DELFT UNIVERSITY OF TECHNOLOGY
Faculty of Civil Engineering and Geosciences
Transport & Planning Section

Course CT4801

Transportation Modeling

prof. dr. ir. P.H.L. Bovy
dr. M.C.J. Bliemer
dr. ir. R. van Nes

Edition: August 2006

Table of Contents

List of symbols

Utility theory
P = probability
N = utility of activity
V = systematic utility component
Z = disutility
K = monetary budget
T = time budget

Trip generation models
T = predicted number of zonal trips
X = zonal explanatory variable
Y = household explanatory variable
Z = personal explanatory variable
N = number of units (households or persons by category)
t = trip rate

Trip distribution models
T = predicted number of trips between zones
F = measure of accessibility
Q = production potential
X = attraction potential
A = number of arriving trips in zone
P = number of departing trips from zone
B = travel performance
c = travel resistance
$f(c)$ = utility value of travel resistance between zones
k = measure for the number of and variability in trip altgernatives
μ = measure of average trip intensity in area

Indices
p = trip purpose
i,j = zone i and j
h = household type
n = person type
a = link
r = route, path
k = optimal route

Terminology

access	=	voortransport
alternative	=	option
arrival	=	aankomst
assignment	=	toewijzing
centroid	=	zwaartepunt van een verkeerszone (meestal ook voedingspunt voor netwerk
connector	=	fictieve verbindingslink tussen centroid en netwerk
departure	=	vertrek
detour	=	omweg
distribution	=	verdeling (naar herkomst en bestemming)
egress	=	natransport
line haul	=	hoofdtransport
means	=	vervoermiddel
modal share	=	aandeel vervoerwijze
modal shift	=	verandering van vervoerwijze
modal split	=	verdeling vervoersomvang naar vervoerwijze
mode	=	means of travel = vervoerwijze
option	=	alternatief
patronage	=	reizigers in openbaar vervoer
ridership	=	patronage
traffic	=	verkeer
travel	=	verplaatsen
transport	=	vervoer
transportation	=	vervoer
transit	=	openbaar vervoer
trip	=	verplaatsing
VOT	=	value-of-time = (reis)tijdwaardering
zone	=	deelgebied van studiegebied

---

lijst van vakwoordenboeken

H. Volker & E. de Wilde
Prisma Vakwoordenboek Transport (4-delig)
Prisma Taal, Uitgeverij Het Spectrum, 1996, ISBN 90.274.4462.5

---

0 Introduction to the course

0.1 Course objectives

The course "Transportation and Spatial Modeling" is a 6 credit course (6 ECTS) and consists of two parts. The module "Spatial Modeling" (1/4 of the study load) will be added to these course notes in the future and will for now be available as separate course notes and handouts. The objective of the module "Transportation Modeling" (3/4 of the study load) is to get insight and practice in the design and use of mathematical models for the estimation of transport demand in the framework of major strategic transportation planning. The course consists of a number of lectures and several exercises in OmniTRANS.

In the lectures the following subjects will be presented:
- the functions of models in the transportation system analysis.
- types of models and their applications.
- theoretical foundations (travel choice theory).
- aggregated models for trip generation, distribution, model split and network assignment.
- disaggregated choice models.
- estimation of model parameters and calibration.

The exercises have two functions:
- getting acquainted with and learning about practice-oriented software that deals with transportation calculations.
- solving a transportation planning problem with the use of the relating model instruments.

0.2 Final attainment level

After studying the course module "Transportation Modeling" the students are expected:

A. To have knowledge of the structure of the modeling analysis process in transportation planning, of the related computational models, their theoretical foundations and their behavioral backgrounds.

B. To have insight into the operation of the quantitative analysis process in transportation planning, in the derivation, the operation and the application possibilities of the different types of transportation models, as well as in the estimation process of model parameters based on travel and traffic observations.

C. To have skills in:
   - building a system description of a transportation network.
   - setting up simple operational mathematical models.
   - applying different types of models for the calculation of the transportation demand.
   - interpreting model results.
   - working with software for transportation calculations.

0.3 Course material

For the course the following material is available. To successfully take the exam, we advice you to study the following material:

1. P.H.L. Bovy, M.C.J. Bliemer & R. van Nes
   Course CT4801: "Transportation Modeling"
   Delft University of Technology, Transportation and Planning Section, 2006
2. P.H.L. Bovy, M.C.J. Bliemer & P.C.H. Opstal
   Course CT4801: "Transportation Modeling – Exercises in OmniTRANS"
   Delft University of Technology, Transportation and Planning Section, 2006
3. Various authors
   Course CT4801: "Spatial Modeling"
   Delft University of Technology, Transportation and Planning Section, 2006
4. Various handouts

It is recommended to study also the following material or use it as a reference book:

5. J. de D.Ortúzar & L.G.Willumsen
   Modeling Transport, 2nd edition
   Wiley, 1994;

For preparation of the exam, there will be questions and answers from previous years available on blackboard (http://blackboard.tudelft.nl).

0.4 Connection to other courses

The figure presents the forward, backward and parallel relations of this course with other courses.

Necessary prior knowledge:
- Spatial and Transport Planning CT2071 (general background)
- Transportation CT3041 (general background)

Preferable prior knowledge:
- Statistics and Operational Research (because of the used techniques)

Forward knowledge:
- Advanced Transportation Modeling and Network Design CT5802
- Dynamic Traffic Management CT5804
- 4th Year Project

Parallel knowledge:
- Traffic Flow Theory and Simulation CT4821
- Data Collection and Analysis CT4831

0.5 Exam regulation

There will be a written and an oral exam at the end of the course "Transportation and Spatial Modeling". The written exam discusses open questions, multiple choice questions and arithmetic questions. The practice consists of several exercises of which the final exercise will be graded in an oral exam. Admission to the written exam will be given after a successful completion of this last exercise. The lecture notes and the manual of the exercises are minimal exam material. The final grade will be a weighted sum of 3/4 of the written exam grade plus 1/4 of the oral exam grade.

1 Introduction

1.1 The role of models in transportation planning

The transport of people and goods from one place to another takes place using transport means such as ships, trains, cars or bicycles. This gives rise to the movement of ships, train traffic, car traffic and bicycle traffic. Transport means, transport services and infrastructural facilities for moving and parked vehicles such as railway tracks, stations, roads, parking lots, bicycle lanes and bicycle parkings combined with organizational (time tables) and control equipment (traffic lights) together make up the transportation system.

Transportation science investigates the characteristics of this system and its various modal subsystems. These characteristics refer to the design, the use, the maintenance and the operations of the system and its elements.

A good operation and performance of the transportation system permanently requires decision making. These decisions are being taken either as part of traffic and transportation policy making or as part of the operation and control of the system. In order to reach good decisions well-founded knowledge is required on the expected impacts of proposed decisions. The required knowledge depends on the type of decision to be taken. The time horizon for which this knowledge is demanded also plays a role.

Civil engineering structures take a long time (often more than ten years) to their final realization. Usually they have a very long lifetime that implies an estimation of its use over a long period of time is necessary. The way of routing a public transport line affects its ridership; a change in the routing may negatively influence its use. In order to keep network changes at a minimum, one needs therefore information about its potential use over a future period of say five years.
Control of a road crossing by an automatic traffic light requires data on waiting and departing cars at certain instances. In such a case, the long term development of car traffic is barely relevant. We may conclude that the type of problem (planning, design, control) determines the required knowledge and its use.

Also the type of situation where a facility is planned influences the type of information that needs to be collected. Extension of the road network in densely populated areas (such as the Randstad) is very difficult; one is forced to look for solutions offered by alternative transport modes. This differs from scarcely populated regions where knowledge about environmental impacts of road construction is of prime importance.

For the sake of sound decisions, factual knowledge about the current use and impacts of transport facilities is necessary. However, also forecasts about future use and impacts are necessary as these may result from autonomous developments or from policy making, such as changes in the transport system (transport fares, line routing, new infrastructure, transport service quality, etc).
In order to make valid forecasts, we need sufficient insight into the factors that will influence the future transportation system (e.g. income, gasoline prices, demographic composition etc). This insight among others can be gained through careful analysis of developments in the past. On the basis of such studies over the past 30 years, a vast body of scientific knowledge has been built up, which is summarized into a transportation theory and made operational in mathematical models. This theory and its models are continually being extended and tested. To this end, empirical observations are necessary with which the parameters of the models are estimated and tested using statistical estimation procedures. The current or future situation then can be calculated using these models.

Models may be used as analytical instruments in a 'what-if' analysis to show the impacts of changes in the system or its environment, or they may be used as design instruments that calculate the optimal design of the system in order to achieve maximum performance. In this course, the focus is on analytical models that can predict the impacts (e.g. level of congestion) of system changes (e.g. new roads).

1.2 Models for transportation analysis

Models are a simplified representation of a part of reality. Their function is to give insight into complex interrelationships in the real world and to enable statements about what (most probably) will happen if changes occur or put in that (part of) reality.
Models in transportation planning are abstract mathematical models, put into the form of (systems of) mathematical equations in which the behavior of a dependent variable Y (e.g. the number of daily train passengers in the Randstad area) can be derived from one or more explaining or independent variables X (e.g. car ownership in Randstad, train fares, etc) and related parameters a.

$$Y = f(a, X) \tag{1.1}$$

The parameters describe the sensitivity of Y to a unit change in X.
We distinguish so-called analytical models and design or programming models.

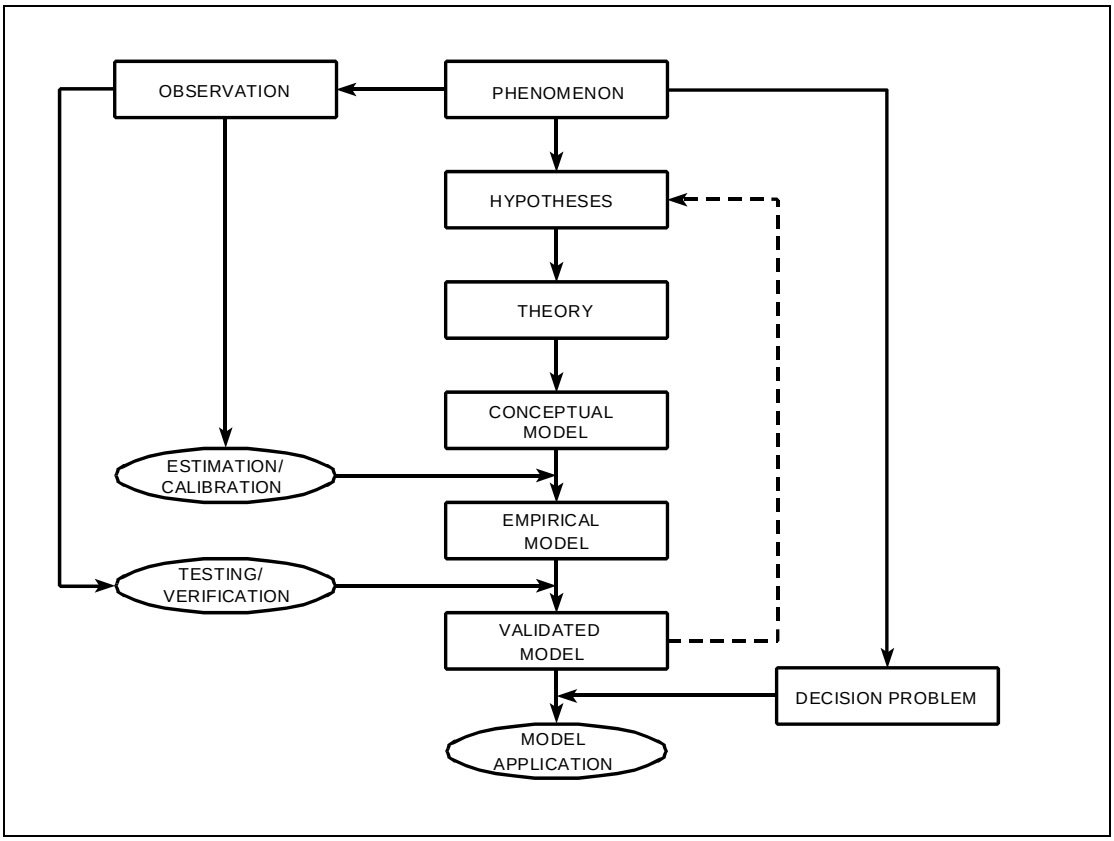


Figure 1.1: *Development process of a model*

The main purpose of analytical models in transportation analysis is enabling 'what-if' calculations. This means the calculation of expected effects in the transportation system if changes (policy measures or interventions) are put in that system (e.g. train fare increase) or if

autonomous changes occur (e.g. population decline). For such applications a validated model is necessary, which means a model of which has been shown with empirical observations that it replicates sufficiently accurate the behavior of the modeled system.

Figure 1.1 describes the process of development of a validated model that is ready for application in a planning problem.

On the basis of observations and reflection hypotheses are formed about the behavior of a part of reality that one likes to represent in a mathematical model for sake of a particular analysis purpose. We call a structured set of interrelated well-defined hypotheses a theory. The translation of such a theory into a quantitative model with quantifiable variables and associated parameters we call a conceptual model. This model is still abstract, its parameters that characterize the system at hand are not yet determined numerically. A possible functional form for such a conceptual model is a linear additive model (1.2):

$$Y = aX_1 + bX_2 + etc. \tag{1.2}$$

In (1.2) Y is the unknown so-called dependent or to be explained variable, while the X_i are the known independent or explaining variables. The unknown parameters a and b may be determined by calibration or estimation using a series of observations about the phenomenon at hand (in this case observations of Y, X_1 and X_2). This results in an empirical model that more or less closely resembles what happens in reality. For example:

$$Y = 0.72X_1 + 3.25X_2 \tag{1.3}$$

If an empirical model has been proven to be able to make valid predictions of the dependent variable Y (using other data than from which the empirical model has been derived) then we call it a validated model.
The development of a validated model in practice mostly does not follow the simple linear process shown in Figure 1.2. It is much more a trial-and-error process with feed back loops to earlier steps in which premature ideas and assumptions are adapted because of insufficient matching of model outcomes and observations from reality.
Typical examples of analytical models are dealt with in Chapters 4 to 6.

The second category of models used in transportation planning are so-called design or programming models. Their purpose is to find those values of a set of design or instrumental variables X that lead to an optimal performance of the modeled system measured by the variable Y. In most cases the possible values for X are restricted to certain ranges due to all kinds of side constraints.

A design/programming model therefore consists of an objective function Y and a set of constraints defined on the X's:

$$\begin{aligned} &\text{maximize} && Y = f(a, X) \\ &\text{subject to} && g(b, X) > 0 \end{aligned} \tag{1.4}$$

Example 1: find the train fare values X_i for different user classes i such that the train patronage Y is maximal. Or: in a given network and with a given travel flow pattern between origins and destinations, determine the traffic loads X_i of all links i that minimize total travel time Y in the network where all X are positive and satisfy flow conservation (flow in equals flow out) at nodes. Example 3: find the set of capacity expansions X_i for all links i of a road network that minimizes total congestion Y subject to a road construction budget constraint Z.
In this course, programming models are used in Chapter 7 on network analysis.

1.3 The use of models in transportation analysis

Transportation models are being used to make predictions and forecasts of future changes in usage of traffic facilities for sake of facility design, control and operation.
Future changes in travel, transport and traffic may be the result of autonomous developments, may be caused by economic, social or spatial policy, or may follow from transportation or traffic measures.

- Autonomous developments result from demographic changes (e.g. migration), increased car ownership, income changes, international economic changes (e.g. oil price). Characteristically, these developments rarely may be influenced by transportation planning.
- Typical for economic policy (e.g. gasoline taxes), social policy (e.g. working hours, labor force participation) and spatial policy (e.g. reduction of agricultural land use) is that these policies may have an impact on transportation but are not developed in the transportation field for transportation related purposes.
- Transportation policy refers to plans designed by transportation professionals to change the transportation system directly (road or rail, public or private, car or bicycle, etc).
- Traffic measures refer to changes in the operations of the traffic system such as traffic lights, parking, public transport services, etc. Decisions at this level mostly are of a technical kind.

However, models also are used to derive potential measures that can influence the transport system, to analyze the effectiveness of measures, and to show side effects of such measures. In addition, transportation models are used to design transport and traffic facilities and services such as the required road capacity, optimization of network structure, design of environmentally friendly road alignments, improved line routings and services in public transportation networks, design of intersection traffic control lights, etc. In the sequel, a number of such applications will be worked out in more detail.

1.4 The function of models in the planning process

Models serve several purposes in transportation analysis. They help among others to gain in a more structural way insight into complex interrelationships. The mathematical equations of the model enable to find out which factors play an important role, and how sensitive the dependent variables are with respect to changes in the independent variables. This 'insight' function of models is an important value on its own, because it allows armchair paper-and-pencil analyses.

A further important function of models is the quantification of expected impacts of proposed plans: the prediction function (see Figure 1.2). In transportation planning alternative plans or actions are developed in order to solve existing or expected problems. The effectiveness of these plans may be measured according to the extent to which these plans fulfill prior established assessment criteria.

Because of the complexity of most problems and related proposed solutions, only models can enable the estimation of the impacts that will result from implementing each of these solutions. In transportation planning examples of such impacts are: traffic loads, travel times, shifts in modal split, congestion level, etc.
In order to arrive at a well-considered decision concerning the solution to be chosen, it is good practice in transportation planning to model and analyze a number of cases (see Figure 1.3). Apart from the proper planning variants these include also a do-nothing (zero base) case and the current situation. In the case of road infrastructure planning (such as to tackle

congestion problems) it is nowadays good practice to analyze at least each of the following three variants:

- environmental variant: this is the plan that can solve the problem with the least environmental damage;
- public transport variant: this plan solves the problem by improving public transport supply and services as far as possible;
- construction variant: the best plan if new road construction is chosen as the solution approach.

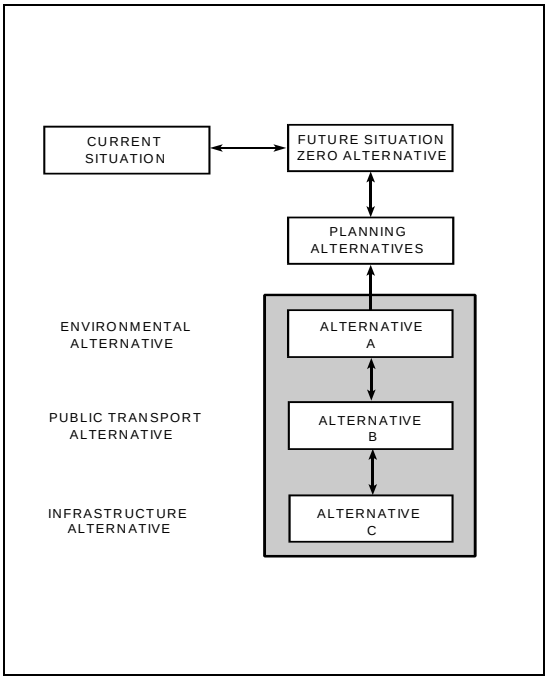


Figure 1.2: *Evaluation of transportation plans using various cases*

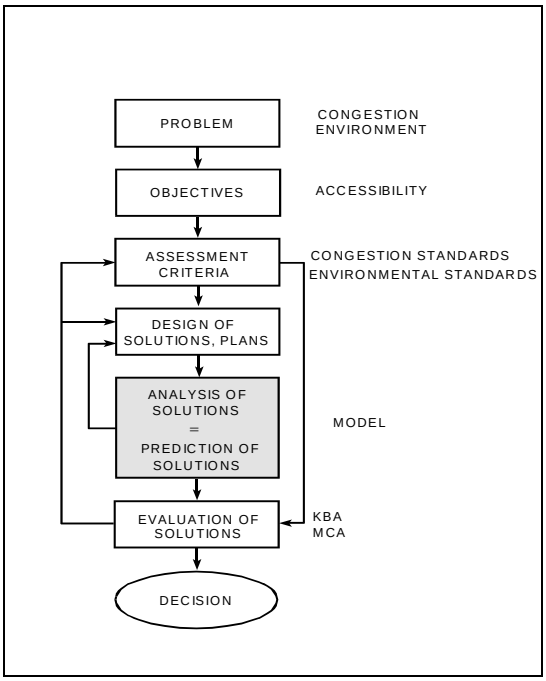


Figure 1.3: *The planning proces*

In order to be able to assess the value and effectiveness of these planning variants, the impacts of these plans will be compared to the corresponding impacts of the do-nothing case and the current situation. For all these cases, model computations are necessary to facilitate comparison of the impacts of interest (congestion, traffic loads, travel times, etc).

Roughly, the set up of such calculations is as shown in Figure 1.4. For each variant a quantitative system description is established. This description in fact gives the values for the independent variables X of the mathematical model. Basically, two classes of variables may be distinguished here, that is the variables that describe the transport system (the networks with nodes and links, their capacities, speeds, etc) and the variables that describe spatial distribution of transport generating activities (inhabitants, employment, retail floor space, etc per zone).

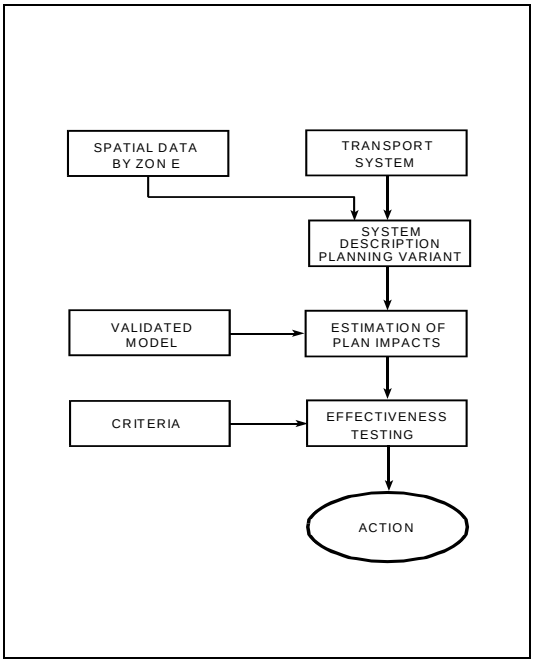


Figure 1.4: *Analysis set up of a plan situation*

1. Description of current situation
   Estimation of current conditions

2. Description of new situation
   Adapting model functions (parameter values)
   Adapting values of independent variables
   Estimation of new conditions

3. Comparison of new and old situation
   Assessment of changes in conditions

Figure 1.5: *The proces of prediction*

1.5 Applications of models in transportation planning

In order to place the following chapters into context, we will give a brief overview of typical applications where a modeling approach may contribute.

