect to sum. Elaboration time is also comparable thus the use of Gamma (or Log-Normal) distributions is preferable for practical applications.

Obtained results have to be considered preliminary and still to receive a confirmation from application to other real systems. Some results seem worth of further research work, such as modelling of the headway as a random variable, the comparison with counted flows, the calibration of the route choice model and link cost functions, and the extension to dynamic process models. In addition, a comparison with the average results obtained through a timetable based (within-day dynamic) approach could be fruitful.

Acknowledgments

Authors wish to thank all the referees who rose several relevant issues helpful to improve the paper (one referee also carefully checked out several spelling errors).

REFERENCES

Abramowiz M., Stegun I. (1970), Handbook of Mathematical Functions with Formulas, Graphs and Matematical Tables", Dover Publications Inc. New York.

Ben Akiva M. and Lerman S. R. (1987). *Discrete Choice Analysis*. MIT Press, Cambridge MA.

Bouzaiene-Ayari B., Gendreau M., Nguyen S. (1995), "On the Modelling of Bus Stops in Transit Networks", Centre de recherche sur les transports, Université de Montréal.

Bouzaiene-Ayari B., Gendreau M., Nguyen S. (1997), "Transit Equilibrium assignment Problem: A Fixed-Point Simplicial -Decomposition Solution Algorithm", Operations Research.

Burrell J.E. (1968), "Multiple route assignment and its application to capacity restraint", In Proceedings o the 4t International Symposium on the Theory of Road Traffic Flow, W. Leutzbach and P. Baron eds.. Karlsruhe, Germany

Cantarella G. E., Cascetta E. (1995), "Dynamic Processes and Equilibrium in Transportation Networks: Towards a Unifying Theory", Trans. Science 31, 107-128.

Cantarella, G.E. (1997), "A General Fixed Point Approach to Multi-mode Multi-user Equilibrium Assignment with Elastic Demand enumeration", Trans. Science.

Cantarella G.E., Binetti M (2000). Stochastic Assignment with Gamma Distributed Perceived Costs. Proceedings of the 6th Meeting of the EURO Working Group on Transportation. Gothenburg, Sweden, September 1998, forthcoming.

Caroti Ghelli F. (1978), *Statistica Bayesiana*, ed. Franco Angeli.

Cascetta E. (1998), *Ingegneria dei Sistemi di Trasporto*, UTET.

Daganzo C. F., Sheffi Y. (1977), "On Stochastic Models of Traffic Assignment", Trans. Science.

Daganzo C. F., Sheffi Y. (1983), "Stochastic Network Equilibrium with Multiple Vehicle types and Asimmetric, indefinite Link Cost Jacobians ", Trans. Science 18, 282-300.

Dial R.B. (1971), "A Probabilistic Multipath Traffic Assignment Model with Obiviates Path Enumeration", Trans. Research.

Fisk C. (1980), "Some Developments In Equilibrium Traffic Assignment Methodology", Trans. Research.

Nguyen S., Pallottino (1988), "Equilibrium Traffic Assignment for Large Scale Transit Networks", EJOR.

Nielsen O. A. (1997), "On The Distributions Of The Stochastic Components In SUE Traffic Assignment Models", In Proceedings of 25th European Transport Forum Annual Meeting, Seminar F On Transportation Planning Methods, Volume II.

Patriksson M. (1994), The traffic assignment problem: models and methods", VNU Science Press, UTRECHT The Netherlands.

Sheffi Y. (1985), Urban Transportation Networks, Prentice-Hall, Englewood Cliffs, New York.

Sheffi Y., Powell W. B. (1982), "An Algorithm For The Equilibrium Assignment Problem With Random Link Times", Networks 12.

Spiess H. (1984), "Contribution à la théorie et aux outils de planification des réseaux de transport urban", Département d'Informatique et de Recherche Opérationnelle, Université de Montrèal.

Wardrop J. G. (1952), "Some Theoretical Aspects of Road Traffic Research", Proc. Inst. Civil Engr. Part II.

Wu J. H., Florian M. (1993), "A Simplicial Decomposition Method for the Transit Equilibrium Assignment Problem", Annals of Operations Research.

Wu J. H., Florian M., Marcotte P. (1994), "Transit Equilibrium Assignment: A Model and Solution Algorithms", Trans. Science.

II
GENERAL TRANSPORT MODELS

Chapter 8

WHEN THE MUSIC'S OVER

Final Results of MUSIC, An EU Project To Design And Implement Traffic Signal Timings Which Meet A Variety Of Transport Goals

Richard Clegg
Networks and Nonlinear Dynamics Group
University of York, York UK
richard@manor.york.ac.uk

Arthur Clune
Networks and Nonlinear Dynamics Group
University of York, York UK
arthur@gridlock.york.ac.uk

Mike Smith
Networks and Nonlinear Dynamics Group
University of York, York UK

Abstract This paper describes the results of the EU funded DGVII project MUSIC (**M**anagement of traffic **USI**ng flow **C**ontrol and other measures). The project was designed to demonstrate on-street the success of new signal control policies which account for the rerouting of traffic. The signal control policy in the MUSIC project used delay-based pricing as a design tool to create a signal policy which reduces network travel time considerably. This combination of pricing and signal control has strong theoretical backing. Computer models were used to create traffic signal timings for three European cities, York (UK), Porto (Portugal) and Thessaloniki (Greece). The timings were designed to meet a variety of targets set by the local authorities in the cities. These targets were not only aimed at reducing congestion but also at helping public transport and increasing pedestrian comfort.

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 133–145.

These signal timings were put in place and the results monitored to assess which targets in each city had been met. The results of the project were striking. While, in Porto, the situation was neither improved nor worsened, in York and Thessaloniki the new timings were a considerable success. In York, the main objective was to reduce bus travel time along one of the city's park and ride routes. The bus travel time was reduced by 30% with no net increase in car travel time. Ridership on this bus route increased considerably during this period. In Thessaloniki, the main aim was congestion reduction and this aim was met on almost all routes measured. In both cities, the MUSIC timings (with slight variations) remain in place as of the time of writing (late July '99).

Key words: Traffic, networks, transport, signals, optimisation

1. Introduction

The MUSIC project has demonstrated that a novel approach to traffic signal control, alone or in combination with other measures, may be used to meet a variety of traffic management goals including:
To reduce the travel times of public transport vehicles;
To improve pedestrian facilities/comfort; and
To reduce delays and stops experienced by vehicles and travellers.

1.1 Designing signal timing plans which account for driver rerouting

It is well known that drivers may change their routes in response to changes in congestion as a result of changes to traffic signal timings. The procedure used in the MUSIC project was designed to account for and attempt to take advantage of that rerouting. By making signal plans which attempt to encourage drivers onto more efficient routes it is hoped that the plans will not only be better at reducing congestion but more stable to change.

2. THE MUSIC APPROACH TO DESIGNING TRAFFIC SIGNAL TIMINGS

The MUSIC approach to designing traffic signal timings may be applied at low cost to any city or town that has an existing network model and traffic signals. The approach has the following six stages:
Stage 1: Agree measurable objectives that quantify some relevant aspects of the local transport policy;
Stage 2: Translate data from an existing network model;
Stage 3: Use an off-line optimisation procedure implemented in software to create new time-of-day traffic signal timing plans which aim to meet the measurable objectives while attempting to take some correct account of travellers' future choices;
Stage 4: Test the new timing plans in the existing network model;
Stage 5: Implement the traffic signal timing plans on-street; and
Stage 6: (optional): Conduct "Before" and "After" studies to assess performance against the objectives agreed in Stage 1.

3. THE MUSIC METHOD

The method used to create the basic signal timing plans was a combination of delay-based road pricing and Smith's policy P_0 [Smith (1979)]. The pricing was used as a design tool to reroute traffic in the simulation onto uncongested routes. Following this, P_0 was used to create signal timings that encouraged traffic to stay on these routes. The simulation was then rerun to assess the benefits of these timings. This use of delay-based pricing as a design tool for signal setting is described by Ghali et al (1994). Since these methods are described in detail in the above references, on the MUSIC web site and in Clegg et al (2000) this paper does not go into the details of the method itself.

Figure 1 shows the process for generating signal timing plans by this method. Once a signal timing plan has been created by this method then any extra road capacity generated can be reallocated to meet the targets set by the local authority where they are not met by the original automatically generated signal timing plan.

4. MODELLING APPROACH

This paper is mainly concerned with the actual results of the project, however, something should be said about the computer models used in the

project. The MUSIC project relied mainly on two models: SATURN [Van Vliet (1995)], and STEER [Clegg et al (1995)]. SATURN is a well-known commercial model based in the UK. STEER was written at York University and aims, amongst other things, to be a useful tool in the signal design process. Because creating computer models of cities is expensive, it was decided that existing network models should be used and translated into the necessary formats for the other models. All three cities had existing SATURN models of their road network although these all needed some expansion and extra calibration and validation as part of the project.

The procedure used in MUSIC was to translate the network into the STEER format to design signal timing plans for the city and then to retranslate those timing plans into the SATURN format in order that assessments could be made in the city authorities own trusted model. It was felt that assessment in two models would provide additional confirmation of the stability of results (or highlight potential problems where models disagreed).

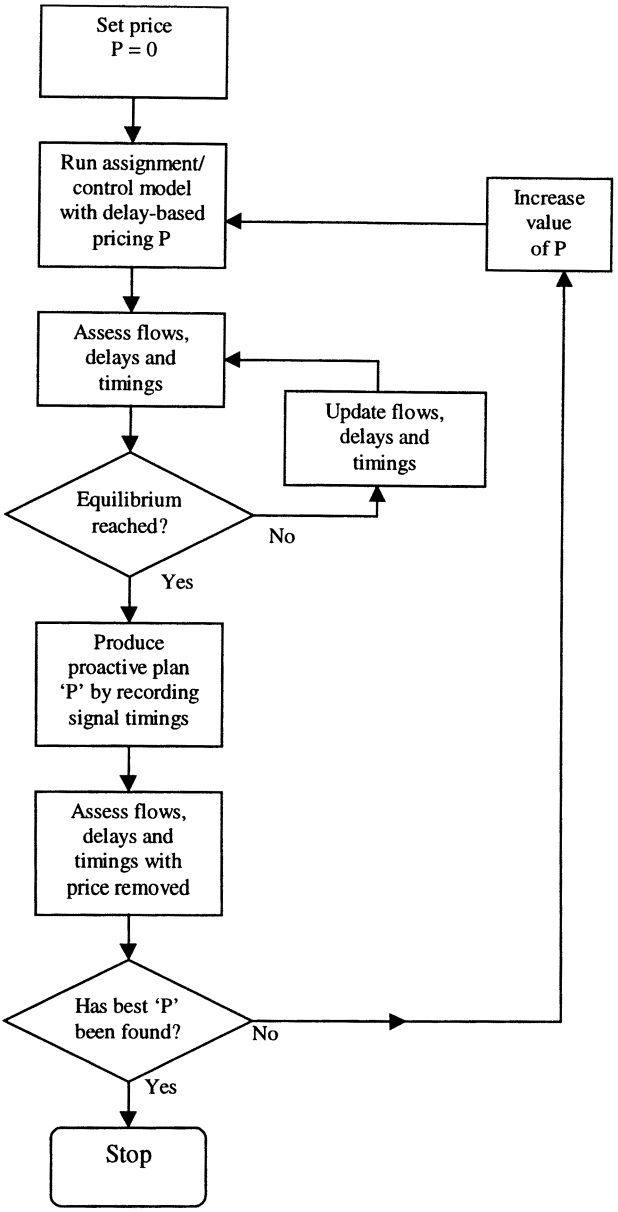

Figure 1: The MUSIC method for designing congestion reducing traffic signal timings

4.1 Notation used in assessment of targets

In this paper, the following notation is used for assessment of targets:

✓ means a target has been met.
~✓ means that a target has been improved but not met (For example, the target was "decrease flow on route by 20%" and the flow only decreased by 18%).
× means that a result has gone the opposite way to the desired target (For example the target was "decrease travel time on route by 10%" and the travel time in fact increased).
×× means that a result has gone the opposite way by a significant amount (more than 15%).
- means that none of the above conditions apply (For example, the target was "Limit the increase in flow to 10%" and the flow has increased by 15% or the target was "Decrease the travel time by 15%" and the travel time has remained unchanged).

5. THE YORK STUDY

The study in York was performed over a smaller area than the studies in the other two cities. MUSIC in York concentrated on changes to one corridor into the city which involved the addition of a bus lane with pre-signals. The operation of the pre-signals is as shown in Figure 2. The signals are bus actuated and allow a bus to bypass a queue of traffic. To work successfully, the signals have to be set to show a considerable amount of red time to general traffic in order to relocate the queue of cars and give advantage to the bus.

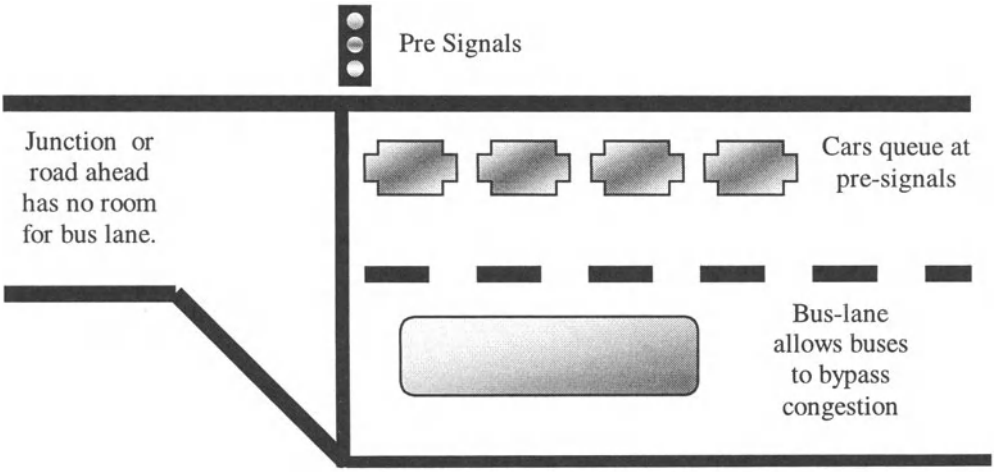

Figure 2: Operation of the pre-signals in York

The objectives set by the local authority for the York study are shown below in Table 1. Hull Road is the west-bound corridor into the city where the changes were made. Melrosegate junction is at the end of this corridor and is extremely congested at peak times. It was hoped that by gating traffic along Hull Road, delays downstream at Melrosegate would also be alleviated. The three targets in M4 (Murton Way, University Road and Tang Hall Lane) were all possible "rat-runs" which could be used if car-travel time on Hull Road were increased as a result of the MUSIC measures. All targets were assessed over the morning peak hour.

Since the number of signals to be changed in the study of York was so small (only two signals), the MUSIC method for creating an initial signal timing plan was felt to be unnecessary. Instead, the approach taken was simply to assess every possible combination of the two signals and to see which best met the objectives for the town. Once the best signal timings in the model had been established, these were implemented in real life.

Measurable	Area	Objective	Target
M1	Hull Road	Bus Travel Time	Reduce by 20%
M2	Melrosegate signals	Delay	Reduce by 50%
M3	Hull Road	Car Travel Time	Limit increase to 10%
M4	Murton Village	Vehicle Flow	Limit increase to 30%
	Tang Hall Lane	Vehicle Flow	Limit increase to 30%
	University Road	Vehicle Flow	Limit increase to 30%

Table1: Objectives for the York Study

		% green time for general traffic at the bus gate					
Area	**Objective**	50%[1a]	**Met?**	50%[1b]	**Met?**	44%	**Met?**
Hull Road	Bus Travel Time	-13%	~✓	-27%	✓	-30%	✓
Hull Road	Car Travel Time	-8%	✓	+2%	✓	-25%[2]	✓
Murton Village	Vehicle Flow	-12%	✓	-14%	✓	-12%	✓
Hull Road (Tang Hall)	Vehicle Flow	-5%	✓	+0%	✓	-9%	✓
University Road	Vehicle Flow	+2%	✓	+5%	✓	+0%	✓
Melrosegate signals	Av. Queue (vehs)	-74%	✓	#		-84%	✓
Melrosegate signals	Max. Queue (vehs)	-71%	✓	#		-80%	✓

Table 2: Results of the York study

#= no data available

The results of the York study was shown in Table 2. To ensure that the benefits were purely a result of the signal timing plans and not the introduction of the bus-lane, the results were compared against the situation where the bus-lane was in place but the signals were set to give generous green time to general traffic. The three columns in the table show how the green time was gradually reduced to the MUSIC settings. The results of the study were somewhat disrupted by the ending of a major road works in the city centre. The column 1a is the last readings before this and the column 1b is the first readings after this event. The figure for car travel time marked 2 is taken from a single day of data and is probably unreliable.

As can be seen, all the targets in York are met by the MUSIC timings. Further, it was discovered that the reliability of park and ride buses had increased considerably as a result. Prior to the MUSIC timings, the standard deviation of morning peak park and ride travel times was 1 minute and 48 seconds. After the MUSIC timings, the standard deviation dropped to only 40 seconds. In this period, the ridership on park and ride buses increased considerably. 25% more passengers used the Hull Road service in the peak hour after the MUSIC timings were increased. While this cannot with certainty be attributed to the MUSIC project, there was no similar increase on the city's other park and ride routes.

It should be noted that in the above figures the target set was delay at the junction. However, it was felt that delay was too difficult to accurately measure at this junction and therefore queue length was used as a proxy for delay.

In conclusion, the York scheme for MUSIC was a considerable success although the scheme was somewhat more limited in scope than in the other two cities. All the targets set by the city council were met. As a result of the MUSIC project, the cities buses had a lower travel time with no (or little) penalty to general traffic. Queuing at a key junction was also reduced. At the time of writing (July '99) the MUSIC timings have been in place for 18 months and still seem to be performing satisfactorily.

6. THE PORTO STUDY

The MUSIC study in Porto took place over a much wider scale. Targets were chosen on fourteen routes in the city and signal timing plans were created for twenty eight junctions. The targets for Porto are shown in Table 3 below.

The "no increase" targets were controls to see whether improvements on one route would be at the expense of another. P1, P2, P3 and P4 were pedestrian areas of the city where the targets were to increase pedestrian comfort by giving them a greater share of the green time on a signal or by reducing the number of vehicles in the area. An initial signal timing plan was created using the MUSIC method previously described. This plan was then adjusted and remodelled to meet the specific targets set by the local authority.

Table 4 (below) shows the final on-street results in Porto. As can be seen, eight targets were met or improved whereas seven were made worse. Overall the Porto study was not considered a great success and the timings were removed two months after their introduction. Several reasons are considered responsible for the problems in Porto. Due to difficulties with the signal controllers, not all of the plans were implemented and the designed offsets to the signals were not put in place. The scheme was not in place for very long and it was felt that traffic was still rerouting as a result of the changes.

Areas	**Objective**	**Target**
MM1,MM2, MM3	Car Travel Time	Decrease by more than 10%
MM1,MM2, MM3	Total Veh. Flow	No more than 10% decrease
MM4,MM5	Car Travel Time	No Increase
P1	Green time to peds	Increase by 5%
P2, P3, P4	Traffic flow	Decrease by more than 10%
RR1,RR2	Bus Travel Time	Decrease by more than 10%
BB1,BB2, BB3	Bus Travel Time	No Increase

Table 3: City objectives for Porto

Area	Objective	Result	Met?
MM1	Car Travel Time	-9.4%	~✓
MM1	Flow	+4.2%	✓
MM2	Car Travel Time	+10.7%	×
MM2	Flow	-7.5%	✓
MM3	Car Travel Time	+16.6%	×
MM3	Flow	-10.4%	✓
MM5	Car Travel Time	-20.0%	✓
P1	Ped. Red. Violations	-30.7%	✓
P2	Flow	+16.3%	××
P3	Flow	+6.2%	×
P4	Flow	-8.0%	~✓
RR2	Bus Travel Time	+22.8%	××
BB1	Bus Travel Time	+5.6%	×
BB2	Bus Travel Time	-4.4%	✓
BB3	Bus Travel Time	+1.6%	×

Table 4: On-street results for Porto

It should be noted that the P1 target to increase the green time for pedestrians was automatically met by changing the signal timings. A new target of reducing pedestrian red-light violations was set to measure the effectiveness of this change.

7. THE THESSALONIKI STUDY

Table 5 below shows the targets for the Thessaloniki and the on-street results achieved. The routes shown are all major routes through the city and many of them are important to public transport. The four delay targets are for key junctions in the city. Thessaloniki was the largest study in MUSIC with 129 signal timing plans introduced in June '98. A variant of the MUSIC timings is in place at the time of writing (July '99)

Area	Objective	Target	Result	Met?
B.Olgas	Travel Time	Reduce by 10%	+9.6%	×
"	Flow	Less than 5% inc.	-2.7%	✓
N.Egnatia	Travel Time	Reduce by 10%	-44.4%	✓
Delfon	Travel Time	Reduce by 10%	-45.2%	✓
"	Flow	Less than 10% inc.	+8.0%	✓
Ag. Demet.	Travel Time	Reduce by 5%	-11.4%	✓
"	Flow	Less than 10% inc.	+1.5%	✓
Lagada	Travel Time	Reduce by 10%	-29.3%	✓
Monastiriou	Travel Time	Reduce by 10%	+3.7%	×
Egnatia	Travel Time	Reduce by 15%	+23.8%	××
Egnatia 2	Travel Time	Reduce by 10%	+49.3%	××
Tsimiski	Travel Time	Reduce by 10%	+24.1%	××
Kountour.	Travel Time	Reduce by 10%	-12.2%	✓
Nikis Av.	Travel Time	Reduce by 10%	-17.5%	✓
Sintri Inter.	Delay	Reduce up to 2 mins	-2.65 min	✓
Sintri 2	Delay	Reduce up to 2 mins	-0.60 min	~✓
YMCA Inter.	Delay	Reduce up to 3 mins	-2.52 min	~✓
YMCA 2	Delay	Reduce up to 2 mins	-2.36 min	✓

Table 5: Thessaloniki objectives and implementation results

The Thessaloniki results were extremely impressive with eleven targets met and two considerably improved set against five made worse. Of these five, it should also be noted that the Egnatia targets (two of the worst results) were on streets where loading and unloading made travel times fluctuate considerable and where, for practical reasons, no changes to timing plans had been made. Overall, it was felt by MUSIC project personnel in Thessaloniki that congestion in the city as a whole had decreased considerably as a result of the project.

The measurements shown in the table were taken one month after the original implementation of the timings. (The MUSIC timings were temporarily removed to test signal timing plans from a new Siemens UTC system. A variant on the MUSIC timings was later implemented in Thessaloniki.)

8. CONCLUSIONS

The MUSIC project proves that considerable benefits can be realised using computer modelled signal timing plans. In two out of three test sites the MUSIC results proved extremely successful. Two things make the MUSIC approach different from more traditional approaches to traffic signal design: Firstly, the MUSIC signal timings are designed to account for and take advantage of driver rerouting; Secondly the signal changes are aimed at wider goals than merely reducing congestion to cars.

The approach used in MUSIC is extremely cost-effective and transferable to any city that has both traffic signals and an existing computer model of the road network. For more information about the project visit the web site on or contact: M.J. Smith, Department of Mathematics, University of York, York, England,YO10 5DD.

REFERENCES

Clegg R.G., Clune A.J. (2000): The MUSIC Project: Urban Traffic Control for Traffic Demand Management to appear in Transportation Research Record.

Clegg R.G., Ghali M.O., Smith M.J. (1995): *Equilibrium/Control results and the approach to a near-equilibrium of a new Dynamic Micro-simulation/Assignment model on a Network Model of York* in Applications of Advanced Technologies in Transportation Engineering, the American Society of Civil Engineers, 568-572.

Ghali, M.O. Smith M.J. (1994): *Comparisons of the performances of three responsive traffic control policies taking drivers' day-to-day route choices into account* Traffic Engineering and Control, 35, 555-560

MUSIC Project web site: http://gridlock.york.ac.uk/music/

Smith M.J. (1979): *A Local Traffic Control Policy which Automatically Maximises the Overall Travel Capacity of an Urban Road Network* Proceedings of the International Symposium on Traffic Control Systems also published in Traffic Engineering and Control (1980), 21, 298-302

Van Vliet, D. (1995): *SATURN 9.2 User Manual*, W.S. Atkins.

Chapter 9

Algorithms for the Solution of the Combined Traffic Signal Optimisation and Equilibrium Assignment Problem

Mike Maher and Xiaoyan Zhang
School of the Built Environment, Napier University, UK

Abstract The combined traffic signal optimisation and equilibrium assignment problem is one in which a traffic engineer tries to optimise the performance of traffic signals while road users choose their routes so as to minimises their travel costs. Two types of solutions can be defined in the combined problem: the *mutually consistent solution* and the *global optimal solution*. The former is a solution at which the two sub-problems are solved simultaneously while the latter is a solution to the bi-level programming formulation of the combined problem. In this paper, we consider the combined signal optimisation and stochastic user equilibrium assignment problem. We present two types of algorithms for the mutually consistent solution and one type of algorithm for the bi-level solution to the problem. The algorithms are tested on a small network to examine their convergence and efficiency.

Key words: Traffic signal optimisation, Stochastic assignment, Bi-level programming problem

1. INTRODUCTION

The combined traffic signal optimisation and equilibrium assignment problem is one in which a traffic engineer tries to optimise the performance

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 147–160.

of traffic signals in a road network while road users choose their routes so as to minimise their travel costs. The input to a signal optimisation (SO) problem consists of link flows in the road network, which comprise the output of a traffic assignment model. A traffic assignment model, on the other hand, requires signal settings as inputs. An equilibrium assignment (EA) model is needed so as to achieve consistency in route choices and to model congestion effects in the network. This can either be a user equilibrium (UE) assignment model or a stochastic user equilibrium (SUE) assignment model. A SUE assignment model is preferable because it accounts for both congestion effects and drivers' differences in route choice.

In the combined SO and EA problem, there is a mutual interaction between the two sub-problems. Each of the two parties (the traffic engineer and the road users) is continuously resolving its own sub-problem given the latest information on the actions of the other party. The process may lead to a *mutually consistent solution*, at which traffic signal settings and link flows are mutually consistent. In this process, traffic signals are optimised for *fixed* link flows. If the traffic engineer knows the road users' route choice behaviour (being such that a SUE is approached), he may optimise signals for an *equilibrium* link flow pattern that is dependent on the signal settings. This results in a *bi-level optimisation problem*, where signal optimisation is the upper-level problem and equilibrium assignment the lower-level problem. We shall consider both the mutually consistent solution and the solution to the bi-level problem, or the *bi-level solution* in this paper.

An iterative algorithm in which the SO and UE problems are solved alternately has been used for the solution of the combined SO and UE problem (Van Vuren and Van Vliet, 1992; Smith and Van Vuren, 1993). This procedure may converge to the mutually consistent solution but convergence is not guaranteed (Fisk, 1984, 1988). Several types of algorithm have been proposed for the solution of the bi-level signal optimisation problem with UE assignment (Sheffi and Powell, 1983; Heydecker and Khoo, 1990; Yang and Yagar, 1995). See Maher and Zhang (1999) for a review for these algorithms. However, these algorithms require repeated UE assignment for direction finding and/or for line search. Using the method of successive averages (MSA) instead of a line search can avoid repeated UE assignment (Sheffi, 1985), but will slow down the convergence of the algorithms. Cascetta *et al.* (1998) considered a combined signal optimisation and SUE assignment. Their algorithm also involves MSA and its variant for step length determination.