1.5.1 Forecasting future developments

In the coming decades the developments in the growth in car and truck use and the stand still or decline of public transport use will remain important questions. For several reasons we need to know whether the growth in car use will continue in the future.
There are a number of societal developments that require a critical look at traffic forecasts. On the one hand we observe an increase in energy prices but also an increase in incomes. However, available leisure time increases. The number of households with young children is declining giving their mothers more opportunities to participate in out-of-home activities. The number of one-person households increases which gives rise to a higher use of cars. Part time employment most probably boosts mobility. Decentralization of housing and work is increasing leading to increasing travel distances and thus car use. Technological innovations are directed at lower energy use. Clean and fuel-efficient engines are developed and cars will be lighter due to the use of plastics. Our prediction models need to be able to estimate the probable impacts of such developments.
Knowledge of the spatial distribution of trips, the modal split, and the spreading of peak periods are necessary for taking sound decisions. Of special importance is to know to what extent traffic during peak hours at bottlenecks will be influenced by the aforementioned developments.

1.5.2 Planning of new or improvement of existing infrastructure

Spatial planning impacts
The geographical location of housing, working and leisure areas influences transport flows. Transportation analysts are asked to calculate the probable transport and traffic implications of spatial planning. This refers e.g. to the consequences of new or enlarged residential areas or changes in employment levels in new business or industrial zones for the size of traffic flows. However, spatial developments on their turn depend on the quality of the transport system. Models are being developed that can indicate the implications of transportation planning actions for spatial changes. Such model applications are important for the calculation of:
- transport flows that are consistent with projections of population and employment;
- impacts of bottlenecks in the road network on spatial developments;
- the influence a good public transport policy can have to counteract further desurbanization of the big cities.

Improvement of the road and railway network
Capacity extension of roads and new roads and railways cause costs in terms of money, noise, exhausts environmental damage, spatial quality, etc. Gradually, the amount of available space declines more and more. Understandably, resistance against building of roads and railways increases. This demands for more attention to environmentally friendly design of facilities for private and public transport. Infrastructures need to be build such that societal costs are minimal. Models are available that help in optimizing infrastructure networks and routings. Because of the strong resistance against new infrastructures, pertinent decisions need to be motivated convincingly which requires very accurate and valid prediction models.

Better utilization of available capacity
Currently, the capacity of roads is underutilized due to among others the lack of information on the part of the drivers about the prevailing traffic conditions, but also due to a poor control of traffic flows. Models can help in calculating route and departure time advises for travelers (e.g. using Variable Message Signs, VMS) such that they will experience and thus cause less congestion. Models can assist in designing traffic lights and ramp metering facilities to control traffic flows such that less congestion will occur. The same holds for the design and

control of buffers in motorways. Also planning and design of special target group facilities such as carpool and truck lanes may be facilitated by advanced traffic flow models.
The planning of maintenance of the roads (where, when, and how) is another issue where models can assist in decision making that minimizes costs and troubles (time losses) to the road users.

Stimulating public transport use
Because of the negative impacts of car use and the societal goal to enable a minimum level of mobility to everyone, sufficient public transport provision is a must in our society. If public transport has to assist in curbing car use it needs to become a serious competitor to the car and thus asks for high quality services. Models can help in designing the set up of networks of public transport lines, can help in calculating the patronage of the services given, and can indicate to what extent new or improved services indeed attract car users.
An important issue is the strong interaction between land use and public transport provision. Public transport stations and stops often are the nuclei of new activity developments and thus influence urban patterns, on the other hand a dedicated land use planning is necessary in order to give sufficient base demand for an efficient supply of services. Modeling this interaction between land use and transportation is at the heart of the transportation profession.

1.6 Course overview

This course deals with the theory and application of transportation models for strategic planning purposes. The models aim at calculating the level of usage of transport facilities: the traffic loads of roads, the patronage of railway lines, the number of transfers at stations, etc. In parallel with these flow levels, other characteristics are calculated with these models: routes, level and duration of congestion, travel times, flows between regions, service quality, etc.
In the calculation process of traffic and passenger flows five steps are distinguished which refer to five different quantities of traffic volume. Figure 1.6 shows an overview. The right column shows topdown the five steps that are distinguished in modeling travel behavior (therefore also called the 5-step model).

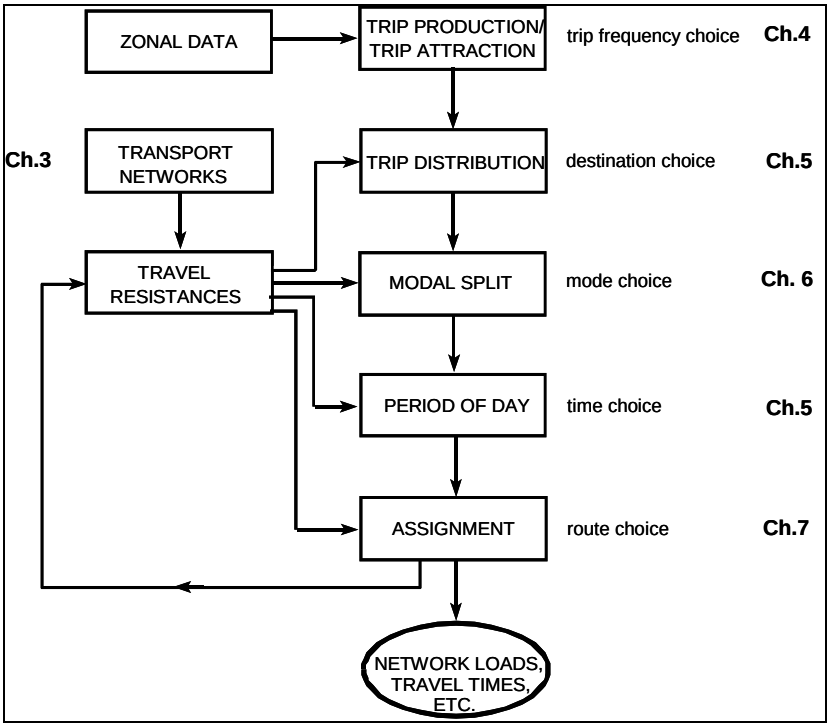


Figure 1.6: *Travel choice model system*

Step 1: trip production and attraction
This submodel describes how often people make trips as a function of their personal and spatial conditions. Summed over all elements (e.g. inhabitants) of a zone of the study area this step results in the numbers of trips (by category) that begin and end in that zone as a function of zonal variables. The models used in this first step are dealt with in Chapter 4.

Step 2: trip distribution
This submodel describes where the trips identified in step 1 go to. It links origins and destinations of trips among others as a function of the travel quality between zones. Summed over individuals this step results in a table giving the number of trips during a certain period between each pair of zones. The models of this step will be treated in Chapter 5. The estimation of trip tables based on observations is discussed in Chapter 8.

Step 3: Modal split
The submodel of this step determines which transport modes the travelers will use for their trips. This depends on the characteristics (travel quality) of the respective competing modes between each zonal pair. The summed individual choices result in the so-called modal split, the distribution of the use of the various modes (car, bicycle, bus, train, etc). This model type is presented in Chapter 6.

Step 4: Period allocation
This submodel determines when trips will be made, in which period of the day (e.g. morning or evening peak period or off-peak).

Step 5: traffic assignment
This last step deals with how travelers choose routes through the networks. By aggregating over individuals this results in traffic loads of routes, sections, and intersections. Chapter 7 gives a detailed account of these models.

Travel resistances or travel quality (of which travel times, costs, distances are a part) play a great dominant role in nearly all travel choices (see arrows in Figure 1.6). Separate models calculate these trip resistance measures from the characteristics of the network elements (capacities, speeds, etc) and from the loads in these network elements (shown in Chapter 3). In principle, these five travel choices are modeled according to the same theoretical principles about traveler choice behavior. This common base is explained in Chapter 2.

The calculations often are not performed in the linear way suggested in Figure 1.6. Some of the choices (such as destination and mode choice) often are combined into one single step, in a so-called simultaneous model. Since travel resistances depend on the level of usage of the network elements (the higher the loads, the higher the resistance), and since this level of usage only is known at the very end of the calculation cycle (assignment step), an iterative calculation procedure is required in order to achieve consistent final results. In practice, this iterative cycle is performed several times for each distinct analysis case.

1.7 References

J. de D. Ortúzar & L.G. Willumsen
Modelling Transport, 2nd Edition
Wiley, 1994

2 A theory of travel choice behavior

2.1 Travel choice

Trip making is the result of individual choice behavior. An individual person decides whether he will leave home to do an activity elsewhere or not, where he will do that activity, how he will travel to that destination, etc. This need not imply that he/she makes all these choices anew everyday, but at a certain moment he/she takes a conscious decision that develops into a habit (habitual behavior). The same applies for a firm that wants to ship goods.

Travel choice behavior is closely related to other (more strategic long-term) choices of the individual person or firm: among others location choice (of home or firm) and mobility choices such as car ownership (for details see course CT3751). The following exposition is limited to travel choices only but the presented theory is equally applicable to other choice types as well.

We distinguish five trip making choices which are hierarchically related to each other:
1. activity choice: whether or not to leave home to do any activity elsewhere (e.g. shopping);
2. destination choice: where to perform the out-of-home activity (city centre? neighboring town? residential neighborhood? etc)
3. mode choice: how to go there and travel back (by foot? bicycle? car? bus? etc);
4. time choice: when depart for the trip? (early morning? etc);
5. route choice: which route to take in the road or public transport network.

The hierarchy means that the characteristics of a trip on a lower level influences the trip choices on higher levels. As an example: the chosen route determines the travel time needed for a trip, which in case of congestion influences the best moment for departing. The traveling period sometimes determines the level of fare one has to pay for public transport that in turn will influence the modal choice between public transport and other modes. Travel durations and travel costs finally are considered by the traveler in his decision where to perform his desired activity.

Characteristic for all these trip choice types is that there are alternatives. Alternatives are mutually exclusive ways (possibilities, opportunities) for making the trip of which only one can be chosen at the same instant. Of the many destinations for an activity, or the many modes or routes to a destination one can use only one. The type of alternative in all these cases of trip making are of a discrete nature: the alternatives are units which cannot be compared on a single scale (bus and bicycle, or routes A and B are simply completely different things). This differs for example from the situation of buying goods where the alternatives often are formed by different quantities of the same product such as buying one or two pounds of sugar.

We assume selfish and rational choice behavior of the travelers taking into consideration the standpoint and information of the decision maker himself. The first assumption means that the traveler decides according to his own personal views and preferences and tries to optimize his own personal situation. Thus, no collective or at the collectivity oriented decisions are assumed. Rational behavior means that the traveler bases his decision on the characteristics of the various alternatives open to him (thus not on the position of the moon or on throwing a dice).

There is not always a choice. Especially in the case of mode use often there is only one possibility (no other options nor alternatives), e.g. going by bicycle for people having no car

or driving license nor having a bus stop near home or school. Travelers having no travel alternatives are called captives, the others are called choice travelers.

The mentioned five travel choice types each constitute separate choice problems which may be solved after each other. This is called a sequential choice process. In many analyses however some choices are taken in common, especially destination and mode choice. These choices are so closely related that one may treat these as a single combined choice. The choice alternatives no longer are separate destinations and separate modes but the alternatives consists of combinations of a destination and a mode. In such cases we speak of simultaneous choices. As an example consider someone living in Delft who wants to make an evening concert visit: alternative A is to go by train to Amsterdam whereas alternative B is to go by tram to The Hague.

Travel behavior differs according to the purpose of the trip: the activity to be done. The distinct travel purposes such as working, shopping, recreation, etc are separate markets with their own laws. The importance of travel time in the trade off of alternatives differs between trips for home-to work trips and shopping. Therefor, in planning practice different choice models are used for different travel purposes.

A final remark. Although the trip is considered as the unit of analysis in this course, choice making often is influenced by the characteristics of the tour, that is of the trip chain from which the trip is a part. This means for example that the mode choice for the home to work trip also is influenced by the characteristics of the work to home leg. Our modeling approach implicitly assumes that such influences are take into account.

2.2 An individual utility theory of travel behavior

We apply the micro-economic utility theory to understand and explain the travel behavior of trip makers. With this theory we can explain very satisfactorily and consistently a fairly broad class of travel and transport phenomena. In addition, this theory enables the derivation of calculation models with which travel choice behavior can be quantified.

Let us take the most important choice, that is to do an out-of-home activity elsewhere, or not. To do this, a chain of at least two trips has to be made. For simplicity sake, in the sequel we speak also in such cases of a trip although often we have to do with a tour.

Assumptions of utility travel choice theory:

1. In the decision to make a trip three utility components play a role:
   - the utility N_i of staying where you are at origin i:
   - the utility N_j of doing an activity of class p at destination j;
   - the disutility (cost) Z_{ij} of traveling between i to j.

   A trip will be made (otherwise said: an out-of-home activity will be performed) only if the net utility is positive:

   $$N_j - N_i - Z_{ij} > 0 \tag{2.1}$$

   For simplicity reason (and without loss of generality) we choose the utility scale of activities such that $N_i = 0$. Also, because we consider the activity classes as independent we omit the index p (purpose) in the sequel.

2. The individual attaches a positive utility N_j to each activity in j. This utility depends on the type of activity p and its characteristics. The total utility N_j is composed of part-

utilities n_{gj} that result from the distinct characteristics g of activity p at location j. Each individual values the activities (and their characteristics) in an idiosyncratic way, that means: according to his own personal preferences:

$$N_j = \sum_g n_{gj} \tag{2.2}$$

3. The traveler attaches to the trip (or trip chain) needed for the activity a disutility Z_{ij}. This disutility depends on the characteristics of the trip or trip chain (mode(s), duration, monetary costs, etc). The total trip disutility is composed of disutilities z_{hij} that result from the distinct characteristics h of the trip or tour (waiting, parking cost, driving time, etc). Each traveler values the trip resistance in his own idiosyncratic way:

$$Z_{ij} = \sum_h z_{hij} \tag{2.3}$$

Figure 2.1 gives an illustration for the case of shopping.
N_x = utility of shopping in x (x is j, k, l respectively)
Z_{ix}^m = disutility of traveling between i and x using mode m.

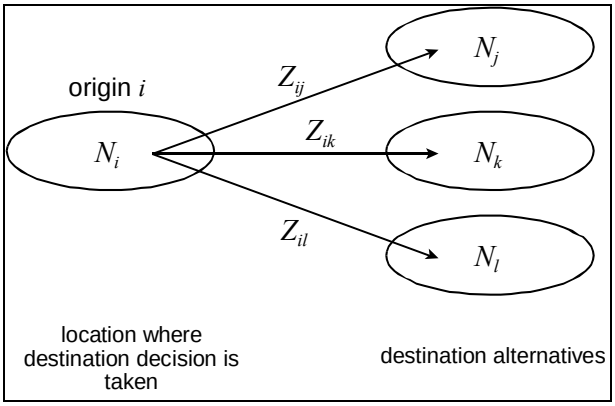


Figure 2.1: *Destination choice (mode is given)*

4. A trip for an activity will only be undertaken if the net utility U_j is strictly positive:

$$U_j = (N_j - Z_{ij}) > 0 \tag{2.4}$$

5. The individual cannot engage in all possible activities because of limitations in available time and money.
   a. There is a limited money budget K equal to his income. This budget K is the sum of all expenditures k_j to activities and trips j.

$$K = \sum_j k_j \tag{2.5}$$

   b. There is a limited time budget T available to out-of-home activities including time needed for related trips. This is the sum of all time expenditures t_j needed for activities and transportation j.

$$T = \sum_j t_j \tag{2.6}$$

6. The traveler chooses the alternative *j* that gives him highest net utility *U* and fits within his time and money budget constraints: the principle of subjective utility maximization. If we denote the chosen alternative with *c* then it holds:

$$U_c = \max(U_j) \tag{2.7}$$

Explanation:
In choosing an alternative option, only the order of the utilities plays a role but not their absolute levels. Because of this there is a great flexibility in choosing the function types for utilities *N* and disutilities *Z* (as long as the order will not be influenced). The utility *N* and disutility *Z* can be described as a (weighted) sum of part-utilities *n* and *z* respectively. This weighted sum expresses the trade-offs the traveler makes between the characteristics *X* and *Y* respectively of an activity or a trip (chain).

$$N_j = \sum_g n_{gj} = \sum_g (\beta_g X_{jg}) \tag{2.8}$$

In case of choosing a shopping centre for example, the individual balances activity aspects such as assortment and quality of products using weighting factors β. He is for example willing to accept a lower quality of a products if it is compensated by a broader assortment from which he can choose.
In trip decision making the traveler makes a trade off between travel aspects such as waiting time, distance, costs and parking service using trade-off factors α. He may be willing for example to pay a higher price if the trip is faster. The specific trade-off between time and money costs via the so-called value-of-time is an important parameter in travel choice behavior (see Section 3.4).

$$Z_{ij} = \sum_h z_{ijh} = \sum_h (\alpha_h Y_{ijh}) \tag{2.9}$$

The balancing of utility *N* and cost *Z* indicates that the individual also makes a trade-off between the characteristics of activities and those of trip making. In shopping travel for example the utility of a larger assortment in a particular centre will be traded off with the extra travel time needed to get there. Because of the existence of limited time and money budgets there is anyhow a trade-off between time and money elements of the activities and associated trips. The money and time expenditure for a concert visit for example are balanced with the costs and travel time of the associated trips.

Note: in these considerations it is assumed that the decision to travel (production/attraction) was already taken and was positive.

Figure 2.2 illustrates this way of utility maximization. If we assume a rational consumer who maximizes his own personal utility, he will decide to choose destination *k* although the activity utility in *j* is larger.

The formulated choice theory is generally applicable to all types of travel choices separately (mode, route, time, etc) or partly combined (e.g. mode and destination).

In the case of trip frequency modeling (also called trip production or trip generation models) per activity type the following explaining variables normally play a role:
1. household composition (type and size)
2. person type (occupation etc)
3. age, gender

4. education
5. car ownership, driving license ownership
6. area type of origin.

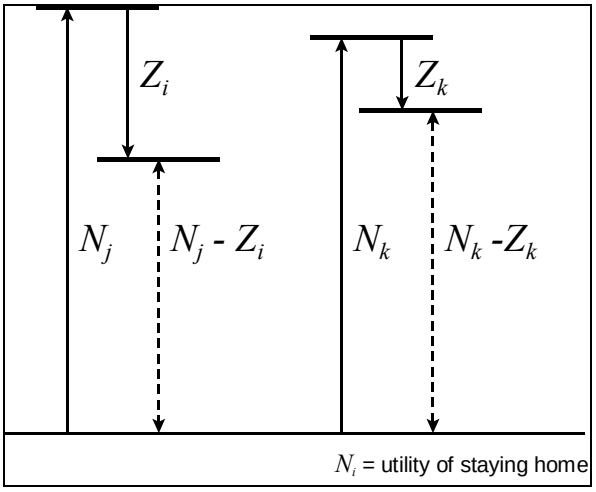


Figure 2.2: *Utility maximization in trip making*

In this case the alternatives from which can be chosen are 0,1,2,3 etc number of trips per activity type (trip purpose) per time unit (day or week). The accessibility quality of the origin areas mostly is not considered as a useful explaining variable. The weighing of variables differs according to activity type.

The destination choice per activity type (also called trip distribution) usually is governed by the following attraction variables of competing destinations:
1. accessibility quality (travel times and costs; available modal alternatives)
2. population size and density
3. employment size and density in various sectors
4. educational supply
5. existence of a city centre

Also here, the relative weighing of factors depends on the activity or trip purpose type.

In modeling mode choice by trip purpose (also called modal split) mostly the following person and trip variables are used (apart from a mode specific constant):
1. travel distance
2. travel time
3. travel money costs
4. parking costs
5. toll costs
6. driving license and car ownership
7. person type (age, gender, etc)

Travel times usually are taken as a weighted sum of those of the distinct trip parts (access, egress, line haul, waiting, transfer, parking etc). In most applications, destination and mode choice are combined into one simultaneous choice decision.
The mode-specific constant expresses the influence of all non-measurable characteristics of each mode, such as comfort.

In the case of route choice (which mostly is not analyzed separately for activity types, but it is for each mode) the following variables are put into the models:
1. travel times (including congestion time loss)

2. travel costs (including tolls)
3. travel distance
4. delays at crossings or transfer points
5. occurrence of congestion

Also here, travel times may be a weighted sum over trip parts (especially in the case of route choice in public transport networks).

With departure time choice the following time-dependent variables play a role in modeling:
1. traveling times (including congestion time loss)
2. traveling costs (including tolls)
3. occurrence of congestion

For further details about definitions of variables and about values of weighing parameters, see for example the documentation of the National Model System, the Randstad Model or the Wolocas model.

2.3 A general travel choice model

The principle of subjective utility maximization (Section 2.2) can be transformed into a broadly and easily applicable computation model with which levels of usage of modes, routes, and networks can be calculated.
A traveler attaches to each alternative i a net utility U_{ip}. This is a weighted sum of valuations of activity utility and travel resistance components:

$$U_{ip} = N_{ip} - Z_{ip} \tag{2.10}$$

Part of this net utility is not observable for the analyst (model maker) because he is unable to know all factors a traveler takes into account nor is he able to measure known influencing factors. Because of this, the analyst defines a separate utility component ε_{ip} (a so-called error term) for each alternative that represents the non-observable utility of alternative i of a traveler p. This error term ε_{ip} follows a probability distribution having as expectation the known observable net utility part V_i. It thus indicates how much the calculated utility V_i deviates with a certain probability from the unknown actual utility U_i:

$$U_{ip} = V_i + \varepsilon_{ip} \qquad E[\varepsilon_{ip}] = 0 \tag{2.11}$$

The observable utility part V_i is a weighted sum of observable characteristics of the considered alternatives, for example travel time components X_k :

$$V_i = \sum_k \beta_k X_k \tag{2.12}$$

where the parameters β_k represent the relative weight of each of the influencing factors X_k. The value-of-time might be for example one of these parameters.