Recently, the authors have developed two efficient algorithms for the solution of the combined SO and UE assignment problem (Zhang & Maher, 1998; Maher & Zhang, 1999): one algorithm for the mutually consistent solution and the other for the bi-level solution. In an example in which the

true solutions can be found by a direct search, it was shown that the algorithms converge to the correct solutions. In this paper, we apply the two solution methods to the combined SO and SUE problem. In addition, we propose another algorithm for the mutually consistent solution. The algorithms will be tested on a small network and two types of solutions will be compared. The formulation of the problem is given in section 2, which is followed by the description and test of the algorithms in sections 3 and 4. The description of the algorithms will be brief; more details can be found in Zhang and Maher (1998) and Maher and Zhang (1999). Conclusions are drawn in section 5.

2. THE PROBLEM FORMULATION AND SOLUTIONS

The most commonly used policy for signal optimisation is to minimise the total journey costs in the network:

$$\begin{aligned} &\underset{\mathbf{s}}{\text{minimise}}\ Z_{SO}(\mathbf{s},\mathbf{v}) = \sum_{a\in A} v_a\, c_a(v_a, s_a) \\ &\text{subject to}\ \ s_a^{\max} \geq s_a \geq s_a^{\min}, \quad a \in A \\ &\qquad \sum_{a\in A_j} s_a = 1, \quad A_j \in A \end{aligned} \tag{1}$$

where $\mathbf{s}=(\ldots, s_a, \ldots)$ is the vector containing green splits; $\mathbf{v}=(\ldots, v_a, \ldots)$ is the vector containing link flows; $c_a(v_a)$ is the cost-flow function for link a; $s_a^{\max}$ and $s_a^{\min}$ are respectively the maximum and minimum allowable green splits for link a, $s_a^{\min}>0$, $s_a^{\max}<1$; and A_j is the set of links heading for the jth signal controlled intersection. If link a is not controlled by a signal, then $s_a^{\max}$, s_a, and $s_a^{\min}$ will all be equal to 1. In this problem, the green splits are the decision variables while link flows are the output from a SUE assignment problem. Given signal settings, $\mathbf{s}$, the SUE assignment problem may be written as (Sheffi, 1985)

$$\begin{aligned} \underset{\mathbf{v}}{\text{minimise}}\ Z_{SUE}(\mathbf{v},\mathbf{s}) = &-\sum_i t_i C_i(\mathbf{v}) + \sum_{a\in A} v_a c_a(v_a, s_a) \\ &- \sum_{a\in A} \int_0^{v_a} c_a(x, s_a)\,\mathrm{d}x \end{aligned} \tag{2}$$

where t_i is the traffic demand between O-D pair i and C_i is the value of the satisfaction function or the expected perceived minimum travel cost between O-D pair i. The expected minimum perceived travel cost $C_i(\mathbf{v})$ is obtained from a stochastic loading based on link flow $\mathbf{v}$. We have included $\mathbf{s}$ in the problem formulation although it is fixed in the SUE problem. We will use $\mathbf{S}(\mathbf{v})$ to denote the optimal solution of problem (1) given $\mathbf{v}$, and $\mathbf{V}(\mathbf{s})$ the optimal solution of problem (2) given $\mathbf{s}$.

A mutually consistent solution, $[\mathbf{s}^{MC}, \mathbf{v}^{MC}]$, of the combined problem can be defined as

$$\mathbf{s}^{MC} = \text{Arg} \min_{\mathbf{s}} Z_{SO}(\mathbf{s}, \mathbf{v}^{MC}) = \mathbf{S}(\mathbf{v}^{MC}) \tag{3a}$$

$$\mathbf{v}^{MC} = \text{Arg} \min_{\mathbf{v}} Z_{SUE}(\mathbf{v}, \mathbf{s}^{MC}) = \mathbf{V}(\mathbf{s}^{MC}) \tag{3b}$$

In other words, $\mathbf{s}^{MC}$ solves the SO problem given $\mathbf{v}^{MC}$, and $\mathbf{v}^{MC}$ solves the SUE assignment problem given $\mathbf{s}^{MC}$. A bi-level solution, on the other hand, is $[\mathbf{s}^{BL}, \mathbf{V}(\mathbf{s}^{BL})]$, where

$$\mathbf{s}^{BL} = \text{Arg} \min_{\mathbf{s}} Z_{SO}(\mathbf{s}, \mathbf{V}(\mathbf{s})) \tag{4}$$

Note the difference between (3a) and (4): the former is solved with *fixed* $\mathbf{v}$ while the latter with *variable* $\mathbf{v}$. It is clear that the mutually consistent solution is also a feasible solution of the bi-level problem. The two types of solutions are generally different and the bi-level solution has a smaller value of the SO objective function than that of the mutually consistent solution. Therefore, the system would perform better at the bi-level solution. By definition, among all the solutions that satisfy SUE conditions, the bi-level solution has the minimum SO objective function value.

3. THE SOLUTION ALGORITHMS

The traffic signal optimisation problem is a special case of the more general network design problem, in which the number of phases, the cycle time, and the offsets of traffic signals are determined. In this paper, we consider signal optimisation for isolated intersections. Thus, given a set of link flows, the SO problem is reduced to several sub-problems of determining the optimal green split for each signal controlled intersection. Each of them may be solved by a standard one-dimensional optimisation

algorithm, such as the Newton method. For the SUE sub-problem, we use the logit-based SUE assignment method developed by Maher (1998). Details of the solution algorithm can be found in the reference cited.

The algorithms for the combined SO and SUE problem are iterative processes. At each iteration, a new solution is calculated based on the current solution. Auxiliary solutions are calculated by solving the SO and SUE sub-problems to provide a search direction. Then an optimal step length is calculated to determine how far to move from the current solution.

3.1 The Mutually Consistent Solution Algorithm: the Dual-Step Algorithm

Suppose at iteration n we have a current solution, $[\mathbf{s}^{(n)}, \mathbf{v}^{(n)}]$. The SO problem is firstly solved to get an auxiliary solution of traffic signals, $\mathbf{s}^*$, using $\mathbf{v}^{(n)}$. Then, the SUE assignment problem is solved to get the auxiliary solution of SUE link flows $\mathbf{v}^*$ for $\mathbf{s}^*$. We then search for *a pair of* optimal step lengths for the two sets of variables respectively in the hyper-plane defined by the three points $[\mathbf{s}^{(n)}, \mathbf{v}^{(n)}]$, $[\mathbf{s}^*, \mathbf{v}^{(n)}]$, and $[\mathbf{s}^*, \mathbf{v}^*]$. Let

$$\mathbf{s}(\alpha)=\mathbf{s}^{(n)}+\alpha(\mathbf{s}^*-\mathbf{s}^{(n)}) \tag{5a}$$

$$\mathbf{v}(\beta)=\mathbf{v}^{(n)}+\beta(\mathbf{v}^*-\mathbf{v}^{(n)}) \tag{5b}$$

Denote the derivatives of the two objective functions along the two directions by respectively $g(\alpha,\beta)$ and $h(\alpha,\beta)$:

$$g(\alpha,\beta)=\frac{dZ_{SO}(\mathbf{s}(\alpha),\mathbf{v}(\beta))}{d\alpha}=\sum_a v_a(\beta)\frac{\partial c_a}{\partial s_a}(s_a^*-s_a^{(n)})$$

$$h(\alpha,\beta)=\frac{dZ_{SUE}(\mathbf{s}(\alpha),\mathbf{v}(\beta))}{d\beta}=\sum_a (v_a(\beta)-u_a(\beta))\frac{\partial c_a}{\partial v_a}(v_a^*-v_a^{(n)})$$

where $[u_a(\beta)]$ are the link flows arising from a stochastic loading based on link flows $[v_a(\beta)]$. Assuming that the two objective functions are quadratic so that the derivatives are linear in the vicinity of the current solution, a pair of optimal step lengths which minimise simultaneously $Z_{SO}(\mathbf{s}(\alpha),\mathbf{v}(\beta))$ and $Z_{SUE}(\mathbf{s}(\alpha),\mathbf{v}(\beta))$ can be found by solving

$$g_{C0} + \alpha(g_{10}-g_{00}) + \beta(g_{11}-g_{10}) = 0$$

$$h_{00} + \alpha(h_{10}-h_{00}) + \beta(h_{11}-h_{10}) = 0$$

where the subscripts refer respectively to values of α and β. Once an optimal pair of step lengths is found, a new solution is given by (5a) and (5b). The stopping criterion can be based on the maximum relative change in the signal splits and link flows at successive iterations:

$$\mathrm{Max}_i \left(|s_a^{(n+1)} - s_a^{(n)}| / s_a^{(n)} , |v_a^{(n+1)} - v_a^{(n)}| / v_a^{(n)} \right) \le \varepsilon$$

where ε is the error tolerance. This stopping criterion will be used in other algorithms described below. In this algorithm, a pair of optimal step lengths is found at each iteration. Therefore, the algorithm is called the *dual-step algorithm.*

An alternative way to determine the search directions is to start with SUE assignment rather than solving the SO problem. Sometimes the order may affect the performance of the algorithm. It could happen that the signal settings approach the optimal sooner than the link flows. In this situation, the current and the auxiliary signal split solutions, $\mathbf{s}^{(n)}$ and $\mathbf{s}^*$, are very close and the derivatives g_{00}, g_{10}, and g_{11} are close to zero. On the other hand, the current and the auxiliary SUE link flows $\mathbf{v}^{(n)}$ and $\mathbf{v}^*$ are much further apart and the derivative $h(\alpha)$ is much larger. Consequently, the step length β will soon approach zero and further iterations will cease to update the solutions. Therefore, in the implementation of the algorithm, each iteration starts with solving the SO problem, which is followed by SUE assignment. Whenever $\mathbf{s}^{(n)}$ and $\mathbf{s}^*$ are very close while $\mathbf{v}^{(n)}$ and $\mathbf{v}^*$ are not, we use $\mathbf{s}^*$ as the new current solution and, after the SUE assignment, do another signal optimisation to get another auxiliary solution. The new auxiliary link flow is far from the new current link flow and can provide a much better search direction.

3.2 The Mutually Consistent Solution Algorithm: the Joint-Step Algorithms

In this section we propose another algorithm for the mutually consistent solution, in which a joint-step is calculated for both signal setting parameters and link flows. The basic idea is to find two solutions that satisfy the SUE conditions: $[\mathbf{s}^{(n)}, \mathbf{V}(\mathbf{s}^{(n)})]$ and $[\mathbf{s}^*, \mathbf{V}(\mathbf{s}^*)]$, and we assume that the SUE conditions are satisfied at the intermediate points on a line between the two points. Then we search for an optimal step length along the line with respect to the SO objective function $Z_{SO}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))$ only, where

$$\mathbf{s}(\alpha)=\mathbf{s}^{(n)}+\alpha(\mathbf{s}^*-\mathbf{s}^{(n)}) \qquad (6a)$$

$$\mathbf{V}(\mathbf{s}(\alpha))=\mathbf{V}(\mathbf{s}^{(n)}+\alpha(\mathbf{s}^{*}-\mathbf{s}^{(n)})) \tag{6b}$$

Denote the derivative of the SO objective function with respect to α by $f(\alpha)$. Then

$$f(\alpha)=\frac{\mathrm{d}Z_{\mathrm{SO}}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))}{\mathrm{d}\alpha}=\sum_{a}V_a(\mathbf{s}(\alpha))\frac{\partial c_a}{\partial s_a}(s_a^{*}-s_a^{(n)})$$

Note that only the partial derivatives with respect to variable **s** are included. This is because the SO problem is unconstrained by the SUE conditions in the mutually consistent solution. See equation (3a). Evaluation of $f(\alpha)$ at any intermediate point requires an extra SUE assignment. To avoid this, we use quadratic interpolation on $Z_{\mathrm{SO}}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))$ or linear interpolation on $f(\alpha)$. This needs only the SUE link flows and derivatives at the current and the auxiliary point with α being 0 and 1 respectively. The SUE link flows at the two points are simply the current and the auxiliary link flows and so no extra SUE assignment is needed. Then, the optimal step size is given by $-f(0)/(f(1)-f(0))$. Once an optimal step size is obtained, the new solution is given by (6a) and (6b), with the latter involving a further SUE assignment.

A variant of this algorithm is to reverse the role of the solution of the two sub-problems: find two points which are optimal with respect to the SO problem, $[\mathbf{S}(\mathbf{v}^{(n)}), \mathbf{v}^{(n)}]$ and $[\mathbf{S}(\mathbf{v}^{*}), \mathbf{v}^{*}]$, and optimise the step length with respect to the SUE objective function, $Z_{\mathrm{SUE}}(\mathbf{v}(\beta),\mathbf{S}(\mathbf{v}(\beta)))$, where

$$\mathbf{v}(\beta)=\mathbf{v}^{(n)}+\beta(\mathbf{v}^{*}-\mathbf{v}^{(n)})$$

$$\mathbf{S}(\mathbf{v}(\beta))=\mathbf{S}(\mathbf{v}^{(n)}+\beta(\mathbf{v}^{*}-\mathbf{v}^{(n)}))$$

The directional derivative of the SUE objective function with respect to β is

$$\frac{\mathrm{d}Z_{\mathrm{SUE}}(\mathbf{v}(\beta),\mathbf{S}(\mathbf{v}(\beta)))}{\mathrm{d}\beta}=\sum_{a}(v_a(\beta)-u_a(\beta))\frac{\partial c_a}{\partial v_a}(v_a^{*}-v_a^{(n)})$$

Evaluation of this derivative requires a solution of the SO problem. Normally, a solution of SO problem may not be as computationally demanding as SUE assignment. Therefore, we may use either a linear interpolation on the derivative, as we did in the above joint-step algorithm, or a more elaborate algorithm, such as the bi-section method, for the line search.

We shall call the first joint-step algorithm the *SO-step algorithm* and the second the *SUE-step algorithm.*

3.3 The Bi-Level Solution Algorithm

This algorithm is similar to the SO-step algorithm in that we firstly find two points that satisfy the SUE constraints: $[\mathbf{s}^{(n)}, \mathbf{V}(\mathbf{s}^{(n)})]$ and $[\mathbf{s}^*, \mathbf{V}(\mathbf{s}^*)]$. We then determine an optimal step size on the line between the two points by minimising $Z_{SO}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))$, where

$$\mathbf{s}(\alpha)=\mathbf{s}^{(n)}+\alpha(\mathbf{s}^*-\mathbf{s}^{(n)}) \tag{7a}$$

$$\mathbf{V}(\mathbf{s}(\alpha))=\mathbf{V}(\mathbf{s}^{(n)}+\alpha(\mathbf{s}^*-\mathbf{s}^{(n)})) \tag{7b}$$

However, in the bi-level formulation, the SO problem is constrained by the SUE conditions. Therefore, the directional derivative of the objective function is

$$\frac{dZ_{SO}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))}{d\alpha}=\sum_a\left[V_a(\mathbf{s})\frac{\partial c_a}{\partial s_a}(s_a{}^*-s_a{}^{(n)})+\left(c_a+V_a(\mathbf{s})\frac{\partial c_a}{\partial V_a}\right)\sum_b\frac{\partial V_a(\mathbf{s})}{\partial s_b}(s_b{}^*-s_b{}^{(n)})\right]$$

Compare this derivative with that in the SO-step algorithm. Here, the partial derivatives include those with respect to $\mathbf{s}$ as well as those with respect to $\mathbf{V}(\mathbf{s})$. The latter is differentiated further with respect to $\mathbf{s}$ by the chain rule. It can be seen that directly minimising $Z_{SO}(\mathbf{s}(\alpha),\mathbf{V}(\mathbf{s}(\alpha)))$ requires repeated SUE assignment and the derivatives of SUE link flows with respect to $\mathbf{s}$. The former is very inefficient while the latter needs special mathematical methods, such as the sensitivity analysis methods by Tobin and Friesz (1988). To overcome the difficulty, we assume that the SUE assignment map $\mathbf{V}(\mathbf{s})$ is approximately linear between the two points. Then we have

$$\mathbf{v}(\alpha)=\mathbf{v}^{(n)}+\alpha(\mathbf{v}^*-\mathbf{v}^{(n)}) \tag{7c}$$

and we can now minimise $Z_{SO}(\mathbf{s}(\alpha),\mathbf{v}(\alpha))$ instead. This is a one-dimensional optimisation problem and can be solved by the bisection method. The derivative of this function with respect to α is now

$$\frac{dZ_{SO}(\mathbf{s}(\alpha),\mathbf{v}(\alpha))}{d\alpha} = \sum_a \left(v_a(\alpha) \frac{\partial c_a}{\partial s_a} (s_a^* - s_a^{(n)}) \right) + \sum_a \left[\left(c_a + v_a(\alpha) \frac{\partial c_a}{\partial v_a} \right) (v_a^* - v_a^{(n)}) \right]$$

The new solution is then given by (7a) and (7b), again, with the latter involving a further SUE assignment.

This algorithm involves an approximation in the optimal step length calculation: the SUE assignment map is linearised over the interval between the current and the auxiliary solution. The interval is generally finite because the auxiliary solution does not in general become closer to the current solution with increasing iteration number. Thus there is no reason to expect that the linearisation will become more and more accurate as the algorithm converges. As a result, the algorithm may converge to some neighbourhood point of the true solution due to the approximation. This problem may be dealt with by reducing the interval between the current and the auxiliary solution of signal settings by, for example, a MSA-type scheme so that the linearisation is made over a smaller and smaller interval. Our experiences have shown that the first few iterations of the bi-level algorithm are very efficient. Therefore, if higher accuracy is desirable, we can introduce the modification after the first few iterations when the solution is close to optimal or when the SO objective function is not reduced at further iterations.

4. NUMERICAL TEST

In this test, the cost function used is a combination of the BPR function (for link travel time) and the signal delay formula by Doherty (1977), that is

$$c_a(v_a) = c_a(0)\left[1 + \mu\left(\frac{v_a}{q_a}\right)^{\gamma}\right] + d_a$$

where $[c_a(0)]$ is uncongested link costs, $[q_a]$ is link capacity, and μ and γ are constants. The signal delay d_a is given by

$$d_a = \frac{T}{2}(1-s_a)^2 + \frac{1980}{q_a s_a} \frac{v_a}{q_a s_a - v_a}, \quad v_a/(q_a s_a) \leq 0.95$$

$$d_a = \frac{T}{2}(1-s_a)^2 - \frac{198.55\times 3600}{q_a s_a} + \frac{220\times 3600 v_a}{(q_a s_a)^2}, \quad v_a/(q_a s_a)>0.95$$

where T is the cycle time. In the following test, the commonly used values μ=1.0, and γ=4.0 in the BPR function will be used. The value of the spread parameter in the logit assignment model is 0.5 and the cycle time T is 90 sec.

The 3×3 grid network shown in Figure 1 is used to test the performance of the algorithms and to compare the two types of solutions. The network has 9 nodes and 24 links. There are 4 centroids (nodes 1, 3, 5, and 7) and 4 O-D pairs (1→5, 3→7, 5→1, 7→3). Node 9 in the middle is a signalised intersection. The uncongested link costs are 20 on all east-west links and 10 on all north-south links; the link capacities are 1000 on all links except on those inner north-south links (links 2→9, 9→2, 6→9, and 9→6), where the capacities are 500. The demand is 880 for all O-D pairs.

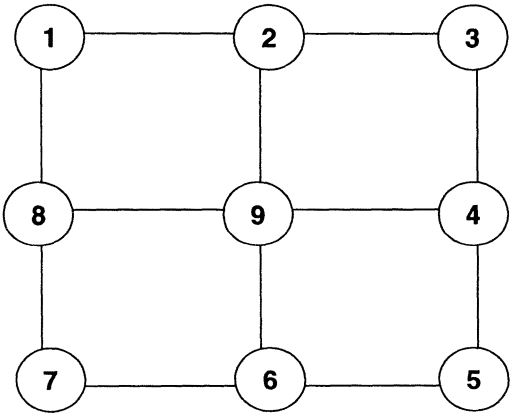

Figure 1. The grid network. All links are two-directional.

Tables 1 and 2 shows the results from different algorithms, including the true solutions of the two types, the number of iterations for the algorithms to converge at the given error tolerance, the solutions when they converge and the solutions at the 20th iteration. The green split solution s_{NS} used in the tables is for the north-south links. The SO objective function is rather flat in the neighbourhood of the solutions. Therefore, we have shown 9 significant digits of the objective function values to demonstrate the accuracy of the proposed algorithms. Three different initial green splits have been tried: 0.3, 0.5, and 0.7 and it was found that the solution from each of the proposed

algorithm at the 20th iteration are the same. The true mutually consistent solution is found by the alternate algorithm mentioned in the introduction and the true bi-level solution by direct search on the green split with increment size of 0.001. The alternate algorithm does not always converge while the direct search algorithm may not be used in a general network involving more than a small number of signal setting parameters. The algorithms are used here merely for the purpose of checking the accuracy of the solution found by the proposed new algorithms. The direct search process has shown that there is only one optimal bi-level solution in this problem.

Table 1. Performance of the algorithms for the mutually consistent solution on the 3×3 grid network with ε=0.001.

Algorithm	$s_{NS}^{(0)}$=0.3		$s_{NS}^{(0)}$=0.5		$s_{NS}^{(0)}$=0.7		Final solution	
	N	$s_{NS}^{(N)}$	N	$s_{NS}^{(N)}$	N	$s_{NS}^{(N)}$	$s_{NS}^{(20)}$	$Z_{SO}^{(20)}$
Alternate	15	0.484	7	0.485	14	0.485	0.484	256300.457
Dual-step	7	0.484	2	0.484	8	0.484	0.484	256300.404
SO-step	3	0.484	2	0.484	3	0.484	0.484	256300.402
SUE-step	6	0.484	2	0.484	5	0.484	0.484	256300.405

Notes: True mutually consistent solution: $s_{NS}^{(50)}$=0.484; $Z_{SO}^{(50)}$=256300.402

Table 2. Performance of the algorithms for the bi-level solution on the 3×3 grid network with ε=0.001.

Algorithm	$s_{NS}^{(0)}$=0.3		$s_{NS}^{(0)}$=0.5		$s_{NS}^{(0)}$=0.7		Final solution	
	N	$s_{NS}^{(N)}$	N	$s_{NS}^{(N)}$	N	$s_{NS}^{(N)}$	$s_{NS}^{(20)}$	$Z_{SO}^{(20)}$
Bi-level	6	0.375	6	0.376	8	0.376	0.376	255318.093
Modified Bi-level	6	0.375	6	0.376	11	0.371	0.370	255315.764

Notes: True bi-level solution: s_{NS}=0.370; Z_{SO}=255315.769

Several points can be seen from Tables 1 and 2. First, all the proposed algorithms converge in relatively few iterations at the given error tolerance and the solution at the 20th iteration is very close or equal to the true solutions. Second, the three proposed algorithms for the mutually consistent solutions perform slightly better than the alternate algorithm in terms of efficiency and accuracy, but the latter does not always converge, as has been mentioned. Among the three algorithms, the SO-step algorithm is more efficient than the other two algorithms. Third, the bi-level algorithm converges to a solution in the close neighbourhood of the true solution and the modified bi-level algorithm converges to the true solution. The modification was introduced after 5 iterations. Finally, and most importantly, the value of the SO objective function (i.e., the total network journey cost) is 0.3% lower at the bi-level solution than that at the mutually consistent solution (calculated as the difference of SO objective function values divided by Z_{SO}^{BL} and multiplied by 100). This reduction in total delay brought about

by the bi-level programming approach is rather limited, yet it is about the highest one could get in this particular example. The amount of reduction depends on how congested the network is.

Table 3. Congestion effects on the 3×3 grid network.

	Mutually consistent solution			Bi-level solution		
Degree of congestion	s_{NS}	Z_{SO}	$v_{(2\to9)}$	s_{NS}	Z_{SO}	$v_{(2\to9)}$
0.89	0.457	402754.0	269.520	0.411	402267.6	239.898
1.14	0.456	781766.6	346.745	0.473	781541.2	360.733

The demand is varied to test the effect of congestion. It was found that the bi-level approach is more effective only when the degree of congestion (average rate of link flows over capacity of all links) is 0.6-0.9. This outcome is similar to that found by Cascetta *et al.* (1998). Outside this range the difference of values of the SO objective function between the two types of approaches is very small. However, the signal settings are quite different (it is possible for different signal settings to result in similar SO objective function values). When the degree of congestion is less than 1.0, s_{NS}^{BL} is smaller than s_{NS}^{MC}, and when the degree of congestion is larger than 1.0, s_{NS}^{BL} is larger than s_{NS}^{MC}. See Table 3 for one such example. In this network, the links in the east-west direction are twice as long as those in the north-south direction; the link capacities on the inner north-south links are half as much as those on all other links. When the network is not saturated, more drivers would choose the shorter routes and pass the signal through the north-south links. By giving a smaller green split in this direction in the bi-level mechanism, more traffic is diverted to the longer links/routes and the overall delay is reduced. When the network is saturated or over saturated, on the other hand, a larger green split is given to the north-south links to attract more traffic to shorter links, although the reduction in total delay by the bi-level approach is not as much as that when traffic is not saturated.

5. CONCLUSION

The problem of combined traffic signal optimisation and SUE assignment has been addressed in this paper. Three types of algorithms have been presented: the dual-step and the joint-step algorithm for mutually consistent solutions and the bi-level algorithm and its modified version for bi-level solutions. The algorithms were tested on a 3×3 grid network including one traffic signal. It has been shown that the algorithms converge to the correct solutions. Further tests with more general networks will be carried out. The main feature of the algorithms presented here is that they use optimal step

lengths; the step lengths can be calculated efficiently without repeated SUE assignment. Therefore, the algorithms can be readily implemented by incorporating the step-finding routines into existing programs for the alternate algorithm. It is also possible to extend the method to deal with more general signal design problems, such as the optimisation of green splits as well as cycle length and offsets.

ACKNOWLEDGEMENT

The funding from the UK Engineering and Physical Science Research Council for the work reported in this paper is acknowledged. We thank Dirk Van Vliet of the Institute for Transport Studies of the University of Leeds for helpful discussions.

REFERENCES

Cascetta, E., M. Gallo and B. Montella (1998). Models and algorithms for the optimisation of signal settings on urban networks with stochastic assignment. *Sixth Meeting of the EUROWG on Transportation*, 9-11 September, 1998, Gothenburg.

Doherty A. R. (1977). A comprehensive junction delay formula. *LTRI Working Paper, Department of Transport.*

Fisk, C. S. (1984). Game theory and transportation systems modelling. *Transportation Research, 18B*, 301-313.

Fisk, C. S. (1988). On combining maximum entropy trip matrix estimation with user optimal assignment. *Transportation Research, 22B*, 69-79.

Heydecker, B. G. and T. K. Khoo (1990). The equilibrium network design problem. In: *Proceedings of AIRO'90 Conference on Models and Methods for Decision Support*, Sorrento, pp 587-602.

Maher, M. (1998). Algorithms for logit-based stochastic user equilibrium assignment. *Transportation Research, 32B*, 539-549.

Maher, M. J. & X. Zhang (1999). Algorithms for the solution of the congested trip matrix estimation problem. *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*, 20-23 July, 1999, Jerusalem, Israel.

Sheffi, Y. (1985). *Urban Transportation networks: Equilibrium Analysis with Mathematical Programming Methods.* Prentice-Hall, Englewood Cliffs, New Jersey.

Sheffi, Y. and W. B. Powell (1983). Optimal signal setting over transportation networks. *Transportation Engineering, 109* (6), 824-839.

Smith M. J. and T. Van Vuren (1993). Traffic equilibrium with responsive traffic control. *Transportation Science, 27*, 118-132.

Tobin R. L. and T. L. Friesz (1988). Sensitivity analysis for equilibrium network flows. *Transportation Science, 22*, 242-250.

Van Vuren T. and D. Van Vliet (1992) *Route Choice and Signal Control.* Athenaeum Press Ltd., Newcastle upon Tyne.