The individual chooses option i from multiple alternatives if:

$$U_{ip} > U_{jp} \quad \forall j \neq i \tag{2.13}$$

thus if:

$$V_i + \varepsilon_{ip} > V_j + \varepsilon_{jp} \quad \forall j \neq i \tag{2.14}$$

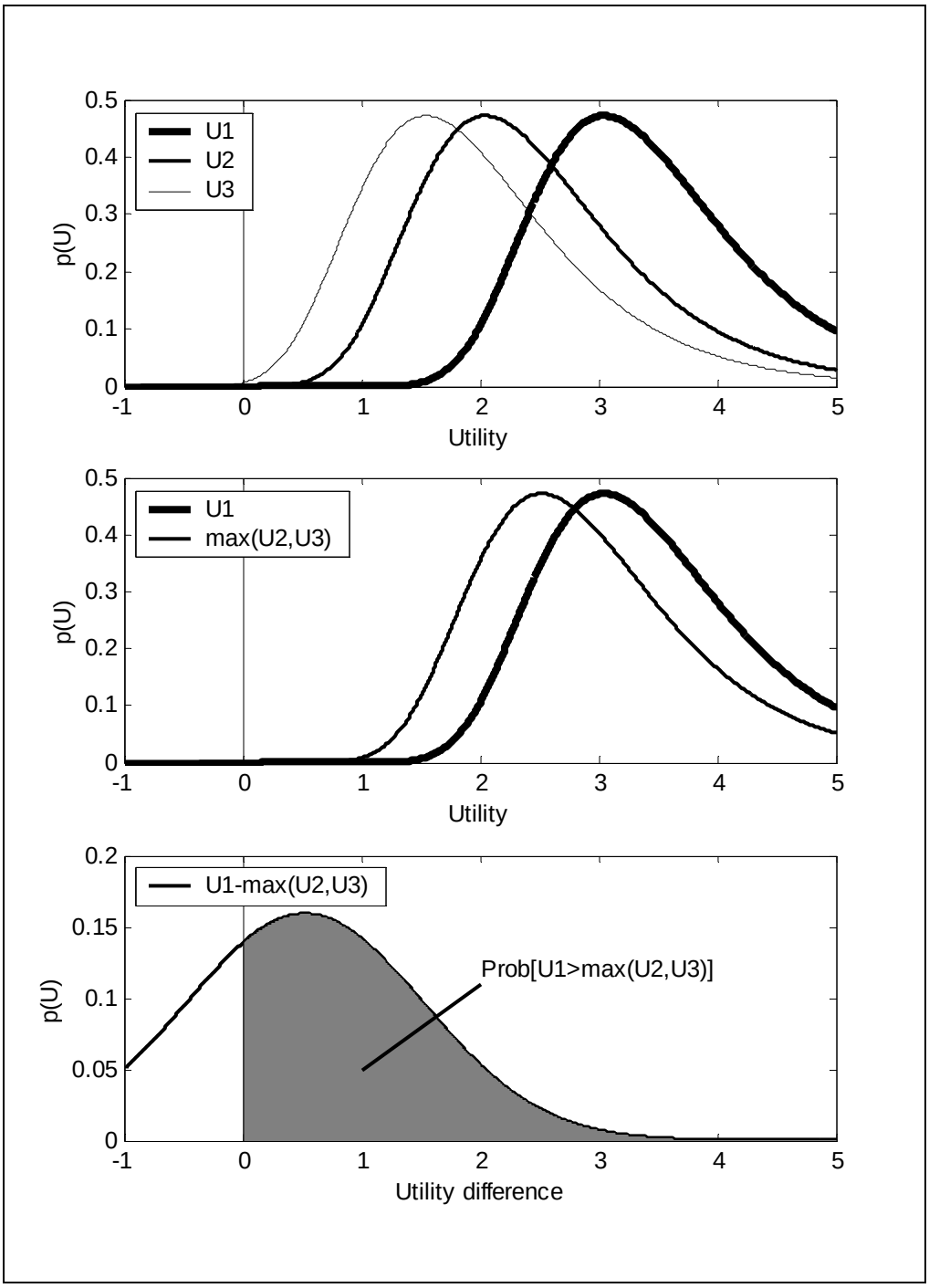


Figure 2.3: *Graphical illustration of the derivation of choice probabilities. The top graph shows the Probability Density Function (PDF) of the subjective utility for three different alternatives. The middle graph shows the PDF of the maximum of utility 2 and 3, and the PDF of the utility of alternative 1. Finally, the lowest graph shows the PDF of the function defined as the difference between the utility of alternative 1 and the maximum of utilities for alternatives 2 and 3. The probability of selecting alternative 1 corresponds with the size of the shaded area.*

2.4 Derivation of logit model

The probability of selecting alternative *i* is given by:

$$\text{Prob}[\text{select}(i)] = \text{Prob}[U_{ip} > \max_{j \neq i}(U_{jp})] = \text{Prob}[V_{ip} + \varepsilon_{ip} > \max_{j \neq i}(V_{jp} + \varepsilon_{jp})] \quad (2.15)$$

If we assume that ε_{ip}, $i = 1,2,...$ are identically and independently distributed according to a Gumbel distribution with scale parameter μ, then it can be shown that:

$$\text{Prob}[\text{select}(i)] = \frac{e^{\mu V_i}}{\sum_j e^{\mu V_j}} \quad (2.16)$$

with:

μ Variance parameter. This parameter depends among others on the units in which the characteristics of alternatives are expressed

This formula is known as the *logit* formula.

The proof of above proposition can be found in [Ben-Akiva and Lerman, 1985], pg106.

Intermezzo: The Gumbel distribution

The cumulative Gumbel distribution is given by:

$$F(\varepsilon) = \text{Prob}[x \leq \varepsilon] = \exp[-\exp[-\mu(\varepsilon - \eta)]] \quad (2.17)$$

The parameters μ and η are referred to as the scale and location parameter. The Gumbel distribution has the following properties:

- The mode equals η
- The mean of the Gumbel distribution equals $\eta + E/\mu$, with E = Eulers constant (~.577).
- The variance is given by: $\pi^2/6\mu^2$.

The Gumbel distribution has the following interesting property:

the expected value of a random variable that is defined as the maximum of a number of Gumbel distributed random variables with equal scale parameters en location parameters $\{\varepsilon_1, \varepsilon_2, ... \varepsilon_n\}$ *equals:*

$$\frac{1}{\mu}\log(\sum_{j=1}^{N}\exp[\mu\eta_j]) + \frac{E}{\mu} \quad (2.18)$$

In the derivation of the logit model, this property is important because it implies that if the utility of two alternatives is Gumbel distributed, the maximum utility of the two is again Gumbel distributed. This leads to the property of the logit model that it computes the joint share of multiple alternatives as that of a hypothetical alternative, of which the expected utility is given by (2.18) (see also Example 7.2).

To give an idea of the Gumbel distribution, we plotted the Gumbel distributed Probability Density Function (PDF) and Normal distributed PDF in one figure. As can be seen, the Gumbel distribution is somewhat similar to the Normal distribution. As can be seen, the Gumbel is not symmetric around its mean, and has almost no probability mass in its left tail.

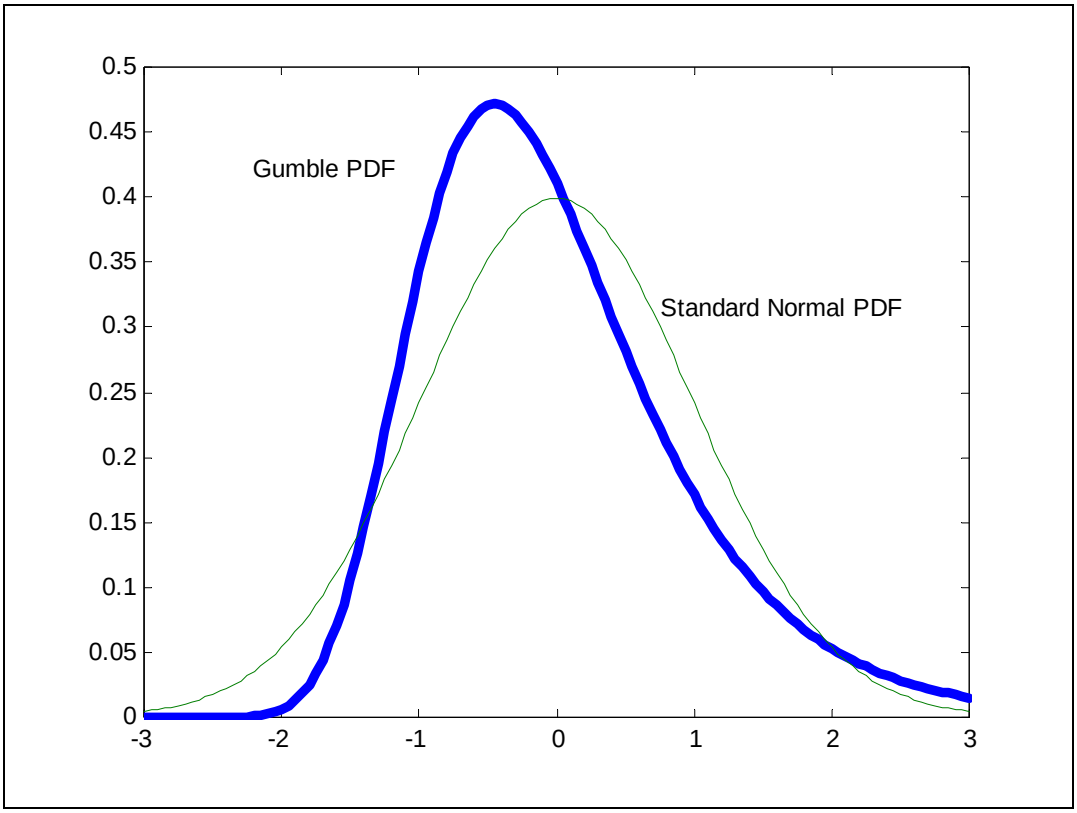


Figure 2.4: *Density function of standard normal distribution and zero mean Gumbel distribution with standard deviation = 1*

-end of intermezzo-

It is quite common to standardize the weight factors β_k in (2.12) for the variance parameter μ. In this case (2.16) changes in:

$$\text{Prob}[select(i)] = \frac{e^{\tilde{V}_i}}{\sum_j e^{\tilde{V}_j}} \tag{2.19}$$

with:

$$\tilde{V}_i = \sum_k \tilde{\beta}_k X_k \ , \ \tilde{\beta}_k = \mu\beta_k$$

Essential properties of the logit choice model are:

- the probability of use of the distinct alternatives depends on their differences in utility;
- the alternative having the highest observed utility (or smallest observed disutility) exhibits the largest probability of use.

If other assumptions about ε or U are taken as a point of departure other types of choice models will result. The logit model however is the most widely applied choice model type and forms the basis for this course, although for route choice analysis the so-called probit model is a better approach (see Section 8.3.2).

This type of choice model is applied at the level of individual travelers and therefore is called a disaggregated model (microscopic). In order to arrive at flows and loads probabilities are summed and grossed up according to sampling fractions.

Disaggregated models first of all are used to find out which factors have a significant influence in individual trip making. Then, these models allow the estimation of parameter values associated with these factors which results in values for the relative importance of each of these factors. Finally, these models can be used in making what-if predictions. To this end, a sample of persons is taken for each of which the choice alternatives and their characteristics are determined. For each person the probabilities of choosing each of his alternatives is calculated. Summation of these probabilities gives the percentage of travelers that will use a particular alternative.

In contrast we have so-called aggregate models (macroscopic) that model the behavior of larger groups of travelers whose behavior is assumed to be identical. Aggregate models work with average values for these groups. Aggregate models are easier in use but give less insight into travel choice behavior and are less accurate in their outcomes. This class of models is dealt with in Chapter 5.

2.5 References

M. Ben-Akiva & S. Lerman
Discrete choice analysis: theory and applications to travel demand
Cambridge, MIT Press, 1985

J.S. Cramer
The logit model for economists: an introduction
London, Edward Arnold, 1990.

3 Transportation system description: networks and data

3.1 Problem introduction

In reality, trips may start and end at every address in the world and may use all available networks, streets, services etc.
In order to solve a particular transportation problem there is no need to describe the system with the highest possible detail. Instead, in order to gain insight into the problem and its solutions, and to make decisions understandable, it often is better to simplify matters. Also to make calculations feasible, the analysis manageable, and the study costs bearable, a simplified description of reality is needed.
To this end, in each transportation study a system description needs to be designed (a so called systems model) giving answers to the following questions:
- which travel markets are relevant (only persons, or also goods? only commuting or also education? etc);
- which geographical area is relevant?
- which transport supply networks should be considered?

The answers to such questions highly depend on the problem at hand: the type of policy measures to be studied, the type of assessment criteria that will be used, etc. In studying a national high-speed rail line, the whole country and at least the direct neighboring countries should be considered. The airline network should be included in the network description but most of the road network need not to be considered in the analysis. In deciding on the set up of a local busline network, the study area need not to be larger than the town in question including its direct surroundings. A detailed description of the networks for car and bicycle travel is needed for the travel analysis.

Because of this high dependence of an adequate system model on the problem and all kinds of governing constraints at hand, designing such a system model is more of an art than a science. It is therefore impossible to give exhaustive detailed guidelines for this system design problem.
In the following we will restrict the treatment to the model design of the physical system: the geographical area and the networks. We will describe the type of activities that are needed in setting up a computerized systems description suitable for modeling analyses of travel demand.

3.2 Study area

The definition of the study area consists of several steps:
- the delineation of the study area
- the subdivision of the study area into zones
- the definition of zone centroids.

Study area delineation
The study area is defined as the area within which transport flows will be analyzed and modeled. It is defined by an imaginary curve on a map. All area outside the curve will not be considered as relevant for the problem at hand. It is assumed that there is no travel demand generated outside the study area. Only within the study area there will be a subdivision into zones and there will be a description of traffic networks.
The size and position of the study area depends on the problem at hand.

In a study of national importance, the whole of Europe most probably will constitute the study area. In a study of local importance, normally the national boundaries delineate the study area, but this depends on the situation at hand. In the case of South Limburg region it is wise to extend the study area into Belgium and Germany because of the strong cross-border interrelationships.

Zonal subdivision
The study area is subdivided into a number of zones, also called traffic zones. These zones represent at a coarser level the origin and destination addresses of individual trips. There is no need to know these addresses at a very detailed level of geographical accuracy, nor is that possible from a cost or data availability point of view. The zonal subdivision implies that travel flows within a zone cannot be analyzed because there is no geographical reference. The level of detail of the subdivision, that means, the number of zones, their individual and average size, their delineation relative to neighboring zones depends on the problem and constraints at hand.
The zones are the geographical units between which the travel flows will be calculated. They are therefor an important geographical and administrative unit in a study for presenting spatial patterns of transportation and for data management.

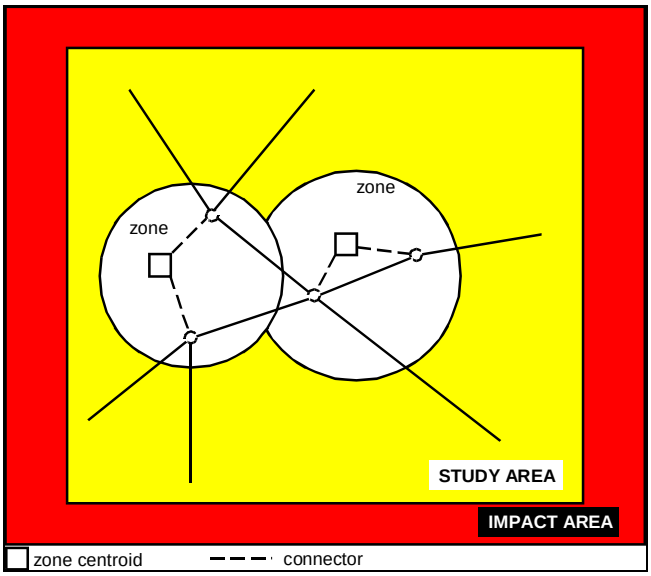


Figure 3.1: *Geographical representation of study area*

In general, the finer the subdivision, thus the more zones, the higher the accuracy in model calculations, but also the higher the costs in data collection, and computation. However, there might be an optimum level of detail because the prediction accuracy of zonal data (e.g. population characteristics) declines with the zonal size.
If public transport is an important issue, the zonal subdivision needs to be extra fine because the use of public transport strongly depends on the characteristics of the local access and egress transport.
Especially modeling the short distance trips will suffer from a coarse zonal system; for long distance trips the error in trip distance or trip time will be limited.

With respect to the delineation of the traffic zones there are a number of existing administrative spatial systems that can be used to define zones. This means that a traffic zone most favorably consists of one or more units of such an existing spatial system. Such systems are:

- municipalities
- census districts
- postal districts
- election districts
- etc.

The characteristics of these spatial systems, consisting of data about the geographical definition of the subareas and about their demographic or socio-economic content are available from various official institutions (Ministry of Transport, Chamber of Commerce, etc). Adopting such data reduces the costs of data collection and system design.

In making municipal transportation plans zone sizes of about 1000 to 2000 inhabitants are advised. For cities of the size of Delft this means about 50 traffic zones. For regional studies a maximum of 500 zones is applied. The Dutch National Model System uses 350 zones.

Zonal form
On the one hand, ideally zones should follow available delineations given by official spatial systems in order to save costs and to increase comparability. On the other hand, traffic zones should have a compact convex form in order to minimize errors in trip distances.

Zonal centroids
A traffic zone is represented by a single point of the zone, called centroid. It is the geographical representation of the zone. It is assumed that all trips start and end in that point. The centroid is part of the modeled transport network. It is a fictitious network node that connects the zone to the surrounding networks. It is linked to the network by so-called connectors which are fictitious links representing the underlying local network not included in the network model (see also Section 3.3). The location of the centroid is chosen such that it is indeed the centre of gravity of the zone, which means that its location minimizes the distance and time errors in geographically representing the individual trip addresses (see Figure 3.2). Interzonal characteristics such as distance or travel times between zones are based on the distances or travel times between the centroids of the zones.

Zonal hierarchy
In most applications a single zoning system is used for all analysis steps. However, different modeling steps may require different zonal systems. Especially the modal choice analysis may benefit from a more detailed spatial description of trips than the other steps. To this end, many studies apply a hierarchy of zones. In applications of the Dutch National Transportation Model, trip characteristics needed for modal and destination choice analysis are established with a 1200 zone system whereas trip production and traffic assignment work with a condensed 350 zone system.

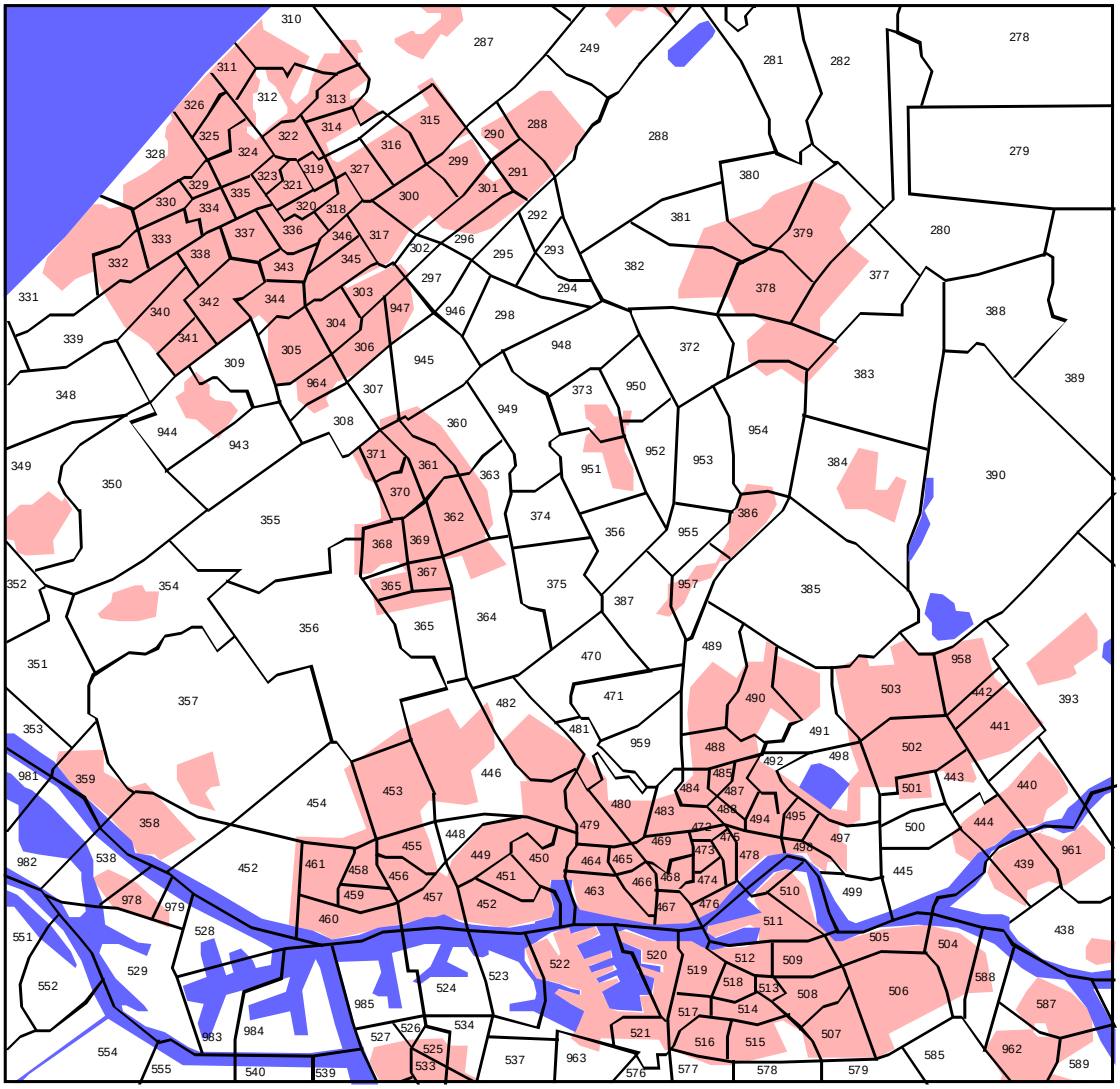


Figure 3.2: *Zonal subdivision of region The Hague-Rotterdam in the Randstadmodel*

3.3 Network description

3.3.1 Network types

In most planning cases, the final aim of the analysis is to know loads of network elements. For correct choice modeling, travel distances and times in the various networks need to be known. For these purposes a computerized description of the various networks (car, bicycle, public transport) is needed that gives the geographical relations within the network as well as enables calculations of trip characteristics such as speed, travel time etc. These networks are simplified representations of the real networks of which the level of detail depends on the problem at hand. Only the networks within the study area need to be modeled.

Such networks consist of nodes and links between nodes. The structure of the networks resembles that of the original real-world network. Zonal centroids are connected to nodes of the modeled network by one or more links.
In the case of private travel (pedestrian, bicycle, car) the modeled network may be directly derived from the physical one by selecting network parts or by aggregating subnetworks into

a single link. Nodes and links of the modeled network correspond directly to physical counterparts.
In the case of public transport the situation is more complex. Apart from a physical network on which public transport vehicles run, we have a line network that defines services and their characteristics such as service type, frequency, capacity, travel time etc. In order to model choices correctly a specific type of network description is needed called a line description. The line network with its stops and transfer points as well as their line characteristics are of prime importance; the underlying physical network is not that important. So, the nodes of a line network description are the stops of the lines, while the links connecting these nodes are the distinct lines available between these stops. Special links are added representing waiting at stops and transferring between lines and stops.

In most analyses today separate analyses are carried out for private and public transport networks. In the near future combined multi-modal networks will become applicable in which these networks are linked using transfer nodes which enable transferring from one type of travel (car, bicycle, bus) to another type (bus, rail). Railway stations are typical multi-modal transfer nodes. Such multi-modal networks enable integrating route and modal choice into one single choice process. A route in a multi-modal network defines automatically the use of the various modes, singly or combined.

3.3.2 Level of network detail

It makes no sense to include all links and nodes of a real network into the system description. It is too costly and it is not necessary in order to solve the problem efficiently. The question then is: how detailed should the description be?
The coarser the modeled network, the less costs for data collection and the quicker analyses will go, but also the less accurate numerical outcomes such as travel times or traffic loads will be. Where the optimum level is depends on the problem at hand: what kind of plans need to be evaluated and what kind of impacts are considered in the assessment criteria?

For the usual cases of area-wide planning, the following rules of thumb should be considered. In modeling travel demand about 75% of demand (in terms of kilometres traveled) should be part of the network analysis. The modeled network should include about 75% of the total network capacity. Because of the hierarchical nature of traffic flow, that means, most travelers try to travel as much as possible on higher order roads, this principle will lead to a sensible reduction in network size. About 20% of the network accounts for about 80% of the traveled kilometres. The modeled network for example need not include residential streets, which as a single road category already forms half of the road network. Groups of residential streets can be represented accurately enough by a single connector link.

The selected network part should be a connected network in which flows are possible between all parts of the study area.

In order to set up a modeled network one should make use of a functional classification of the real network. Each transportation network can be divided into a number of subnetworks or layers having a distinct transportation function. Each link in the network can be attached a functional class according to the degree it serves a flow or access purpose for the trips on the link. One can distinguish about ten such functional classes. Examples are:
- motorway (100% flow function)
- urban motorway
- arterial
- collector
- residential street (5% flow, 95% access)

– residential cul-de sac (0% flow, 100% access).

After having defined the functional classification for the study area network at hand (which mostly is already available) the selection of the modeled network works topdown. First, the top functional class is selected completely and the percentage of total capacity selected is calculated. Then, the next level is included and the selected capacity value is determined. The question now is when to stop.

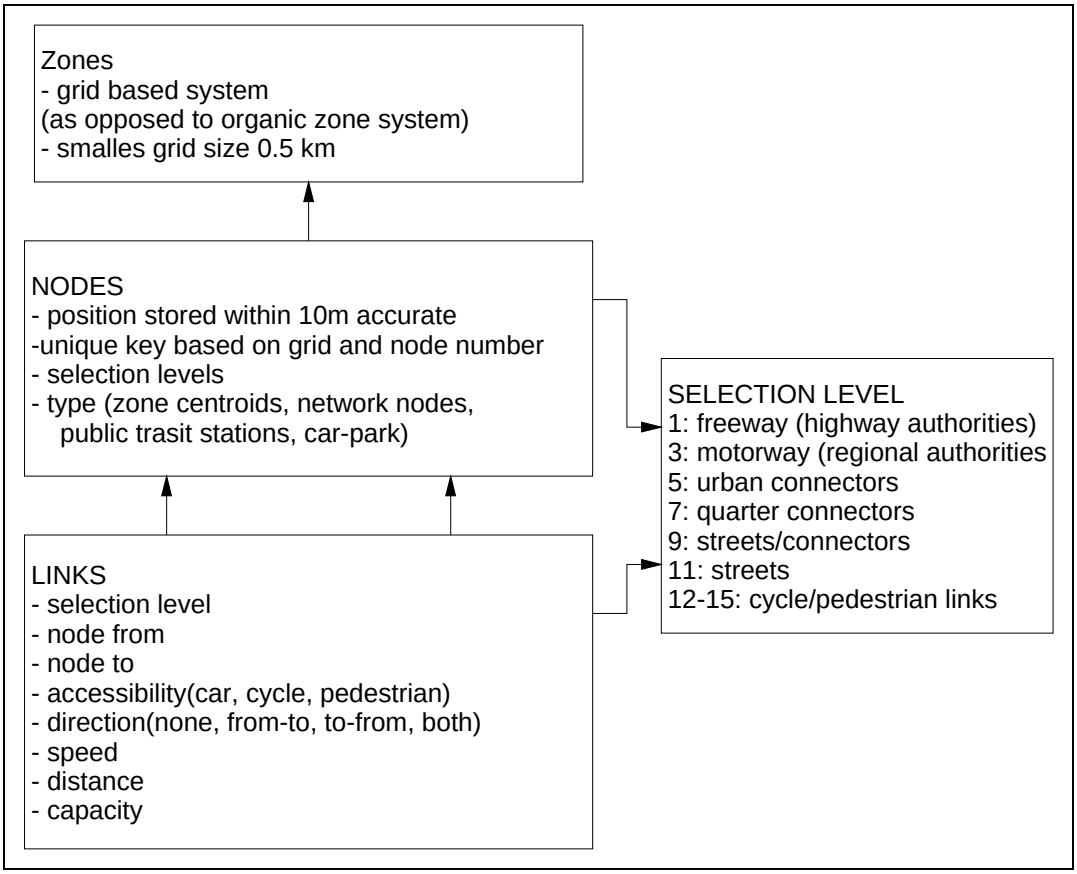


Figure 3.3: *Basis Network (Rijkswaterstaat the Netherlands, in use since 1977, currently being upgraded*

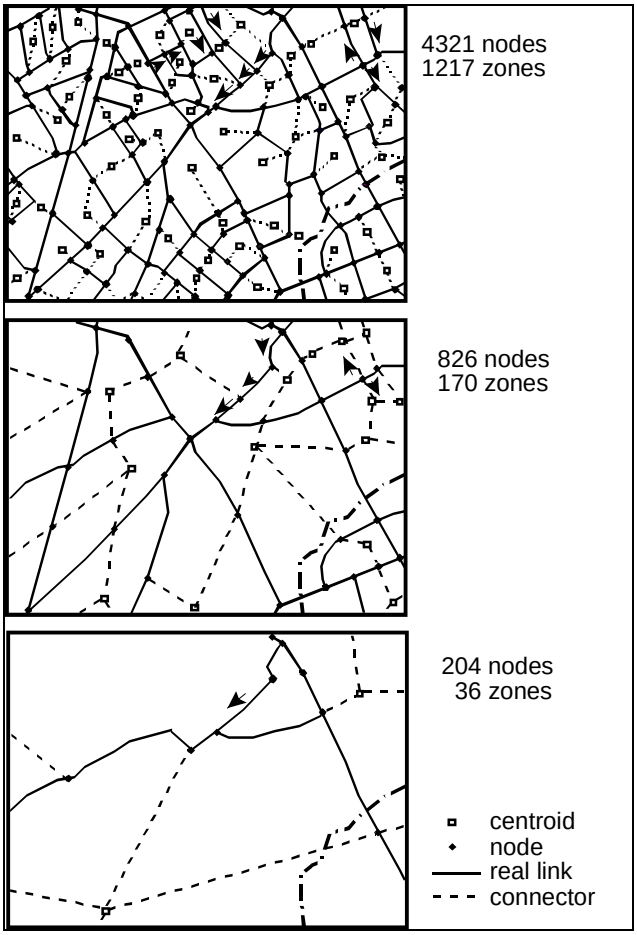


Figure 3.4: *Description of a road network at three levels of spatial detail (Eindhoven). Source: Bovy & Jansen [1981]*

At least all those network parts should be included for which detailed load figures are asked. In order to be accurate enough, the next lower functional layer of the network should be included as well because these offer routing alternatives to the studied links. So, if one is interested in arterials, one should include collector roads in the modeled network as well. If one is interested in the phenomenon of rat-running one is forced to model nearly the complete network. To be accurate enough for other purposes as well, about 75% of network capacity should be part of the modeled network.

Figure 3.4 gives an illustration of network selection by showing three modeled descriptions of the same real network (Eindhoven): fine, moderate, and coarse level of detail. The mid level network, having only one-fifth of the number of nodes compared to the fine level, performs best: analysis cost and computing time are only one-fifth but accuracy is nearly the same as with the fine network.
In urban networks, the nodes are the critical elements. The more nodes the more input data are required. In some software packages the number of links joining in a node is limited. This may require the definition of auxiliary nodes in order to represent reality correctly. Of special importance is the way of node coding, this means, whether or not turning movements at nodes are specified explicitly. For an illustration see Figure 3.6.

Zonal centroids are part of the modeled network. These are the entry and exit points of the trips. Centroids are connected to the network by connector links. Centroids may be connected to existing nodes or to dedicated auxiliary nodes specially introduced in the links. This way of

centroid connecting depends on the software package at hand. In order to achieve sufficient accuracy in modeling, the centroid should have connections in all relevant directions.

Network coding
The modeled network consists of nodes and links (also called arcs). The nodes represent physical intersections or auxiliary nodes or centroids. Centroids are a special category of nodes where shortest routes start and end. Links represent physical links or auxiliary connections such as connectors. Bi-directional physical links are represented by two uni-directional links.
Zonal centroids are part of the modeled network. These are the entry and exit points of the trips. Centroids are connected to the network by connector links. Centroids may be connected to existing nodes or to dedicated auxiliary nodes specially introduced in the links is way of centroid connecting depends on the software package at hand. In order to achieve sufficient accuracy in modeling, the centroid should have connections in all relevant directions.

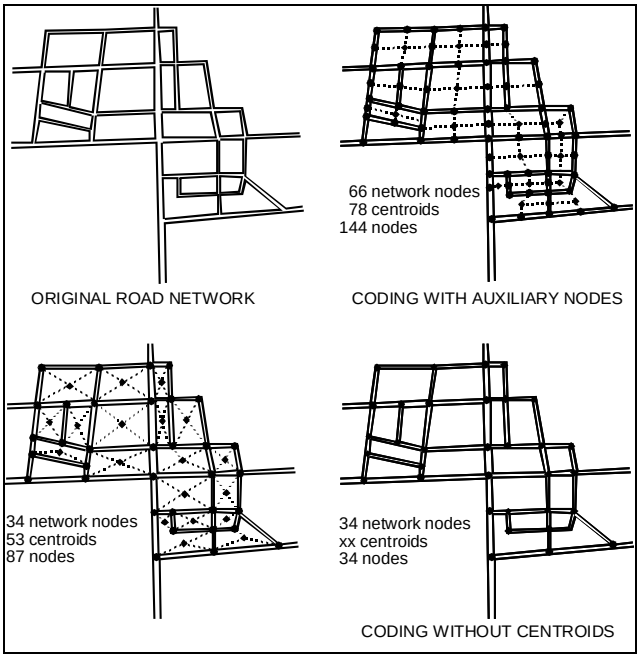


Figure 3.5: *Examples of possible specifications of road network structure*

In larger intersections travel time losses are caused due to waiting at the intersection entries. In addition, the travel time losses at the distinct turning movements may differ significantly. These travel time losses significantly influence the route choice of travelers. To model these travel times accurately enough it may be necessary to introduce in the modeled network description special links that represent these turning movements.

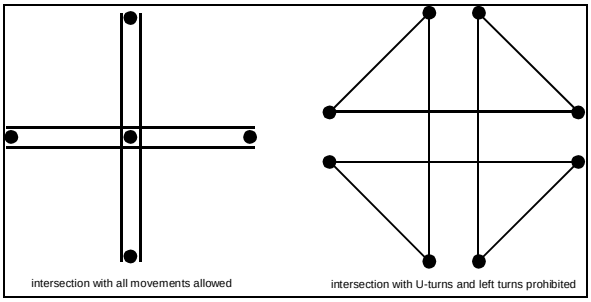


Figure 3.6: *Forms of network coding*

3.4 Travel resistance

Generalized time or cost functions
Performing activities and traveling requires spending time and money. Analogously to micro-economic theory, transportation theory adopts the following two equations as constraints to travel choice behavior:

money budget

$$\sum_n y_{np} k_n = K_p \tag{3.1}$$

time budget

$$\sum_n y_{np} t_n = T_p \tag{3.2}$$

where:
y_{np} = number of trips of person p for activity type (purpose) n
k_n = costs (including travel costs) of performing activity n
t_n = time (including travel time) needed for performing activity n
K_p = money budget (equal to income) of person p;
T_p = time budget of person p.

If the personal income during time period T is denoted with INK_p, then it holds

$$K_p = INK_p \cdot T_p \tag{3.3}$$

implying that (3.1) also may be written as:

$$\sum_n y_{np} \frac{k_n}{INK_p} = T_p \tag{3.4}$$

If individuals maximize their activity utility within their time and budget constraints, it can be shown that this behavior implies the trading off of time and money. This means that both constraints may be combined into a single weighted summation.
Weighted summation leads to the equation:

$$\sum_n y_{np} (t_n + \gamma \frac{k_n}{INK_p}) = (1+\gamma) T_p \tag{3.5}$$

The travel resistance Z_n to performing an activity n thus is:

$$Z_n = t_n + \gamma \frac{k_n}{INK_p} \tag{3.6}$$

Z_n is called generalized time.

If we apply this to a zonal interchange *i-j* the following equation results

$$Z_{ijv} = t_{ijv} + \gamma \frac{k_{ijv}}{INK} \tag{3.7}$$

in which:

Z_{ijv} = generalized time between zones i and j with mode v;
t_{ijv} = travel time between zones i and j with mode v
k_{ijv} = travel cost between zones i and j with mode v;
γ = a parameter mostly proportional to income ($\gamma = 3$)

The ratio INK/γ is called value-of-time (VOT). This is the amount of money that travelers are willing to spend in order to save one unit of travel time.
We may expect that additional factors will influence the travel resistance, such as e.g. physical effort in bicycling.

Time is a scarce good and thus valuable. In making travel choices, travelers make a trade off between time and money: time is valued in money. Car drivers try to save parking costs by walking a longer distance to their destination. In the case of road tolling (such as formerly applied in the Benelux tunnel and Zeeland bridge) a lot of car drivers prefer making a detour to save out-of-pocket expenses. The extent to which they do this depends on their value-of-time which in turn depends on their income or their gross hourly rate (for business men). Table 3.1 shows value-of-time figures currently applied by the Dutch Ministry of Transport in their transportation studies.

Before-and-after studies of road traffic tolling show that the car driver is willing to make a detour of about 5 kilometres (or willing to drive extra 5 minutes) in order to save Hfl. 3.50 toll. From these studies it can be derived that one hour in-vehicle time on average is valued at about 8 to 10 guilders (in 1988). The average value-of-time of road freight shipments amounts to about Hfl. 63 (≈ € 29) per hour (in 1992).

income group (averaged over all modes)	trip purpose		
in guilders/month (gross)	commuting	business	other
< 2500	9.20	19.80	7.30
2501 – 4000	9.70	27.80	8.20
4001 – 6000	13.00	37.90	9.30
> 6000	13.40	48.20	11.40
all groups	11.30	37.50	8.70
income group (averaged over all purposes)	trip mode		
in guilders/month (gross)	car	train	bus/tram
< 2500	8.70	6.80	5.00
2501 – 4000	9.80	8.10	6.00
4001 – 6000	13.50	9.70	7.20
> 6000	17.60	12.70	10.10
all groups	12.00	8.90	6.50
mode (averaged over all income groups)	purpose		
	commuting	business	other
car	11.40	37.60	9.10
train	11.60	33.00	7.90
bus/tram	9.50	32.90	5.60
all groups	11.30	37.50	8.70

Table 3.1: *Base value-of-time values for evaluation, in guilders (of 1988) per hour [Source: Ministry of Transport]*

The value-of-time of travelers also plays a role in assessing the yield of infrastructure investments. Such investments normally lead to travel time gains; using the aforementioned value-of-time figures these travel time gains can be expressed in money terms and balanced with the investment costs.

Explanation of Table 3.1:
The values in the table only apply to persons aged 16 or older; income figures refer to households; one has to take account of real price increases. The VOT-figures give the distribution over three variables: trip purpose, mode and income. In the table two-way splits are shown averaged over the third variable. The VOT-figures are derived from revealed travel behavior. Travel time losses generally are valued higher than travel time gains because of distortions of time plans.

Objective, perceived and modeled resistance
In determining travel resistance we need to distinguish objective, perceived and modeled resistance.

Objective resistance or costs (X_{nk}) can be observed objectively by measurement rules. Time and distance traveled by car or bus for example.

Subjective or perceived resistance or costs result because of individual weighing of objective costs. Decisions of travelers are made on the basis of perceived costs. The assessment of resistance is related to personal characteristics and depends on observational errors in personal estimates of times and distances. Perceived time is judged according to the extent a person can dispose freely of his time, thus whether one has a job or not, whether one is caring for children or not, whether a person goes to school or not.

Modeled resistance or costs (β_{nk}, X_{nk}) are used in a model analyzing travel behavior. The parameter values are derived from observations on actual behavior using estimation and testing models. From empirical analyses using disaggregated estimation models it has been found that time is perceived very differently. Especially with respect to public transport trip making it appears that access and egress times (that is the time from the door at the origin address to the vehicle, respectively the time from the vehicle to the door at the destination address), waiting times and transfer times are weighed 2 or 3 times higher than the in-vehicle travel time (see course notes CT3751).
Also the waiting times in the car at intersections is valued higher than the in-car driving time (Hamerslag, 1979).

3.5 Shortest path calculation

Introduction.
The calculation of shortest paths in a network is an essential step in each transportation analysis. This calculation is done very many times in each analysis so that its calculation time is predominant in the analysis duration.
The shortest paths serve several functions within the transportation analysis.

- Shortest paths between nodes or zonal centroids give the travel resistance of trips in a network (expressed in simple or generalized time, in distance, or in costs). These trip resistances are used in modeling choice behavior (mode, destination, time and route choice).
- Shortest paths also are used in the modeling step of assigning traffic flows between zones to the network in order to find link loads.

Very many algorithms exist to find shortest paths in networks. We may distinguish:
- tree builder algorithms
- matrix algorithms

The tree-builder algorithms search sequentially for each origin node (or centroid) the shortest route to all other nodes (and thus also destination centroids). These algorithms are first introduced by Moore (1957). A special class are the once-through algorithms due to Dijkstra (1959).
Matrix algorithms search simultaneously shortest routes from all origin (centroid) nodes to all destination (centroid) nodes of the network.
In general, in larger transportation networks, tree builder algorithms are more efficient: they need less computer memory less calculation time. On the contrary, matrix algorithms are easy to program.

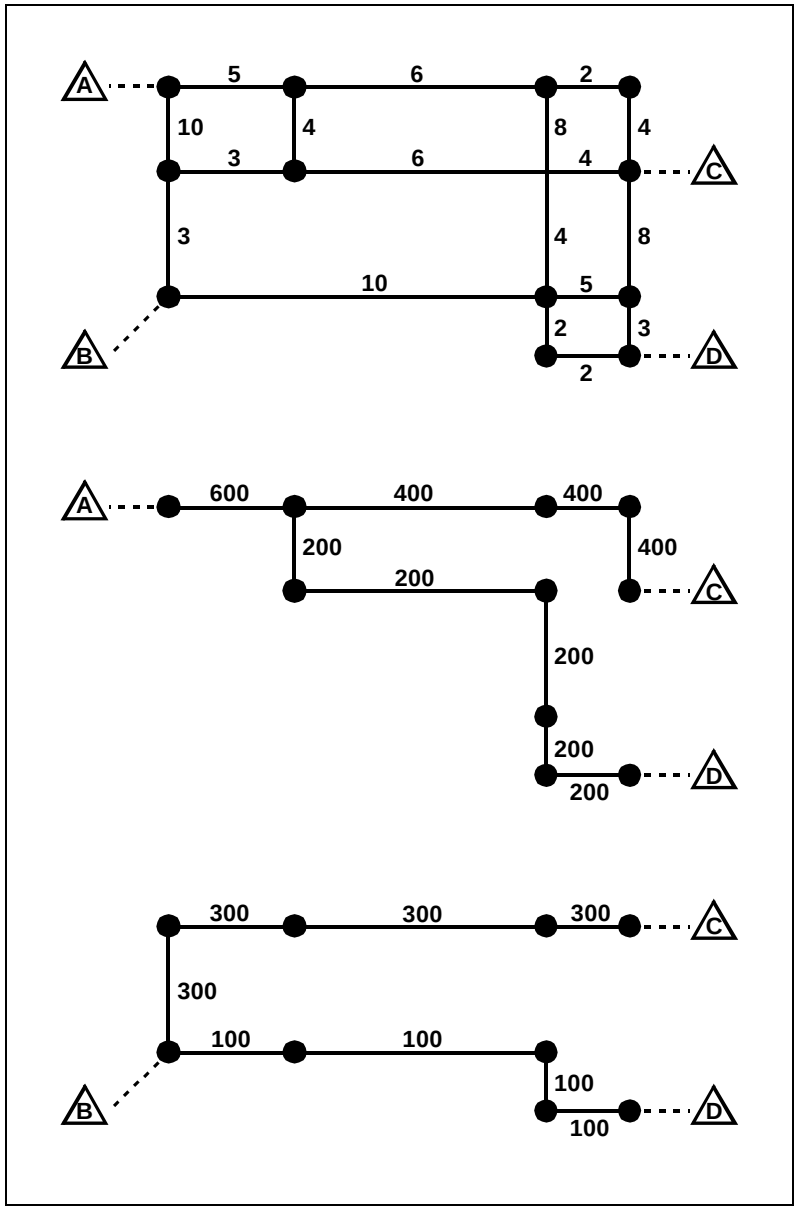


Figure 3.7: *Illustration of shortest path search*

Tree builder algorithm.
The network is described by nodes (index n) and links (index k). The resistance of a link is represented by a generalized time z_k.

To each node n a label is attached consisting of 3 components: q_{in}, m_{in} and α:
- q_{in} is the smallest resistance from the origin node i to node n found after each computation step;
- m_{in} is the node number of the last node in the shortest route, the so-called backnode;
- α_n is an (0,1) indicator showing whether still computations have to be carried out with this node n or not, that is, whether the node is active (0) or non-active (1).