Yang H. and S. Yagar (1995). Traffic assignment and signal control in saturated road networks. *Transportation Research, 29A,* (2), 125-139.

Zhang, X. and M. Maher (1998). An algorithm for the solution of bi-level programming problems in transport network analysis. In: *Mathematics in Transport Planning and Control* (J. D. Griffiths, Editor), Elsevier, 177-186.

Chapter 10

PROCEDURES FOR DESIGNING NETWORK CONTROLS

Results from two Iterative Methods

Janet Clegg
Networks and Nonlinear Dynamics Group
University of York, York UK
Jc12@york.ac.uk

Yanling Xiang
Networks and Nonlinear Dynamics Group
University of York, York UK
Yanling@gridlock.york.ac.uk

Abstract The paper introduces a mathematical optimisation method which could be implemented into software in order to assist the transport planner in his decision making. The method optimises some objective whilst moving towards equilibrium. This objective need not be total travel time, the aim may be to reduce bus journey time for example. The control parameters can be signal timings, prices or capacity expansions.

Key words Traffic, networks , transport, optimisation, linear programming

Introduction

Urban Transportation is at a cross-roads; with changing targets ahead, and an expanding plethora of increasingly sophisticated controls the transport planner faces a daunting task.

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 161–174.

The target for the planner in the past was simply to minimise total delay, but nowadays targets can be much more complex; for example reducing bus travel time without increasing car travel time by too much or holding traffic back at some point on the network in order to avoid large queues further downstream.

The planner has complete responsibility for devising strategies, including pricing levels, likely to be successful when tested (on computers and in reality) against these new and changing targets. Vast computational resources are available for the assessment of the strategies once the planner has devised them, but at present the software is not capable of assisting in suggesting optimal strategies. A natural progression, therefore, is to develop software which can suggest optimal strategies to the planner when he is presented with a particular target.

The design of optimal or near optimal strategies, including the prices to be charged and signal timings to be implemented is far more difficult than assessing any given option. To find the best possible strategy the planner would need a superhuman insight and the task is becoming increasingly difficult as controls become more sophisticated, road networks become more complex and congestion increases. An effective and proven mathematical optimisation methodology implemented within helpful and easy-to-use software is an important tool which is currently absent from the transport planners tool-box. These mathematical tools are now (and only just now) becoming available to allow this development to actually take place. Thus now, for the first time, it is perhaps possible to implement a really helpful decision support system.

The solution methods proposed here were partially inspired by cone-fields introduced in Smale (1976). Smale introduced dynamical systems whose solution trajectories were not uniquely defined, they merely move in roughly the right direction, rather than exactly the right direction; and argued that such dynamical systems may be more appropriate for the study of the evolution of economic systems. The solution trajectories in these systems have their direction of motion at each point confined to be within a cone, instead of being confined to be in a precise direction.

This paper extends the cone projection method introduced in Battye et al (1998) and Clegg and Smith (1998). These cone projections are similar to those discussed in Zhang and Nagurney (1995) and Nagurney and Zhang (1996, 1998). However there appear to be some very substantial differences. The main difference is that throughout we are concerned with finding an appropriate direction to move controls.

One important element of the equilibrium modelling which arises is that the "equilibrium objective function" introduced by Beckmann et al (1956) is changed to one which allows asymmetries. The dynamics in this paper exploit Lyapunov methods (see Lyapunov (1907)).

1. FORMULATIONS OF THE PROBLEM

The formulation will be developed initially for a very simple network and this will hopefully make the notation in the case of a more general network clearer. Consider the network in Fig 4, which consists of just one origin and one destination with two links between them constituting two possible routes to the destination. Suppose the optimisation problem is to enlarge the capacities of these two links in order to reduce the total cost which would include travel time and also the cost of expansion. Let x_1 , x_2 be the flows along routes one and two; c the least cost of reaching the destination from the origin; y_1, y_2 the capacity expansions; $h_1(x_1, y_1)$, $h_2(x_2, y_2)$ the cost functions for travelling along routes one and two respectively and w be the demand. The objective function $Z = cw + y_1 + y_2$ is the total cost which we want to minimise.

Naturally it is required that $w = x_1 + x_2$. Also, to satisfy the equilibrium condition (more costly routes carry no flow, see Wardrop (1952)), the following inequalities must be satisfied

$c - h_i(x_i, y_i) \leq 0$ and if $c - h_i(x_i, y_i) < 0$ then $x_i = 0$ (for i=1,2).

The demand constraint can be written in a similar way to these equilibrium constraints as follows

$$w - x_1 - x_2 \leq 0 \quad \text{and if} \quad w - x_1 - x_2 < 0 \quad \text{then} \quad c = 0.$$

Let the vectors $\mathbf{z} = (x_1, x_2, c)$, $\mathbf{p} = (y_1, y_2)$ and the functions $f_1 = h_1(x_1, y_1) - c$, $f_2 = h_2(x_2, y_2) - c$, $f_3 = x_1 + x_2 - w$, then the problem becomes that of finding $\mathbf{z}$ and $\mathbf{p}$ such that for a given demand w the objective Z is minimised and the following inequalities

$$\begin{aligned} -z_i &\le 0 \\ -f_i &\le 0 \\ \sum z_i f_i &\le 0 \end{aligned} \qquad (1)$$

are satisfied.

The above example does not involve any junctions which naturally create delays. In a more complex network the vector $\mathbf{z}$ would include flows, delays and costs; the functions $f_i(\mathbf{z},\mathbf{p})$ (where the vector $\mathbf{p}$ includes control variables for example green times or prices) would represent capacity restrictions, demand constraints and equilibrium constraints. Suppose we have some objective function, $Z(\mathbf{z},\mathbf{p})$ which we require to minimise (the objective could be, for instance, bus travel time) subject to equilibrium and possibly some constraints $g_j(\mathbf{p}) \le 0$ on the vector of controls $\mathbf{p}$. This more general bilevel programming problem may be written in terms of inequalities as follows.

Minimise Z subject to

$$\begin{aligned} -z_i &\le 0, \\ -f_i(\mathbf{z},\mathbf{p}) &\le 0, \\ \sum z_i f_i(\mathbf{z},\mathbf{p}) &\le 0, \\ g_j(\mathbf{p}) &\le 0 . \end{aligned} \qquad (2)$$

Alternatively by defining the function $x_+ = \max(0,x)$, the above inequalities will be satisfied provided the function $E(\mathbf{z},\mathbf{p})$ is zero where E is given by

$$E(\mathbf{z},\mathbf{p}) = \sum_i (-z_i)_+^2 + \sum_i (-f_i)_+^2 + \sum_j (-g_j)_+^2 + (\sum z_i f_i)_+^2 . \qquad (3)$$

In this case the bilevel programming problem becomes that of minimising Z such that E is zero. Note that E will be zero only when all four terms in the expression are zero.

Any $(\mathbf{z},\mathbf{p})$ is called feasible if $g_j(\mathbf{p}) \leq 0$ and $-z_i \leq 0$. A feasible $(\mathbf{z},\mathbf{p})$ which also satisfies the constraints (or E=0) is called a feasible equilibrium.

2. OUTLINE OF THE TWO ITERATIVE PROCEDURES

2.1 Method 1

This first method is based on the formulation of the problem using equation (3) and is referred to as the cone projection method. It ensures convergence to an equilibrium and also seeks to simultaneously reduce the objective function, Z. The general form of the cone projection method is to continually follow a polygonal path which at each step follows a direction which moves towards equilibrium, while doing the best for the objective function Z.

In the cone projection method the function $E(\mathbf{z},\mathbf{p})$ given by (3) is split into two functions such that $E = E_1 + E_2$. Note that there is not a unique way of performing this split but that the method is to be extended so that the function E is split into many separate functions such that each one contains just one term from the function E. The functions E_1 and E_2 are measures of dis-equilibrium; they will be non-zero and positive away from equilibrium and both zero if and only if equilibrium has been reached. The function Z is to be minimised subject to both E_1 and E_2 being zero. We require to travel in the direction of a vector δ which reduces E_1 and E_2 towards zero whilst reducing Z, or reducing any increase in Z.

Naturally -grad E_i (the steepest descent direction) is the direction which reduces E_i most rapidly, for i=1,2. Consider the cone, $C = H_1 \cap H_2$, consisting of the intersection of the two half-spaces of locally non-increasing E_i, $H_i = \{\delta;\ \delta \cdot (\text{grad } E_i) \leq 0\}$. Each set H_i contains vectors which will either reduce E_i or cause no increase, provided a small distance is travelled along the direction of the vector. Since C is the intersection of the H_i sets, travelling a short distance in any direction

contained in C will not increase either of the functions E_i. The method is essentially to follow at each point the direction

$$- \text{ grad } (E_1+E_2) + \text{Proj}_C(- \text{ grad } Z), \tag{4}$$

where $\text{Proj}_C(-\text{ grad } Z)$ is -(the gradient of Z) projected onto the cone $C = H_1 \cap H_2$. The first term of equation (4) uses steepest descent directions, - grad E_i, to reduce E_1+E_2; whereas the second half of (4) attempts to reduce Z, but instead of taking the direction - grad Z this gradient is projected onto the cone of locally non-increasing E_i. The direction takes account of the bilevel nature of the problem since reducing E always gets priority. Once $E=E_1+E_2$ is reduced below some initial target value, say $E \leq \varepsilon$ (where $\varepsilon > 0$) and also Z is approximately minimised within the region $E \leq \varepsilon$, ε is reduced and the process repeated. The method ensures that reaching equilibrium always gets priority.

The ε introduced here may be utilised to ensure that the direction travelled remains continuous by including an extra term involving -grad Z which attempts to reduce Z more quickly when $E<\varepsilon$. Following Battye et al (1998) and Clegg and Smith (1998), assuming all gradient vectors have been normalised to be unit length, the new direction can be written

$$(E/\varepsilon - 1)_+ [- \text{grad } (E_1 + E_2)] + (1 - E/\varepsilon)_+ (- \text{grad } Z) + E/\varepsilon \, \text{Proj}_C (- \text{grad } Z). \tag{5}$$

2.2 Method 2

Method two deals with the problem in the form given in equation (2), that is a set of inequalities. For simplicity in notation, let $\mathbf{x} = (\mathbf{z},\mathbf{p})$ and suppose we define a sequence of vectors $\mathbf{x}^k$ such that as k increases $\mathbf{x}^k$ moves closer to equilibrium and simultaneously attempts to reduce the objective, Z, as much as possible. Let $d\mathbf{x}^k = \mathbf{x}^k - \mathbf{x}^{k-1}$ so that $d\mathbf{x}^k$ is the direction to travel in order to obtain the subsequent

vector $\mathbf{x}^k$ in the iterative procedure knowing the vector $\mathbf{x}^{k-1}$. Ideally we would require that the new vector $\mathbf{x}^k$ satisfies the inequalities

$$\begin{aligned} -x_i^k &\le 0 \\ -f_i(\mathbf{x}^k) &\le 0 \\ \textstyle\sum_i x_i^k f_i(\mathbf{x}^k) &\le 0 \end{aligned}$$

Using $\mathbf{x}^k = \mathbf{x}^{k-1} + d\mathbf{x}^k$ and letting $h(\mathbf{x}) = \sum_i x_i f_i(\mathbf{x})$ these inequalities can be approximated by the following set of inequalities

$$\begin{aligned} -x_i^k &\le 0 \\ -f_i(\mathbf{x}^{k-1}) - f_i'(\mathbf{x}^{k-1}) \,.\, (\mathbf{x}^k - \mathbf{x}^{k-1}) &\le 0 \\ h(\mathbf{x}^{k-1}) + h'(\mathbf{x}^{k-1}) \,.\, (\mathbf{x}^k - \mathbf{x}^{k-1}) &\le 0 \end{aligned}$$

which are of the form

$$\begin{aligned} -x_i^k &\le 0 \\ -c_i - a_i \,.\, \mathbf{x}^k &\le 0 \\ c + a \,.\, \mathbf{x}^k &\le 0 \end{aligned} \qquad (4)$$

where

$$\begin{aligned} c &= h(\mathbf{x}^{k-1}) - h'(\mathbf{x}^{k-1}) \,.\, \mathbf{x}^{k-1}, \\ c_i &= f_i(\mathbf{x}^{k-1}) - f_i'(\mathbf{x}^{k-1}) \,.\, \mathbf{x}^{k-1}, \\ a &= h'(\mathbf{x}^{k-1}), \\ a_i &= f_i'(\mathbf{x}^{k-1}) \end{aligned}$$

are constants since $\mathbf{x}^{k-1}$ is known. Since these approximate constraints are now linear in $\mathbf{x}^k$ the problem of minimising Z subject to constraints (4) can be treated as a linear programming problem. The simplex method is used to obtain the vector $\mathbf{x}^{k*}$ which minimises Z subject to the linearised constraints and the vector $\mathbf{x}$ is updated using the formula

$$\mathbf{x}^k = \mathbf{x}^{k-1} + \delta\,(\mathbf{x}^{k*} - \mathbf{x}^{k-1})$$

where δ is the step size in the iterative procedure. A suggested method for choosing δ could be a line search in the direction of $\mathbf{x}^{k*} - \mathbf{x}^{k-1}$. The starting value $\mathbf{x}^0$ for the iterative process can be chosen using part guess work; for example letting the initial flows along routes be the demand for that origin-destination pair divided by the number of

different routes; or for green times to be chosen by equally distributing the green time between all stages.

3. APPLICATION OF THE TWO METHODS

Method 1 has already been applied to the very simple network in Figure 3, in this paper it is applied to the more complex network considered by Hai Yang (1996). The network involves just one origin and destination pair (from node 1 to node 6) but it has a total of five possible routes and two sets of signalised junctions situated at nodes 4 and 5.

Let x_r be the flow along route r, v_i the flow along link i, b_i the bottle-neck delay at the exit of link i, s_i the saturation flow on link i, t_i^0 the free-flow time on link i, λ_i the green time for the signal at the exit of link i (λ_i =1, i=5...9), c the least cost of travelling from origin to destination and ρ the demand. The five routes pass through links (5,1,7), (5,2,8), (6,3,7), (6,4,8) and 9 and the control variables in this case are the green times, λ_i. Table 1 lists some data used within this problem.

Table 1 Input data for network two

Link no.	1	2	3	4	5	6	7	8	9
v_i	x_1	x_2	x_3	x_4	x_1+x_2	x_2+x_4	x_1+x_3	x_2+x_4	x_5
s_i	50	50	50	80	100	100	100	100	200
t_i^0	4	2	3	4	4	3	4	4	15

The following delay formula (taken from Hai Yang (1996)) has been used

$$t_i = t_i^0\{ 1 + (v_i/\lambda_i s_i)^2 \}$$

The aim is to minimise total travel time, $Z = \sum_i x_i(t_i + b_i)$, subject to the following inequalities

$$-f_i \le 0,\ -g_j \le 0,\ -h \le 0,\ \sum f_i b_i + \sum g_j x_j + ch \le 0,\ \lambda_1+\lambda_3 \le 1,\ \lambda_2+\lambda_4 \le 1,$$

where

$$f_i = \lambda_i s_i - v_i,\quad g_j = \sum_{(k \text{ on route } j)} (b_k + t_k) - c,\quad h = \sum x_i - \rho .$$

We use functions E_1 and E_2 defined by:

$$E_1 = \sum_i (-f_i)_+^2 + \sum_i (-g_j)_+^2 + h_+^2 + (1-\lambda_1-\lambda_3)_+^2 + (1-\lambda_2-\lambda_4)_+^2 + \sum(\lambda_i - 0.95)_+^2 \sum(0.05-\lambda_i)_+^2$$

$$E_2 = (\sum f_i b_i + \sum g_j x_j + Ch)_+^2 + \sum_i (-x_i)_+^2 + \sum_i (-b_i)_+^2 + (-c)_+^2.$$

Using method 1 the "optimal" total travel time for demand $\rho = 100$ was found to be Z=1377.7 with control green times $\lambda_1 = 0.395$, $\lambda_2 = 0.285$. Figure 1 displays the value of $E = E_1+E_2$ as iterations proceed and Figure 2 shows the value of the objective function Z. Note that in Figure 1 the size of E increases occasionally over iteration numbers. This is because the iterative procedure is at this point attempting to reduce Z.

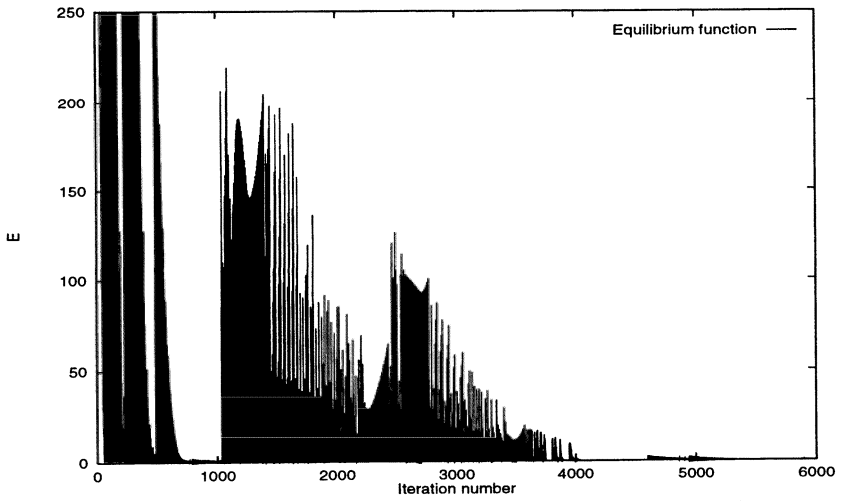

Figure 1 Equilibrium function E displayed through iterations

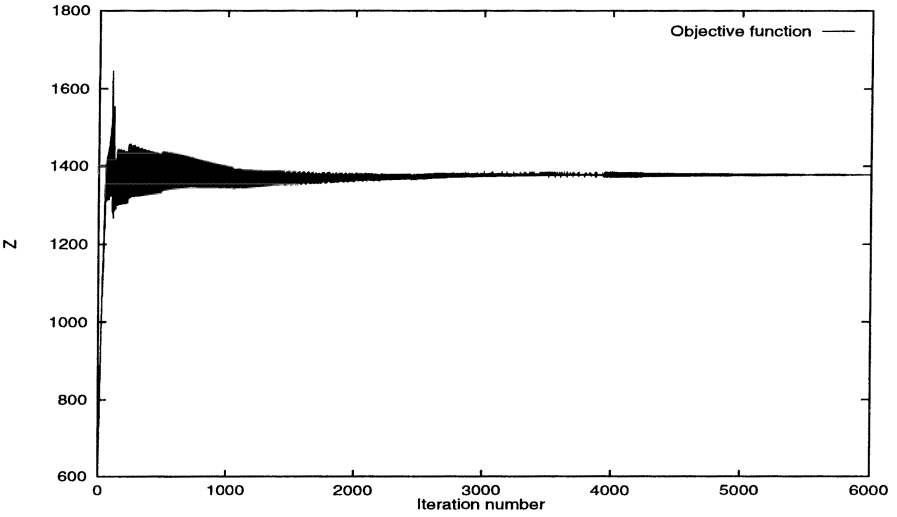

Figure 2 The value of the objective function through iterations

Figure 3 A Simple network

Method 2 has been applied to the more simple network shown in Figure 3. Here we consider a simple two-route network, taken from Marcotte (1988). The network has one origin and one destination with two links between the two nodes. The aim is to minimise total cost by choosing capacity expansions y_i along the links. The flows along the links are represented by x_1, x_2; capacity expansions are y_1, y_2 and congestion functions are given by

$$f_1(x_1, y_1) = 2x_1/(1+y_1) \quad \text{and} \quad f_2(x_2, y_2) = 8x_2/(1+y_2).$$

We define c to be the least cost of reaching the destination and ρ to be the demand. Within this problem the control variables are the y_i. We aim to minimise the following objective function representing total cost (congestion cost and construction cost) by choosing optimal capacity expansions

$$\begin{aligned} \text{Minimise} \quad & Z = \rho c + y_1 + y_2 \\ \text{subject to} \quad & c - f_1 \leq 0 \\ & c - f_2 \leq 0 \\ & \rho - x_1 - x_2 \leq 0 \\ & x_1 f_1 + x_2 f_2 - \rho c \leq 0. \end{aligned}$$

The first two inequalities ensure that drivers follow their cheapest route, the third inequality ensures that the demand, ρ, is correct and the last inequality ensures equilibrium. Method 2 outlined in this paper produced an optimal value for the objective function of Z=1.578 for demand ρ=1 which agrees with that quoted in Marcotte (1988). Varying demands (ρ =0..2) have been examined; the optimal values of Z are depicted in Figure 4 and Figure 5 displays the values of the control variables, y_1,y_2 as ρ increases from 0 to 2. Note that in this particular problem the optimal value of y_2 is zero and therefore is not visible in Figure 5.

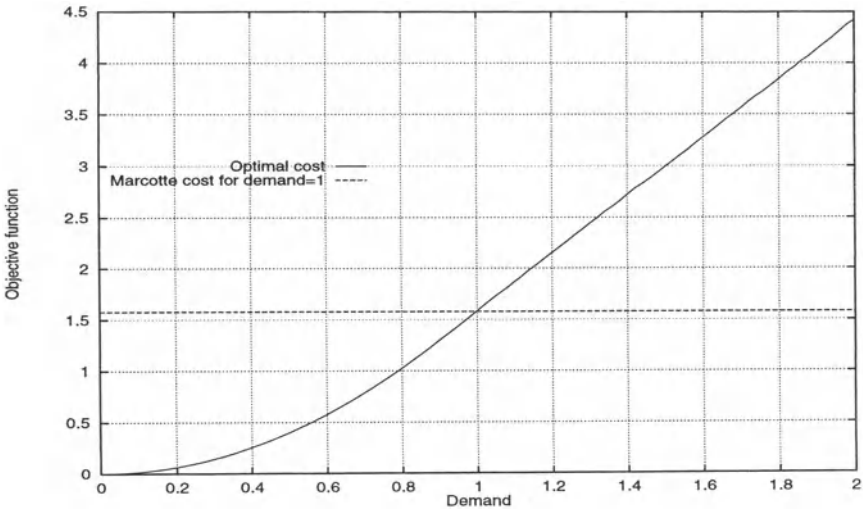

Figure 4 Optimised total travel time varying with demand

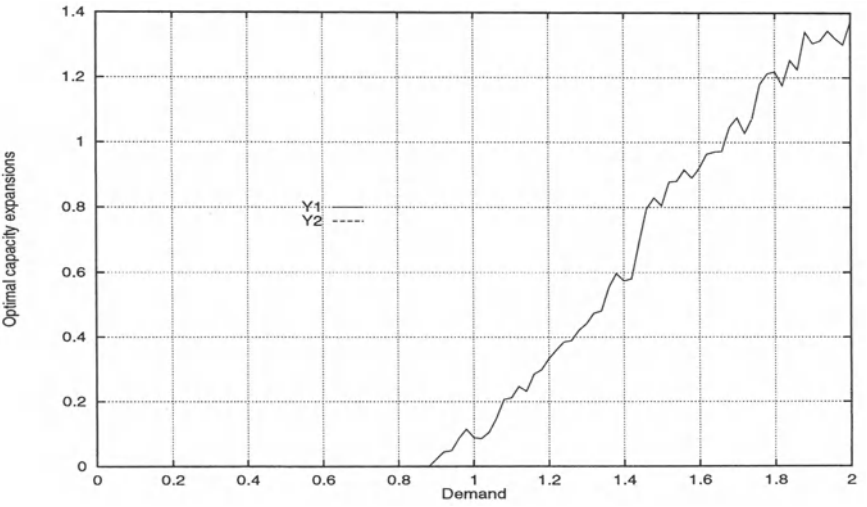

Figure 5 Values of the control parameters for optimal cost

4. CONCLUSIONS

In this paper a mathematical formulation has been developed for the problem of optimising a traffic network whilst ensuring equilibrium. Two different methods of iterating towards an optimal solution have been described; the cone projection method and a linear programming technique. Both these methods have been tested on simple networks and the results look promising.

The linear programming method is new in this field and the results in this paper are the first whereas the cone projection method has been tested on other networks. From this initial test of the linear programming method it appears to behave better than the cone projection method. It converges to the solution faster and reaches this solution in a more continuous manner (without the oscillations visible in Figure 1).

The large number of iterations seen in Figure 1 is impractical, but the fact that so many iterations have been applied in this test does not mean that this number would be required if the method were applied to a full size network. In this paper an extremely small step size was used in the iterative process mainly because it was considered better to use one which was too small than one too large. Also in practical situations the algorithm would not have been left running for so long, since the accuracy of the results would have been satisfactory after much fewer iterations. There is further work to be done on various aspects of the methods, for example better techniques for choosing an efficient step length.

Work is at present being done towards implementing the linear programming method to the network of York city. Hopefully once this is completed the performance of the network with new optimal signal timings will be compared to its performance with the timings which are implemented in York at present in order to evaluate the benefits of the method. Provided the method does succeed in producing signal timings which are beneficial, it could be implemented into software used by the traffic planners in order to assist them in their decision making.

REFERENCES

Battye A, Smith M J and Xiang Y (1998), "The cone projection Method of Designing Controls for Transportation Networks", Proc.3rd IMA International Conf. on Math. in Transport Planning and Control, Math. in Transport Planning and Control, pp 29 – 37.

Beckman M, McGuire C B and Winsten C B (1956), "Studies in the economics of transportation", Yale University Press, New Haven, CT.

Clegg J and Smith M J (1998), "Bilevel optimisation of Transportation Networks", Proc. 3rd IMA International Conf. on Math. in Transport Planning and Control, Math. in Transport Planning and Control, pp 29 – 37.

Lyapunov A M (1907), "Probleme general de la stabilite de mouvement", Ann. Fac. Sci. Toulouse 9, pp 203-274. Reprinted in Ann. Math. Stud. No 12, 1949.

Marcotte P (1988), "A note on a bilevel programming algorithm by Leblanc and Boyce", Transpn. Res. B, Vol 22B, No 3, 233-237.

Nagurney A and Zhang D (1996), "Projected dynamical systems and variational inequalities with applications", Kluwer Academic Publishers, Boston, Massachussetts.

Nagurney A and Zhang D (1998), "Network equilibria and disequilibria", Equilibrium and Advanced Transportation Modelling, (Editors Marcotte P, Nguyen S), Kluwer Academic Publishers, Massachussetts, pp 201-243.

Smale S (1976), "Exchange processes with price adjustment", Journal of Mathematical Economics, 3, pp 211-226.

Wardrop J G (1952), "Some theoretical aspects of road traffic research", Proc. Institution of Civil Engineers II, 1, 235-278.

Yang H (1996), "Equilibrium network traffic signal setting under conditions of queuing and congestion", Appl. advanced technologies in transportation engineering, Proc.4th International Conference, American society of civil engineers, pp 578-582.

Zhang D and Nagurney A (1995), "On the stability of projected dynamical systems", Journal of Optimisation Theory and Applications, 85, pp 97-124.

Chapter 11

APPROACH TO CONGESTION OPTIMUM TOLL IN TRAFFIC NETWORKS

Manuel A. Gomez-Suarez
University of A Coruna
mago@udc.es

Luis P. Pedreira-Andrade
University of A Coruna
lucky@udc.es

J.Antonio Seijas-Macias
University of A Coruna
jasm@udc.es

Abstract In this article we study the pricing of transportation in congestion highways. The traditional approach to tolls consider the introduction of marginal tolls into networks as a solution to congestion. Recently, some authors consider another type of tolls, minimum-revenue tolls, where they get the same effect than with marginal tolls by means of tolling only a few links in the network. In [2] an algorithm using a link-path formulation for a multicommodity assignment problem is presented. We use minimum revenue tolls using a link-node formulation, for only one single commodity and multiple O-D pairs, with a better performance of the algorithm.

Keywords: Optimum toll, traffic assignment, pricing in transportation

1. INTRODUCTION

Great volumes of traffic flow in traffic networks have produced the existence of a new phenomenon in the traffic stream: congestion. Con-

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 175–186.

gestion is associated to a higher level cost for the use of the network as a consequence of increasing time needed to cross the links. The reaction of planners to this situation has followed two schemes, basically: the first, improving the capacity of the network in order to reduce the cost function and then minimize the effect of congestion, and the second, pricing for the use of the links where congestion appears. A solution would be a toll that produces a new flow vector in the network with lower levels of congestion at the links. Economic literature on tolls [11] establishes that the optimum toll is the marginal social cost toll (marginal toll). Users of the link have to bear the private cost associated to cross the link, and an additional cost (toll) representing the social cost that new users impose over the rest of the users that are crossing the link. Toll is calculated as a portion of time and monetary value is estimated by means of a value of time. This approach to optimum toll has some problems: most important being the fact that, at the optimal solution, every link in the network is tolled. But introduction of tolls in links has an extra cost (costs associated to establishing barriers and other systems to pick up the toll), and it is possible that the cost of tolls be greater than the revenue collected at some links.

In this work we consider a double problem of programming. In a first instance we calculate the traffic flow vector that optimizes the network, and the second problem is calculating the alternative toll vector associated to minimum costs. We compare our formulation with a very similar one from Bergendorff et al. in [2] and with the proposed algorithm of Dial [3]. The article is compound in the following sections. In section 2 we introduce notations and definitions in order to establish the problem to optimize. Section 3 studies the construction of the valid toll set where we choose the optimum toll set. The network optimization procedure is considered in section 4. Finally, we briefly discuss the conclusions of our article in the last section.

2. DEFINITION OF THE PROBLEM

Let $DG = (N, A)$ be a directed graph that represents a network with a set of nodes N and a set of links between them, A. x is the flow vector of the links, and b the demand/supply vector associated to every node in the network. If $b_i < 0$, then node i is a destination node; if $b_i > 0$ then i is an origin node, and i is an intermediate node if $b_i = 0$. All users move from origin nodes to destination nodes, according to a fixed demand/supply vector b. We consider the link-node matrix G, which is

a $card(N)$ x $card(A)$ matrix.[1] Elements of G are:

$$g_{ij} = \begin{cases} 1, \text{ if link } j \text{ is incident from node } i. \\ -1, \text{ if link } j \text{ is incident to node } i. \\ 0, \text{ otherwise.} \end{cases} \tag{11.1}$$

Then, we establish the feasible flow vector set of the network as

$$F = \left\{ x \in IR^{|A|} : Gx = b, x \geq 0 \right\}. \tag{11.2}$$

Every link has associated a cost function $c : A- > A$ that establishes a cost for using the link. This cost is valued as the time a car needs to cross the link, and it is a function of the flow crossing it. Assumptions about the cost function are:

Assumption 1: Strictly monotonic function: c is a strictly monotonic function if it verifies:

$$(c(x^1) - c(x^2))^T (x^1 - x^2) > 0, \forall x^1, x^2 \geq 0, x^1 \neq x^2. \tag{11.3}$$

Assumption 2: Nonlinear and separable function: cost associate to each link is a nonlinear function that depends exclusively on the flow that crosses the link.

Assumption 3: SS - convex function: c is an S - convex function if $c(x)^T x$ is a convex function. When this function is strictly convex then c is an SS - convex function.

The traffic assignment problem is defined as user optimal and system optimal [12]. Both problems are:

User Optimal:

$$\begin{array}{ll} \min\limits_{x \in F} & \sum_{i \in A} \int_0^{x_i} c(w) dw \\ \text{s.t:} & Gx = b, \\ & x \geq 0. \end{array} \tag{11.4}$$

System Optimal:

$$\begin{array}{ll} \min\limits_{x \in F} & c(x)^T x \\ \text{s.t:} & Gx = b, \\ & x \geq 0. \end{array} \tag{11.5}$$

User optimal may be considered an inequality variational problem [9], then a x^* vector is a solution to user optimal if and only if

$$c(x^*)^T (x - x^*) \geq 0, \qquad \forall x \in F. \tag{11.6}$$

In general, there is no coincidence between the optimal flow vector to both problems. Then, we have two different solutions as a consequence of the approach to the traffic network that we were managing.

2.1 MARGINAL SOCIAL COST AND TOLLS

Toll functions try to set a function, relating the level of the toll, as the flow / capacity ratio of the roadway varies. In [2] the authors state the principle of toll pricing:"The tolls imposed should be such that the resulting tolled user equilibrium problem has at least one solution and every such solution is an untolled system optimal solution". The traditional approximation to the variable toll is based on the difference between short-run marginal cost and short-run average cost; these are SRMC and SRAC [10]. We can define the marginal toll as the marginal social cost:$(\partial c_i/\partial x_i)x_i, \quad \forall i \in A$, and the toll vector is

$$\pi = \nabla c(x^*)^T x^*, \tag{11.7}$$

where x^* is the optimal flow vector. Now, we reformulate user-optimal (11.4) incorporating marginal tolls (11.7) in our objective function. Next, we establish the lemma:

Lemma 1 Let $x^* \in F$ be a system-optimal solution, then x^* is a solution to a user-optimal problem with tolls, that is

$$c(x^* + \pi)^T (x - x^*) \geq 0, \quad \forall x \in F. \tag{11.8}$$

Marginal tolls may be used to obtain an optimal solution for the two programming problems. But this solution calculates a toll different from zero for every link, therefore we have as many tolls as links. In this situation, when the network is big enough there is an excessive number of tolls, making them the solution to the congestion problem not practical. When we use the BPR function as the link cost function, marginal tolls are greater in those links with a greater value of the tf parameter (minimum time to cross the link with flow zero). The links with worst tf value have greater tolls, and as a consequence, with these marginal tolls, congestion is not reduced.

Lemma 2 Let x^* be a solution to a user-optimal problem with tolls, then

1. $\pi \geq 0$, and $\pi_i = 0$ if and only if $x_i = 0$, $\forall i \in A$.
2. When we use a BPR function:

$$c_i(x) = tf_i \left(1 + \alpha \left(\frac{x_i}{K_i} \right)^{\beta} \right), \tag{11.9}$$

where α and β are parameters and tf is the time to cross the link without flow and K is the capacity of the link, then $\pi_i \geq \pi_j$ if $tf_i \geq tf_j$.

Proof 1 This lemma is a direct consequence of the definition of marginal tolls. Accordingly, marginal toll is establishing by (11.7). Then, using the BPR function (11.9), tolls result:

$$\pi_i = tf_i \alpha \beta \left(\frac{x_i}{K_i} \right)^{\beta}, \tag{11.10}$$

is only zero if flow in the link is null. Moreover, this function is a direct function of the tf parameter. Then for two links with the same value for the rest of parameters, the value of toll is greater in that parameter with a greater value of tf.

Introduction of marginal tolls in a network produces two perverse effects: the first one, every link has a toll. The second effect is that the marginal social cost tolls do not reduce congestion because they penalize links with greater cost. We need to study if there is possible to establish a toll vector equivalent and associated to a optimal flow vector.

3. CONGESTION TOLL

Set X is settled by the flow vectors that are the solution to user traffic assignment problem with tolls, and S^* is the set of system-optimal vectors:

$$X = \left\{ x \in F : (c(x) + \pi)^T (x' - x) \geq 0, \forall x' \in F \right\}, \tag{11.11}$$

$$S^* = \arg\min \left\{ c(x)^T x : x \in F \right\}. \tag{11.12}$$

Since every toll vector is such that the flow vector associated is a solution to the problem of system traffic equilibrium assignment without tolls, and to the user equilibrium assignment with toll, then

$$\emptyset \neq X \subseteq S^*. \tag{11.13}$$

Every vector $\pi \geq 0$ that satisfies (11.13) is a valid toll vector, and then, we define the toll vector set T

$$T = \{ \pi \geq 0 : \emptyset \neq X \subseteq S^* \}. \tag{11.14}$$

In [2] it is proved that every element in X is a member of set S^*, and X is a non-empty set. Then, every vector greater or equal than zero in X is considered a valid flow vector and we define the valid toll set; for a flow vector $x^* \in S^*$ system-optimal we can establish the valid toll set that guarantees that x^* is a solution to user-optimal with tolls, as

$$W(x^*) = \{ \pi \geq 0 : x^* \in X \}. \tag{11.15}$$

Under assumptions (A1-A3), the optimal solution to the system-optimal problem is a global solution (see [4],[5]), then X contains only one element, but the set $W(x^*)$ may contain several valid toll vectors. In order to characterize the valid toll set we use the Karush-Kuhn-Tucker (KKT) conditions [5], and $W(x^*)$ is constructed as indicated in the next theorem:

Theorem 1 For $x^* \in S^*, W(x^*)$ is a convex polyhedron constructed with vectors π of the linear inequality system with variables π and λ:

$$(c(x^*) + \pi) \geq G^T \lambda, \tag{11.16}$$

$$(x^*)^T (c(x^*) + \pi) = b^T \lambda. \tag{11.17}$$

Proof of this theorem comes directly from the Karush-Kuhn-Tucker (KKT) conditions of optimality in the system optimal problem [5]. Then, we construct toll valid set T as union of polyhedrons W.

Theorem 2 (Bergendorff et al., 1997) If c is a strictly monotonic function, then

$$T = \bigcup_{x^* \in S^*} W(x^*). \tag{11.18}$$

Proof 2 See [2].

When S^* is a set with only one element, then system-optimal is unique and is guaranteed by (11.5), and then we can formulate the next corollary:

Corollary 1 Let c be a strictly monotonic and SS-convex function, there is an only solution x^* to the traffic assignment problem, and the valid toll set is $W(x^*)$.

Both theorems and corollary guarantee that the convex polyhedron is non-empty, and then, we can calculate a set of valid toll vectors associated to optimal flow vector.

4. NETWORK OPTIMIZATION

Network optimization is a two-level programming problem. The optimal solution is a vector (x^*, π^*) that solves the problem in two non-parallel phases:

1. First phase: Traffic Assignment.

$$\begin{aligned} \min_{x \in F} \quad & x^T c(x) \\ \text{s.t:} \quad & Gx = b, \\ & x \geq 0 \end{aligned} \tag{11.19}$$

2. Second phase: Optimum Toll.

$$\begin{array}{ll} \min\limits_{(\pi \in W(x^*),\lambda)} & \pi^T x^* \\ \text{s.t:} & (c(x^*)+\pi) \geq G^T\lambda, \\ & (x^*)^T(c(x^*)+\pi) = b^T\lambda, \\ & \pi \geq 0. \end{array} \tag{11.20}$$

This problem is called minimum-revenue congestion pricing because the objective of the second phase is minimizing the total revenue of tolls. Solving this problem we get a solution that reduces the number of links which have been tolled at the optimal solution. Then, we reduce the associated cost of establishing tolls.[2] The first phase problem is a classical traffic assignment problem and we can compute it by traditional methods of nonlinear programming (Frank-Wolfe, Active-Set, Gradient-project,...).[3] The second phase problem is a linear programming problem with constraints. We use the flow vector solution from the first phase as the linear coefficients of the objective function. In order to simplify the problem, we could eliminate from the network the links with flow zero (if they exist), then we would have to update the vector and matrix of the problem, that correspond with the eliminated links. The solution to the second phase produces tolls that avoid the perverse effects of marginal tolls, and reduce the number of link-tolls, and the total revenue. We define a free toll path as a path between the origin node i and destination node j, where the links that belong to the path have minimum-revenue toll equal to zero. Next propostion shows that there exists always a free-toll link.

Proposition 1 Let $\pi \in W(x^*)$ be a minimum-revenue toll vector, where x^*, $(x^* \geq 0)$ is a solution to the user-optimal problem with tolls, then there is a free toll path between every origin node and every destination node.

Proof 3 The Lagrangian function is

$$\begin{aligned} L(\pi,\lambda,\mu,\phi,\varphi) = &-\pi^T x^* + \mu^T(c(x^*)+\pi-G^T\lambda)+ \\ &+\phi(x^{*T}(c(x^*)+\pi)-b^T\lambda)+\varphi^T\pi \end{aligned} \tag{11.21}$$

The Karush-Kuhn-Tucker conditions are:

$$\frac{\partial L}{\partial \pi} = -x^{*T}+\mu^T+\phi x^{*T}+\varphi^T = (\phi-1)x^{*T}+\mu^T+\varphi^T = 0 \tag{11.22}$$

$$\frac{\partial L}{\partial \lambda} = -\mu^T G^T - \phi b^T = 0 \tag{11.23}$$

$$c(x^*) + \pi \geq G^T \lambda, \;\; \mu \geq 0, \;\; \mu^T (c(x^*) + \pi - G^T \lambda) = 0 \tag{11.24}$$

$$x^{*T}(c(x^*) + \pi) = b^T \lambda \tag{11.25}$$

$$\pi \geq 0, \;\; \varphi \geq 0, \;\; \varphi^T \pi = 0 \tag{11.26}$$

Now, we prove that there exists an index i such that $\pi_i = 0$.

Suppose on the contrary that $\pi > 0$. The complementary slackness condition then implies that $\varphi = 0$. Hence, condicion (11.22) simplifies to

$$\mu = (1 - \phi) x^* \tag{11.27}$$

If there exists an index j such that $x_j^* > 0$, and $\mu_j = 0$ then $\phi = 1$ which in turn implies that $\mu = 0$. Hence, there are the following cases:

Case 1. $\mu = 0$. Hence, (11.23) implies that $\phi b = 0$. If $\phi = 0$, this is possible only if $x^* = 0$, which a contradiction with the hypothesis. If $b = 0$, (11.25) implies that $x^{*T} \pi = 0$, so that if there exists an index j such that $x_j^* > 0$, then $\pi_j = 0$, contradicting the assumption.

Case 2. $\mu \geq 0$ and $\mu \neq 0$. In this case, $\phi \neq 1$, since if $\phi = 1$ we have already seen that $\mu = 0$. Hence, $\mu_j > 0$ if and only if $x_j^* > 0$, and $\mu_j = 0$ if and only if $x_j^* = 0$. Let $J \subset IN$ be defined by $i \in J$ if and only if $\mu_j > 0$. Then, if h_J denotes the vector formed by joining the components of the vector h indexed by J, (11.24) implies that

$$(c(x^*) + \pi)_J = (G^T \lambda)_J \tag{11.28}$$

>From (11.23),

$$\begin{aligned} -\mu^T G^T \lambda - \phi b^T \lambda = -\mu_J^T (G^T \lambda)_J - \phi b^T \lambda = \\ = (\phi - 1) x_J^{*T} (c(x^*) + \pi)_J - \phi b^T \lambda \\ = (\phi - 1) x^{*T} (c(x^*) + \pi) - \phi b^T \lambda = \\ = (\phi - 1) b^T \lambda - \phi b^T \lambda = -b^T \lambda = 0 \end{aligned}$$

Hence, (11.25) implies that

$$x^{*T}(c(x^*) + \pi) = 0 \tag{11.29}$$

which in turn implies that $\pi_J = 0$, contradicting the assumption.

Hence, the assumption that $\pi > 0$ is incorrect, so that there exists an index i such that $\pi_i = 0$.

The former proposition supposes that we have links with toll zero. Another consequence is that there is no direct relation between toll and the parameters of the cost function. We eliminate the perverse effect of marginal social cost tolls.

4.1 COMPLEXITY OF THE MINIMUM-REVENUE TOLL

One of the most common algorithms in linear programming is the simplex method. The worst-case behavior of the simplex method determines that the overall costs are $O(m^4+nm^2)$ arithmetic operations, where m is the number of constraints of the problem and n is the number of variables (see [1]). This complexity number is polynomial in m and n.[4] Obviously, calculus of the worst-case behavior is based on dense-matrix problems, but network problems are on sparse matrix, generally matrix G is very sparse (only 3 or 4 non-zero elements at each row), and costs are lower than $O(m^3 + nm)$.

The number of variables and constraints in problem (11.20) are: number of variables $= card(A) + card(N)$, that is, the number of variables obtained by adding the number of links (that are the components of vector π) and the number of nodes (that determines the components of vector λ). The number of constraints=$card(A) + 1$, where $card(A)$ indicates the number of inequality constraints, and there is one equality constraint in order to complete the feasible region. Let $card(A) = l$ and $card(N) = k$ be the number of links and nodes of the network, respectively. Then if our problem has $(l + k)$ variables and $(l + 1)$ constraints, the costs of the simplex method are $O((l+1)^3 + (l+k)(l+1))$, that is a polynomial time when we use $O(l + 1)$ iterations, and exponential time if it requires $O\begin{pmatrix} l+k \\ l+1 \end{pmatrix}$ iterations.

Bergendorff's algorithm in [2] uses a node-link incidence matrix. The number of variables is equal to the number of links plus the number of pairs of Origin-Destination nodes (O-D nodes), and the number of constraints is equal to the number of paths between origin and destination nodes that determine the number of inequality constraints, plus one equality constraint. Normally, in large-size networks, the number of paths

is much greater than the number of nodes, and the costs of the simplex method are $O((p+1)^3 + (p+1)(l+h))$, where p is the number of paths and h the number of pairs of O-D nodes in the network. Then, for large-size networks

$$O((l+1)^3 + (l+k)(l+1)) \leq O((p+1)^3 + (p+1)(l+h)). \quad (11.30)$$

At the example in [2] the data of the network are: number of links, 18; number of nodes, 9; number of paths, 48; and number of O-D pairs, 4. Then, the complexity of the simplex algorithm is $O(19^3 + 27 * 19)$ in our approach, that is a lower operational cost than $O(49^3 + 49 * 22)$, considering the former approach of the authors.

5. CONCLUSION

Some authors have treated the problem of minimum-revenue toll in a optimum flow network [7] and [8]. Both authors study algorithms that solve the problem. Dial's algorithm has complexity O(card(A)), a very good value, but it is only valid for single origin-destination networks. We adapt a reformulation of the algorithm in [2] in order to improve the cost complexity of the former algorithm. Nevertheless, the result is not as good as in [3], but our approach is valid for multiple O-D networks.

Consequences of calculating the solution to the optimum toll problem is a valid toll vector that minimizes the revenue of tolls. A second advantage for this solution is that the number of links with tolls is less than with the marginal toll solution. Another consequence is reducing the number of tolls with very low values that are not feasible in practice and we eliminate the perverse effect of the parameter tf. There is no direct relation between the parameter and the toll.

Optimum toll problem is a linear programming problem with several inequality constraints and one equality constraint. There are a lot of algorithms for linear programming, among the more typical is the simplex algorithm [1]. Studies on their computational complexity show that the simplex algorithm is more sensitive to variations in the number of constraints than in the number of variables. In [6] the complexity of simplex algorithm is, in practice, linear, with factor 3-6, for the number of constraints, and sublinear for the number of variables. In our approximation to the problem, the number of variables is the sum of the number of links plus the number of nodes in the network, and number of inequality constraints is the number of links plus one. For [2] the number of variables is only the number of links, the number of inequality constraints is equal to the number of paths between origin node adding the number of pairs of O-D nodes. In large networks the number of paths is bigger than the number of links. Then, our algorithm introduces a slight improvement

of the complexity of the problem. We study an example of the problem and calculate the solution to a network example that has been studied in [2]. Complexity of their solution is greater than our proposal.

Normally, linear programming methods are very dependent on the number of constraints. Thus, we must choose the expression of the problem involving a lower number of constraints.

Notes

1. card ( ) , represents cardinality of the set ().
2. See [2] and [3].
3. See [9] for a wide study about these techniques.
4. This value is calculated supposing that we perform $O(m)$ iterations; in the worst-case if we need $\binom{n}{m}$ iterations then the simplex method could be an exponential algorithm.

References

[1] Ahuja, R.K, T.L.Magnanti and J.B. Orlin. Network Flows. Prentice-Hall, New Jersey, 1993.

[2] Bergendorff, P., D.W. Hearn and M.V. Ramana. Congestion Toll Pricing of Traffic Networks. In Panos M. Pardalos, Donald W. Hearn and William H. Hager, editors, Network Optimization, pages 51-71, Springer-Verlag, Lectures Notes in Economics and Mathematical Systems, Berlin Heidelberg New York Tokyo, 1997.

[3] Dial , R.B. Minimal-revenue congestion pricing part I: A fast algorithm for the single-origin case. Transportation Research Part B, 33:189-202, 1999.

[4] Fletcher, R. Practical Methods of Optimization. J. Wiley and Sons, New York, 1987.

[5] Gill, P.E., W. Murray and M. Wright. Practical Optimization. Academic Press, New York, 1981.

[6] Gill, P.E., W. Murray and M. Wright. Numerical Linear Algebra and Optimization. Addison-Wesley, New York, 1991.

[7] Hearn,D. and Ramana,M. Solving Congestion Toll Pricing Models. In Patrice Marcotte and Sang Nguyen, editors, Equilibrium and Advanced Transportation Modelling, pages - , Kluwer Academic Publishers, 1998.

[8] Larsson,T. and M. Patriksson Side constrained traffic equilibrium models - Traffic Management Through Link Tolls. In Patrice Marcotte and Sang Nguyen, editors, Equilibrium and Advanced Transportation Modelling, pages - , Kluwer Academic Publishers, 1998.

[9] Patriksson, M. The Traffic Assignment Problem. VSP, Utrecht, The Netherlands, 1994.

[10] Shah, A. Optimal Pricing of Traffic Externalities: Theory and Measurement. International Journal of Transport Economics, 17(1):3-19, 1990.

[11] Small, K. Urban Transportation Economics. Harwood Academic Press, Chur, Switzerland, 1992

[12] Wardrop, J. Some Theoretical Aspects of Road Traffic Research. Proc. Inst. Civ. Eng. Part II, 1:325-378, 1952.

Chapter 12

A Dynamic Network Loading Model for Simulation of Pollution Phenomena

Mauro Dell'Orco
Department of Highways and Transportation
Polytechnic of Bari
dellorco@dvt005.poliba.it

Abstract It is well known that traffic flows, vehicles speed and acceleration closely concern traffic pollution. Therefore, calculating these characteristics as precise as possible is really relevant when dealing with such a phenomena. Usual tools in computing the values of traffic characteristics are traffic assignment models. A relevant component of these models are networks loading models allowing to calculate link flows from path flows. Existing networks loading models can be divided into aggregate and disaggregate models (microsimulation models). The latter ones allow the car-following and then a precise calculation of traffic parameters but they need considerable computing resources. In this paper, a mesosimulation model has been developed to study the flows propagation on the network. Since the proposed model is disaggregate as for flow characteristics and aggregate as for links performances, it does not need great computing resources in calculating vehicular speed and acceleration. Therefore, utility of this model entirely appears when dealing with simulation of traffic pollution.

Keywords Traffic pollution, network loading models

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 187–201.

1. INTRODUCTION

Pollution phenomena have become a relevant issue of transportation studies, since in major urban areas traffic congestion has led to an often unsustainable pollution level. Therefore, optimisation of traffic flows closely concerns minimisation of traffic pollution.

Usually the models for simulation of traffic pollution are made up by (fig. 1):

a demand sub-model;
a supply sub-model;
a demand/supply interaction sub-model, or traffic assignment model;
a flow propagation sub-model;
a pollutants emission and diffusion sub-model.

The first four models deal more closely with the study of transportation

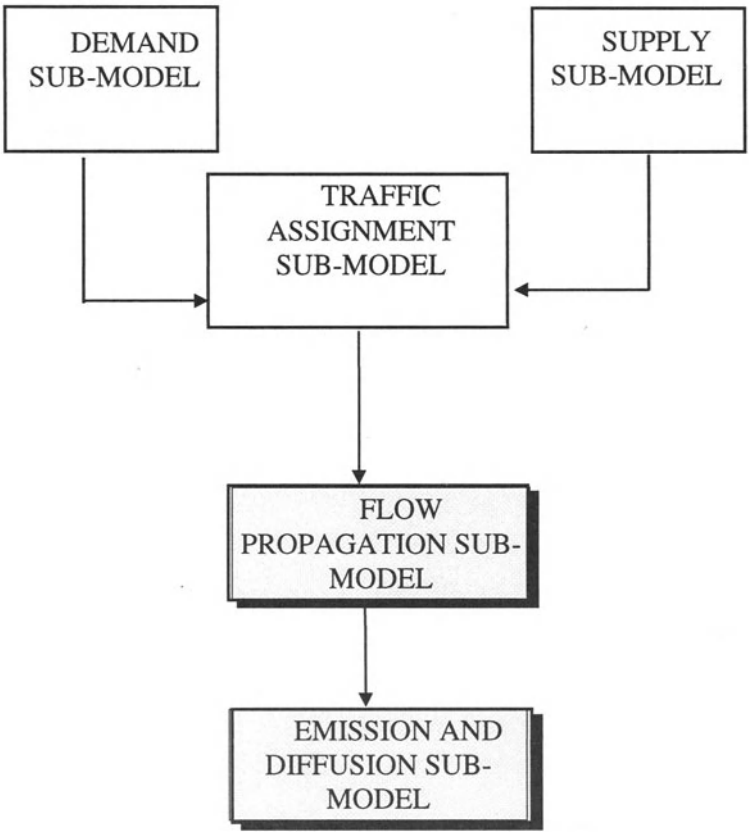

Fig. 1 – Models for simulation of traffic pollution

systems. This paper focuses on flow propagation models which are a relevant component of traffic assignment models, allowing to calculate link flows from path flows. In particular a model which takes into account the speed variability over time has been developed. Through this model it is possible to calculate instantaneous values of vehicle speed and acceleration and then more realistic values of pollutants emission.

2. FLOW PROPAGATION MODELS

Depending on the hypotheses on demand in the reference period, flow propagation models can be divided into static and dynamic models. Static models are specified through a linear relation between link flows and path flows such as:

$$f_a = \Sigma_k \delta_{ak} h_k$$

where
f_a is the flow on link a;
$\delta_{ak} = 1$ if link a belongs to path k, 0 otherwise;
h_k is the flow on path k.

Dynamic flow propagation models are defined through non-linear relations which are generally hard to manipulate. Nevertheless, they give more realistic speed values and then they appear more right for studies of pollution phenomena. At present dynamic models are still closely investigated. Different approaches are followed in literature depending on whether the link performances are expressed in aggregate or disaggregate way (fig. 2).

	AGGREGATE MODELS	DISAGGREGATE MODELS
IMPLICIT VEHICLES TRACING	**MACROSIMULATION**	
EXPLICIT VEHICLES TRACING	**MESOSIMULATION**	**MICROSIMULATION**

Fig. 2 - Approaches to flow propagation

Disaggregate models (microsimulation models) are very sophisticated ones. They can describe the single vehicle movements, even overtaking or

parking manoeuvres, but require considerable computing resources. Moreover, the level of their results is right for projects but it is too disaggregate for planning purposes, which need more aggregate results. Many microsimulation models have been proposed in literature to study in particular junctions and control strategies (HUTSIM, INTEGRATION, NETSIM, SIMIR, SIMNET, etc...).

Aggregate models can be divided, in turn, depending on whether vehicles movements are implicitly or explicitly traced. In the former case macrosimulation models are involved. They consider the traffic as a continuous fluid subject to congruence and continuity of flow constraints (Lighthill and Whitham, 1955; Isaksen and Payne, 1972; Rathi, Lieberman and Yedlin, 1987; Ross, 1988). More recently Daganzo (1994) has proposed another space-continuous macrosimulation model. This model describes by means of "cell transmission" the dynamic evolution of traffic over a freeway network, possibly having three-legged junctions. The description is consistent with the hydrodynamic analogy.