For each centroid or origin node i the following computation steps are carried out:
1. (Initialization) all nodes are labelled (B,0,1) where B is a very large number. The null means that no backnode has been found yet.
2. The origin node is labelled (0,0,0)
3. It is checked whether at least one node is active (α=0). At the first step this is the origin node. If no active node is found the calculations for the origin node i are completed.
4. Next, one of the active nodes is chosen (n)
5. Next it is determined which nodes (k) are connected to the active node.
6. For each of these nodes (k) it is determined whether the shortest path is via the active node. This is the case if :

   $q_{in} + z_k < q_{ik}$

   Node k then is labeled with components ($q_{in}+z_k,n,0$).
   This means that this node is attached the new shorter resistance, the backnode number is changed, and this node now becomes active.
7. The value of α of node n is set to one, thus is made passive.
8. The computation continues with step 3.

The Moore-type algorithms may select an active node in various ways:
- according to the sequence it is put into an auxiliary table
- taking the node with lowest node number, or
- randomly.

The once-through algorithms however select the node with the smallest resistance from the origin node. With the once-through algorithm each node becomes active only once. This saves computation time. However, calculation time is needed to find out which node is 'closest' to the origin node. This calculation time often is larger than the extra computation time if nodes become active several times during tree building.

As input to the calculations the network has to be organized into special data structures. Also the output, being the shortest route trees of which one per origin centroid, need to be organized in special data structures. Many alternative ways of organizing these data are possible offering different opportunities for further application such as the reconstruction of alternative routes to the shortest one using the backnode information in the shortest tree. For an elaboration of these data possibilities as well on search algorithms, see Ajuha et al (1993).

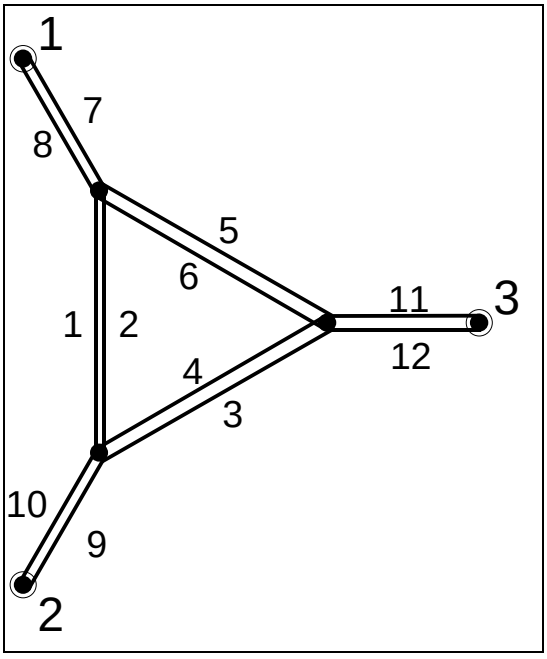
3.6 Assignment map

There are a number of ways to store the results of a shortest path computation. A very compact way is the *shortest path tree*. The shortest path tree is a vector with a length that corresponds with the number of nodes in a network. This vector contains for each node its *backnode*: this is the first node that is encountered on the path from the current node to the origin of the shortest path tree.

In one shortest path tree all the information that is needed to reconstruct the paths from one origin to all network nodes can be stored. Hence, if one wants to store the all-to-all shortest path information for a network with N nodes, then N shortest path trees with N elements each are needed.

Albeit very compact, the shortest path tree is not the most suitable form to store paths, when one needs to make computations. In this case it is much more common to use an *assignment map*. The assignment A is a table with a height that corresponds to the number of network links and a width that corresponds with the number of routes. If element a, r of assignment map A equals 1, i.e. $A(a,r) = 1$, this indicates that route r traverses arc a.

If we consider the network shown in Figure 3.8, with origin/destination nodes 1, 2 and 3, then the corresponding assignment map is given by the matrix A. The columns in this matrix correspond with the shortest-paths for OD-pairs 1-2, 1-3, 2-1, 2-3, 3-1 and 3-2.
The assignment map may be used for various types of analysis, such as:
- computing the total flow on a link
- determining the routes that pass over a link
- determining the routes that pass over a particular combination of links
- determining the links that a route consists of
- determining the contribution of a particular route-flow to a link flow



$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Figure 3.8: *Example network with 6 nodes and 12 links. Nodes 1,2 and 3 function as origin and destination nodes*

3.7 References

M.G.H. Bell & Y. Iida
Transportation network analysis
New York, John Wiley & Sons, 1997

G.R.M. Jansen
Kortste route zoeken in netwerken
Verkeerskunde

R.K. Ajuha, T.L. Magnanti & J.B. Orlin
Network flows: theory, algorithms and applications
Prentice Hall, Englewood Cliffs, 1993

4 Trip generation modeling

4.1 Introduction

The trip generation stage of the classical transport model aims at predicting the total number of trips produced in the zone and attracted by it respectively for each zone of the study area. This has been usually considered as the problem of answering a question such as: *how many trips* originate at each zone? However, the subject has also been viewed sometimes as a *trip frequency* choice problem: how many shopping (or other purpose) trips will be carried out by this person type during an average week? This is usually undertaken using discrete choice models and it is then cast in terms like: what is the probability that this person type will undertake zero, one, two or more trips with this purpose per week?

Sections 4.2 to 4.5 of this chapter were derived from the book 'Modelling Transport' by J. de Ortuzar and L.G. Willumsen, while some of the examples were derived from the manual 'Travel Demand Modeling with TRANSCAD 3.0' by Caliper Corporation.
Section 4.2 starts by defining some basic concepts and will proceed to examine some of the factors affecting the generation and attraction of trips (Section 4.3). Then we will review the main modeling approaches. There are three primary tools that are used in modeling trip generation:

- *Regression Models* (Chapter 4.4):
  Two types of regression are commonly used. The first uses data aggregated at the zonal level, with average number of trips per household in the zone as the dependent variable and average zonal characteristics as the explanatory variables. The second uses disaggregated data at the household or individual level, with the number of trips made by a household or individual as the dependent variable and the household and personal characteristics as the explanatory variables.
- *Cross-Classification* (Chapter 4.5):
  Cross-Classification methods separate the population in an urban area into relatively homogenous groups based on certain socio-economic characteristics. Then, average trip production rates per household or individual are empirically estimated for each class. This creates a lookup table that may be used to forecast trip productions.
- *Discrete Choice Models* (Chapter 4.6):
  Discrete choice models use disaggregated household or individual level data to estimate the probability with which any household or individual will make trips. The outcome can then be aggregated to predict the number of trips produced.

All three approaches can be applied to estimate the zonal trip generation using different units of analysis (zones, households or person) (Table 4.1). We will start with considering zonal and household-based linear regression trip generation models, giving some emphasis to the problem of non-linearities which often arise in this case. We will also address the problem of aggregation (e.g. obtaining zonal totals), which has a trivial solution here precisely because of the linear form of the model. Then we will move to cross-classification models, where we will examine not only the classical category analysis specification but also more contemporary approaches including the person category analysis model. Finally discrete choice methods will be described.

Analysis Unit	Linear Regression	Cross-Classification	Discrete Choice
Zone	x (Section 4.4.1)		
Household	x (Section 4.4.2)	x (Section 4.5.1)	x (Section 4.6)
Person	x	x (Section 4.5.3)	x (Section 4.6)

Table 4.1: *A classification of trip generation modeling approaches*

4.2 Classification of trips

4.2.1 Trip purpose

It has been found in practice that better trip generation models can be obtained if trips by different purposes are identified and modeled separately. In the case of home-based trips, five categories have been usually employed:

- trips to work;
- trips to school or college (education trips);
- shopping trips;
- social and recreational trips;
- other trips.

The first two are usually called compulsory (or mandatory) trips and all the others are called discretionary (or optional) trips. The latter category encompasses all trips made for less routine purposes, such as health, bureaucracy (need to obtain a passport or a certificate) and trips made as an accompanying person. Non-home-based trips are normally not separated because they only amount to 15-20% of all trips.

4.2.2 Time of day

Trips are often classified into peak and off-peak period trips; the proportion of journeys by different purposes usually varies greatly with time of day, see Table 4.2.

Purpose	AM Peak (%)	Off Peak (%)
work	52.12	12.68
education	35.06	4.96
shopping	1.54	11.35
social	0.79	5.40
health	1.60	2.74
bureaucarcy	3.89	18.35
accompanying	2.09	2.14
other	0.19	0.73
return to home	2.72	41.65

Table 4.2: *Example of trip classification*

The morning (AM) peak period (the evening peak period is sometimes assumed to be its mirror image) is usually taken between 7:00 and 9:00 and the representative off-peak period between 10:00 and 12:00.

4.2.3 Person type

This is another important classification, as individual travel behavior is heavily dependent on socioeconomic attributes. The following categories are usually employed:

- income level (e.g. nine strata in the Santiago survey);
- car ownership (typically three strata: 0, 1 and 2 or more cars);
- household size and structure (e.g. six strata in most British studies).

It is important to note that the total number of strata can increase very rapidly and this may have strong implications in terms of data requirements, model calibration and use.

4.3 Factors affecting trip generation

In trip generation modeling we are typically interested not only in person trips but also in freight trips. For this reason models for four main groups (i.e. personal and freight, trip productions and attractions) are usually required. In what follows we will briefly consider some factors which have been found important in practical studies.

Definitions
In determining the level of trips entering or leaving a zone during a certain period (day or peak hour) it is useful to distinguish between home-based and non-home-based trips. The reason is that modeling of trip making behavior is easier that way.
Home-based trips are trips starting or ending at the home. Non-home-based trips have other types of addresses as their origins and destinations. Both types of trips are organized in chains or tours having two or more trips, see Figure 4.1. All trips in this figure are home-based trips, except for trip (f) from zone *j* to zone *i* which is a trip from work to a shop.

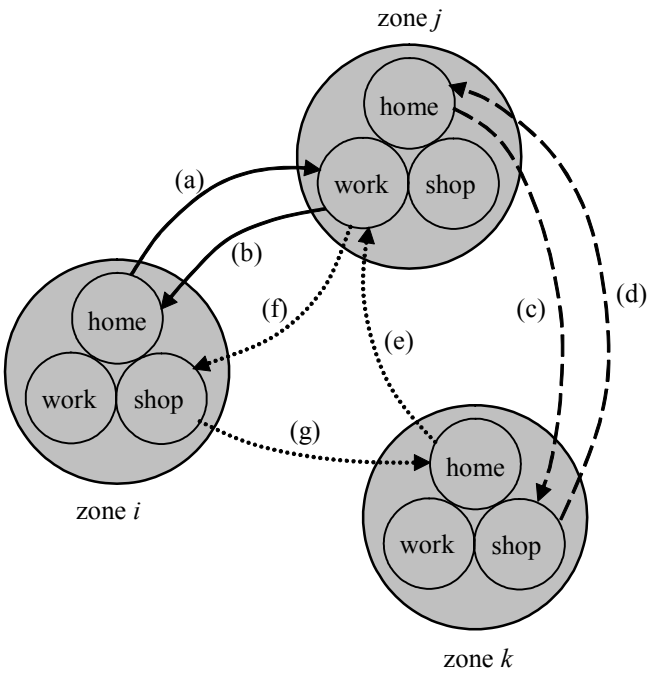


Figure 4.1: *Trip productions and attractions versus origins and destinations*

Literature defines the home-end of the trip as the production, as it is the household and its activities that produces all trips. Consequently, the non-home-end is defined as the attraction. For non-home-based trips the origin is the production and the destination is the attraction. Trip production is described using household characteristics, while attraction can be explained using variables reflecting the intensity of activity types, such as employment, education, floor space, etc. However, since we have non-home-based trips as well, production is also partly described using such activity-based variables as well.

Note that given these definitions for production and attraction, they can be either departure or arrival. A home-to-work trip is a departure from home, while a work-to-home trip is an arrival at home. Similary, a home-to-work trip is an arrival at work, while a work-to-home trip is a

departure from work. For example, trips (b), (c), and (f) are departures from zone *j*, while trips (a), (d), and (e) are arrivals at zone *j*. The terms "productions" and "attractions" therefore are technically only referring to the factors that determine the behavior. For transportation network modeling we are mainly interested in the total departures from a zone and the total arrivals at a zone (which contain home-based and non-home-based trips), see Table 4.8 for an example. These zonal departures and arrivals are often also referred to as total production and attraction (also in these lecture notes, where we will predominantly focus on home-based trips), although total departures and arrivals would be a better term.

4.3.1 Personal trip productions

The goal of trip production is to estimate the total number of trips, by purpose, produced or originating in each zone. Trip production is performed by relating the number or frequency of trips to the characteristics of the individuals, of the zone, and of the transportation network.

The following factors have been proposed for consideration in many practical studies:
- income;
- car ownership;
- household structure;
- family size;
- value of land;
- residential density;
- accessibility.

The first four have been considered in several household trip generation studies, while value of land and residential density are typical of zonal studies. The last one, accessibility, has rarely been used although most studies have attempted to include it. The reason is that it offers a way to make trip generation elastic (responsive) to changes in the transport system.

4.3.2 Personal trip attractions

In many ways, estimating trip attractions is similar to estimating trip productions because the problem is the same: predicting the number of trips attracted by relating the number or frequency of trips to the characteristics of the individuals, the zone, and the transportation network. Thus, the methods described in the trip production paragraph -- cross-classification, regression, and discrete choice -- may also be used to estimate the number of trips attracted to a zone.

In production models, estimates are primarily based on the demographics of the population within a zone. For attraction models, the variables that have been found to have the best explanatory power are those based on characteristics of the land use, such as office and retail space or the employment levels of various sectors. As with production models, characteristics of the transportation network are rarely used, which means that the models cannot reflect impacts on trip attractions from changes in accessibility. Also similar to production models, information on the work trip is relatively easy to acquire from such sources as the census or locally initiated surveys. Thus, models of work trip attractions should always be estimated directly using data from the study area, instead of applying models based on national averages or based on another study area.

Regression models are often used to estimate trip attractions because of the high correlation between the number of trips made and explanatory variables such as employment and office/retail space (particularly for work trips). Cross-classification can also be used for trip attraction, in which the classification is usually based on the employment sectors, and

sometimes employment density. However, the difficulty in collecting the disaggregated data on which to generate the cross-classification table (e.g. it's much easier to collect a statistically valid sample of households than of offices or retail shops) has made cross-classification a rarely used tool for trip attractions. The difficulty of collecting disaggregated data for attraction has also limited the use of discrete choice, although logit models could be applied at the aggregate level.

4.3.3 Freight trip productions and attractions

Freight trips normally account for few vehicular trips; in fact, at most they amount to 20% of all journeys in certain areas of industrialized nations, although they can still be significant in terms of their contribution to congestion. Important variables include:

- number of employees;
- number of sales;
- roofed area of firm;
- total area of firm.

To our knowledge, neither accessibility nor type of firm have ever been considered as explanatory variables; the latter is curious because it would appear logical that different products should have different transport requirements.

4.4 Regression analysis models

Regression methods[1] can be used to establish a statistical relationship between the number of trips produced and the characteristics of the individuals, the zone, and the transportation network.

Two types of regression models are commonly used. The first uses data aggregated at the zonal level, with average number of trips per household in the zone as the dependent variable and average zonal characteristics as the independent (explanatory) variable. The second uses disaggregated data at the household or individual level, with the number of trips made by a household or individual as the dependent variable and the household and personal characteristics as the independent variables.

The best situation is when data for the study area are available that include relevant independent variables (e.g. socio-economic and accessibility factors) and data on frequency of trips for various trip purposes. In this case, you can estimate a regression model that is specifically made for the study area instead of transferring models from another area.

Intermezzo: linear regression analysis

Parameter estimation

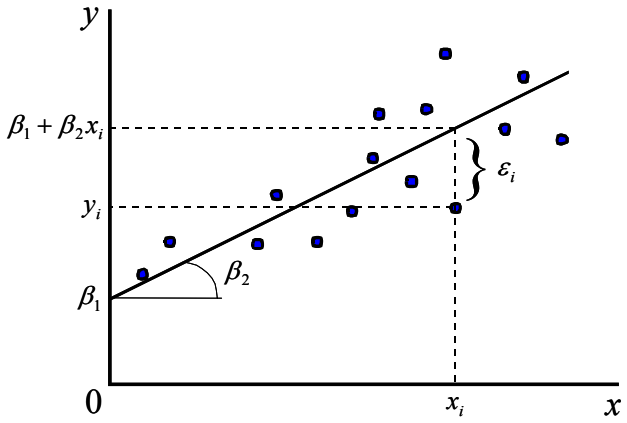
Consider a data set consisting of observations y_i ($i = 1,...,n$) of the explained (endogenous) variable y which are known to depend on explanatory (exogenous) variables $x_1,...,x_k$ for which we have observations $x_{1i},...,x_{ki}$ ($i = 1,...,n$).

Applying regression in this context means that the parameters $\beta_1,\ldots,\beta_k$ are estimated in the following linear model: y_i

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad \text{for } i = 1, \ldots, n$$

where ε_i represents the error term and β_1 represents the intercept (note that $x_{1i} = 1$ for all i).

[1] For an explanation of the principle of regression see Ortùzar & Willumsen, *Modelling Transport*, 1994, Chapter 4.2.1



Consider the case where $k = 2$, hence the model simplifies to: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$.
If we plot all the observations (x,y) (see figure), then regression of x on y comes down to fitting the "best" line through the dots representing these observations.
We define "best" as the case in which the sum of squared distances between observed y_i 's and model predictions $\hat{y}_i = \beta_1 + \beta_2 x_i$ is minimal, i.e., we determine β_1 and β_2 such that

$\sum_{i=1}^{n} (y_i - (\beta_1 + \beta_2 x_i))^2$ is minimal.

Statistical properties

Once we have determined the optimal values for $\beta_1, \ldots, \beta_k$, denoted by $\hat{\beta}_1, \ldots, \hat{\beta}_k$, we may want to check whether or not these estimates are significantly different from 0. Therefore we test the null-hypothesis H_0: $\hat{\beta} = 0$ against the alternative hypothesis H_1: $\hat{\beta}_p \neq 0$ $(p = 1,\ldots,k)$.

First, we make the following (classical) assumptions:

- $x_1,\ldots,x_k$ are non-random
- ε_i's are independent and normally distributed with mean 0 and unknown variance σ^2.

Further, we need to determine the level of significance α, which can be defined as the probability of erroneously rejecting the null-hypothesis. A common value is $\alpha = 0.05$.

Finally, the standard error $se(\hat{\beta}_p)$ should be computed for each parameter. Then we can use the test-statistic $T_p = \hat{\beta}_p / se(\hat{\beta}_p)$ which follows a t-distribution with $n - k$ degrees of freedom.

We reject the null-hypothesis if $|T_p| > t_{n-k}^{\alpha/2}$. Otherwise, we assume that β_p equals 0.

Software packages like SPSS, TRANSCAD, Matlab and spreadsheets can be used for linear regression analysis. For more details, consult Wonnacott and Wonnacott (1980).

- end of intermezzo -

4.4.1 Zonal-based multiple regression model

In this case an attempt is made to find a linear relationship between the number of trips produced or attracted by zone and average socioeconomic characteristics of the households in each zone. The following are some interesting considerations:

1. Zonal models can only explain the variation in trip making behavior between zones. For this reason they can only be successful if the inter-zonal variations adequately reflect the real reasons behind trip variability. For this to happen it would be necessary that zones not only have a homogeneous socioeconomic composition, but also represent an as wide as possible range of conditions. A major problem is that the main variations in person trip data occur at the intra-zonal level (within zones).

2. *Role of the intercept.*
   One would expect the estimated regression line to pass through the origin; however, large intercept values (i.e. in comparison to the product of the average value of any variable and its coefficient) have often been obtained. If this happens the equation may be rejected; if on the contrary, the intercept is not significantly different from zero, it might be informative to re-estimate the line, forcing it to pass through the origin.
3. *Null zones.*
   It is possible that certain zones do not offer information about certain dependent variables (e.g. there can be no home based trips generated in non-residential zones). Null zones must be excluded from analysis; although their inclusion should not greatly affect the coefficient estimates (because the equations should pass through the origin), an arbitrary increment in the number of zones which do not provide useful data will tend to produce statistics which overestimate the accuracy of the estimated regression.
4. *Zonal totals versus zonal means.*
   When formulating the model the analyst appears to have a choice between using aggregate or total variables, such as trips per zone and cars per zone, or rates (zonal means), such as trips per household per zone and cars per household per zone. In the first case the regression model would be:

$$T_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + ... + E_i \tag{4.1}$$

   whereas the model using rates would be:

$$t_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + ... + e_i \tag{4.2}$$

   with

   $t_i = T_i/N_i$; $x_i = X_i/N_i$; $e_i = E_i/N_i$ and N_i the number of households in zone i.

   Both equations are identical, in the sense that they seek to explain the variability of trip making behavior between zones, and in both cases the parameters have the same meaning. Their unique and fundamental difference relates to the error-term distribution in each case; it is obvious that the constant variance condition of the model cannot hold in both cases, unless N_i was itself constant for all zones i.

   Now, as the aggregate variables directly reflect the size of the zone, their use should imply that the magnitude of the error actually depends on zone size; this *heteroscedasticity* (variability of the variance) has indeed been found in practice. Using multipliers, such as $1/N_i$, allows heteroscedasticity to be reduced because the model is made independent of zone size. In this same vein, it has also been found that the aggregate variables tend to have higher intercorrelation (i.e. multicollinearity) than the mean variables. However, it is important to note that models using aggregate variables often yield higher values of R^2, but this is just a spurious effect because zone size obviously helps to explain the total number of trips. What is certainly unsound is the mixture of means and aggregate variables in a single model.

The various difficulties encountered with zonal regression models (dependence on zone size, zonal boundaries, spurious correlations, etc.) have led to the use of models based on the true behavioral units: households or persons.

4.4.2 Household-based regression model

Intra-zonal variation may be reduced by decreasing zone size, especially if zones are homogeneous. However, smaller zones imply a greater number of them and this has two consequences:

- more expensive models in terms of data collection, calibration and operation;
- greater sampling errors, which are assumed non-existent by the multiple linear regression model.

For these reasons it seems logical to postulate models which are independent of zonal boundaries. At the beginning of the 1970s it was decided that the most appropriate analysis unit in this case was the household (and not the individual); it was argued that a series of important interpersonal interactions inside a household could not be incorporated even implicitly in an individual model (e.g. car availability, that is, who has use of the car). We will challenge this thesis in Section 4.5.3.