Other approaches to flow propagation are based on space-discrete macrosimulation models (Merchant and Nemhauser, 1978; Carey, 1987; Wie et al., 1994). Two distinct formulations are allowable:

exit link function formulation;

travel time formulation.

In the first formulation exit flow w is assumed depending on the number n of vehicles on the link, whatever are their positions:

$$w = W(n)$$

In particular Merchant and Nemhauser proposed the following state equation:

$$n_{i+1} = n_i + (u_i - W(n_i))\cdot\Delta t$$

where

i denotes the individual period (i = 0, 1, ...I);

n_i is the number of vehicle on the link within the period i (i = 0, 1, ...I);

u_i is the flow entering the arc in the period i;

$W(n_i)$ is the exiting flow.

This model does not guarantee explicitly the respect of FIFO rule and has some drawbacks which lead to unrealistic results. In fact at the beginning of simulation the outflow has immediately a positive value, so the corresponding travel time is zero. Moreover the last traveller will never reach the end of the link (see example in the comparison section).

In the second formulation travel time $\tau(t)$ is assumed depending on the number n(t) of vehicles on the link at time t

$$\tau(t) = T(n(t))$$

and assigned to each vehicle entering the link at time t.

A linear formulation for T(n(t)) has been suggested by Friesz et al. (1993). Also for this model the respect of FIFO rule is not guaranteed; in some cases it has been ensured through a set of additional constraints (Janson, 1995) or by means of a particular travel time formulation (Astarita, 1996. In this paper, a detailed analysis of Dynamic Network Loading models is provided).

When vehicles movements are explicitly traced the case deals with mesosimulation models. Such models consider the traffic neither a continuous fluid (as in macrosimulation models) nor single vehicles (as in microsimulation models) but a sequence of "packets" of vehicles. A packet is a set of vehicles leaving at the same time and following the same path. Then *a priori* knowledge of vehicles paths is needed. Since packet size is arbitrary, such models allow also to simulate a single vehicle movement.

If vehicles are uniformly distributed inside each packet the packets are "continuous" (see for example Di Gangi 1992). Otherwise, if all vehicles belonging to a packet are grouped and represented by the head of the packet itself, the packets are "point-packets" (Leonard and Gower, 1989; Cascetta, Cantarella and Di Gangi, 1991).

Regarding continuous packets models it has already been demonstrated that the hypothesis of uniform distribution (in space or in time) of vehicles inside the packet leads to results which have internal inconsistencies (Dell'Orco, 1997). To avoid them, the effective distributions of vehicles within the packets should be known. However, the specification of these distributions would produce some functions hard to be handled.

As for point-packets, the hypothesis that all vehicles belonging to a packet are represented by the head of the packet could influence the external consistency, that is the capacity of the model to represent the phenomenon correctly. However, this influence is reduced as time interval or link length approaches to 0. In the following a point-packet model will be described and applied to study the flow propagation on a test network.

3. HYPOTHESES AND DEFINITIONS

Let K be the set of feasible paths on a network. A "packet" (j, k) is the set of vehicles leaving at the same time and following the same path $k \in K$. Packets are "point-packets" and their movement will be studied for discrete time intervals $[t-\Delta t, t]$, $[t, t+\Delta t]$.

Speed will be assumed equal for all packets running at the same time on the same link ($\partial V/\partial s = 0$).

Moreover, let:

$m_{j,k}$ be the total number of vehicles belonging to the packet (j, k);

V_i^t be the speed on the link i at time t, common for all vehicles on the link;

a_i^t be the acceleration during the interval$[t, t+\Delta t]$, constant and common for all vehicles on the link i;

$n_{j,k}^{t,i}$ be the number of vehicles belonging to the (j, k) on the link i at time t;

d_i be the lenght of link i;

$s_{j,k}^{t,i}$ be the position on the link i at time t of the head of the packet (j, k); the value of $s_{j,k}^{t,i}$ is 0 if the packet (j, k) is not on the link i;

N_i^t be the total number of vehicles on the link i at time t.

It is obvious that:

$$N_i^t = \sum_{k \in K} \sum_{j \le t} n_{j,k}^{t,i} \tag{1}$$

4. THE PROPOSED MODEL

According to the logical sequence of fig. 1 the proposed model is downstream the assignment model and upstream the emission and diffusion model. It gets the path flows from the former and gives link flows, speed and acceleration to the latter. Actually the aim of this paper is to study the flow propagation on the networks and not traffic assignment. Therefore, calling K_{OD} the set of feasible paths between the origin O and the destination D, the complete knowledge of the paths (link sequence and path flows) $k \in K_{OD}$ $\forall O,D$ is required.

The model will be explained at first for a single link and then it will be extended to the flow propagation on a network.

The proposed model is a mesosimulation model with point-packets whose relevant characteristic is that speed is assumed equal for all vehicles on the same link, but variable over time. All vehicles belonging to the same packet are assumed to occupy on the link the same position of the head of the packet (fig. 3).

Because of these assumptions the speed at time t on the link i depends on number of vehicles on the same link i at the same time t:

$$V_i^t = V\left(N_i^t\right). \tag{2}$$

This means that the speed of each packet is influenced by what happens *behind* it. The arrival of a packet on a link increases the number of vehicles present on the link, which in turn decreases the speed of all packets present

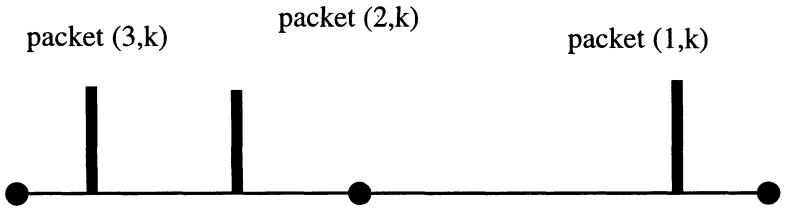

Fig. 3 – Point-packets sequence

on the link.

According to the hypotheses $n_{j,k}^{t,i}$ depends only on $s_{j,k}^{t,i}$:

$$n_{j,k}^{t,i} = \begin{cases} 0 & \text{if} \quad s_{j,k}^{t,i} = 0 \quad \text{or} \quad s_{j,k}^{t,i} > d_i \\ m_{j,k} & \text{if} \quad s_{j,k}^{t,i} > 0 \quad \text{and} \quad s_{j,k}^{t,i} < d_i \end{cases} \tag{3}$$

Therefore, for a single link the model is:

$$a_i^t = \left(V_i^{t+\Delta t} - V_i^t \right) / \Delta t$$

$$s_{j,k}^{t+\Delta t,i} = s_{j,k}^{t,i} + V_i^t \Delta t + \tfrac{1}{2} a_i^t \Delta t^2$$

$$n_{j,k}^{t,i} = \begin{cases} 0 & \text{if} \quad s_{j,k}^{t,i} = 0 \quad \text{or} \quad s_{j,k}^{t,i} > d_i \\ m_{j,k} & \text{if} \quad s_{j,k}^{t,i} > 0 \quad \text{and} \quad s_{j,k}^{t,i} < d_i \end{cases} \tag{4}$$

$$N_i^{t+\Delta t} = \sum_{k \in K} \sum_{j \le t} n_{j,k}^{t+\Delta t,i}$$

$$V_i^{t+\Delta t} = V(N_i^{t+\Delta t})$$

The model (4) is a fixed-point problem which has been resolved through the Method of Successive Average (MSA). Let:

$$s_{j,k}^{t+\Delta t,i} = s(V_i^{t+\Delta t});\ n_{j,k}^{t,i} = n(s_{j,k}^{t,i});\ N_i^{t+\Delta t} = N(n_{j,k}^{t+\Delta t,i})$$

Thence

$$V_{i,m+1}^{t+\Delta t} = 1/m \cdot V(N(n(s(V_{i,m}^{t+\Delta t})))) + (m-1)/m \cdot V_{i,m}^{t+\Delta t}$$

where $V_{i,m}^{t+\Delta t}$ is the value of the speed $V_i^{t+\Delta t}$ at iteration m.

To extend this model on a network, additional constraints are needed such as:

flow conservation in the nodes. Assuming a generic node r the sum u_r of vehicles entering the node must be equal to the sum w_r of vehicles leaving the same node:

$$u_r = w_r$$

$$u_r = \sum_{i \in BW_r} n_{j,k}^{t,i} \Big| s_{j,k}^{t,i} > d_i \tag{5}$$

$$w_r = \sum_{i \in FW_r} n_{j,k}^{t,i} \Big| s_{j,k}^{t-\Delta t,i} = 0, s_{j,k}^{t,i} > 0$$

where BW_r and FW_r are, respectively, the set of links terminating (backward star) and originating (forward star) at node r;

interaction among converging flows. Let Q_{BWr} be the sum of outflows over BW_r and C_{FWr} be the overall capacity of FW_r. Then, since the total entering flow must be not greater than capacity of downstream links:

$$Q_{BWr} \le C_{FWr} \tag{6}$$

The competition among different flows entering a junction is resolved through the supply division rules. The overall supply can be divided among entering links according to proportions based on geometric characteristics of the links, such as their width or number of lanes:

$$c_i = \beta_i \cdot C_{FW_r} \qquad (i \in BW_r) \tag{7}$$

where c_i and β_i are, respectively, the supply on the link i ($i \in BW_r$) and the proportion assigned to the same link. Since the β_i are proportions it is required that:

$$\sum_{i \in BW_r} \beta_i = 1 \tag{8}$$

If $c_i \geq (n_{j,k}^{t,i} \mid s_{j,k}^{t,i} > d_i)$ no additional constraints are needed; otherwise, let K_r be the subset of BW_r for which $(n_{j,k}^{t,i} \mid s_{j,k}^{t,i} > d_i) > c_i$. Then (5) turn into (Astarita, 2000):

$$u_r = w_r = \frac{\beta_i}{\sum_{a \in K_r} \beta_a} \cdot \sum_{a \in K_r} n_{j,k}^{t,a} \mid s_{j,k}^{t,a} > d_a + \sum_{b \notin K_r} n_{j,k}^{t,b} \mid s_{j,k}^{t,b} > d_b \tag{9}$$

The extended model is therefore made up by (4) and (9). Since this is not an assignment model but a flow propagation model, paths are not modifiable and are previously assigned to each packet. Consequently, packets on the same link at the same time t are labelled according to different paths and they are added up to calculate the instantaneous density of the link.

With respect to other existing mesosimulation models, this model does not average instantaneous values of speed and acceleration and therefore allows a more precise calculation of vehicles movement.

5. THE EMISSION MODEL

Several emission models exist in literature. Most of them (see for example CORINAIR model and Horowitz's model) calculate emission by mean values of speed and acceleration in a standard urban driving cycle. Through the proposed model instantaneous traffic characteristics are obtained.

Therefore, an emission model allowing to find traffic pollution as a function of this provided outcome could be used together with this model.

For classes of instantaneous speed and for classes of speed by acceleration, the MODEM model (Joumard et al., 1994) provides instantaneous emission of four pollutants (CO, HC, NO_x and CO_2) concerning vehicles with engine capacity less than 1400 cm^3 to more than 2000 cm^3. The study selected 150 vehicles to represent a valid sample of emission of the 1995 European car fleet in terms of vehicle technology, emission standards or vehicle production year and engine capacity. The considered standards are ECE R15/03 and ECE R15/04 for gasoline-engined vehicles and ECE R15/04 for controlled three-way catalyst vehicles. In this paper only CO emissions have been considered as an example. The following Table 1 (André M. and Pronello C., 1997) shows instantaneous values of emissions for vehicles without catalyst as a function of speed classes and speed·acceleration classes.

Classes of instant. speed	Classes of speed ·acceleration (instantaneous values in m/s·m/s^2)						
	<–15	–15 to –10	–10 to –5	–5 to 0	0 to 5	5 to 10	10 to 15
≤ 0				395			
0 - 10	396	397	422	469	597	777	888
10 - 20	371	373	385	529	730	958	1068
20 - 30	396	402	490	661	858	997	1178
30 - 40	415	509	601	969	875	1052	1221
40 - 50	446	505	581	675	879	1055	1276
50 - 60	466	587	685	736	857	1006	1278
60 - 70	590	756	694	698	856	940	1310
70 - 80	622	685	701	752	791	926	1476
80 - 90	845	935	847	815	797	919	1198

Table 1. - Instantaneous CO emissions (g/h) for vehicles without catalyst (ECE 15/04); engine capacities 1.4 - 2 litres

6. NUMERICAL EXAMPLE

In this section results obtained applying the proposed flow propagation model together with MODEM model have been compared with those obtained calculating emission by a model which averages speed and acceleration values of a standard urban driving cycle. The models have been applied on the simple test network shown in fig. 4. The paths, that is the sequences of links and their lengths, are in Table 2. Their *a priori* knowledge

is required since the proposed model is applied apart from search of feasible paths. Because of simplicity of test network, the set K of feasible paths can be found immediately without search algorithms. Namely, the paths are:

$k_1 = \{1\text{-}4, 4\text{-}3, 3\text{-}5\}$

$k_2 = \{1\text{-}2, 2\text{-}5\}$

$k_3 = \{1\text{-}2, 2\text{-}3, 3\text{-}5\}$

$k_4 = \{5\text{-}1\}$

In the table 2 classical path-links incidence matrix has been slightly modified to take into account the arrangement of links into paths. This expedient, together with a labelling system, allows the splitting of exiting packets among links of forward star. Origin and destination are the nodes 1 and 5.

				PATHS			
	N.		d [m]	k_1	k_2	k_3	k_4
LINKS	1	**1-2**	119	0	1	1	0
	2	**1-4**	147	1	0	0	0
	3	**2-1**	158	0	0	0	0
	4	**2-3**	179	0	0	2	0
	5	**2-5**	123	0	2	0	0
	6	**3-5**	112	3	0	3	0
	7	**4-3**	178	2	0	0	0
	8	**5-1**	155	0	0	0	1
	9	**5-3**	158	0	0	0	0
	10	**5-4**	149	0	0	0	0

Table 2. – Path–Links Incidence Matrix

In the reference period (1 hour) sinusoidal flows with maximum amplitude of 1800 veh/h run from origin to destination nodes along paths having capacity in between 2700 and 5000 veh/h. Time interval Δt has been assumed equal to 8 s.

In the following figures there are some results concerning outflows on links belonging to two significant paths.

It is worth noting that the flow on link 3-5 is the greatest one, since this link

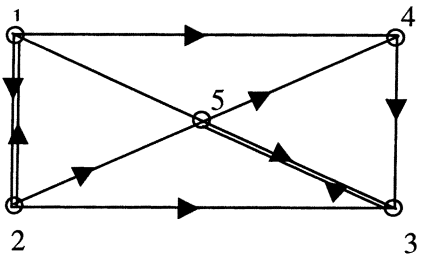

Fig. 4 – Test network

belongs to both path k_1 and path k_3.

Using MODEM emission model, CO instantaneous emissions on the network have been calculated. They have been then compared with those obtained using a model (Horowitz, 1982) with mean rates of emissions for different kinds of movement (acceleration, deceleration, uniform movement and queue). Fig. 7 shows the results of comparison for the link 3-5. It is not surprising that values of emission obtained through Horowitz's model are lower since averaging instantaneous values of emission reduces the effects of congestion.

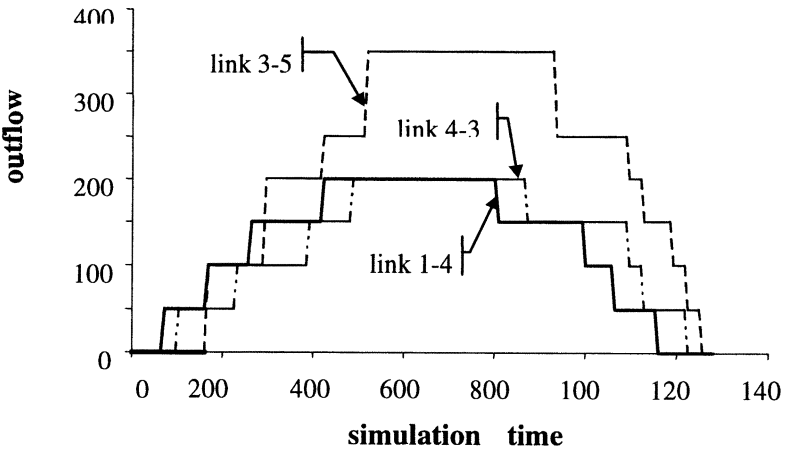

Fig. 5 - outflows on path k_1

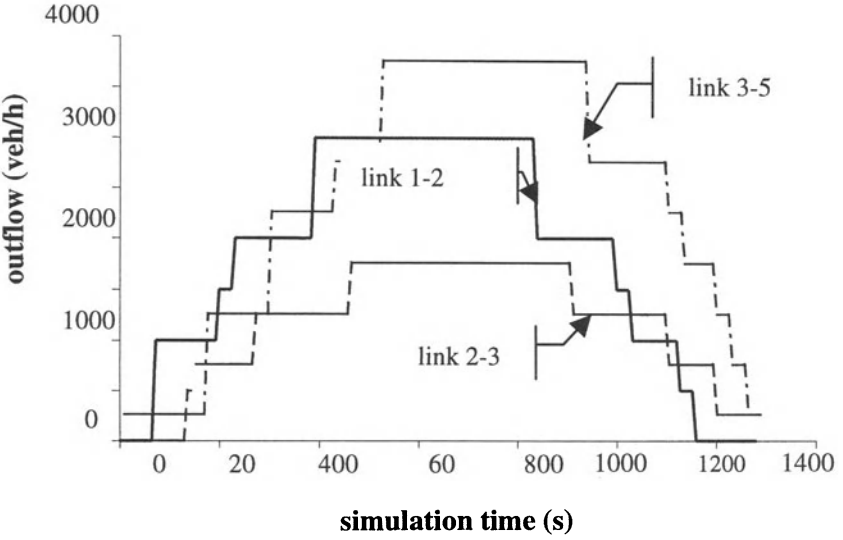

Fig. 6 - outflows on path k_3

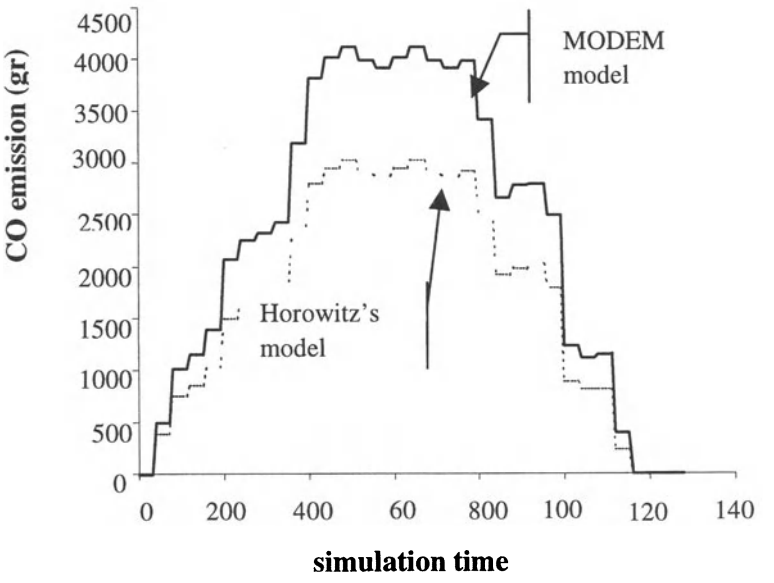

Fig. 7 - Comparison between MODEM and Horowitz's models

7. CONCLUSIONS

Calculating pollutant emissions for different traffic conditions is a relevant issue in traffic pollution studies. Usually, existing pollution models use mean values of pollutant emission for different types of movement, although instantaneous values are available as a function of speed and acceleration. This is because Dynamic Network Loading models do not provide with vehicles acceleration (macrosimulation models) or need great computational resources (microsimulation models) to do it.

In this paper a mesosimulation model has been proposed with hypotheses of discrete packets and accelerated movement of vehicles. The model needs small computing resources and provides with results which can be effectively used in calculating instantaneous emission of pollutants.

Comparing these results with those obtained using a model which has mean rates of emissions for different kinds of movement (acceleration, deceleration, uniform movement and queue), it is evident that more realistic values of emission are obtained through proposed model, since averaging instantaneous values of emission the effects of congestion are reduced.

References

André M., Pronello C. (1997), «Relative influence of acceleration and speed on emissions under actual driving conditions», *International Journal of Vehicle Design* 18 No.3/4, 340-353

Astarita V., (1996), «A continuous time link model for dynamic network loading based on travel time functions», *Transportation and Traffic Theory* (ed. Lesort), Pergamon, Oxford 79-102

Astarita V., (2000), "La modellizzazione delle intersezioni nei modelli dinamici di simulazione del traffico" in "Metodi e Tecnologie dell'Ingegneria dei Trasporti" – (G. E. Cantarella and F. Russo Eds.) - Laboratorio di Analisi dei Sistemi di Trasporto – Reggio Calabria - Collana Trasporti Franco Angeli

Carey M. (1987), «Optimal time-varying flows on congested networks», *Operations Research* 35 No.5, 58-69

Cascetta E., Cantarella G.E., Di Gangi M. (1991), «Evaluation of Control Strategies through a Doubly Dynamic Assignment Model», *Transportation Research Rec.* N.1306

Cascetta E., Cantarella G.E. (1993), «Modelling dynamics in transportation networks: State of the art and future developments», *Simulation Practice and Theory,* 65-91

Dell'Orco M. (1997), *Sviluppo di un modello di mesosimulazione per il caricamento dinamico delle reti* - 6° Convegno SIDT - Bologna.

Di Gangi M. (1992), *Continuous-flow approach in dynamic network loading*, Second International CAPRI seminar on Urban Traffic Networks, Compendium Vol.I

Horowitz J. (1982), *Air quality analysis for urban transportation planning,* MIT Press

Joumard R., Jost P., Hickman A.J., Hassel D. (1994), *Hot passenger car emission modelling as a function of instantaneous speed and acceleration,* 3rd International Symposium Transport and air pollution, INRETS (preprints).

Isaksen L., Payne H.J. (1972), *Simulation of freeway traffic control system,* Simulation Council Proc. 2, 35-42. Simulation Councils, La Jolla , California

Leonard D.R., Gower P., Taylor N.B. (1989), «CONTRAM: Structure of the model», *TRRL Research Report 178,* Crowtorne

Lighthill M.J., Whitham G.B. (1964), «An Introduction to Traffic Flow Theory», *H.R.B. Special Report 79,* Washington, D.C., 8-35

Merchant D.K., Nemhauser G.L. (1978), «A model and an algorithm for the dynamic traffic assignment problems», *Transportation Science* 12 (3), 183-207

Rathi A. K., Lieberman E. B., Yedlin M. (1987), «Enhanced FREFLO program: Modelling of congested environments.», *Transportation Research Record* 1112, 61-71

Ross P. (1988), «Traffic dynamics», *Transportation Research* 22B N.6

Wie B.W., Tobin R.L., Friesz T.L. (1994), «The Augmented Lagrangian Method for Solving Dynamic Network Traffic Assignment Models in Discrete Time», *Transportation Science* Vol. 28 N.3, August 1994

Chapter 13

STATED PREFERENCE STUDY OF MODE CHOICE IN THE HELSINKI METROPOLITAN AREA

Jari Kurri
Helsinki University of Technology, Transportation Engineering
P.O.Box 2100, FIN-02015 HUT, Finland
jari.kurri@hut.fi

Juha Mikola†
Helsinki University of Technology, Transportation Engineering

Nina Karasmaa
Helsinki University of Technology, Transportation Engineering

Abstract The main purpose of the study was to prove stated preference data for the updating of the travel demand models for the Helsinki metropolitan area. The second goal of the research was to estimate value of time and other trade-offs related to the preferences of the people living in the area. The study showed that carefully designed and tailored mail-back questionnaires can be used as a source of data. Nevertheless, the combined use of revealed and stated preference information turned out to be quite difficult since it is nearly impossible to define the variables of stated preference exercise in the same way as the variables used in mode choice models based on actual choices. Therefore more attention should be paid to the quality of data gathered in mobility surveys.

*Paper presented at the 7th EURO Working group meeting on transportation, 2–3 August 1999, Helsinki, Finland, and submitted for review and publication

†Now Nokia Research Center

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 203–223.

Keywords: Travel demand models, logit models, model transferability, value of time

1. INTRODUCTION

The purpose of the study [1] was to give basic information for the updating of the Helsinki metropolitan area travel demand models, and to collect new evidence about the preferences of the people living in the area. The idea with updating existing models or transferring previously estimated models to a new application context is to reduce or eliminate the need for a large data collection and model development effort in the application context. However, the usefulness of a transferred model depends on the degree to which it can provide valid information about the behaviour or phenomenon of interest in the application context. In the present study we examined the validity of stated preference (SP) data, or information about hypothetical choices, for the updating of mode choice models that have been estimated with information about actual choices, or revealed preference (RP) data.

The present research is also one step in the series of stated preference studies carried out or analyzed by Helsinki University of Technology, and partly financed by Helsinki Metropolitan Area Council YTV, which is responsible for the maintenance of the model system in the area. Although primary interest has been in the applications, these studies have provided an excellent way to get into practical and methodological questions and problems with SP methods. In addition, we have used SP techniques successfully for the estimation of the value of travel time savings that can be derived from the parameters of mode choice models as well as from route or abstract mode choice models. Besides questions related to transferability of models and estimation of value of time, we concentrate here on some other factors as well that have an effect on mode choice.

The Helsinki metropolitan area consists of four cities: Helsinki (546,000 inhabitants), Espoo (205,000 inhabitants), Vantaa (174,000 inhabitants), and Kauniainen (8,000 inhabitants). The city centre of Helsinki is located in a peninsula in the Gulf of Finland, and the metropolitan area forms a half circle around it with a radius of 25 to 30 kilometres (land area 764 km^2). In the city centre (peninsula, 12 km^2) there are about 100,000 workplaces and 65,000 inhabitants. The total number of workplaces in the metropolitan area is about 480,000.

The public transport system in the metropolitan area consists of bus and tram traffic, three railway lines for commuter trains, and one metro line. Buses dominate in public transport. About 60% of all public transport trips are made by bus. Of the 2.5 million daily trips made by the inhabitants of the area, about 47% are made by car, 28% by public trans-

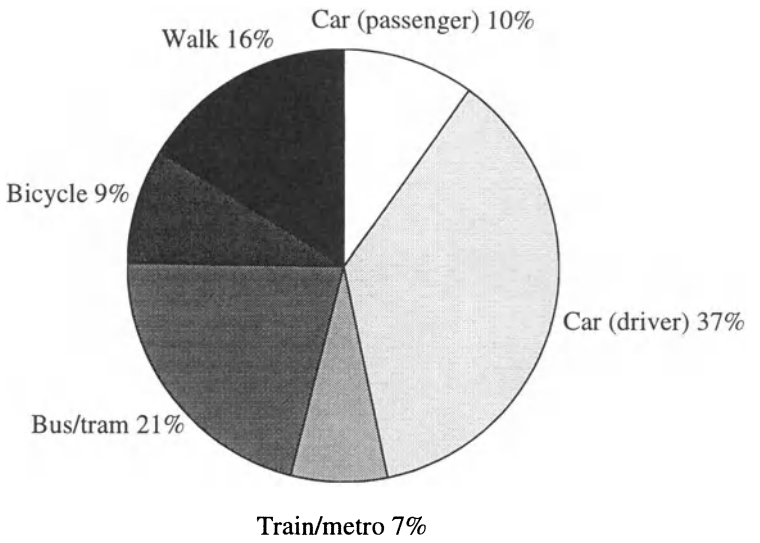

Figure 13.1 Modal split in the Helsinki metropolitan area in 1995

port, and 25% by bicycle or on foot (Figure 13.1). The share of public transport decreased continuously during the 1970's and 1980's as a result of growing car ownership. During the first part of the 1990's, however, the average car ownership decreased due to a deep economic depression in Finland. The present car ownership is about 345 cars per 1,000 inhabitants, and about 60 per cent of all households have at least one car. Nevertheless, over 60% of the trips to or from the city centre are made by public transport.