In a household-based application each home is taken as an input data vector in order to bring into the model all the range of observed variability about the characteristics of the household and its travel behavior. The calibration process, as in the case of zonal models, proceeds stepwise, testing each variable in turn until the best model (in terms of some summary statistics for a given confidence level) is obtained. Care has to be taken with automatic stepwise computer packages because they may leave out variables which are slightly worse predictors than others left in the model, but which may prove much easier to forecast.

Example 4.1:
Consider the variables trips per household (t), number of workers (X_1) and number of cars (X_2). Table 4.3 presents the results of successive steps of a step-wise model estimation. Assuming a large sample size, the appropriate number of degrees of freedom ($n - 2$) is also a large number so the t-values may be compared with the critical value 1.645 for a 95% significancy level on a one-tailed test (we know the null hypothesis is unilateral in this case as t should increase with both X_1 and X_2).

The third model is a good equation in spite of its low R^2. The intercept 0.91 is not large (compare it with 1.44 times the number of workers, for example) and the explanatory variables are significantly different from zero (N_0 is rejected in all cases). The model could probably benefit from the inclusion of other variables.

Step	Equation	R^2
1	$t = 2.36\ X_1$	0.203
2	$t = 1.80\ X_1 + 1.31\ X_2$	0.325
3	$t = 0.91 + 1.44\ X_1 + 1.07\ X_2$	0.384

Table 4.3: *Example of stepwise regression*

No. of cars	Number of workers in household			
	0	1	2	3 or more
0	0.9/0.9	2.1/2.4	3.4/3.8	5.3/5.6
1	3.2/2.0	3.5/3.4	3.7/4.9	8.5/6.7
2 or more	-	4.1/4.6	4.7/6.0	8.5/7.8

Table 4.4: *Comparison of trips per household (observed/estimated) per household class.*

An indication of how good these models are may be obtained from comparing observed and modeled trips for some groupings of the data (see Table 4.4). This is better than comparing totals because in such case different errors may compensate and the bias would not be detected. As can be seen, the majority of cells show a reasonable approximation (i.e. errors of less than 30%). If large biases were spotted it would be necessary to adjust the model parameters; however, this is not easy as there are no clear-cut rules to do it, and it depends heavily on context.

Example 4.2:
The following trip frequency model (see Table 4.5) is an example of a household-based trip production model. It is a linear regression model using dummy variables. It calculates the weekly number of trips by individual households. The level of (weekly) trip making appears to depend on:
- size of the household
- life cycle
- highest education in the household
- structure of the household
- number of driving license owners

Income and car ownership appear not to be significant as explanatory variables. This is most probably a result of the unspecific nature of the models, that is no distinction between trip purposes.
The model has been estimated using data from the Dutch Longitudial Mobility Panel [see Bovy & Kitamura, 1986]. On average, a Dutch household makes about 50 trips a week. The largest contribution is given by the number of household members: 22.5 trips per person per week. The higher the educational level the more trips are made with in the extreme a difference of 14.1 trips per week (10.1– – 4.0).
The model performs extremely well with it explained trip variance of 90%.

Model:

$$Y = \sum_k \alpha_k X_k \tag{4.3}$$

Y: weekly # household trips (all purposes, all modes)

Variable		Parameter value
# members household		22.5
lifestyle	couple, no kids, male < 35 year	5.6
	couple, no kids, male 35-64 year	-2.8
	couple, no kids, male > 64 year	-5.6
education	elementary school	-4
	intermediate vocational or higher general preparatory school	3.4
	higher vocational school	9.4
	university	10.1
# driver license		3.9
structure	# adults	-5
	# children < 6 year	1.9
	# children 6-12 year	3.5
Goodness-of-fit $R^2 = 0.9$		

Table 4.5: *Regression model (with dummy variable) of household weekly trip production (Source: Bovy & Kitamura, 1986)*

- end of example -

4.4.3 The problem of non-linearities

As we have seen, the linear regression model assumes that each independent variable exerts a linear influence on the dependent variable. It is not easy to detect non-linearity because apparently linear relations may turn out to be non-linear when the presence of other variables is allowed in the model. Multivariate graphs are useful in this sense; the example of Figure 4.2 presents data for households stratified by car ownership and number of workers. It can be seen that travel behavior is non-linear with respect to family size.

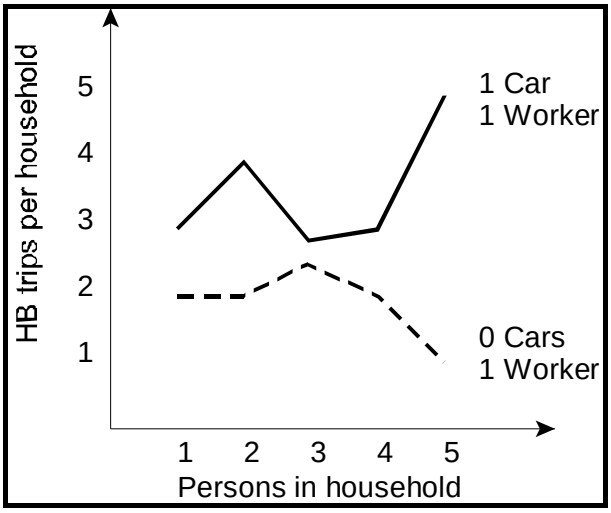


Figure 4.2: *An example of non-linearity*

It is important to mention that there is a class of variables, those of a qualitative nature, which usually shows non-linear behavior (e.g. type of dwelling, occupation of the head of the household, age, sex). In general there are two methods to incorporate non-linear variables into the model:

1. Transform the variables in order to linearize their effect (e.g. take logarithms, raise to a power).
   However, selecting the most adequate transformation is not an easy or arbitrary exercise, so care is needed; also, if we are thorough, it can take a lot of time and effort;
2. *Use dummy variables.*
   In this case the independent variable under consideration is divided into several discrete intervals and each of them is treated separately in the model. In this form it is not necessary to assume that the variable has a linear effect, because each of its portions is considered separately in terms of its effect on travel behavior. For example, if car ownership was treated in this way, appropriate intervals could be 0, 1 and 2 or more cars per household. As each sampled household can only belong to one of the intervals, the corresponding dummy variable takes a value of 1 in that class and 0 in the others. It is easy to see that only (n - 1) dummy variables are needed to represent n intervals.

Example 4.3:
Consider the model of Example 4.1 and assume that variable X_2 is replaced by the following dummies:

Z_1, which takes the value 1 for households with one car and 0 in other cases;
Z_2, which takes the value 1 for households with two or more cars and 0 in other cases.

It is easy to see that non-car-owning households correspond to the case where both Z_1 and Z_2 are 0. The model of the third step in Table 4.3 would now be:

$$\begin{aligned} t &= 0.84 + 1.41X_1 + 0.75Z_1 + 3.14Z_2 \\ R^2 &= 0.387 \end{aligned} \qquad (4.4)$$

Even without the better R^2 value, this model is preferable to the previous one just because the non-linear effect of X_2 (or Z_1 and Z_2) is clearly evident and cannot be ignored. Note that if the coefficients of the dummy variables were for example, 1 and 2, and if the sample never contained more than two cars per household, the effect would be clearly linear. The model is graphically depicted in Figure 4.3.

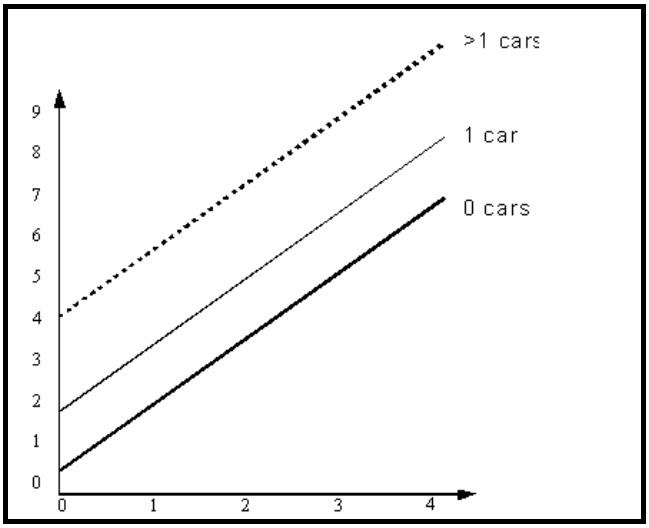


Figure 4.3: *Regression model with dummy variables*

- end of example -

4.4.4 Obtaining zonal totals

In the case of zonal-based regression models, this is not a problem as the model is estimated precisely at this level. In the case of household-based models, though, an aggregation stage is required. Nevertheless, precisely because the model is linear the aggregation problem is trivially solved by replacing the average zonal values of each independent variable in the model equation and then multiplying it by the number of households in each zone. However, it must be noted that the aggregation stage can be a very complex matter in non-linear models.

Thus, for the third model of Table 4.3 we would have:

$$T_i = N_i(0.91 + 1.44V_{1i} + 1.07V_{2i}) \tag{4.5}$$

where T_i is the total number of home based trips in the zone, N_i is the total number of households in it and V_{ji} is the average value of variable X_j for the zone i.

On the other hand, when dummy variables are used, it is also necessary to know the number of households in each class for each zone; for instance, in the model of Example 4.3 we require:

$$T_i = N_i(0.84 + 1.41V_{1i}) + 0.75N_{1i} + 3.14N_{2i} \tag{4.6}$$

where N_{hi} is the number of households of (car ownership) class h in zone i.

This last expression allows us to appreciate another advantage of the use of dummy variables over separate regressions. To aggregate the models, in the first case, it would be necessary to estimate the average number of workers per household (X_1) for each car-ownership group in each zone, and this may be complicated.

4.5 Cross-classification or category analysis model

Cross-classification methods of calculating productions separate the population in an urban area into relatively homogenous groups (households or persons) based on certain socio-economic characteristics. For example, one may classify households in an area by both family size (1, 2, 3, 4, ≥ 5 persons/household) and by auto ownership (0, 1, ≥ 2 autos/household), which results in 15 classes (see Table 4.6). Average trip-production rates (the estimated number of trips that will be taken by a household or individual) are empirically derived from either disaggregated or aggregate data sets for each of the classes. In the example above, 15 average trip rates would be derived.

Once trip rates are known for each class, these trip rates are usually applied to each zone.

Family Size	0 cars	1 car	≥2 cars
1			
2			
3			
4			
≥ 5			

Table 4.6: *Example Cross-Classification Table*

Each zone may be subdivided into a few classes by using the proportion of households or persons within a zone that have a certain characteristic. Using this method, more than one average trip rate is used to estimate productions for any one zone. For example, a zone may be divided into households without cars and households with cars. In this case, 2 average trip rates will be applied to each zone.

$$T_i^p = \sum_h N_{hi} t_h^p \tag{4.7}$$

4.5.1 The household-based category model

The method is based on estimating the response (e.g. the number of trip productions per household for a given purpose) as a function of household attributes. Its basic assumption is that trip generation rates are relatively stable over time for certain household stratifications. The method finds these rates empirically and for this it typically needs a large amount of data; in fact, a critical element is the number of households in each class. Although the method was originally designed to use census data in the UK, a serious problem of the approach remains the need to forecast the number of households in each strata in the future.

Let t_h^p be the average number of trips with purpose p (and at a certain time period) made by members of households of type h. Types are defined by the stratification chosen; for example, a cross-classification based on m household sizes and n car ownership classes will yield mn types h. The standard method for computing these cell rates is to allocate households in the calibration data to the individual cell groupings and total, cell by cell, the observed trips $\breve{T}_h^p$ by purpose group. The rate t_h^p is then the total number of trips in cell h, by purpose, divided by the number of households N_h in it. In mathematical form it is simply:

$$t_h^p = \frac{\tilde{T}_h^p}{N_h} \tag{4.8}$$

The 'art' of the method lies in choosing the categories such that the standard deviations of the frequency distributions depicted in Figure 4.4 are minimized.

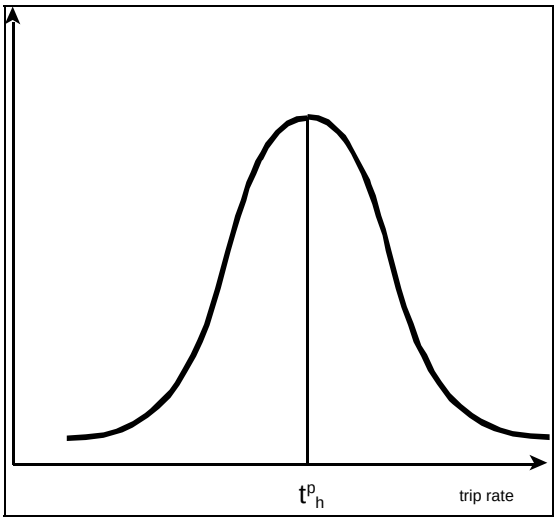


Figure 4.4: *Trip-rate distribution for household type*

The method has, in principle, the following advantages:
1. Cross-classification groupings are independent of the zone system of the study area.
2. No prior assumptions about the shape of the relationship are required (i.e. they do not even have to be monotonous, let alone linear).
3. Relationships can differ in form from class to class (e.g. the effect of changes in household size for one or two car-owning households may be different).

And in common with traditional cross-classification methods it has also several disadvantages:
1. The model does not permit extrapolation beyond its calibration strata, although the lowest or highest class of a variable may be open-ended (e.g. households with two or more cars and five or more residents).
2. Unduly large samples are required, otherwise cell values will vary in reliability because of differences in the numbers of households being available for calibration at each one. Accepted wisdom suggests that at least 50 observations per cell are required to estimate the mean reliably.
3. There is no effective way to choose among variables for classification, or to choose best groupings of a given variable; the minimization of standard deviations hinted at in Figure 4.4 requires an extensive 'trial and error' procedure which may be considered infeasible in practical studies.

Model application at zonal level

Let us denote by h the household type (i.e. with and without a car), be N_{hi} the number of households of type h in zone i. With this we can write the trip productions with purpose p by household type h in zone i, T_i^p, as follows:

$$T_i^p = \sum_h N_{hi} t_h^p \tag{4.9}$$

To verify how the model works it is possible to compare these modeled values with observed values from the calibration sample. Inevitable errors are due to the use of averages for the t_h^p; one would expect a better stratification (in the sense of minimizing the standard deviation in Figure 4.4) to produce smaller errors.

There are various ways of defining household categories. The first application in the UK (Wootton and Pick 1967), which was followed closely by subsequent applications, employed 108 categories as follows: six income levels, three car ownership levels (0, 1 and 2 or more cars per household) and six household structure groupings, as in Table 4.7. The problem is clearly how to predict the number of households in each category in the future.

Group	No. employed	Other adults
1	0	1
2	0	2 or more
3	1	1 or less
4	1	2 or more
5	2 or more	1 or less
6	2 or more	2 or more

Table 4.7: *Example of household structure grouping*

The following Table 4.8 is an example of a Dutch personal trip generation model consisting of a production and an attraction part. The table includes trip rates per unit for the evening peak hour. Separate models are given for three spatial settings (agglomerations, towns and rural) and for two trip purposes each, namely home-to-work and other.

evening peak			production		attraction		total
			labour-force	inhabitant	employment		
					retail	other	
agglomeration	**home-work**	**arrivals**	0.2282				∑
		departures				0.2626	∑
	other	**arrivals**		0.1259	0.3224	0.0257	∑
		departures		0.0633	0.3173	0.2626	∑
town	**home-work**	**arrivals**	0.2169				∑
		departures				0.2435	∑
	other	**arrivals**		0.1633	0.5934	0.0323	∑
		departures		0.895	0.6127	0.2435	∑
rural	**home-work**	**arrivals**	0.236				∑
		departures				0.2674	∑
	other	**arrivals**		0.13	0.8417	0.0372	∑
		departures		0.0744	0.7039	0.2674	∑

Table 4.8: *Personal trip rates used in the Randstadmodel [Source: Randstadmodel 1994]*

Production trip rates for home-to-work are based on working persons, whereas all other production trip rates refer to inhabitants. Trip attractions are based on employment by type (retail and other).
The trip rates were estimated using various data sources (OVG). Calculation of total origins and destinations per zone is performed by horizontally adding the various production and attraction components.

4.5.2 Estimation of trip rates by multiple class analysis (MCA)

MCA is an alternative method to define classes and test the resulting cross-classification which provides a statistically powerful procedure for variable selection and classification. This allows us to overcome several of the disadvantages cited above for other types of cross-classification methods.

Consider a model with a continuous dependent variable (such as the trip rate) and two discrete independent variables, such as household size and car ownership. A grand mean can be estimated for the dependent variable over the entire sample of households. Also, group means can be estimated for each row and column of the cross-classification matrix; each of these can be expressed in turn as deviations from the grand mean. Observing the signs of the deviations, a cell value can be estimated by adding to the grand mean the row and column deviations corresponding to the cell. In this way, some of the problems arising from too few observations on some cells can be compensated.

Example 4.4:
Table 4.9 presents data collected in a study area and classified by three car-ownership and four household-size levels. The table presents the number of households observed in each cell (category) and the mean number of trips calculated over rows, cells and the grand average.

Household size	0 car	1 car	2+ cars	Total	Mean
1 person	28	21	0	49	0.47
2 or 3 persons	150	201	93	444	1.28
4 persons	61	90	75	226	1.86
5 persons	37	142	90	269	1.9
Total	276	454	258	988	
Mean trip rate	0.73	1.53	2.44		1.54 (= grand mean)

Table 4.9: *Number of households per cell and mean trip rates for a particular purpose*

As can be seen, the values range from 0 (it is unlikely to find households with one person and more than one car) to 201. Although we are cross-classifying by only two variables in this simple example, there are already four cells with less than the conventional minimum number (50) of observations required to estimate mean trip rate and variance with some reliability. We would like to use now the mean row and column values to estimate average trip rates for each cell, including that without observations in this sample. We can compute the deviation (from the grand mean) for zero cars as 0.73 - 1.54 = -0.81; for one car as 1.53 - 1.54 = -0.01, and for two cars or more 2.44 - 1.54 = 0.90; similarly, we can calculate the deviations for each of the four household size groups as: -1.07, -0.26, 0.32 and 0.36. If the variables are not correlated with these values we can work out the full trip-rate table; for example, the trip rate for one person household and one car is 1.54 - 1.07 - 0.01 = 0.46 trips. In the case of one person and no car, the rate turns out to be negative and equal to -0.34 (1.54 -1.06- 0.82); this has no meaning and therefore the actual rate is forced to zero.

Table 4.10 depicts the full trip-rate table together with its row and column deviations.

Household size	Car ownership level			
	0 car	1 car	2+ cars	Deviations from grand mean
1 person	0	0.46	1.37	-1.07
2 or 3 persons	0.46	1.27	2.18	-0.26
4 persons	1.05	1.85	2.76	0.32
5 persons	1.09	1.89	2.8	0.36
Deviations	-0.81	-0.01	0.9	

Table 4.10: *Trip rates calculated by multiple classification*

- end of example -

Contrary to standard cross-classification models, deviations are not only computed for households in, say, the cell one person-one car; rather, car deviations are computed over all household sizes and vice versa. Thus, if interactions are present these deviations should be adjusted to account for interaction effects. This can be done by taking a weighted mean for each of the group means of one independent variable over the groupings of the other independent variables, rather than a simple mean (which would in fact be equivalent to assuming that variation is random over the data in a group). These weighted means will in general tend to decrease the sizes of the adjustments to the grand mean when interactions are present. Nevertheless, the cell means of a multiway classification will still be based on means estimated from all the available data, rather than being based on only those items of data falling in the multiway cell.

Apart from the statistical advantages, it is important to note that cell values are no longer based on only the size of the data sample within a given cell; rather, they are based on a grand mean derived from the entire data set, and on two (or more) class means which are derived from all data in each class relevant to the cell in question.

Example 4.5:
Table 4.11 provides a set of rates computed in the standard category analysis procedure (i.e. by using individual cell means). These values may be compared with those of Table 4.10.

Two points of interest emerge from the comparison. First, there are rates available even for empty cells in the MCA case. Second, some counterintuitive progressions, apparent in Table 4.11 (e.g. the decrease of rate values for 0 and 1 car-owning households when increasing household size from 4 to 5 or more), are removed in Table 4.10. Note that they could have arisen by problems of small sample size at least in one case.

Household size	Car ownership level		
	0 car	1 car	2+ cars
1 person	0.12	0.94	
2 or 3 persons	0.6	1.38	2.16
4 persons	1.14	1.74	2.6
5 persons	1.02	1.69	2.6

Table 4.11: *Trip rates calculated using ordinary category analysis (Section 4.5.1)*

- end of example -

4.5.3 The person-category approach

This is an interesting alternative to the household-based models discussed above. This approach offers the following advantages:

1. A person-level trip generation model is compatible with other components of the classical transport demand modeling system, which is based on tripmakers rather than on households;
2. It allows a cross-classification scheme that uses all important variables and yields a manageable number of classes; this in turn allows class representation to be forecast more easily;
3. The sample size required to develop a person-category model can be several times smaller than that required to estimate a household-category model;

4. Demographic changes can be more easily accounted for in a person-category model as, for example, certain key demographic variables (such as age) are virtually impossible to define at household level;
5. Person categories are easier to forecast than household categories as the latter require forecasts about household formation and family size; these tasks are altogether avoided in the case of person categories. In general the bulk of the trips are made by people older than 18 years of age; this population is easier to forecast 15 to 20 years ahead as only migration and survival rates are needed to do so.

The major limitation that a person-category model may have relates precisely to the main reason why household-based models were chosen to replace zonal-based models at the end of the 1960s; this is the difficulty of introducing household interaction effects and household money costs and money budgets into a person-based model.

Variable definition and model specification

The estimation of person-based trip rates per person type follows the same line as explained before with respect to households (MCA). Model development entails the following stages:

1. Consideration of several variables which are expected to be important for explaining differences in personal mobility. Also, definition of plausible person categories using these variables;
2. Preliminary analysis of trip rates in order to find out which variables have the least explanatory power and can be excluded from the model. This is done by comparing the trip rates of categories which are differentiated by the analyzed variable only and testing whether their differences are statistically significant;
3. Detailed analysis of trip characteristics to find variables that define similar categories. Variables which do not provide substantial explanation of the data variance, or variables that duplicate the explanation provided by other better variables (i.e. easier to forecast or more policy responsive) are excluded. The exercise is conducted under the constraint that the number of final categories should not exceed a certain practical maximum (for example, 15 classes).

For this analysis the following measures may be used: the coefficient of correlation R_{jk}, slope m_{jk} and intercept a_{jk} of the regression $t_j^p = a_{jk} + m_{jk} t_k^p$. The categories j and k may be treated as similar if these measures satisfy the following conditions:

$R_{jk} > 0.900$
$0.75 < m_{jk} < 1.25$
$a_{jk} < 0.10$

These conditions are quite demanding and may be changed.