2. METHODS

2.1 STATED PREFERENCE DATA

Stated preference (SP) techniques refer to a number of different approaches all of which use people's statements of how they would respond to different hypothetical situations presented to them (people state their preferences) [2]. There are a number of different ways to collect stated response data, including ranking, choice and rating. Choice data are the most natural as they are stated response counterpart to revealed preference data. In the present study, a certain kind of choice or ranking method was used. In this method, two or more mode alternatives are presented to the respondent at a time, and the respondent is asked to set the alternatives in order of preference. The ranking data can be transformed into expanded choice data by assuming that the first alternative

is chosen over the second and third best, and that the second best is chosen over the third best. Nevertheless, we utilized only the first choice. If there are only two alternatives, ranking and choice methods give of course the same information.

Unlike traditional revealed preference (RP) data, SP data do not describe actual behaviour. Therefore, the consistency of SP data with actual market behaviour is always questionable. On the other hand, RP data have many deficiencies that SP data do not have. For instance, measurement of explanatory variables for RP models often involves many simplifying assumptions such as spatial and temporal aggregation whereas in SP data the values perceived by the respondent equal the values presented to the respondent and used in modelling. Moreover, by choosing appropriate experimental design, the researcher can avoid some of the typical drawbacks of RP data: (multi)collinearity (two or more variables are linearly correlated so that it is not possible to isolate their effects), and lack of variance (in reality travel times, for example, may not vary sufficiently).

In order to minimise various biases inherent in SP exercise, it is important that the hypothetical situations presented to the respondents are as realistic as possible and that the attribute values relate to respondents' present experience. We used data which were gathered in two stages between 1995 and 1997. At the first stage in the autumn of 1995, over 3,000 persons sampled from census data were interviewed by telephone. The questions concerned socio-economic information and trips made during the interview day (Monday to Friday). The respondents were also asked if they are willing to take part in a follow-up study, i.e., the present stated preference study.

At the second stage, between March and August of 1997, a questionnaire including stated preference questions was sent by post to those who had told that they are willing to take part in the follow-up study. The questionnaires were returned by post. Travel time and cost and other variables included in the stated preference exercise were tailored to one of the trips the respondent reported at the first stage. Furthermore, the names and types of origin and destination as well as departure time of the selected trip were printed out on the first page of the questionnaire. Nevertheless, as there were one and a half years between the first and second stages, the hypothetical choice situation may have been too abstract for many respondents. Although it was almost impossible for the respondents to recall the trip, this should not have been a problem for regular trips between home and work, or other trips made frequently. Unfortunately the respondents were not asked how often they made the

trip. The second-stage questionnaire was sent to 1,471 persons, of which 888 (60%) returned it.

The SP experiment was arranged in two parts. The first part was common to all the respondents and dealt with general factors that have an influence on mode choice. The second part, however, was totally different for car drivers, public transport users, and the rest of the respondents. Thus there were actually four different SP exercises. The second part of the questionnaire for present car drivers was a part of an EU 4th framework research (DG VII Transport) project named Transprice where a common SP survey was carried out in eight European cities ([3],[4]). The second part for present public transport users dealt with different kinds of transfers, and the effects of availability of seat and low-floor buses. The present paper deals with this second part for public transport users as well as the first, common part.

2.2 MODE ALTERNATIVES AND VARIABLES

In SP studies, hypothetical alternatives are characterised by various levels of attributes presented to the respondents. Table 13.1 shows the attributes (variables) and the levels of the attributes used in the first (common) part of the study. There were five mode alternatives, of which two or three were presented to each respondent. The mode alternatives studied were car (as driver), train or metro, bus or tram, bicycle, and walk. The set of available mode alternatives was generated separately for each trip and each respondent, although some simple rules were also utilized. For example, the walk alternative was considered unavailable if the walking distance was more than about five kilometres. The mode that was actually chosen and reported in the first stage of the study was of course always included in the set of available alternatives.

Ten choice tasks between the selected mode alternatives were presented to each respondent. The respondents were asked to set the alternatives in order of preference, or to choose the best alternative if there were only two alternatives. The levels of the attributes characterising the alternatives were varied according to an experimental design. Absolute values were shown although most of the values were calculated as percentage change with respect to the base values. The base values equalled or were at least close to the present values.

Public transport trips often include several stages with different vehicles. In the present SP study, the maximum number of stages was three, that is, up to two transfers were possible. Moreover, the bus (or tram) mode did not, by definition, include any stages by metro or train.

This means that the possible combinations of the vehicles for the bus alternative were

BUS|TRAM[+BUS|TRAM[+BUS|TRAM]]

where | denotes 'or' and [·] means that the mode in square brackets is optional. The rail mode (train or metro), on the other hand, could include all public transport modes: train, metro, bus and tram. Of course, at least one stage of the trip was by train or metro. The possible combinations of the vehicles for the train mode were, therefore,

TRAIN|METRO[+BUS|TRAM|TRAIN|METRO[+BUS|TRAM|TRAIN|METRO]]

BUS|TRAM+TRAIN|METRO[+BUS|TRAM|TRAIN|METRO]

BUS|TRAM+BUS|TRAM+TRAIN|METRO.

The number of transfers and succeeding stages and vehicles were chosen separately for each trip and each respondent.

2.3 EXPERIMENTAL DESIGN

The experimental design of the common part of the study was not based on any particular statistical procedure such as fractional factorial design. Instead, the level of each variable was first chosen randomly, based on the levels presented in Table 13.1, and then quite complex tests were performed to make sure that the question provide new information about the preferences of the respondent. For example, dominated alternatives were excluded, so that the exercise would not be too boring for the respondents. This was not as straightforward as it sounds. For instance, in-vehicle time for bus cannot be directly compared with in-vehicle time for car. Nevertheless, actual choice can be used to resolve this problem. There were 3 to 11 variables, depending on which alternatives were presented to the respondent.

The experimental design of the second part of the study, directed to present public transport users, was based on fractional factorial design. There were two questions comprising of four alternatives with three attributes. Both questions included total travel time and travel cost as attributes. The third attribute was the availability of a seat or the type of a bus (low-floor or normal), respectively.

3. MODELLING

Multinomial logit models were used to analyze the responses (for more information on discrete choice models and standard logit mode choice models with linear-in-parameters utility functions, see e.g. [5], or [6].

Table 13.1 The attributes and the levels of the attributes for the first part of the study

Alternative Variable	Number of levels	Levels (percentage change or absolute value)
Car (driver)		
In-vehicle time (min)	3	−10%, 0%, +10%
Driving cost (FIM)	3	−20%, 0%, +20% 0%, +20%, +40%
Parking charge (FIM)	2	usually 0 and 4 (FIM)
Local train or metro Bus or tram		
In-vehicle time (min)	3	−20%, 0%, +20%
Walking time (min)	2 (1)	−20% and/or +20%
Headway(s) (min)	2	present, half/double
Fare (FIM)	3	based on single and seasonal ticket
Bicycle		
Cycling distance (km)	2	0%, +10%
Road conditions for cycling	2	"along bicycle paths", "on roadways"
Walk		
Walking distance (km)	1	no change (present value used)

Unlike classical regression analysis, qualitative or discrete choice models are suitable for cases where the dependent variable is not continuous and quantitative but can have only a finite number of different qualitative values. Furthermore, random utility models such as logit model make use of an abstract construct called utility which represents the relative attractiveness of each alternative. The higher the utility U_{in} of alternative i and observation n is, the higher is the probability $P_n(i)$ to choose that alternative. The utility is assumed to consist of two additive parts, a deterministic or measurable part V_{in}, and a stochastic part ϵ_{in} ($U_{in} = V_{in} + \epsilon_{in}$). The choice probabilities are, according to the principle of utility maximisation,

$$\begin{aligned} P_n(i) &= \Pr\left(U_{in} \geq U_{jn}, \forall j \neq i, j = 1, \ldots, m\right) = \\ &= \Pr\left(V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}, \forall j \neq i, j = 1, \ldots, m\right), \end{aligned}$$

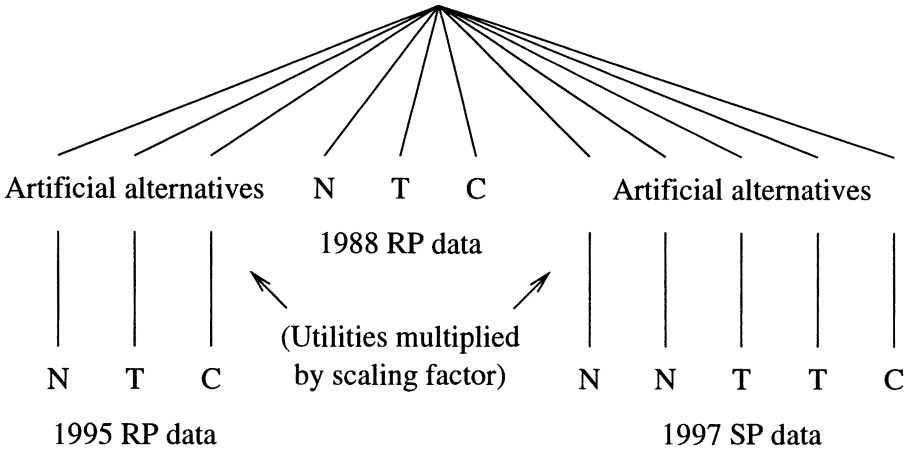

Figure 13.2 An example of the artificial tree structure used in the combination of different data sets. N stands for non-motorised modes walk and bicycle, T for public transport, and C for car

where m is the number of alternatives. If all ϵ_{in} across all alternatives i and all observations n are identically and independently distributed according to an extreme value distribution named after Gumbel, it can be proved that the probability to choose alternative i is

$$P_n(i) = \frac{e^{V_{in}}}{\sum_{j=1}^{n} e^{V_{jn}}} \quad (i = 1, \ldots, m).$$

It should be noted that stated preference data include several choices from each respondent, and therefore the stochastic terms are not actually independent.

The utility functions mostly used are linear with respect to parameters to be estimated:

$$V_{in} = \beta_1 x_{in1} + \beta_2 x_{in2} + \cdots + \beta_k x_{ink},$$

where β_j is the coefficient for variable x_j, x_{inj} is the value of variable x_j for alternative i and observation n, and k is the number of variables in the model. If the utility of alternative i does not include coefficient β_j, we can assume that $x_{inj} = 0$. There is no constant term but, instead, the utility functions for all alternatives except one include so called alternative specific constants.

3.1 COMBINATION OF DATA SETS

The present travel demand models in the Helsinki metropolitan area are based on revealed preference data that were gathered in the autumn

of 1995 and that form the basis of the first stage of the present stated preference study. The data consist of trips made by 3,082 persons that were interviewed by telephone. The sample size was quite small, and therefore attempts have been made to utilize the data and models from similar mobility survey in 1988 [7] (6,733 interviews) as well as the data of the present stated preference study. The motive for the use of additional information was simple: it is quite expensive to gather large amount of data that are needed to estimate the models.

There are different ways to make use of data from different time (temporal transferability) or from different source (e.g., revealed preference vs. stated preference information). The simplest approach would of course be to combine the data without any scaling (perhaps corrected for inflation etc.), or even to use models that are estimated with data from other context. Less naive methods include various scaling approaches as well as Bayesian updating where only the estimates of the parameters and estimates of the covariance matrices of the estimators of the parameters are needed. These methods have been tested with the revealed preference data gathered in the Helsinki metropolitan area in 1981 and 1988 [8], and in 1988 and 1995 [9]. In the present study we combined three data sets (1988 RP, 1995 RP and 1997 SP) according to a method that recognizes the fact that the data from different sources may include random variation of different extent and therefore the variances of the utility functions may differ. The basic idea of the method used [10] is to scale the utility functions of the data sets in such a way that the variances are equal. The purpose of this procedure is merely to combine different data sets and hence provide new information of the preferences that are implicitly assumed homogenous. Therefore, this method is perhaps not the best choice if preferences are assumed to change over time. Nevertheless, in the present study we assumed that the changes in preferences are so small that 1988 RP data could be used. Furthermore, this data set acted as a base, and the role of 1995 RP and 1997 SP data were to provide new information for the updating of the models estimated with the 1988 RP data.

Let us suppose that there are two data sets to be combined, one consisting of revealed preference information and the other of stated preference information. The method can easily be generalized to more than two dimensions. Let U^{RP} and U^{SP} be the utility functions for revealed and stated preference data, respectively. The utility functions are of the following form:

$$\begin{aligned} U^{RP} &= \beta \mathbf{x}^{RP} + \alpha \mathbf{w}^{RP} + \epsilon^{RP} \\ U^{SP} &= \beta \mathbf{x}^{SP} + \gamma \mathbf{z}^{SP} + \epsilon^{SP} \end{aligned}$$

where vector $\mathbf{x}$ includes those variables that are in common for both data sets, and $\mathbf{w}$ and $\mathbf{z}$ the variables that exist only in the RP or SP data, respectively. The unknown coefficients of the utility functions are included in β, α, and γ, and ϵ denotes the unknown Gumbel-distributed random terms.

In order to make the variances of ϵ^{RP} (or U^{RP}) and ϵ^{SP} (U^{SP}) equal, the SP utilities are multiplied by an unknown correction factor μ defined by

$$\mu^2 = \frac{\text{Var}\,(\epsilon^{RP})}{\text{Var}\,(\epsilon^{SP})}.$$

If the random components ϵ^{RP} and ϵ^{SP} follow Gumbel distribution and are identically and independently distributed (IID) then the combined data set also has an IID Gumbel distributed random noise. The final SP utility function is then

$$U^{SP} = \mu(\beta \mathbf{x}^{SP} + \gamma \mathbf{z}^{SP} + \epsilon^{SP}).$$

The multiplication of the SP utility function with the unknown variance factor μ makes the utility function non-linear in parameters. Therefore, maximum likelihood estimation of the model parameters with standard logit model estimation programs cannot be used directly. Nevertheless, at least Alogit software (by Hague Consulting Group) can be used, provided that a special kind of an artificial tree structure is generated [10]. The tree structure used in the present study is shown in Figure 13.2. The alternatives of the base context (1988 RP data) are given as such, whereas each mode alternative of the 1995 RP and 1997 SP data is given structured below a dummy alternative. The dummy alternative has only one variable, the logsum of the actual alternative at the lower level. If the coefficients of the logsum variables of the dummy alternatives are forced to be equal ($\theta_{\text{RP 1988}}$ or $\theta_{\text{SP 1997}}$), then the θ values correspond to the μ factors of the two contexts providing additional mode choice information.

4. TRAVEL DEMAND MODELS

4.1 MODEL SYSTEM

The present travel demand model system in the Helsinki metropolitan area is a traditional four-stage model:

- trip generation
- distribution (choice of destination)

- modal split (choice of mode)
- assignment.

The models have been estimated with revealed preference data, such as the data that form the basis of the first stage of the present stated preference study. Trip generation "model" is currently based on simple cross-tabulation of the data. The most important categorization is the division according to the person's access to a car. The models for trip distribution and modal split, on the other hand, are multinomial logit models, or nested logit models [5] with combined mode and destination choice models. For trip assignment, or network loading, a standard multipath equilibrium assignment model (EMME/2 system) has been used. The Helsinki metropolitan area has been divided into 117 zones. There are separate models for four trip purpose groups: home-based work trips, home-based school trips, other home-based trips, and non-home-based trips. Mode choice models include three mode alternatives: car (passenger or driver), public transport, and non-motorised mode (walk or bicycle). We concentrate here on home-based work trips.

In the present study we were concerned with nested logit models of combined mode and destination choice. The present stated preference study was aimed at providing new information for the updating of these travel demand models. This aim was not fully achieved, however, mainly because of the fact that it was not possible or reasonable to define the alternatives in the same way as those used in RP models, or to use exactly the same variables. Travel forecasting in the model system is based on the use of disaggregate models with zonal means.

4.2 VARIABLES

The mode choice models include five coefficients and two alternative specific constants. Car and public transport modes have generic coefficients for total travel time and travel cost. In addition, there are two additional parameters, one for the number of transfers for public transport mode, and the other for the number of cars in household, included in the utility function of car. The non-motorised mode, consisting of walk and bicycle, has only one explanatory variable: natural logarithm of distance.

The total travel time includes walking, waiting and in-vehicle times. The time components and the number of transfers have been calculated from the results of traffic assignment with EMME/2 program. This means that the number of transfers as well as travel time are real-valued. Travel costs for public transport are average values based on the ticket types used by the persons interviewed. Travel costs for car are out-

of-pocket costs, including parking costs. Travel costs for 1988 revealed preference data were multiplied by 1.26, to take account of inflation.

4.3 MANIPULATION OF THE SP DATA

In order to use stated preference information, gathered in the present study, for the updating of the present travel demand models, a number of problems had to be solved. Firstly, there was a problem with different model structures: how to exploit SP data with mode choice information only while the model system consists of combined mode and destination choice models. Theoretically, the best way is to estimate the parameters of both models simultaneously. Nevertheless, it is also possible to estimate them sequentially, although this may result in some loss of information.

Secondly, the current travel demand models include three mode alternatives (car, public transport, and non-motorised modes walk and bicycle), and all alternatives are available for all observations. In particular, this means that people without a driving licence and living in families with no cars are assumed to be able to choose car mode, and even with the same probability as people actually having a car available. The SP data, on the other hand, are comprised of five mode alternatives, and all mode alternatives were not available for every respondent. For example, it would not be a good idea to offer a choice with car as an alternative to a person that hardly ever had the opportunity to use car, either as a driver or as a passenger. Besides, it is quite difficult to find variables that would explain why people sometimes choose taxi (very expensive but fast) or car as passenger (perhaps without paying anything). Therefore, only car as driver was included in the stated preference exercise. Moreover, this mode was assumed to be available only to those respondents that had a driving licence and a car.

The non-motorised modes walk and bicycle are actually completely distinct mode alternatives that have been aggregated for modelling purposes only. This means that the non-motorised modes should be separated in the stated preference exercise. Furthermore, it is not reasonable to regard walk as an available mode for trips that are longer than about five kilometres. Now, if both car and walk alternatives were unavailable for a particular respondent, there would be only two alternatives left, public transport and bicycle. For most of the people, however, cycling is really not an option. Therefore, to make sure that there were at least two feasible mode alternatives for each respondent, the public transport alternative was divided into two: rail (local train or metro) and bus (or tram). Still, there were areas that train and metro network do not cover.

In order to reduce the five mode alternatives of the stated preference exercise to three alternatives, the two public alternatives were combined in such a way that the utility functions for both alternatives were identical, that is, the coefficients were the same but the values of the variables were usually different. The same model specification applied to walk and bicycle, too. In addition, the stated preference exercise included an additional dummy variable that is not used in the model system, namely road conditions for cycling (along bicycle paths, or on roadways), the purpose of which was to make the choice task more interesting for the respondents. It was assumed that bicycle path is the normal case, and cycling along roadways brings additional (dis)utility.

5. RESULTS AND FINDINGS

5.1 MODEL UPDATING

Table 13.2 presents the mode choice models estimated with revealed and stated preference data. The estimates of the parameters are quite well in accordance with our expectations regarding the signs (e.g., time and cost coefficients should be negative) and relative values of the coefficients (e.g., time coefficient divided by cost coefficient, that is, value of time). In order to combine data from different sources, here revealed and stated preference information, the ratios of two coefficients, or a trade-off between two factors, should not be too different when calculated from models with different data. Unfortunately, this prerequisite was not satisfied since the value of total travel time, derived from the model with stated preference information only (SP95), is much higher than the values from the models with revealed preference information only (RP88 and RP95). It seems that the difference in values is due to the cost coefficients: the coefficients of revealed preference models suggest that people are more sensitive to changes in travel cost than could be inferred from hypothetical choices. Therefore, combining the data is perhaps not the best way to make use of the stated preference information.

In addition to value of time, Table 13.2 shows how many minutes in total travel time one transfer corresponds. Nevertheless, it should be remembered that the number of transfers was generated in different ways depending on the type of the data. As regards stated preference data, the number of transfers was always an integer between 0 and 2. As to revealed preference data, on the other hand, the number of transfers was real-valued, and the maximum value was more than 2. Therefore, stated preference data gave more negative coefficients than revealed preference data.

Table 13.2 The coefficients of mode choice models for home-based work trips (N stands for non-motorised modes, T for public transport, C for car)

Variable		Modes (NTC)	Source of data				
			RP88	RP95	SP95	RP95 + SP95	RP88 + RP95 + SP95
tot. travel time	(min)	TC	−0.0254	−0.0382	−0.0389	−0.0479	−0.0429
travel cost	(FIM)	TC	−0.1752	−0.1853	−0.0726	−0.1199	−0.1353
transfers		T	−0.3179	−0.2981	−0.4704	−0.5179	−0.3925
ln(distance)	(km)	N	−2.836	−2.847	−1.894	−2.716	−2.867
SP95 ln(dist.)[a]	(km)	N			−0.468	−0.6239	−0.6511
cars per hh		C	1.206	1.129		1.018	1.159
RP88 dummy	(0/1)	N	1.742				1.483
RP88 dummy	(0/1)	T	0				0
RP88 dummy	(0/1)	C	−1.681				−1.931
RP95 dummy	(0/1)	N		1.511		1.280	1.604
RP95 dummy	(0/1)	T		0		0	0
RP95 dummy	(0/1)	C		−1.787		−1.965	−1.928
SP95 dummy	(0/1)	N			1.742	2.808	3.306
SP95 dummy	(0/1)	T			0	0	0
SP95 dummy	(0/1)	C			−0.126	0.1930	0.5441
θ RP88							1
θ RP95						1	0.9686
θ SP95						0.7198	0.6770
Value of time (FIM/h)		TC	8.7	12.4	32.1	24.0	19.0
One transfer in min.		TC	12.5	7.8	12.1	10.8	9.2
Observations			4774	1993	3942	5938	10714
ρ^2(c)			0.2162	0.2149	0.1296	0.1533	0.1780

[a] additional parameter for cycling along roadways

Separate mode specific constants or dummies were estimated for each data set. It follows from the properties of logit models that one mode specific constant is not identifiable, and therefore the constant for public transport was fixed to zero. The car dummy for stated preference data (SP95) is much greater than the corresponding estimates for revealed preference data sets since the SP car alternative was available only for those respondents that had a driving licence and a car available. Furthermore, the models with SP data included an additional parameter for cycling along roadways, multiplied by the logarithm of cycling distance.

5.2 VALUE OF TIME AND OTHER TRADE-OFFS

5.2.1 Value of in-vehicle time. According to the present stated preference study, the value of in-vehicle time is about 38 Finnish marks (6.4 euros) per hour. This value, as well as most of the other values derived from the mode choice models of the present stated preference study, is quite well in accordance with the values from a number of other SP studies conducted in the Helsinki metropolitan area ([11], [12]). Nevertheless, these values are somewhat higher than the average value that is in use in socio-economic analysis of transport sector investments in Finland, which is about 4.7 euros per hour and per person, or about 7.4 euros per hour and per car. In Finland, travel time savings account for about 70 per cent on average of the measured benefits of road projects, and rather less than half of the benefits in investments in urban public transport. Since the share of travel time savings is already quite high, it seems unlikely that the unit value of time will be made higher, based on the results of stated preference studies. Furthermore, higher value of (in-vehicle) time would benefit car traffic more than public transport. This might be a problem as there seems to be general agreement on the fact that socio-economic cost-benefit analysis, at least in the way it is used in Finland today, does not take account of all benefits of investments in public transport or non-motorised modes. The key question here is how out-of-vehicle time, such as waiting or walking time, is treated.

5.2.2 Income effects. In addition to trip-related factors such as trip purpose and travel mode, the value of time is affected by income and other socio-economic factors. From a theoretical point of view, income effects play a very important role in socio-economic evaluation as well as in travel demand forecasting, particularly in the long run in the presence of inflation. Traditionally, travel cost variables in mode choice models have been divided by income, implicitly assuming that value of time is proportional to income. Figure 13.3 gives the value of total travel time, derived from a mode choice model with income-specific cost coefficients, as a function of household income. The ratio of value of time to income seems to be almost constant, with the exception of lower and higher income brackets. We have found the same kind of a relationship in several other studies with stated preference data, even with personal income brackets, and in-vehicle time instead of total travel time. On the other hand, models estimated with the revealed preference data, on which the present stated preference exercise is based, seem to result in values of time that have no relationship at all with income.

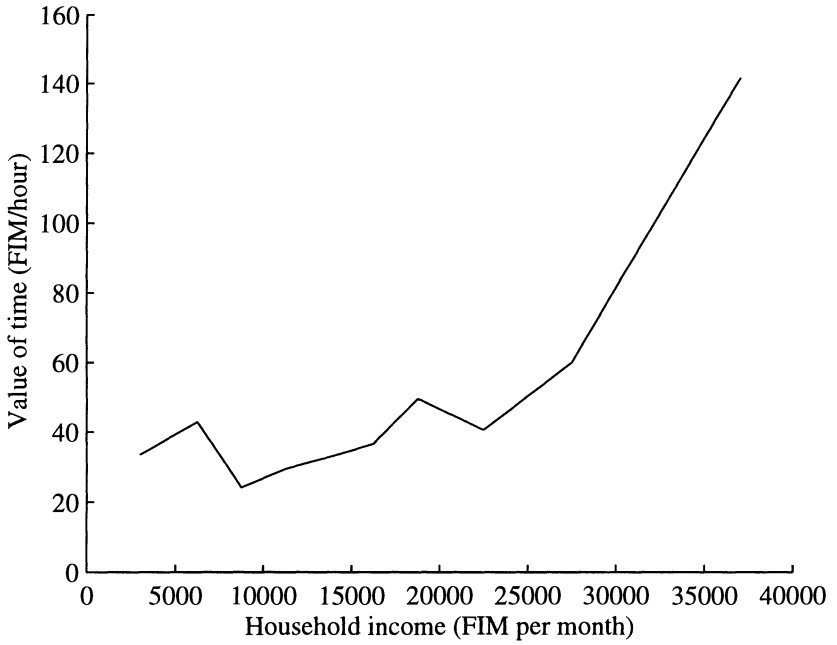

Figure 13.3 The effect of household gross income on the value of total travel time

5.2.3 Effect of trip purpose. The value of time is clearly affected by the purpose of the trip. According to the mode choice models, the value of time for commuter trips, or home-based work trips, is about 5.4 euros per hour, which is lower than the value for other trips, 8.2 euros per hour. This contrasts with the usual presumption that commuters value their time more highly than people making trips not related to work. Perhaps the most important reason for this unexpected result, a reason not directly related to differences in preferences, is the fact that the trips, on which the SP exercise was based, were actually made one and a half years before the SP study. Work trips are made almost every day, and the choice task was therefore not too abstract. For the other purpose groups, however, the case might be quite the opposite. This may have had an influence on the choices made by the respondents. Nevertheless, it should be kept in mind that value of time is a trade-off between time and money. While time pressures related to work trips may raise the value of time, the fact that work trips are made frequently may have an opposite effect. People are perhaps more willing to pay for travel time savings if the trip is made seldom.

5.2.4 Modal effects. Mode of travel certainly has an effect on the value of time. For example, the average value of time for car mode is usually higher than that of public transport. Nevertheless, the differences in the values stem from various factors, not all of which are mode-related. Firstly, socio-economic characteristics, income in particular, have an effect on mode choice, and therefore the users of different modes value their time differently. Secondly, if the values of time were corrected for income and other socio-economic effects, the values for car users could still be higher. Faster modes, such as car compared with public transport, could attract those with higher values of time. Finally, trips are usually nuisances rather than something that people wish to have more, and the inconvenience of making a trip may depend on the mode used. For instance, driving a car is stressful on one hand, but comfortable on the other. All of these factors, socio-economic characteristics, self-selectivity, and comfort, have some implications to the use of values of time in socio-economic evaluation. For example, it could be argued that time savings for persons attracted to an improved mode should be valued at the rate appropriate to the mode from which they are transferring.

According to a mode choice model with no allowance for the actual choice, value of total travel time was

- car 24 FIM per hour
- local train or metro 42 FIM per hour
- bus or tram 32 FIM per hour
- bicycle 50 FIM per hour
- walk 49 FIM per hour.

If the total travel time for public transport modes is divided into in-vehicle and out-of-vehicle time, the model gave

- car 15 FIM per hour
- public transport
  - in-vehicle time 19 FIM per hour
  - out-of-vehicle time 50 FIM per hour
- bicycle 47 FIM per hour
- walk 49 FIM per hour.

The respondents were further divided into modal groups, based on the mode they actually chose (RP choice). Separate models for each group gave the following values of total travel time:

- car users 34 FIM per hour
- train and metro passengers 32 FIM per hour
- bus and tram passengers 30 FIM per hour

- bicycle users 42 FIM per hour
- pedestrians 17 FIM per hour.

It should be noted that the number of walk and bicycle trips were quite small so that we cannot draw any definitive conclusions. Furthermore, travel times for walk and bicycle are calculated afterwards assuming average speed of 4 (walk) and 15 kilometres per hour (bicycle). No travel cost were attached to walk and bicycle.

5.2.5 Walking, waiting, and transfers. According to the present study, public transport waiting time is about 1.6 times as uncomfortable as in-vehicle time, and walking time is 3.3 times as uncomfortable as in-vehicle time. One transfer in public transport is equivalent to 6–10 minutes of in-vehicle time, or about 0.6 euros per trip. This may be unrealistically high, as compared with average travel cost. Here, it should be pointed out that the number of transfers was not varied across the ten choice options presented to each respondent. Moreover, the total waiting time was not explicitly shown to the respondents. Instead of waiting time, headway of each public transport vehicle was presented. The total waiting time was assumed to be half of the sum of the headways. This certainly had an effect on the results. For instance, many other studies have suggested that waiting time is more uncomfortable than walking time. To sum up, it is possible that the choice was so complicated that the results are somewhat biased.

5.2.6 Availability of seat. According to the models, the value of a seat in a bus is 5.30 Finnish marks (0.9 euros) per hour. Travel time standing is over 70 per cent more uncomfortable than travel time sitting. People under 30 years gave a much lower valuation to a seat than people over 65.

5.2.7 Low-floor buses. The importance of a low-floor bus is considerably lower than that of availability of seat. In fact it seems that young and middle-aged people find low-floor buses even worse than normal buses. This contrasts with the present policy in the area, according to which all new buses are low-floor (to be more specific, the bus operators in the area are encouraged to use low-floor buses since, under the competitive route bidding system, routes, timetables, fares and minimum vehicle standards are pre-specified, and the acquisition of new low-floor buses is, apart from the price offered, one of the few means to win the competition). Of course, low-floor buses are designed to be fully-accessible to all passengers, including parents with young children in buggies as well as the elderly and people with disabilities. In the Helsinki metropolitan

area, one adult is entitled to travel free of charge when accompanying a small child in a pram or push-chair. Nevertheless, for average passenger there are several reasons for preferring normal buses to low-floor buses. The disadvantages of low-floor buses include fewer seats, rearward facing seats, and forward facing seats succeeding rearward facing seats in such a way that there is no room for legs. On the other hand, passengers get benefit from improved operational efficiency from faster boarding and alighting.

6. CONCLUSIONS

The present study, as well as other stated preference researches conducted in the Helsinki metropolitan area have shown that carefully designed and tailored mail-back questionnaires can be used as a source of data. Nevertheless, the need for customising to actual trip means that it may be impossible to carry out the study at one and the same time. We have used two-stage studies where the purpose of the first stage is to collect (revealed preference) information about actual trips, whereas the second stage consists of a stated preference exercise. This method has been a successful way to gather information that is needed for estimation of value of time and other trade-offs.

Since stated preference information is not based on actual behaviour, the validity of the results is always somewhat questionable. One of the reasons to be concerned about the validity is the fact that stated and revealed preference methods often give different results. For instance, it has been argued that stated preference data always give lower values of time than revealed preference data. This wisdom is not correct, however. It is true that the design of stated preference exercise affects the results, and therefore the results may be biased, but it is equally important to keep in mind that the way the revealed preference data are gathered has an effect, too. The measurement of explanatory variables for RP models is not an easy task, and many simplifying assumptions have to be made. Stated preference methods have the advantage that the values of the variables perceived by the respondent equal the values presented to the respondent and used in modelling. Revealed preference data, on the other hand, are usually based on some kind of engineering judgement and typically suffer from spatial and temporal aggregation error. The variables used in stated preference exercise must be of better quality since the values are presented to humans. For example, it would be absurd to present to the respondents a public transport choice option with 1.5 transfers although this kind of engineering values are routinely used in mode choice models based on revealed preference data.

One of the objectives of the present study was to gather data for the updating of the travel demand models for the Helsinki metropolitan area. This goal turned out to be quite difficult to achieve, for various reasons. For example, it is quite difficult to use the same variables for both stated preference exercise and travel demand models aimed at forecasting travel behaviour 20 or 30 years ahead. The explanatory variables for forecasting models are defined in such a way that it is relatively easy to predict their value in the future. In particular, this means that disaggregate variables or values at individual level should be avoided. The variables used in stated preference exercise, on the other hand, should be tailored or customised to actual trips. More sophisticated forecasting methods that operate on individual level variables only, such as sample enumeration or microsimulation, require more detailed data that are expensive to collect.

Acknowledgments

The authors wish to thank Helsinki Metropolitan Area Council YTV for access to data and financing the projects.

References

[1] Mikkola J. (1997). Factors affecting the choice of travel mode in the Helsinki metropolitan area (in Finnish). Master's thesis, Helsinki University of Technology, Department of Engineering Physics and Mathematics, Applied Mathematics. 72 pp. (also Publications of the Helsinki Metropolitan Area Council (YTV) C 1997:10. Helsinki.)

[2] Pearmain D., Swanson J., Kroes E., and Bradley M. (1991). Stated preference techniques : a guide to practice. Second edition. Steer Davies Gleave and Hague Consulting Group. 109 pp.

[3] Pursula M. (1997). The TransPrice trans European VOT study. Value of time seminar, Institute of Transport Economics TØI, Oslo, Norway, 21–22 May 1997. 5 pp.

[4] Ghali, M. O., Pursula, M., Milne D., Keränen M., Daleno M., and Vougioukas M. (1997). Assessing the impact of integrated trans modal urban transport pricing on modal split. The 25th European Transport Forum Annual Meeting. Proceedings of Seminar E, Transportation Planning Methods, Volume I. PTRC Education and Research Services Ltd, Volume P 414. Brunel University, England, 1–5 September 1997. pp. 341–352.

[5] Ben-Akiva, M. E. and Lerman, S. R. (1985). Discrete Choice Analysis. Theory and Applications to Travel Demand. MIT Press, Cambridge,

Massachusetts. 390 pp.

[6] Ortúzar, J. de D. and Willumsen L. G. (1994). Modelling Transport. Second edition. John Wiley & Sons, Chichester, West Sussex, England. 439 pp.

[7] Pursula, M. and Kanner, H. (1992). Traffic forecasting in the Helsinki metropolitan area transportation study 1988. Transportation Research Record No. 1357. Washington, D.C. pp. 29–39.

[8] Karasmaa, N. and Pursula, M. (1997). Empirical studies of transferability of Helsinki metropolitan area travel forecasting models. Transportation Research Record No. 1607. Washington, D.C. pp. 38–44.

[9] Karasmaa, N. (1998). Temporal transferability of the Helsinki metropolitan area travel demand models (in Finnish). Publications of the Helsinki Metropolitan Area Council (YTV) B 1998:3. Helsinki. 159 pp.

[10] Bradley, M. and Daly, A. (1997). Estimation of Logit Choice Models Using Mixed Stated Preference and Revealed Preference Information. 6th International Conference on Travel Behaviour. Quebec, Canada, 1991. In Understanding travel behaviour in an era of change. (P. Stopher and M. Lee-Gosselin, eds.). Elsevier Science Ltd, UK, pp. 209–232.

[11] Pursula M. and Kurri J. (1996). Value of time research in Finland. PTRC Value of Time Course and Seminar. Easthampstead, London, 28–30 October 1996. 14 pp.

[12] Pursula M. and Weurlander M. (1999). Modelling Level of Service Factors in Public Transportation Route Choice. Transportation Research Record No. 1669. Washington, D.C. 1999. pp. 30–37.

Chapter 14

EFFECTS OF DATA ACCURACY IN AGGREGATE TRAVEL DEMAND MODELS CALIBRATION WITH TRAFFIC COUNTS

Michele Ottomanelli
Dept. of Highways and Transportation - Polytechnic of Bari (Italy - EU)
tel: +39 080 5963380, fax: +39 080 5963329
ottomanelli@dvt005.poliba.it

Abstract This paper concerns with aggregate calibration of urban travel demand model parameters from traffic counts. A bi-level sequential Non-linear Generalised Least Square Estimator (NGLS) has been proposed to calibrate a travel demand model. The first aim was to find out the effects of the required input data accuracy assumptions on the model calibration. The second aim was to show the possibility to improve model link flows estimation performance even if the starting demand model was properly calibrated by using expensive disaggregate data. An experimental analysis was carried out on a real middle-sized town: the model was calibrated and validated under different "a priori" assumptions on data accuracy level of the starting data. The employed data were a traffic counts set and a maximum likelihood starting estimate of the travel demand model parameters.

Key words: NGLS estimator, travel demand models, calibration, traffic counts

Introduction

In urban transportation planning activities it is often necessary to determine the link traffic flows on the road network. The reliability of the link flows estimate depends on the reliability of the Origin-Destination travel

M. Pursula and J. Niittymäki (eds.),
Mathematical Methods on Optimization in Transportation Systems, 225–247.

demand matrix (O-D matrix) estimate, as well as the supply and traffic assignment models.

The O-D matrices in a given area could be estimated by using two different methodological approaches: the direct estimation method or the mathematical models estimation (indirect estimation). In both approaches very expensive surveys must be conducted in order to collect the necessary input data. Sometimes (e.g. in small urban areas, or when few resources are available), the great cost of the survey might compel the practitioner to conduct travel demand estimation using the resources and the poor available information by applying "very pragmatic" methods, which lack in theoretical consistence.

In recent years researchers have made many efforts in order to propose effective methodologies which provide "better and better" travel demand estimates by using information *cheap, easy and immediate to collect.*

Both mathematical consistency and practitioners' needs can be satisfied with several methodologies which suitably employ, as input data, the aggregate information contained in traffic counts (TC) measured on a set of network links.

In this direction, great attention has been given to methodologies which use TC for updating and/or correcting O-D matrices. These methodologies, following different approaches, can also use together with TC other available information (Van Zuylen *et al*, 1980; Cascetta, 1984; Maher, 1983; Yang *et al*, 1992; Bell, 1983; Cascetta *et. al,* 1986;1997).

On the contrary, few works concerning the calibration of travel demand models (TDM) have been proposed in literature even if this approach provides, at the same cost (at least), a mathematical tool for both the O-D matrix estimation and its variations forecasting, with respect to socio-economic and transport system attributes changes (Tamin *et al.*, 1989; Willumsen, 1981; Cascetta, 1986; Cascetta and Russo,1997). In fact, economical, practical and theoretical advantages could be reached at the same time by using the TC-information to calibrate aggregate TDM.

In particular, from the economical point of view, TC information use results doubly convenient. Actually, we find "direct" convenience related both to the easy survey organising and to the cheap data-collecting (especially with automatic counters). We obtain also a "derived" economic convenience represented by the economic benefit deriving from the re-use of existing data, such as outdated surveys, outdated demand estimate, outdated models etc.

Moreover, these methodologies have a lot of practical purposes:

- Aggregate transportation demand model calibration;
- Combined O-D matrix and transportation demand model parameters estimation;

- Transportation demand model parameters updating or adjustment;
- Transportation demand models transferability (spatial and temporal) .

These aims could be achieved under different theoretical approaches and assumptions that depend on the nature of the available information too. Consequently, it is possible to consider more estimators of a model parameters vector.

For a detailed discussion on the theoretical framework of these methodologies see Cascetta (1986; 1998) and Cascetta & Russo (1997).

Even if they are referred to simplified models or to small test networks, the few but good results showed in literature about the application of these methodologies have been the starting point for this research.

This paper shows some results of a current research and it deals with the robustness and the statistical performances of Non-linear Generalized Least Square (NGLS) estimator for the calibration of a system of urban travel demand model using TC. The available input data are: a starting estimate of model parameters vector $\hat{\beta}$ and a set of traffic counts (TC) f^{obs}.

The application of the NGLS estimator needs an "a priori" assumption about the "quality" (accuracy) of the starting data ($\hat{\beta}$ and TC).

This subjective assumption could considerably influence model calibration and therefore the link flows estimate.

Aim of this paper is to put in evidence, through an experimental analysis on a real size network, the following elements:

- the influence of different data accuracy assumption on the demand and on the link flows estimation;
- the possibility, under correct assumptions, to improve the model forecasting capability, even if the starting model parameters have been obtained through a Maximum-Likelihood (ML) calibration based on data collected through expensive and time consuming disaggregate surveys.

In the following section the theoretical background of the aggregate parameter estimation problem will be shortly described, followed by the problem formulation, the proposed experimental analysis methodology and the results. The paper ends with the conclusion and further researches .

1. THEORETICAL BACKGROUND OF THE ESTIMATION PROBLEM

A Travel Demand Model can be used to simulate, within a fixed study area, the average number of trips and their main characteristics, such as:

- the period h
- the purpose of the trip s;

- the origin o and the destination d of the trips;
- the mode used m;
- the path k, consisting of a sequence of links belonging to the network

The opportunity of moving the other characteristics of the trips depends on the user's choice. The user's decisions making can be modelled on the basis of the random utility theory (Ben Akiva & Lerman, 1987).

According to this theory, it is possible to predict the probability that user chooses one of the available alternative j on the basis of the following assumptions:

- each user *i* (or class of users) examines all the available alternatives *j* (e.g. destination, mode, etc) belonging to his/her choice set I_i;
- user associates to each alternative j a *perceived utility* U_j;
- the utility U_j is a random variable, specified as follows

$$U_j(\mathbf{SE}, \mathbf{T}) = V_j(\mathbf{SE}, \mathbf{T}) + \xi_j \qquad \text{with } E[\xi_j]=0 \text{ and } E[U_j]=V_j \ \forall j \in I_i$$

it consists of a deterministic (measurable) component V_j and a random error term ξ_j due to several causes (missing attributes, perception errors, etc.). The vectors **SE** and **T** are constituted by socio-economic and level of service variables (attributes) that are representative of user, land use and transportation system;

- users chooses the alternative with maximum utility.

Further to the stated assumptions the probability $p_i(j)$ that a user *i* chooses the alternative *j* is given by the probability that the perceived utility U_j is the higher then the utility of any other alternatives.

This can be formalised in the following probability choice function:

$$p_i(j) = \Pr[\xi_k - \xi_j \geq V_j - V_k \ \forall j \in I_i] \qquad \forall i, k$$

Thus, the choice probability depends on both the specification of the deterministic (systematic) utility V_j and the distribution of the random residual ξ_j.

Usually the systematic utility is specified as a linear model of the attributes by means of a parameters vector:

$$V_j(\mathbf{SE}, \mathbf{T}, \beta) = \Sigma_k \beta_k X_k = \beta_1 X_1 + \beta_2 X_2 + ... \qquad \forall j \in I_i$$

Once both systematic utility and random residual distribution are specified, it is possible to determine the functional form of the choice function $p_i(j)$. For example, if we assume each of the random residuals independently and identically Gumbel variates, then the choice probability is given by the widely known Multinomial Logit model:

$$p(j) = \exp(V_j) / \Sigma_{j'} \exp(V_{j'}) \qquad \forall j \in I_i$$

A system of travel demand models is constituted by many sub-model and allows to forecast many choice dimension, such as trip generation, trip distribution, modal split, path choice and so on.

After that the functional form of systematic utility and distribution of random terms have been defined then model has been specified and it is necessary to calibrate it. Thus, it is necessary to define an estimator of the parameters β_k of the variables. The functional form of the parameter estimator depend on many elements, such as the sampling technique, the available data, the statistical approach, and so forth.

To specify a travel demand model parameter estimator it is necessary to determine the relationships between the vector β of the unknown parameters and the available sources of information: a starting estimate $\hat{\beta}$ of the vector β and a set $\mathbf{f}^{obs}$ of TC.

The travel demand model (TDM) could be defined as the functional link between these elements. Indeed, a TDM is a mathematical function which gives, through the vector of parameters β, the Origin-Destination travel demand vector **d** for the given urban transportation system.

Thus, it can be formalized as follows (Cascetta, 1998):

$$\mathbf{d} = \mathbf{d}(\mathbf{SE}, \mathbf{T}, \beta) \qquad (1)$$

where **SE** and **T** are respectively the given sets of socio-economic and transport system attributes, representing the land use and the level of service of the supply transportation system. For a fixed urban system the previous relationship could be written as:

$$\mathbf{d} = \mathbf{d}(\beta)$$

In this case **d** is only a function of the parameters vector.

In the appendix typical partial-share system of demand model has been described.

1.1 RELATIONSHIP BETWEEN THE UNKNOWN PARAMETERS AND THE STARTING ESTIMATE

Let $\hat{\beta}$ be a starting estimate of the "true" unknown vector β, consisting of N_β components β_i. Hence, assuming the vector $\hat{\beta}$ as the determination of a random variable and the i-*th* parameter $\hat{\beta}_i$, it can be written as follows:

$$\hat{\beta}_i = \beta_i + \sigma_i \qquad (\text{with } i=1,2,...,N_\beta) \tag{2}$$

where σ_i is the random error in the estimation, assumed to be a random variable distributed with zero mean and variance-covariance matrix **Z**.
In practical application matrix **Z** is assumed to be diagonal.
The evaluation and the meaning of the variances of the random terms $\hat{\beta}$ depend on the statistical approach to the problem and on the origin of the starting estimate. Further discussion will follow.

1.2 RELATIONSHIP BETWEEN THE UNKNOWN PARAMETERS AND THE TRAFFIC COUNTS

The second source of information is a set of traffic counts. Traffic counts constitute an aggregate and experimental information to this problem.

The vector of traffic counts $\mathbf{f}^{obs}$ contains traffic flows observed (measured) on a set of links of all modal networks (car, transit, pedestrian etc.).

Let **d** be the vector of the "true" travel demand in a given urban area. Because of the errors due to traffic assignment model and therefore to the assignment matrix estimation $\hat{\mathbf{M}}$, the "true" traffic link flow vector **f** is different from the $\mathbf{f}^{sim}$ link flows vector estimated through assignment model. If ε^{ass} is the random error, in the flows estimate (assignment error) we assume:

$$\mathbf{f} = \mathbf{f}^{sim} + \varepsilon^{ass} = \hat{\mathbf{M}}\mathbf{d} + \varepsilon^{ass} \tag{3}$$

where f_l^{sim} is the l-th component of the simulated link flows vector $\mathbf{f}^{sim}$ which is given by:

$$f_l^{sim} = \sum_{od} m_{l,od}\, d_{od}$$

where:
- $m_{l,od}$ is the generic element of the assignment matrix $\hat{\mathbf{M}}$ and it represents the fraction of O-D flow d_{od} contributing to link flow f_l
- d_{od} is the O-D travel demand flow between the origin o and the destination d.

In addition, even if **d** has been estimated through a properly specified and calibrated system of models, the travel demand vector $\mathbf{d}(\beta)$ differs from the "true" one: the error is measured with the random term ε^{mod}.

Under this assumption, the above discussion can be formalised as follows:

$$\mathbf{d} = \mathbf{d}(\beta) + \varepsilon^{mod} \tag{4}$$

The above formulation summarises and highlights that even if we introduce in the TDM the "true" vector of the parameters, the model gives only the mean value of the random variable O/D-vector.

Owing to errors that occur during the counting operation, errors due to the time variability of the travel demand and in the users' path choice, the traffic counts are different from the "true" link flows. This difference can be assumed as a random error ε^{mis} and the vector of traffic counts can be considered as a random variable with mean value equal to the vector of the "true" link traffic flows. It can be summarized as follows:

$$\mathbf{f}^{obs} = \mathbf{f} + \varepsilon^{mis} \tag{5}$$

Consequently, from the stochastic relations (4), (3) and (5) we obtain the stochastic relation between the traffic counts vector and the TDM parameters:

$$\mathbf{f}^{obs} = \mathbf{M}\hat{}\,\mathbf{d}(\beta) + \varepsilon \tag{6}$$

where the term ε summarizes all the considered random errors discussed above. It will be distributed with mean $E[\varepsilon]=\mathbf{0}$ and variance-covariance matrix $E[\varepsilon\,\varepsilon^T]=\mathbf{W}$.

An experimental formula proposed by Cascetta, Nuzzolo and Velardi (1986) could be employed to evaluate the variance of the error term ε.
This formula is referred to road networks assignment and it gives the coefficient of variation value of the assignment error ε on each car network link:

$$Cv = 0.9 \exp(-0.0011\, f_l^{obs}) \tag{7}$$

where f_l^{obs} is the observed traffic volume on the l-th link of the network.

2. GENERAL FORMULATION OF THE ESTIMATION PROBLEM

The equations (2) and (6) constituted system of stochastic equations that can be used to specify different kinds of estimators of the model parameters vector β. Hence, in general, the transportation demand model calibration problem (TDMC) using traffic counts can be formulated as an operational research problem whose solution β^* is an estimator of the parameters vector.

In fact, β^* could be defined as the vector that minimises the "distance" between the unknown vector β and its starting estimate $\hat{\beta}$ together with the "distance" of the vector $\mathbf{f}^{sim}$ of the traffic link flows, obtained by assigning the travel demand vector $\mathbf{d}(\beta)$ to the network, from the vector of the traffic counts $\mathbf{f}^{obs}$.

In its general form the TDMC problem can be formulated as follows:

$$\beta^* = \text{argmin } Z(\beta) = \text{argmin } [z_1(\beta, \hat{\beta}) + z_2(\mathbf{f}^{sim}, \mathbf{f}^{obs})] \quad (8)$$

subjected to $\beta \in S_\beta$

where S_β represents the set of the feasible solutions for β and it can be constituted by a set of analytical and/or informal constraints to the problem (8) (e.g. Value of Time constrain, congruence in the sign of the parameters etc.).

The "distances" to minimise in the (8) are measured by the functions z_1 and z_2. The formal specification of these functions depends on the statistical approach to the problem, in particular on the assumptions made for the probability distribution of random errors σ_i and ε_l and on the statistical meaning that the analyst gives to the available information.

3. A CLASSIFICATION OF MODEL PARAMETERS ESTIMATORS FROM TRAFFIC COUNTS

This section propose, two different classifications of the estimators that use TC for TDM calibration in order to give a concise overview on the problem and for better placing this work in the general framework.

The first classification of the estimators has been based on the available data. The second one depends on the assumption made on the origin of the available data and on the statistical assumption of the random error terms.

3.1 THE ESTIMATORS FORMULATION AND THE AVAILABLE DATA

In general, the available data could have different origins. Actually, it is possible to have and use aggregate information (cordon counting, screen-line counting, O-D demand flows counting, ect.) and/or disaggregate information (e.g. data collected with SP and/or RP sampling surveys).

Thus, it is possible to have the following kinds of estimators:

a) Aggregate Estimators (by using only aggregate data)
b) Mixed Estimators (by using both aggregate and disaggregate data)

Within the aggregate estimator category we can specify estimators for model parameters calibration only or for both the parameters and the O-D demand matrix estimation. Therefore we have the following two sub-categories of estimators:

a1. Direct estimators (for aggregate model parameters calibration only)

$$\beta^* = \text{argmin}\,[z_2(\mathbf{M}^{\wedge}\mathbf{d}(\beta), \mathbf{f}^{\wedge})]$$

a2. Combined estimators (for O-D matrix and model parameters estimation)

$$\beta^*, \mathbf{d}^* = \text{argmin}\,[z_1(\mathbf{d}, \mathbf{d}^{\wedge}) + z_2(\beta, \beta^{\wedge}) + z_3(\mathbf{M}^{\wedge}(\beta)\mathbf{d}, \mathbf{f}^{\wedge})]$$

subject to $\beta \in S_\beta$ and $\mathbf{d} \in S_d$

where:

- β are the parameters of the path choice model
- $\mathbf{S}_d$ is the set of the feasible solutions for $\mathbf{d}$
- $\mathbf{d}^{\wedge}$ is the observed vector of O-D travel demand (target)

Similarly, the category of mixed estimators can be divided in two sub-sets: the simultaneous and the sequential estimators. This distinction depends on the way of mixing the aggregate information with the disaggregate ones. In the simultaneous sub-category two sources of information are used at the same time and we have only one optimisation problem to solve.

A general formulation of mixed-simultaneous estimator can be written as follows:

b1. Mixed-Simultaneous estimation

$$\beta^* = \text{argmax}\{\Sigma_i \ln p_i[j(i)](\beta) + z_2[\mathbf{M}^{\wedge}\mathbf{d}(\beta), \mathbf{f}^{\wedge}]\}$$

where the first term is a Maximum-Likelihood (ML) estimator in which $p_i[j(i)]$ is the probability (assumed to be statistically independent) that the user (individual) *i*, belonging to the observed sample, chooses the transportation alternative j(i) of his/her choice set. The information on the choices j derives from sampling disaggregate surveys and the specification of the probability depends on the choice model to calibrate.

The proposed estimation method combines both disaggregate and aggregate data, but the parameters calibration is carried out by using the information on two separate levels.

In the first level the disaggregate data are used to calibrate the model through an ML estimator. In the second one the available TC are used to correct and/or improve with, cheap aggregate data, the previous parameters estimate. The method could be considered as a parameter estimator and itself be defined as mixed-sequential estimator.

b2. Mixed-*Sequential* estimation (bi-level)

1st level) disaggregate estimation

$$\beta^{ML} = \text{argmax } \ln L(\beta) = \text{argmax } \{\Sigma_i \ln p_i[j(i)](\beta)\} \qquad (9)$$

2nd level) aggregate estimation (by using traffic counts)

$$\beta^* = \text{argmin}[z_1(\hat{\beta}, \beta) + z_2(\hat{\mathbf{M}}\,\mathbf{d}(\beta), \hat{\mathbf{f}})]$$

where $\hat{\beta} = \beta^{ML}$ (obtained by solving the 1st level problem) and where both the optimisation problems are subject to $\beta \in S_\beta$.

3.2 THE ESTIMATORS AND THE STATISTICAL APPROACH

As above mentioned, the mathematical formulation of z_1 and z_2 depends on the following elements:

a) assumption on the nature (origin) of the starting information sources (experimental, non-experimental, aggregate, disaggregate);
b) assumptions on the probability distribution of the random error terms ε and σ .