Model application at the aggregate level

Let t_n be the trip rate, that is, the number of trips made during a certain time period by (the average) person in category j; t_n^p, is the trip rate by purpose p. T_i is the total number of trips made by the inhabitants of zone i (all categories together). N_i is the number of inhabitants of zone i, and α_{ni} is the percentage of inhabitants of zone i belonging to category n. Therefore the following basic relationship exists:

$$T_i^p = N_i \sum_n \alpha_{ni} t_n^p \tag{4.10}$$

Example 4.6:

Another Dutch example of a cross-classification trip rate model is WOLOCAS. This model predicts trip productions for new residential areas. The trip rates refer to person types classified by sex, car ownership, occupational status, education and age (see Tables 4.12-4.14). The trip production model

distinguishes three trip purposes. The daily trip rates were estimated using the continuous National Dutch Mobility Survey. In applying this model estimates are required for the demographic size and composition of the new residential areas. The trip rates only apply to persons over 12 years old. [Source: Wolocas, 1990].

Trip purpose WORK		**education**		
		low	**medium**	**HBO/univ**
working man	with car	1.7	1.743	1.702
	without car	1.656	1.656	1.656
working woman	with car	1.369	1.331	1.331
	without car	1.268	1.31	1.31

Table 4.12: *Personal daily trip production rates for "work", split by person type.*

Trip purpose SERVICES		**age**	
		12 - 18 year	**> 18 year**
man	with car		0.656
	without car	1.983	0.917
woman	with car	-	1.111
	without car	1.983	1.201
average		1.983	

Table 4.13: *Personal daily trip production rates for "services", split by person type.*

Trip purpose OTHER		**age**	
		12 – 40 year	**> 40 year**
working	with car	1.185	0.779
	without car	0.974	0.98
not-working	with car	1.158	1.442
	without car	1.512	0.98

Table 4.14: *Personal daily trip production rates for "other", split by person type.*

- end of example -

4.6 Discrete choice methods

Since individuals choose whether to make specific trips, discrete choice models such as binary logit can be used to predict trip production. With binary logit, the probability that an individual will choose to make one or more trips (as opposed to not traveling) can be expressed as:

$$P_n(1+) = \frac{1}{1+e^{\beta(x_{0n}-x_{1n})}}$$
$$P_n(0) = 1 - P_n(1+) \qquad (4.11)$$

where:

$P_n(0)$ = the probability that a person n will make no trip
$P_n(1+)$ = the probability that a person n will make one or more trips
β = the vector of coefficients that is estimated by the model
x_{1n} = the vector of explanatory variables in person n's utility of making one or more trips
x_{0n} = the vector of explanatory variables in person n's utility of not making a trip

From the estimated coefficients, you can see how the explanatory variables will impact the probability with which an individual will make a certain trip.

In addition, you can aggregate the disaggregated probabilities to obtain the proportion of the population that will take this type of trip, and thus generate the aggregate number of trips produced by a zone.

Interpreting the results of a logit model
A model yielded the following results:

Parameter	Estimate	t-Stat
Constant	-0.474	-2.4
Sex	0.267	2.1
Age	-0.047	-19.1
Married Female	0.314	2.8
Married Male	1.594	12.4
Fem w/ Child<6	-1.742	-11.4
Education	0.211	13.8

Table 4.15: *Results of a logit model predicting trip making probability*

with a goodness-of-fit ('Adjusted Rho Squared') of 0.22. The variables Sex, Age, Married Female, Married Male, Fem w/ Child < 6 are all dummy variables (i.e. equal to 1 or 0). Note that all of the coefficients are significant at a 95% confidence level (t-statistic > 2).

This logit model predicts the probability with which an individual will make a work trip according to the following equation:

$$P(\text{trip}) = \frac{1}{1 + e^{0.474 - 0.267(\text{Sex}) + 0.047(\text{Age}) - 0.314(\text{MarFem}) - 1.594(\text{MarMale}) + 1.742(\text{Chld<6}) - 0.211(\text{Educ})}}$$
$$P(\text{no trip}) = 1 - P(\text{trip}) \qquad (4.12)$$

From the estimated coefficients, you can see how the explanatory variables will impact the probability with which an individual makes a trip to work. For example, the coefficient for education (0.211) suggests that, all else being equal, people with more education are more likely to make work trips than those with less education. Note that none of the signs from the model seem unreasonable. You can also use the equation above to calculate how a change in an explanatory variable will impact a person's probability of making a work trip. For example, a person with a high school diploma (Educ = 10) who has a 50% probability of making a work trip, would, all else being equal, have a 70% probability of making a work trip if he or she had a Bachelor's Degree.

Example 4.7:
The Dutch National Transport Model System is fully based on disaggregated discrete choice models of travel demand. The trip frequency submodel predicts the probability that an individual with known characteristics will make no, one or more round tours for specific trip purposes on an average day. In later stages, tours are decomposed into trips.
The unit of analysis is a person, member of a household with known characteristics. This trip production model distinguishes 5 trip purposes, further divided by person type. In total, eight different models are distinguished.
The model has a two-stage structure:
– in step 1, the model predicts the probability of making no versus one or more tours, thus a binary-choice model.

– in step 2, another binary choice model predicts the conditional probability of making one tour versus two or more tours given the fact that tours are being made. This process is repeated several times.

Model 2 is applied consecutively in the cascade.
This second model is identical in all these steps (but is different from model 1). It is a so-called stop/repeat-model.
Both models are binary logit models with linear utility functions including dummy variables describing personal and household characteristics.
In practical applications, instead of working with probabilities, the expected number of tours is calculated.

This expected number of tours (*ENT*) per purpose may be calculated using:

$$ENT = \frac{P_r(1+)}{1 - P_r(R)} \tag{4.13}$$

where $P_r(1+)$ is the probability outcome of model 1, and $P_r(R)$ is the probability of making additional trips which is the outcome of model 2.

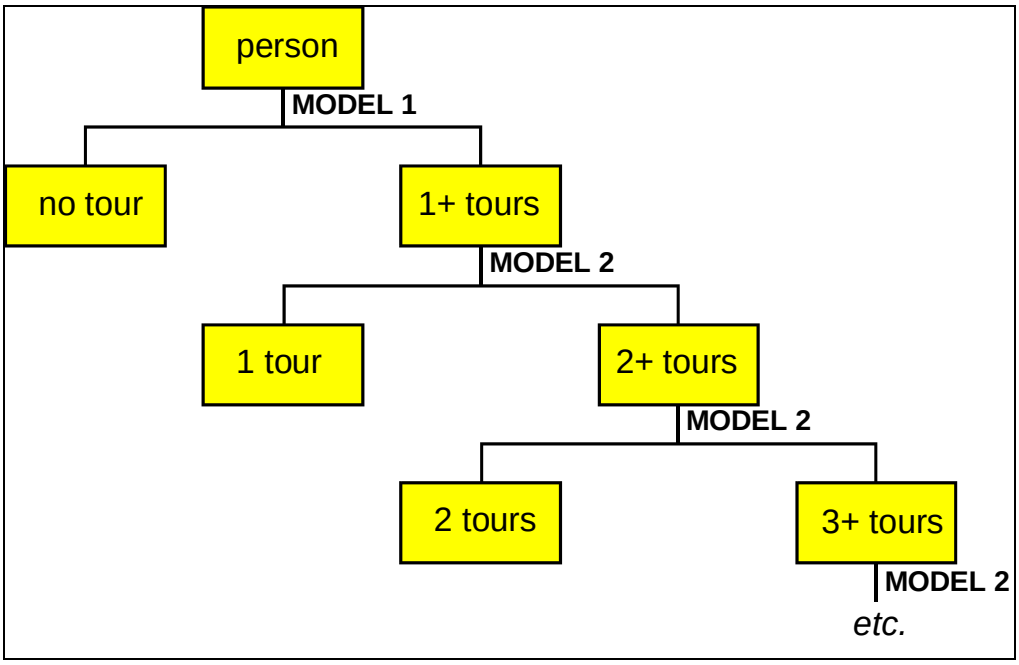


Figure 4.5: *Structure of disaggregated trip choice models for trip production in the Dutch National Transport Model System [Source: LMS, 199?]*

- end of example -

4.7 Trip balancing

It might be obvious to some readers that the models above do not guarantee, by default, that the total number of trips originating (the origins O_i) at all zones will be equal to the total number of trips attracted (the destinations D_j) to them, that is the following expression does not necessarily hold:

$$\sum_i O_i = \sum_j D_j \tag{4.14}$$

The problem is that this equation is implicitly required by the next sub-model (i.e. trip distribution) in the structure; it is not possible to have a trip distribution matrix where the total number of trips (T) obtained by summing all rows is different to that obtained when summing all columns.

The solution to this difficulty is a pragmatic one which takes advantage of the fact that normally the trip generation models are far 'better' (in every sense of the word) than their trip attraction counterparts. The first normally are fairly sophisticated household-based models with typically good explanatory variables. The trip attraction models, on the other hand, are at best estimated using zonal data. For this reason, normal practice considers that the total number of trips arising from summing all origins O_i is in fact the correct figure for T; therefore, all destinations D_j are multiplied by a factor f given by:

$$f = \frac{T}{\sum_j D_j} \tag{4.15}$$

which obviously ensure that their sum adds to T.

Several procedures can be used to balance trip productions and attractions in which productions and attractions from several trip purposes can be balanced in one step. The procedure offers the following methods for balancing:

- *Hold Productions Constant*
  Productions are held constant and the attractions are adjusted so that their sum equals the sum of the productions.
- *Hold Attractions Constant*
  Attractions are held constant and the productions are adjusted so that their sum equals the sum of the attractions.
- *Weighted Sum of Productions and Attractions*
  Both productions and attractions are adjusted so that their sums equal the user specified weighted sum of productions and attractions.
- *Sum to User Specified Value*
  Both productions and attractions are adjusted so that their sums equal a user specified value.

REGRESSION MODELS		
level	**equation**	**notes**
zonal	$T_i^p = \sum_k \alpha_k^p X_{ik}$	T_{ip} = zonal production of trips for purpose p X_{ik} = k^{th} zonal explanatory variable of zone i
household	$T_i^p = \sum_h N_{hi} \sum_k \beta_{hk}^p Y_{hk}$	Y_{hk} = k^{th} household explanatory variable for household type h N_{hi} = no. of households of type h in zone i
person	$T_i^p = \sum_n N_{ni} \sum_k \gamma_{nk}^p Z_{nk}$	Z_{pk} = k^{th} personal explanatory variable for person type p N_{ni} = no. of persons of type n in zone i
CROSS CLASSIFICATION		
level	**equation**	**notes**
household	$T_i^p = \sum_h N_{hi} t_h^p$	t_h^p = trip rate per household of type h
person	$T_i^p = \sum_n N_{ni} t_n^p$	t_n^p = trip rate per person of type n
DISCRETE CHOICE MODELS		
level	**equation**	**notes**
household	$T_i^p = \sum_h N_{hi} t_h^p \sum_{x=0}^{\max} x P_{hp}(x)$	$P_{hp}(x)$ = probability that a random household from householdgroup h will make x trips (x = 0,1,2,..,max) for purpose p
person	$T_i^p = \sum_n N_{ni} t_n^p \sum_{x=0}^{\max} x P_{np}(x)$	$P_{np}(x)$ = probability that a random person from persongroup n will make x trips (x = 0,1,..,max) for purpose p

Table 4.16: *Summary of model specifications for trip generation models*

4.8 Summary

To summarize the various approaches to trip generation modeling, Table 4.16 offers the prediction formulae to be used for each approach. We define the following:

p = trip purpose index;

i = zone index; T_i^p

T_i^p = predicted number of trips for zone i (origins or destinations, productions or attractions) for purpose p;

α, β, γ = trip generation parameters at zonal, household, and personal level respectively.

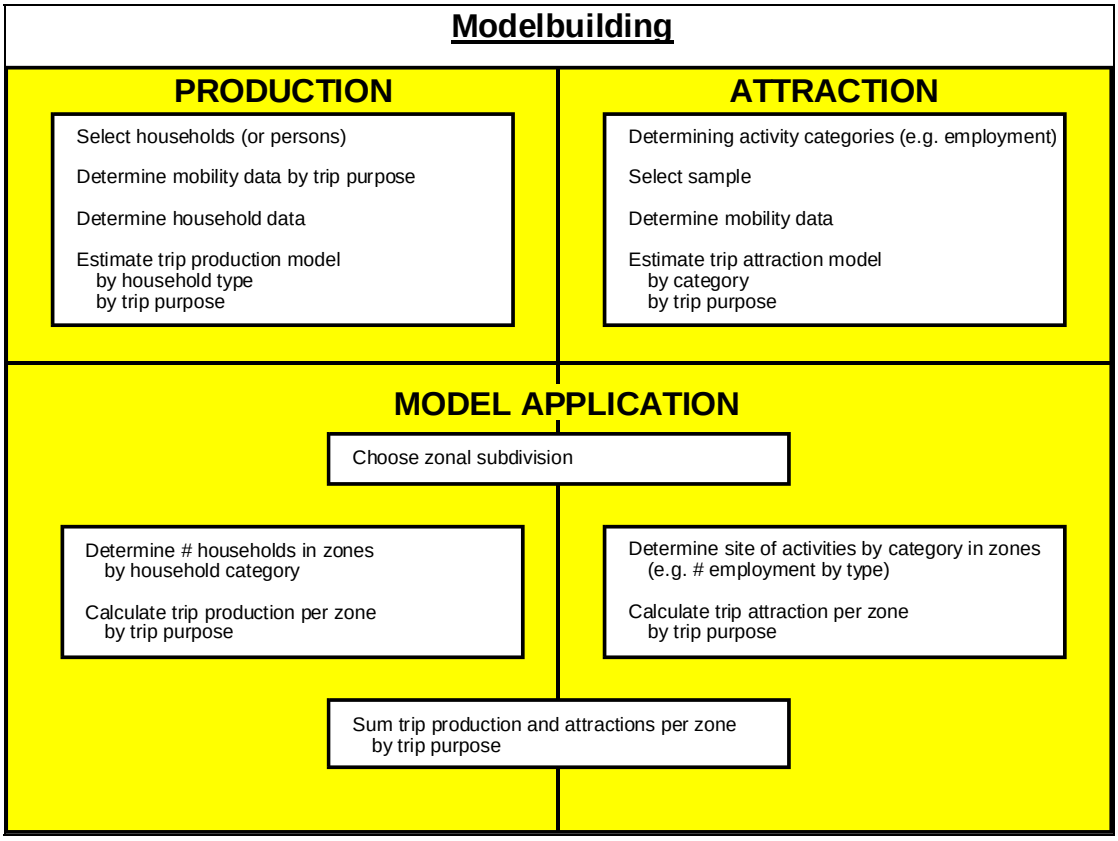


Figure 4.6: *Summary of procedure to calculate origins (departure) and zonal destinations (arrivals)*

4.9 References

P.H.L. Bovy & R. Kitamura
Invloed van paneluitval op ritfrequentieschattingen
CVS Proceedings, 1986

DHV
Het Randstadmodel. Verkeer en vervoer in de Randstad
Technische rapportage en toetsing
Amersfoort, DHV, September 1994

INRO-TNO (Verroen & Van der Vlist)
WOLOCAS2: hoofdlijnen en technische documentatie
Delft, INRO, April 1993

Rijkswaterstaat, Dienst Verkeerskunde
Landelijk Model Systeem voor verkeer en vervoer
Rijkswaterstaat, Rotterdam, 1990

T.H. Wonnacott & R.J. Wonnacott
Regression; A second course in statistics
Wiley, New York, 1980

5 Trip distribution models

5.1 Introduction

After all available and relevant information on the number of trips departing or arriving in each zone has been collected, the next step in transport modeling is to *distribute* these trips over origin-destination (OD) cells. This can be done either at *disaggregated* or at *aggregate* level. In disaggregated trip distribution, destination choice proportions are simulated using information on individual characteristics. In aggregate trip distribution individual characteristics are not considered. In this course we will only discuss aggregate trip distribution models.

The point of departure to the aggregate trip distribution step are the margins of the origin-destination (OD) table computed in the trip generation step. This implies estimates are available for either:
- the number of trip departures,
- the number of trip arrivals,
- the number of trip departures and the number of trip arrivals,
- none.

Which case is true depends on the purpose of the study and the availability of data and leads to separate trip distribution models. Figure 5.1 schematizes the trip distribution problem for the case where both the number of trip departures and the number of trip arrivals are known: the objective is to forecast an OD-table based on estimates of future productions and attractions and measurements of current OD-flows, or measurements of the generalized cost of each trip.

Origins	Destinations				$\sum_j T_{ij}$
	1	2	...	n	
1					P1
2		?			P2
...					...
m					Pm
$\sum_i T_{ij}$	A1	A2	...	An	

Figure 5.1: *Doubly constrained trip distribution*

Two basic categories of aggregate trip distribution methods predominate in urban transportation planning:

- The first basic category of aggregate trip distribution methods is based on the gravity model (Sections 5.2 - 5.6). For gravity models, typical inputs include one or more flow matrices, an impedance matrix reflecting the distance, time, or cost of travel between zones, and estimates of future levels of productions and attractions. The gravity model explicitly relates flows between zones to interzonal impedance to travel.

  The gravity model was originally motivated by the observation that flows decrease as a function of the distance separating zones, just as the gravitational pull between two objects decreases as a function of the distance between the objects. As implemented for planning models, the Newtonian analogy has been replaced with the hypothesis that the trips between zones *i* and *j* are a function of trips originating in zone *i* and the relative

attractiveness and/or accessibility of zone *j* with respect to all zones.

Modern derivations of the gravity model show that it can be understood as the most likely spatial arrangement of trips given limited information available on zonal origin totals, zonal destination totals, and various supporting assumptions or constraints about mean trip lengths (Ortúzar en Willumsen, 1994).

Many different measures of impedance can be used, such as travel distance, travel time, or travel cost. There are also several potential impedance functions used to describe the relative attractiveness of each zone. Popular choices are the exponential functions typically used in entropy models. As an alternative to impedance functions, one can use a friction factor lookup table (essentially a discrete impedance function) that relates the impedance between zones to the attractiveness between zones.

Prior to applying a gravity model, one has to calibrate the impedance function. Typically, calibration entails an iterative process that computes coefficients such that the gravity model replicates the trip length frequency distribution and matches base year productions, or attractions, or both (see Chapter 9).

- The second basic category of aggregate trip distribution are the growth factor methods (Section 5.7). These involve scaling an existing matrix (called base matrix) by applying multiplicative factors (often derived from predicted productions and/or attractions) to matrix cells.

Practical Issues
Some of the classical growth factor methods do not take into account any information about the transportation network, and thus cannot reflect impacts of changes in the network. This may be reasonable for very short-term forecasts, but it is invalid for medium to long-term forecasts for which the network has changed, or to forecast scenarios that include changes in the network. Since most transportation planning involves analysis of transportation networks, gravity models or more sophisticated destination choice models should be used.

In aggregate analysis, the choice of the impedance function should be based upon the mathematical properties of the function and the data distributions to be modeled. In practice, the selection of the functional form of the model should be based on the shape of the measured trip length distribution; consequently, examination of empirical trip length distributions is an important input into the decision process. Both smooth impedance functions and discrete functions (i.e. friction factors) can be used, as well as hybrid functions combining functions or utilizing smoothed discrete values (see Section 5.6).

Distribution models should be estimated and applied for several trip purposes. The rationale for this is that both the alternatives and individuals' willingness to travel differ greatly by purpose.

5.2 Derivation of the gravity model

The gravity model in its general form states that the number of trips between an origin and destination zone is proportional to the following three factors:
- a factor for the origin zone (the production ability)
- a factor for the destination zone (the attraction ability)
- a factor depending on the travel costs between origin and destination zone

Mathematically this is summarized as follows:

$$T_{ij} = \mu Q_i X_j F_{ij} \tag{5.1}$$

with:

T_{ij} = number of trips from zone i to zone j
Q_i = production ability of zone i
X_j = attraction ability of zone j
F_{ij} = accessibility of j from i (depends on travel costs c_{ij})
μ = measure of average trip intensity in area

In this chapter this model will be referred to as the general trip generation model; it will be made more specific for cases where extra information is available, such as the number of arrivals or departures.

It can be seen that above model is in line with intuitively clear symmetry assumptions: if two possible destination zones have similar attraction abilities and are equally accessible from an origin zone, there is no reason to expect that more trips will be made from that origin zone to the first destination zone than to the second destination zone.

The model can be formally underpinned using the utility theory described in Section 2.2 as follows.

According to utility theory, decision makers aim at maximizing their perceived net utilities. Utility is derived from activities. To maximize utility, in general multiple types of activities are needed during a day, e.g. working and living.

Individual utility U_{ijp} of making a trip from origin i to destination j for a specific homogeneous travel purpose (e.g. home to work) is:

$$U_{ijp} = U_i + U_j - f(c_{ij}) + \varepsilon_{ijp} \tag{5.2}$$

U_i = average utility of origin bound activity in i
U_j = average utility of destination bound activity in j
$f(c_{ij})$ = utility value of travel resistance (cost) between i and j
ε_{ijp} = individual error term, accounting for misperceptions, taste variation and non-modeled attributes

Define

$$V_{ij} = U_i + U_j - f(c_{ij}) \tag{5.3}$$

Now we can write

$$U_{ijp} = V_{ij} + \varepsilon_{ijp} \tag{5.4}$$

If we assume that the error term ε_{ijp} is Gumbel distributed with scale parameter b (logit-assumptions, see Section 2.3), then for each decision maker the probability that he will opt for a trip from zone i to zone j equals:

$$P_{ij} = \frac{\exp(bV_{ij})}{\sum_{ij} \exp(bV_{ij})} = \frac{1}{k} \exp(bU_i) \cdot \exp(bU_j) \cdot \exp(-bf(c_{ij})) \tag{5.5}$$

with

$$k = \sum_{ij} \exp(bV_{ij}) \tag{5.6}$$

P_{ij} = probability that an individual will make a trip from i to j
b = scale parameter in Gumbel distribution
k = a measure for the number of and variability in trip alternatives. The larger k, the more choice opportunities for a traveler.

With P travelers, the expected number of trips between i and j amounts to:

$$T_{ij} = \mu Q_i X_j F_{ij} \tag{5.7}$$

which is the general trip distribution model with:

Q_i = production potential $= \exp(bU_i)$ (5.8)
X_j = attraction potential $= \exp(bU_j)$ (5.9)
F_{ij} = accessibility of j from $= \exp(-bf(c_{ij}))$ (5.10)
μ = measure of average trip intensity in area $= P/k$

Taking this general trip distribution model as starting point, we can formulate various derived models depending on additional constraints imposed on the model, especially on the number of trip arrivals and departures in the zones (see Table 5.1).