On the basis of these assumptions it is possible to specify three classes of estimators β^* for the vector of the parameters, that are:

- Maximum-Likelihood Estimators (ML);
- Generalized Least Square Estimators (GLS);
- Bayesian Estimators (B)

The reader can find in literature a deeper theoretical discussion on these categories of estimators. In particular, in Cascetta & Nguyen (1986), with reference to O-D matrix estimation, the specifications of the function z_2 that depends on the distributional assumption of the random residuals ε, are described. Besides, Cascetta & Russo (1997) paid more attention to the functional form of the z_1 distance and to the general formulation of the TDMC problem.

With respect to the estimation of O-D matrix using TC, a lot of works proposed in literature (Cascetta, 1984; Cascetta and Nguyen, 1986; Di Gangi, 1989; Ortùzar and Willumsen, 1994) showed that the GLS estimators class could be the most suitable and robust for real-size applications.

Conversely, very few works have been proposed in literature which consider the TDMC problem with applications to real size networks (Cascetta and Russo, 1997) and/or they concern with calibration of simplified demand models (Tamin *et al*, 1989). However, these works showed the potential robustness properties of the GLS estimators class to solve the TDMC problem too.

4. NON-LINEAR GENERALISED LEAST SQUARE ESTIMATOR AND THE ACCURACY OF DATA

In this paper, a Non-Linear Generalised Least Square Estimators (NGLS) has been proposed to estimate the unknown vector β of a TDM coefficients. In particular, this specification has been proposed for the second level problem, while an ML estimator has been used in the first one.

Keeping the same nomenclature used in the previous sections, an NGLS estimate $\beta^*=\beta^{NGLS}$ of the vector β can be obtained by solving the following problem:

$$\beta^*= \beta^{NGLS}=\arg\min\{[(\beta-\hat{\beta})\mathbf{Z}^{-1}(\beta-\hat{\beta})]+[(\mathbf{f}^{obs}-\mathbf{M}\hat{\mathbf{d}}(\beta))\mathbf{W}^{-1}(\mathbf{f}^{obs}-\mathbf{M}\hat{\mathbf{d}}(\beta))]\}$$

$$\text{subject to } \beta\in \mathbf{S}_\beta$$

Even if no distributional assumptions on random residuals are needed, GLS estimation requires the evaluation of the covariance matrices $\boldsymbol{Z}$ and $\boldsymbol{W}$.
By assuming these matrices as diagonal , it is possible to have a simplified form of the NGLS estimator. Even though theoretically questionable, this assumption is frequently used for practical applications. In fact, from the practical point of view, it leads to a great reduction of the computational effort with negligible improvements in the estimation results. Under this assumption, the NGLS estimator can be formulated in the following form that meets both practical and theoretical needs:

$$\beta^{NGLS} = \underset{\beta\in S\beta}{\arg\min}\left[\sum_i \frac{(\beta_i-\hat{\beta}_i)^2}{\mathrm{var}(\sigma_i)}+\sum_l \frac{(f_l^{obs}-\sum_{od} m_{l,od}d_{od}(\beta))^2}{\mathrm{var}(\varepsilon_l)}\right] \quad (10)$$

Obviously, the observed link flows should be uncorrelated or negatively correlated and should have the lowest variance as possible. This means that the selected counting section must be located on links with high traffic volume and along a screen-line.

On the contrary of other kinds of estimators, NGLS estimator have at least two advantages. First, as mentioned above, it needs no explicit distributional assumption about the stochastic elements (5) and (6). The second advantage is the possibility to state explicitly the accuracy of each data by means of the values of the variances $var(\sigma_i)$ and $var(\varepsilon_i)$ contemplated in problem (10).

The variance of the link flow error can be evaluated by computing the coefficient of variation by means of the experimental formula (7).

The method to evaluate the variance $var(\sigma_i)$ depends both on the statistical approach and on the meaning and origin of the starting estimate of β:

a) it could be determined analytically if $\hat{\beta}$ were previously estimated through an ML estimator, as in (9);

b) it could be assumed "a priori" from the analyst, if the starting estimate were assumed as an exogenous non-experimental information (i.e. outdated estimate, transferred parameters etc.).

In the first case, it is possible to determine an approximate expression of dispersion matrix **Z** of the ML estimator (9). For large sample, the elements of **Z** can be evaluated by means of the negative inverse of the Hessian of the log-likelihood function, computed at point β^{ML}(Judge *et al.*, 1985).

The second case is the most recurrent in practical applications. When very few resources and data are available to estimate the travel demand in a certain area, practitioners try to specify and calibrate some kinds of models starting from the available data, such as model parameters that are outdated or that have been transferred from other "similar" areas. In this approach the starting estimate of the parameters must be considered as a "a priori" non experimental information source. Consequently, the variance $var(\sigma_i)$ of each terms β_i represents the analyst's confidence in the available starting estimate or the expected "quality" level of it.

In this context, transportation engineers convey that TC are suitable, robust and cheap experimental information to use for correcting the starting available data; in fact, sometimes TC are used as deterministic constraint to adjust O-D matrices with GLS-estimators.

This means that the analyst assumes a great accuracy for the TC (low variance value of ε_l).

In this paper we have pointed out the effects of these subjective "a priori" assumptions on data accuracy (or quality) on the demand models calibration by using the NGLS estimator (10).

In order to isolate the effects of the assumption on TC accuracy, a high "quality" (low variance) starting estimate $\hat{\beta}$ of parameters vector has been required. Thus, the starting vector $\hat{\beta}$ has been obtained by the specification and calibration of the considered system of models with a Maximum-Likelihood estimator based on data collected by means of disaggregate sampling surveys (road side interview, house hold interviews) (ELASIS, 1993; Cascetta *et. al.*; 1993).

4.1 STATEMENT OF THE EXPERIMENTAL ANALYSIS

In order to investigate on the effects of different data accuracy assumptions on the demand model forecasting performance, an experimental analysis has been carried out.

The parameters vector of the considered system of travel demand models (see appendix) have been estimated for a great number of starting data accuracy assumptions. Then, each calibrated models has been validated with respect to the link flows estimation performance.

Owing to hypothesis of exogenous origin of the starting information, in the problem (10), the "a priori" accuracy assumptions are represented by the variance of the stochastic items (2) and (6).

Let the random terms, σ and ε, be distributed with zero mean:

$$E[\sigma_i]=E[\varepsilon_l]=0 \qquad \text{with } i=1,2,...,N_\beta \text{ and } l=1,2,...,NC$$

where NC are the traffic counts used in the (10) to estimate the N_β parameters of the model. Since the variance of a random variable can be evaluate as function of the coefficent of variation Cv, the variance of the random terms in the (2) and (6) can be written as:

$$var[\sigma_i]=(CvB \times E[\hat{\beta}_i])^2 \tag{11}$$

$$var[\varepsilon_l]= (CvF \times E[f_l^{obs}])^2 \tag{12}$$

where CvB and CvF are, respectively, the coefficients of variation of the starting estimate $\hat{\beta}$ and TC.

The CvF could be calculated as function of the coefficient of variation by means of the experimental formula (7).

Following the proposed approach an accuracy assumption on the available data has been defined through the couple (CvB, CvF): each one represents the analyst's "a priori" level of confidence to the data used in the estimation of β.

As result of the problem (10), an estimate β^* of the model parameters vector has been obtained for each accuracy assumption. The influence of the accuracy assumption on the parameters estimation can be formalised writing the estimate as function of the couple (CvB, CvF):

$$\beta^* = \beta^{NGLS}(\text{CvB, CvF}) \tag{13}$$

Consequently, the travel demand vector $\mathbf{d}^*$ depends on the accuracy assumption too, as it has been estimated as function of the parameters vector (13) through the calibrated model:

$$\mathbf{d} = \mathbf{d}(\text{CvB, CvF}) = \mathbf{d}[\mathbf{SE, T}, \beta^{NGLS}(\text{ CvB, CvF})] \tag{14}$$

By means of a Stochastic User Equilibrium (SUE) model each travel demand vector (14) has been assigned to the network. Thus, for each accuracy assumption the vector of estimated link traffic flow $\mathbf{f}^{sim}$ has been attained:

$$\mathbf{f}^{sim} = \mathbf{f}^{sim}(\text{CvB, CvF}) = \mathbf{M}^{\wedge}\,\mathbf{d}\,[\mathbf{SE, T}, \beta^{NGLS}(\text{CvB, CvF})] \tag{15}$$

In order to evaluate the statistical performance of the estimator (10) the link traffic flow vector $\mathbf{f}_{in}^{sim}$ relative to the network assignment of the starting demand vector $\mathbf{d}(\mathbf{SE, T}, \beta^{\wedge})$ has been estimated too. Therefore, we have the starting link flows vector:

$$\mathbf{f}_{in}^{sim} = \mathbf{M}^{\wedge}\,\mathbf{d}(\mathbf{SE, T}, \beta^{\wedge}) \tag{16}$$

Each calibrated model has been validated by comparing the estimated link traffic flows and the starting link flows vector with the observed link flows set. To reach this aim some statistics and indicators have been evaluated.

5. EFFECTS OF DATA ACCURACY: MODELS VALIDATION

For each accuracy assumption (CvB, CvF) the calibrated demand model has been validated to find out the effects of the accuracy assumption on its forecasting performance. The simulated car-network link flow vectors have

been compared to two sets of observed link flows. The first observed flows set (hold-out-sample) is constituted by traffic counts that are not used for the model calibration. The second one is constituted by the all available TC, that are the old-out-sample and the traffic counts used for calibrating the model.

Let NV be the number of the observed link flow. For each model calibration the statistical performances of the estimator (10) have been evaluated by means of the following statistics:

Mean Square Error

$$MSE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}] = [\Sigma_l(f_l^{obs} - f_l^{sim})^2 / NV]$$

Root Mean Square Error

$$RMSE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}] = \{(MSE)^{1/2} / [(\Sigma_l (f_l^{obs}) / NV]\}$$

Mean Absolute Error

$$MAE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}] = \Sigma_l|(f_l^{obs} - f_l^{sim})| / NV$$

Relative Mean Absolute Error

$$RMAE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}]=[\Sigma_l|(f_l^{obs} - f_l^{sim})|/ f_l^{obs}] / NV$$

In addition, proportional reduction of the statistics have been evaluated with respect to their starting values:

$$\Delta MSE\% = \{MSE(\mathbf{f}_{in}, \mathbf{f}^{obs}) - MSE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}]\}/MSE(\mathbf{f}_{in}, \mathbf{f}^{obs})$$

$$\Delta RMSE\%=\{RMSE(\mathbf{f}_{in}, \mathbf{f}^{obs}) - RMSE[\mathbf{f}^{sim}(CvF,CvB), \mathbf{f}^{obs}]\}/RMSE(\mathbf{f}_{in}, \mathbf{f}^{obs})$$

$$\Delta MAE\% =\{MAE(\mathbf{f}_{in}, \mathbf{f}^{obs}) - MAE[\mathbf{f}^{sim}(CvF, CvB), \mathbf{f}^{obs}]\}/MAE(\mathbf{f}_{in}, \mathbf{f}^{obs})$$

$$\Delta RMAE\%=\{RMAE(\mathbf{f}_{in}, \mathbf{f}^{obs}) - RMAE[\mathbf{f}^{sim}(CvF,CvB), \mathbf{f}^{obs}]\}/RMAE(\mathbf{f}_{in}, \mathbf{f}^{obs})$$

where $\mathbf{f}_{in}$ is the starting estimate of the link flows vector and $\mathbf{f}^{obs}$ the vector of the observed link flows.

Morover, the total link flow and the average link flow have been evaluated with respect to the observed network links to measure the performance in total demand level estimation.

6. THE NUMERICAL PROCEDURE

A large number of accuracy assumptions have been considered: the coefficients of variation values CvF and CvB have been varied into the range [0.01 - 0.4] and [0.01 – 0.6] respectively, with steps 0.01 and 0.1.

To evaluate the CvF the experimental formula (7) has been also used: it gives the values of the CvF for each item of the calibration TC set.

The observed link flows used in the validation procedure (hold-out sample) have been drawn at random from the available traffic counts set. The other TC have been used as input data in the problem (10). The sets of traffic counts employed both for calibrating and validating the demand models have been described in Tab. 1.

Tab. 1 – Available Traffic Count sets

	Car Network link	Transit Network link
Calibration TC	42	159
Validation TC (old-out-sample)	10	

A computer program based on a projected gradient algorithm has been developed to solve the problem (10) and implement the analysis. For each couple (CvB, CvF), the outputs of the developed program are the vector of the estimated model parameters and the estimated modal (car and transit) O-D matrices. The input data considered are: a vector of parameters $\hat{\beta}$, the vector of traffic counts $\mathbf{f}^{obs}$, the vectors of socio-economic and level-of-service attributes and the assignment matrix $\hat{\mathbf{M}}$. Link traffic flows estimation has been attained by the module T-Road of the commercial MT-Model package. It is based on a SUE-Probit assignment model. The developed program has been implemented on a PC with Pentium/133Mhz CPU with RAM 32Mbyte and needed an average time of 270sec to run.

6.1 APPLICATION TO A REAL SIZE NETWORK

The numerical procedure has been implemented on a real size network and with real data. The network is referred to the town of Salerno, middle-size town of Southern Italy with roughly 150.000 inhabitants.

The considered area has been divided into 53 traffic zone and modelled by means of 53 internal-centroids. The relationship with the outside areas has been represented through 9 external-centroids. The starting value of total estimated demand flow was about 23.000 vehicles (morning peak hour).

The network characteristics has been described in the Tab. 2 and Tab. 3.

Tab. 2 – Area zoning

	total	external	internal

centroids	62	9	53

Tab. 2 – Network characteristics

	nodes	links	
		total	traffic counts
Road Network	555	1134	52
Transit Network	563	2131	199

To each accuracy assumption, the estimator (10) has been applied and the correspondent travel demand vector **d** has been estimated by the model (14). The transportation demand system in the area has been simulated through a system of mathematical models based on the random utility theory (see appendix). It is constituted by three sub-models: trips emission, trips distribution and modal split sub-models.

Model calibration needed the estimation of 45 parameters belonging to the unknown vector β for each data accuracy assumption.

Each estimated demand vector has been assigned to estimate the link traffic flows.

6.2 ANALYSIS RESULTS

To put in evidence the effects of TC accuracy assumptions on the calibration, a high accurate starting parameters vector has been considered (Cascetta et al., 1993). Intuitively, as the starting parameters vector is quite accurate, if congruent assumptions have been made on the TC accuracy level then no great variation in the total demand flow estimation and negligible variation in Value-of-Time (VoT) have to be expected. This means that the starting estimate $\hat{\beta}=\beta^{ML}$ must have low values of CvB. On the contrary wrong assumptions on the (CvF, CvB) would show biased estimate both in values of the parameters and demand estimate. In fact, this analysis shows that reliable assumptions for the accuracy of $\hat{\beta}$ is CvB=0.01÷0.02. Moreover, for every value of CvF, biased estimate of total flow demand has been obtained by assuming low confidence in the starting $\hat{\beta}$ (CvB greater than 0.1).

Keeping CvB in the correct range, it is possible to investigate the TC accuracy assumption effects.

Two main effects originated by the "a priori" assumptions have been found out.

With respect to the right value of CvF, the overestimation of confidence level in TC (i.e. CvF low value) leads up to 50% of under-estimation error in the total demand flow and considerable variation in the VoT.

Actually, assuming high values for accuracy level of TC, means to assume a wrong deterministic constraints to the problem (8), while it is originally stochastic.

For fixed values of CvB, lower errors in the total demand flow estimation have been obtained when the confidence level in TC information is underestimated. The expected small variations of VoT have resulted from CvF greater than 0.1. These results can be explained considering that random error ε in the (10), summarises all errors related to traffic flows measurement, demand and assignment modelling and not only to traffic counts.

Same results and considerations can be found out in the work by Di Gangi (1989) but with respect to O-D matrix estimation problem in which the effects of the assignment errors are stronger.

Improvements in link flows estimation have been obtained by assuming correct values of the data accuracy ($0.3 \leq CvF \leq 0.4$ and $0.01 \leq CvB \leq 0.02$). The results of the model validation with respect to the hold-out-sample have been reported in Tab.4.

Tab. 4 – Car-Network link flows estimation statistical performance (hold-out-sample)

	CvF	*0.3*	*0.3*	*0.4*	*0.4*	*(exper.)*	*(exper.)*
	CvB	*0.01*	*0.02*	*0.01*	*0.02*	*0.01*	*0.02*
Statistics	*Starting values*		best				
MAE total flow	447	333	103	376	210	436	355
ΔMAE% total flow		25.50	***76.96***	15.88	53.02	2.46	20.58
MAE	170.70	168.10	169.90	169.20	168.60	168.00	164.30
ΔMAE%		1.52	0.47	0.88	1.23	1.58	***3.75***
RMAE	0.24	0.24	0.23	0.23	0.24	0.23	0.23
ΔRMAE%		1.33	1.73	2.68	1.32	2.73	***4.17***
MSE ($x10^{-3}$)	50.37	48.13	45.40	50.01	46.41	53.12	50.52
ΔMSE%		4.46	***9.87***	0.55	7.86	-5.45	-0.31
RMSE	0.271	0.26	0.26	0.27	0.26	0.28	0.27
ΔRMSE%		2.26	***5.06***	0.27	4.01	-2.69	-0.15

The same statistics have been computed by using both the hold-out-sample and all the available traffic counts (Tab. 5).

With respect to the starting values, it has been possible to improve all the computed statistics. In particular, the MSE statistics of the hold-out-sample have been improved up to 10% (Tab. 4). Small negative values of ΔMSE% and ΔRMSE% have been calculated assuming CvF values computed by means of the (7) that could be used only for car-network assignment, while in reality it has been used for transit-network TC.

Tab. 5 – Car-Network link flows estimation statistical performance (all available TC)

	CvF	*0.3*	*0.3*	*0.4*	*0.4*	*(exper.)*	*(exper.)*
	CvB	*0.01*	*0.02*	*0.01*	*0.02*	*0.01*	*0.02*
Statistics	*Starting values*		best				
MAE total flow	4499	3964	2955	4268	3391	4498	4163
ΔMAE% total flow		11.89	***34.32***	5.13	24.63	0.02	7.47
MAE	223.52	220.04	216.44	221.38	217.56	220.27	216.63
ΔMAE%		1.56	***3.17***	0.96	2.67	1.45	3.08
RMAE	0.43	0.43	0.43	0.43	0.43	0.43	0.42
ΔRMAE%		0.70	1.39	0.85	1.27	1.46	***2.49***
MSE(x10^{-3})	78.16	74.52	68.90	76.53	71.45	76.86	73.78
ΔMSE%		4.66	***11.85***	2.09	8.59	1.67	5.61
RMSE	0.206	0.20	0.19	0.20	0.20	0.20	0.20
ΔRMSE%		2.36	***6.11***	1.05	4.39	0.84	2.85

Even if no explicit constraints have been applied to the problem (10), the correct estimation of each parameter sign has been obtained. This confirm the robustness of the NGLS estimator.

7. CONCLUSIONS, METHODOLOGICAL REMARKS AND FURTHER DEVELOPMENTS

In this paper a bi-level sequential NGLS estimator has been proposed to calibrate a travel demand model with traffic counts data. This estimator category provides high performances under economical, computational and theoretical aspects. Under correct assumptions on the quality of the available information, the proposed method could be used to combine successfully "a priori" information on the starting estimate in the calibration procedure.

This is not an easy task since the procedure is affected by the subjective analyst's confidence in the information sources. By means of an experimental analysis the influence of available data accuracy assumption on the link traffic flows estimation have been investigated. It has been showed that TC information can be effectively used to improve an existing estimate of demand model parameters even if obtained through a Maximum Likelihood estimator based on disaggregate data. These numerical results have strengthened the more theoretical findings on GLS estimators that showed that by combining survey data with traffic counts (or other aggregate data) a reduction of ML estimate variance can be obtained (Judge, *et al.,* 1985).

From the methodological point of view these results lead to some relevant items that have to be pointed out:

- it is possible to reduce and/or avoid the high cost of the necessary disaggregate sampling surveys to implement the traditional model calibration or direct demand estimation methods;

with respect to the O-D matrix estimation methodologies:

- it is possible to provide, at least at the same cost, a calibrated mathematical model for travel demand forecasting;
- the computational effort is lower since that the number of variables to estimate is inferior.

Taking in to account the size of the network, the associated real data and the type of calibrated system of models the obtained results are very encouraging.

Further development of this work will deal with:

- the aggregate calibration of path choice model parameters with a fixed-point formulation of the problem (8);
- the effect of spatial selection of traffic count sections (field measurement design);
- the economical assessment of the savings by updating outdated surveys and data (i.e. old estimates of O-D matrices and model parameters) from TC.

Acknowledgements

Author would like to express his sincere thanks to Prof. E. Cascetta, for his inputs and suggestions and to Proff. G. E. Cantarella and A. Nuzzolo for their helpful tutoring during the PhD course.

References

Bell, M.G.H. (1983). The estimation of an Origin-Destination matrix from traffic counts. *Transportation Science*, 31: 198-217

Ben Akiva, M., Lerman, S.R. (1987). *Discrete Choice Analysis*. MIT Press, Cambridge, MA

Cascetta, E.(1984). Estimation of trip matrices from traffic counts and survey data: a generalized leas squares estimator. *Transportation Research,* 18B: 289-299

Cascetta, E. (1986). A Class of Travel Demand Estimator using Traffic Flows. *CRT Publication No. 375*. Universitè de Montreal, Montreal, Canada

Cascetta, E., Nguyen S. (1986). A unified framework for estimating or updating Origin-Destination matrices from traffic counts". *Transportation Research,* 22B:437-455

Cascetta E., Nuzzolo, A., Velardi, V. (1993). A system of models for the evaluation of integrated traffic planning and control policies. *Workshop Integration Problems in Urban Transportation Planning and Management Systems*. Capri, 28-29 October 1993

Cascetta E., Nuzzolo, A., Velardi, V. (1986). Un'analisi sperimentale dei modelli di assegnazione alle reti urbane di trasporto privato. *Proc. of IV National Conference of PFT-CNR*, Torino (Italy)

Cascetta, E. (1998). *Teoria e Metodi dell'Ingegneria dei Sistemi di Trasporto.* UTET, Italy

Cascetta, E., Russo, F. (1997). Calibrating Aggregate Travel Demand Model with Traffic Counts: Estimators and Statistical Performance. *Transportation,* 24: 271:293

Di Gangi, M. (1989). Una valutazione delle prestazioni statistiche degli estimatori della matrice O/D che combinano i risultati di indagini e/o modelli con i conteggi di traffico. *Ricerca Operativa*, no. 51: 23-59

ELASIS-CSST (1993). Un sistema integrato di supporto alle decisioni per la gestione della mobilità e per la pianificazione dei trasporti urbani. Quaderno n. 26

Judge, G., Griffiths, W. (1985). *The Theory and Practice of Econometrics.* 2nd Ed. J. Wiley, NY

Maher, M.J. (1983). Inference on trip matrices from observations on link volumes: A bayesian statistical approach. *Transportation Research,* 17B: 435-447

Ortùzar, J.de D., Willumsen, L.G. (1994). *Modelling Transport*, 2nd Ed., J. Wiley & Sons Ed.

Russo, F., Iannò, D. (1991). Prestazioni statistiche degli estimatori dei parametri di modelli di domanda con i conteggi di flussi. Internal Report n. 2, DIMET, University of Reggio Calabria, Italy

Tamin, O.Z., Willumsen, L.G.(1989). Transport demand model estimation from traffic counts. *Transportation*, 16: 3-26

Van Zuylen, J. G., Willumsen, L.G. (1982). The most likely O-D matrix estimated from traffic counts. *Transportation Research*, 14B: 281-293

Willumsen, L.G. (1981). Simplified transport models based on traffic counts. *Transportation,* 10: 257-278

Appendix: The Calibrated Demand Model

The calibrated travel demand system of models is a traditional partial-share model. It is constituted by the following three sub-models: emission model, trip distribution model and modal split model. Its structure can be considered as behavioural since the first sub-model is descriptive, while the other two are based on the random utility theory.

The demand model, for the purpose s and the period h, has been formalised as follows:

$$d_{od}(s,m,h) = d_o(s,h)\, p(d/o,s,h)\, p(m/o,d,s,h)$$

where $d_{od}(s,m,h)$ is the average number of trips from the origin *o* to the destination *d* made by the users in the period *h*, for the purpose *s*, by means of mode *m*.

The sub-models have been specified as follows:

Trip emission model (*index per category* type) gives the number of trips which start from the origin *o*, for the purpose *s,* in the period *h*:

$$d_o(s,h) = \Sigma_c[N(o,c) \times TS(c,s)]$$

where *c* is the users category, N(o,c) is the number of users resident in the zone *o*, belonging to category c, TS(c,s) is the number of trips made by the user in the period *h*, for the purpose *s*.

Trip distribution model (Logit type). It gives the probability that users moving from *o* for the purpose *s,* choose the destination *d*:

$p(d/o,s) = \exp(V_{od}) / \Sigma_{d'} \exp(V_{od'})$ with systematic utility $V_{od} = \Sigma_k \beta_k X_{(o,d)k}$

Modal split model (Logit type) gives the probability that user chooses mode m to go in the destination *d*:

$p(m/o,d,s,h) = \exp(V_{odm}) / \Sigma_{m'} \exp(V_{odm'})$ with systematic utility $V_m = \Sigma_k \beta_k X_{(m)k}$

Thus, the calibrated travel model has been specified as follows:

$$d_{od}(s,h,d,m) = d_o(s,h) \{[\exp(V_{od}) / \Sigma_{d'} \exp(V_{od'})][\exp(V_{odm}) / \Sigma_{m'} \exp(V_{odm'})]\}$$

Four home-based trip purposes (*s*) have been considered:
Home to Work (H-W),
Home-to-other Constrained Purposes (H-CP),
Home-to-other Non-Constrained Purposes (H-NCP),
Home-to-School (H-S)

The modal choice set is constituted of five alternatives (m):

I_i = {Car-Driver, Car-Pool, Bus, Motorcycle, Pedestrian}

Attributes ($X_{(m)k}$ and $X_{(o,d)k}$) and parameters β of the models have been summarised in the following tables.

Trip generation model: Purposes, parameters and attributes

purpose attributes	H-W	H-CP	H-CNP	H-S
nos. employees	β_1	β_4	β_7	β_{10}
modal choice logsum	β_2	β_5	β_8	β_{11}
SZ (same zone)	β_3	β_6	β_9	β_{13}

Modal Choice Model: parameters and attributes – (H-W)

attribute mode	time	cost	parking distance	ASA mode	destination distance
car-driver	β_{13}	β_{14}	β_{17}	β_{18}	
car-pool	β_{13}	β_{14}		β_{19}	
bus	β_{13}	β_{14}		β_{22}	
motorcycle	β_{13}	β_{14}		β_{16}	
pedestrian	β_{20}				β_{21}

Modal Choice Model: parameters and attributes – (H-CP)

attribute mode	time	cost	ASA mode	Dummy age>35	destination distance
car-driver		β_{23}	β_{26}		
car-pool		β_{23}	β_{27}		
bus		β_{23}	β_{30}		
motorcycle		β_{23}	β_{24}	β_{25}	
pedestrian	β_{28}				β_{29}

Modal Choice Model: parameters and attributes – (H-NCP)

attribute / mode	time	cost	ASA mode	parking distance	destination distance
car-driver		β_{31}	β_{33}	β_{34}	
car-pool		β_{31}	β_{35}		
bus		β_{31}	β_{38}		
motorcycle		β_{31}	β_{32}		
pedestrian	β_{36}				β_{37}

Modal Choice Model: parameters and attributes – (H-S)

attribute / mode	time	cost	ASA mode	destination distance
car-pool	β_{39}	β_{40}	β_{42}	
bus	β_{39}	β_{40}	β_{45}	
motorcycle	β_{39}	β_{40}	β_{41}	
pedestrian	β_{43}			β_{44}