	Departures unknown	**Departures known**
Arrivals unknown	Direct Demand (Section 5.3)	Origin Constraint (Section 5.4.1)
Arrivals known	Destination Constraint (Section 5.4.2)	Doubly Constraint (Section 5.5)

Table 5.1: *Types of distribution models according to imposed constraints on arrivals and departures*

5.3 Direct demand model

In this case, no additional trip constraints are imposed, so this model equals the general trip distribution model:

$$T_{ij} = \mu Q_i X_j F_{ij} \tag{5.11}$$

with:
Q_i = production potential of i
X_j = attraction potential of j
F_{ij} = accessibility of j from i
μ = measure of average trip intensity in area

The production and attraction potentials may be derived from population, area, number of jobs, etc. Both the *numbers* of departures (productions) and arrivals (attractions) are

unknown. They are determined endogenously. The resulting flows are estimated solely on the potentials of zones *i* and *j* and the impedance between them.

Although the direct demand model is easy to implement, a disadvantage of the direct demand model is that it predicts a large number of trips per unit of analysis (e.g. person) for particularly accessible origin zones (zones that have many high attraction potential zones nearby). This is not realistic under all circumstances. For example, the number of home-work trips per person will in general not increase, even if many job opportunities are close-by. For this reason this method is rarely used in practice.

5.4 Singly constrained trip distribution model

5.4.1 Origin constrained

In the origin constrained trip distribution model, the number of trip departures P_i are imposed as a set constraints on the general trip distribution model:

$$\sum_j T_{ij} = P_i \tag{5.12}$$

where P_i is the known number of trips departing from zone *i*, which is determined exogenously (for example estimated using a trip generation model). Combining this with the general trip distribution model

$$T_{ij} = \mu Q_i X_j F_{ij} \tag{5.13}$$

we can write

$$\sum_j T_{ij} = \sum_j (\mu Q_i X_j F_{ij}) = \mu Q_i \sum_j (X_j F_{ij}) = P_i \tag{5.14}$$

Solving for Q_i yields

$$Q_i = \frac{P_i a_i}{\mu} \tag{5.15}$$

where a_i is defined as $a_i = \dfrac{1}{\sum_j (X_j F_{ij})}$

Combining (5.13) and (5.15) gives the origin constrained distribution model

$$T_{ij} = P_i \frac{X_j F_{ij}}{\sum_j (X_j F_{ij})} = a_i P_i X_j F_{ij} \tag{5.16}$$

with:

a_i = balancing factor
P_i = number of trips departing from zone *i*
X_j = attraction potential of zone *j*
F_{ij} = accessibility of zone *j* from zone *i*

The origin constrained trip distribution model is therefore a *proportional model* that splits the given trip numbers originating in *i* over the destinations *j* in proportion to their relative accessibility and utility opportunity.

Although different definitions of accessibility may be used, the factor $\sum_j (X_j F_{ij})$ is often referred to as the *accessibility* of zone *i*. By dividing the total number of departures by the accessibility of a zone, we avoid the phenomenon that causes the total number of departures from an origin zone to increase if it is close to zones with high attraction. Of course, if a *destination* is highly accessible, i.e. many origin zones with high production abilities are close-by, this will still result in a large number of trip *arrivals*. This might be an unwanted effect if the attraction ability is based on, e.g., the number of jobs in a zone.

Equation (5.16) shows that the absolute levels of X_j and F_{ij} are not essential in this proportional model. If we multiply each of these variables by an arbitrary constant factor, this would not affect the model outcomes. This characteristic leaves a lot of freedom in specifying both variables.

A practical example of an origin constrained model is the WOLOCAS model. This model is designed to predict the impact of the development of new residential areas (e.g. VINEX). While the number of trips originating from these areas is estimated using detailed trip generation models, no explicit estimate is made of the number of trip arrivals. Instead, attraction abilities are supplied for different trip purposes based on the number of jobs (home-work), the number of facilities (shopping), and the number of inhabitants (other trip purposes).

[see also: Ortúzar & Willumsen, 1990, p.136]

5.4.2 Constrained to destinations

In analogy with the derivation of the origin constrained trip distribution model, the *destination constrained* trip distribution may be derived. The internal trip numbers are constrained to exogenously given arrivals. The number of arriving trips A_j in *j* is known (e.g. by using a separate trip generation model), which implies:

$$\sum_i T_{ij} = A_j \quad \forall j \tag{5.17}$$

Skipping the derivation (which is analogous to the derivation in Section 5.4.1), the model is given below:

$$T_{ij} = A_j \frac{Q_i F_{ij}}{\sum_i (Q_i F_{ij})} = b_j Q_i A_j F_{ij} \tag{5.18}$$

with:

b_j = balancing factor = $\dfrac{1}{\sum_i (Q_i F_{ij})}$
Q_i = production potential of *i*
A_j = number of trips arriving at zone *j*
F_{ij} = accessibility of *j* from *i*

The destination constraint trip distribution model is therefore a *proportional model* that splits the given trip numbers arriving at j over the origins i in proportion to their relative accessibility and utility opportunity.

[see also: Ortúzar & Willumsen, 1990, p.136]

In this case, also, the model outcomes are not sensitive to the absolute levels of Q_i and F_{ij} (see equation (5.18)). Multiplying both variables with an arbitrary constant does not change the results. This characteristic gives considerable freedom in defining these variables.

5.5 Doubly constrained trip distribution model

The doubly constrained model arises if both the number of trip departures and the number of trip arrivals are imposed on the general trip distribution model. The derivation of the doubly constrained trip distribution model is as follows. We again start with the general trip distribution model:

$$T_{ij} = \mu Q_i X_j F_{ij} \tag{5.19}$$

Now we have two sets of constraints in that the numbers of arrivals and departures in the zones are exogenously given. Thus, the number of arriving trips A_j at j is known and the number of departing trips P_i from i is known (again, e.g., using separate models). This yields

$$\sum_j T_{ij} = P_i \tag{5.20}$$

and

$$\sum_i T_{ij} = A_j \tag{5.21}$$

Hence,

$$\sum_j T_{ij} = \sum_j (\mu Q_i X_j F_{ij}) = \mu Q_i \sum_j (X_j F_{ij}) = P_i \tag{5.22}$$

and

$$\sum_i T_{ij} = \sum_i (\mu Q_i X_j F_{ij}) = \mu X_j \sum_i (Q_i F_{ij}) = A_j \tag{5.23}$$

Solving for Q_i and X_j:

$$Q_i = \frac{P_i a_i}{\mu} \tag{5.24}$$

and

$$X_j = \frac{A_j b_j}{\mu} \tag{5.25}$$

where a_i and b_j are balancing factors for the trip constraints, defined by

$$a_i = \frac{1}{\sum_j (X_j F_{ij})} \tag{5.26}$$

and

$$b_j = \frac{1}{\sum_i (Q_i F_{ij})} \tag{5.27}$$

Hence, the doubly constrained trip distribution model now is:

$$T_{ij} = \frac{1}{\mu} a_i b_j P_i A_j F_{ij} \tag{5.28}$$

The parameter μ may be included in the estimated values for a_i and b_j resulting in:

$$T_{ij} = a_i b_j P_i A_j F_{ij} \tag{5.29}$$

with:
- a_i = balancing parameter
- b_j = balancing parameter
- P_i = number of trips departing at zone i
- A_j = number of trips arriving at zone j
- F_{ij} = accessibility of zone j from i

Whereas the trip distribution can be computed directly with the non-constrained and the singly constrained trip distribution models (provided sufficient input data are available), this is not the case with the doubly constrained trip distribution model. If all input data are available (i.e. the number of departures, P_i, the number of arrivals, A_j, and the values of the distribution function, F_{ij}), equations (5.24) - (5.27) define the coefficients a_i and b_j in model (5.29) in an implicit manner. To determine these coefficients, an iterative procedure may be used. The following example illustrates such a procedure.

Example 5.1: trip distribution using a doubly constrained model
Consider a study area consisting of two zones. The following data on population and labor are available:

zone	inhabitants	jobs
1	1000	300
2	800	200

It should be emphasized that in this case the number of inhabitants was determined more accurately than the number of jobs. From national data it follows that the number of work-related trips is on average 0.25 per person per day. The number of work-related trips arriving in a zone is 0.8 for each job. The travel resistance may be assumed to be equal for all OD-pairs in this example.

Questions:
(a) Formulate the doubly constrained distribution model and define its variables.
(b) Compute the trip distribution using the doubly constrained trip distribution model.

Answers:
(a) $T_{ij} = a_i b_j P_i A_j F_{ij}$

For the definition of the variables, see above.
(b) If the travel resistance is equal for all OD-pairs, distribution function values equal to 1 may be used; i.e. $F_{ij} = 1$. The trip generation is obtained as follows:

$$\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} = 0.25 \begin{pmatrix} 1000 \\ 800 \end{pmatrix} = \begin{pmatrix} 250 \\ 200 \end{pmatrix} \text{ and } \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = 0.8 \begin{pmatrix} 300 \\ 200 \end{pmatrix} = \begin{pmatrix} 240 \\ 160 \end{pmatrix}$$

Now the values of the balancing factors a_i and b_j need to be computed. This is done in an iterative way. As an initializing step we fill a tableau with the known values of P_i, A_j, and F_{ij}:

from zone	to zone 1	2	total	P_i	factor
1	1	1	2	250	
2	1	1	2	200	
total	2	2	4	450	
A_j	240	160	400		
factor					

The total number of trip departures (450) does not match the total number of arrivals (400). The first step is therefore to balance them. This is done by multiplying the trip arrivals with a factor 450/400 (because the number of departures is more accurately known):

from zone	**to zone 1**	**2**	**total**	**P_i**	**factor**
1	1	1	2	250	125
2	1	1	2	200	100
total	2	2	4	450	
A_j	270	180	450		
factor					

The following step is factor each row in order to match the row totals (departures):

from zone	**to zone 1**	**2**	**total**	**P_i**	**factor**
1	125	125	250	250	
2	100	100	200	200	
total	225	225	450	450	
A_j	270	180	450		
factor	270/225	180/225			

And each column in order to match the column totals (arrivals):

from zone	**to zone 1**	**2**	**total**	**P_i**	**factor**
1	150	100	250	250	1
2	120	80	200	200	1
total	270	180	450	450	
A_j	270	180	450		
factor	1	1			

In this case there is no use in performing more iterations as this would not alter the solution anymore. The table above gives the requested trip distribution.

[see also: Ortúzar & Willumsen, 1990, p.136]

- end of example -

5.6 Distribution functions

As can be deduced from Example 5.1, knowledge of distribution functions is essential when applying trip distribution models. The distribution function F (also referred to as deterrence function) represents the relative willingness to make a trip as a function of the generalized

travel costs c_{ij}. In general such a function will be a monotonously decreasing function of travel costs.
Usually different distribution functions are used, depending on the trip purpose and the attributes of the trip maker. The difference can be attributed to the fact that different categories of a population (students, housewife, employees) value travel resistance in a different way, due to differences in monetary and time budgets. A requirement to apply these different distribution functions of course is that the trip production and attraction can be estimated by category. In the Netherlands most trip distribution models distinguish between different trip purposes (e.g. work, business and others), while some also distinguish between different categories of travelers (e.g. car available, and no car available).

A typical example of the shape of distribution functions for various travel modes is shown in Figure 5.2. All functions decrease monotonously. The distribution function for the bicycle exceeds the others for small travel distances, but decreases to approximately zero at 10 km. The public transit curve is smallest initially, but barely decreases. At a distance of approximately 50 kilometres the public transit curve exceeds the car curve.

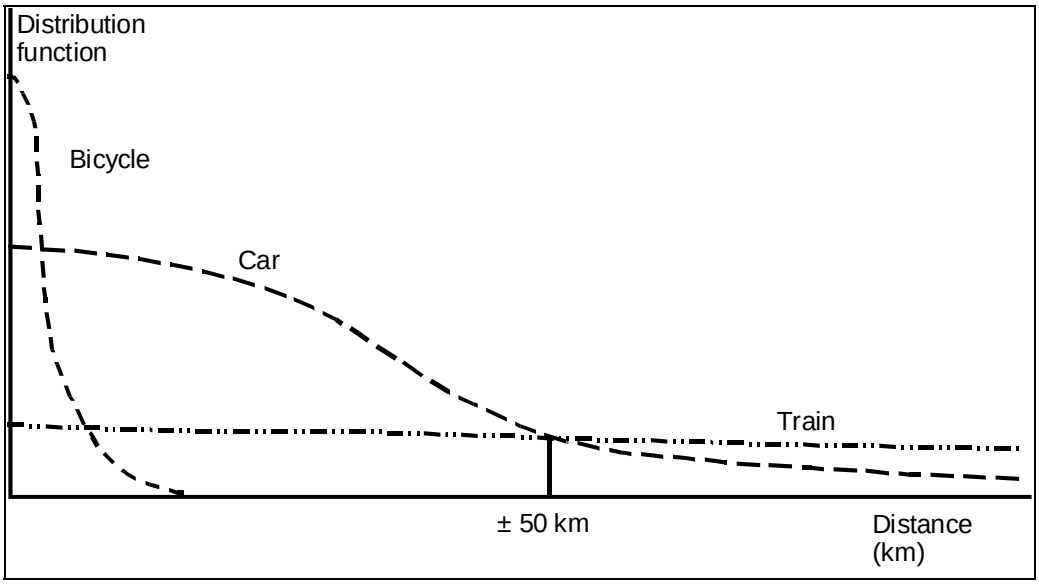


Figure 5.2: *Example of distribution function for various modes*

5.6.1 Mathematical requirements

Distribution functions are usually estimated on the basis of empirical data. In this procedure the following theoretical requirements may be imposed:

1. The function should be decreasing with generalized travel time; a larger travel time should lead to diminished willingness to make a trip:

   $$F(c_{ij}) \geq F(c_{ij} + \Delta c) \text{ if } \Delta c > 0$$

2. The expression $\int_0^\infty F(c)cdc$ is finite.

   This requirement implies that a limited number of trips originate from each zone, even if the study area is not bounded. If this requirement is not met, the number of trips depends

on the boundaries of the study area, preventing model parameters from being transferable to other studies.

Power functions (see next section) with an exponent less than or equal to 2 do not meet this requirement.

3. The fraction $\frac{F(ac_{ij})}{F(c_{ij})}$ depends on the value of c_{ij}.

   This requirement expresses that if the travel costs decrease with a constant factor, this has an impact on the trip distribution.

4. Fixed absolute changes should have a diminishing relative impact on the willingness to make a trip:

$$\frac{F(c_{ij}+\Delta c)}{F(c_{ij})} > \frac{F(c_{ij}+A+\Delta c)}{F(c_{ij}+A)} \text{ if } A>0$$

   The exponential distribution function (see next section) does not meet this requirement.

Remark – Practical versus mathematical requirements:
Although these requirements stem from intuitively clear arguments, some of them are ignored in practice in order to obtain a better fit with empirical data. For example: when using a unimodal car-trip distribution model (only car trips modeled), a best fit with empirical data is obtained if a distribution function is chosen that doesn't meet the first requirement. Instead of decreasing over the full travel distance domain, a typical distribution function in such a unimodal model first increases, reaches a maximum between 1 and 3 kilometres, and then decreases. This reflects the idea that for short distances the willingness to make a trip by car is low, due to competition of the bicycle (see Figure 5.2). It should be noted that a more elegant solution would be to take both travel modes into account simultaneously as is discussed in the chapter on mode choice.

5.6.2 Continuous distribution functions

Over time, different mathematical forms of distribution function have been proposed. The following overview is by no means complete, but is quite representative for Dutch practice. Note that all distribution functions are written with an index *m* (mode choice). If only one mode is considered, this index may be omitted.

Power:

$$F_{ijm}(c_{ijm}) = c_{ijm}^{-\alpha_m} \tag{5.30}$$

This function is believed to be relatively accurate for large travel distances or costs, and less so for small distances. It is rarely used in practice. If α_m=2, the Newtonian model is obtained. Although the name may suggest differently, in transport planning the name 'gravity model' is used for trip generation models using a wide range of distribution functions, including the Newtonian.

Exponential:

$$F_{ijm}(c_{ijm}) = \alpha_m \exp(\beta_m c_{ijm}) \tag{5.31}$$

A counterintuitive property of this function is that a fixed absolute increase in travel time results in a fixed relative decrease in the (modeled) willingness to make a trip. Therefore the function is not believed to be accurate when the range of traveled distances in the study area exceeds 15 km.
Nevertheless, this function appears quite often in theoretical research due to its nice mathematical properties. Most derivations of the trip distribution model in the literature result in a model with an exponential distribution function.

Top-exponential (Tanner):

$$F_{ijm}(c_{ijm}) = \alpha_m c_{ijm}^{\gamma_m} \exp(\beta_m c_{ijm}) \tag{5.32}$$

Lognormal:

$$F_{ijm}(c_{ijm}) = \alpha_m \exp(\beta_m \ln^2(c_{ijm} + 1)) \tag{5.33}$$

Top-lognormal:

$$F_{ijm}(c_{ijm}) = \alpha_m c_{ijm}^{\gamma_m} \exp(\beta_m \ln^2(c_{ijm} + 1)) \tag{5.34}$$

Log-logistic:

$$F_{ijm}(c_{ijm}) = \frac{\mathrm{MAX}^m}{1 + \exp[\beta_m + \gamma_m \log(c_{ijm})]} \tag{5.35}$$

Functions of this type are used in the WOLOCAS model.

The top-exponential and top-lognormal distribution functions are used in practice as an alternative for the exponential and lognormal functions. They ignore the requirement that a distribution function should be monotonously decreasing (see Figure 5.3). This improves the fit with empirical data if a unimodal model is used (see remark above).

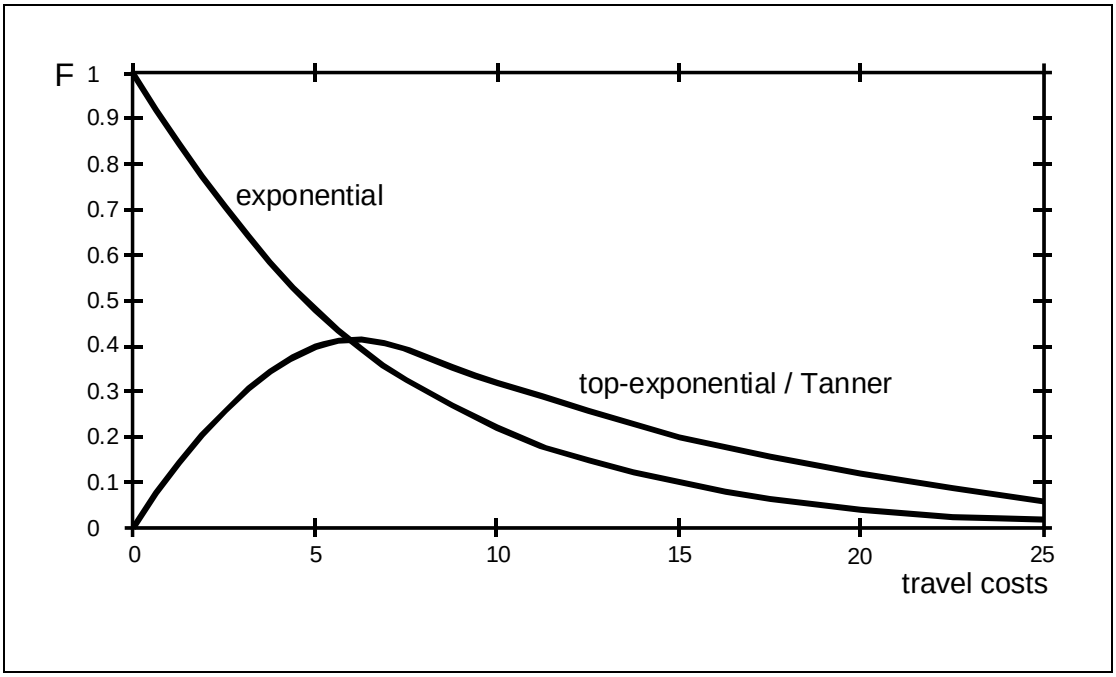


Figure 5.3: *Exponential distribution functions (parameters a = 1, β = -0.15), and top- exponential distribution function (parameters a = 0.25, β = -0.15, γ = 0.75)*

5.6.3 Discrete distribution functions

As an alternative to the continuous distribution, a discrete or piecewise constant distribution function may be used. The mathematical form of this function is:

$$F_{ijm}(c_{ijm}) = \sum_{k=1}^{K} F_{ijm}^{k} d^{k}(c_{ijm}) \tag{5.36}$$

with

$$F_{ijm}^{k} \geq 0 \text{ and } d^{k}(c_{ijm}) \in \{0,1\}$$

where k is the cost bin, K is the number of cost bins (e.g. 10), and $F_{ijm}{}^{k}$ is the value of the distribution function for cost bin k. The function $d^{k}(c_{ijm})$ is the membership function which is 1 if c_{ijm} lies in cost bin k and zero otherwise.

This function defines a fixed distribution function value for each cost-bin. A property of this approach is that no assumptions on the shape of the distribution function are imposed. In Figure 5.4 an example of a discrete distribution function is shown.

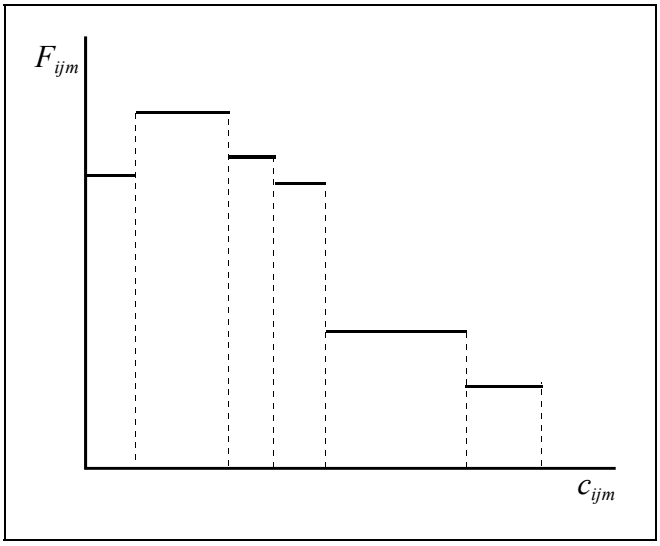


Figure 5.4: *Example discrete distribution function*

5.7 Growth factor models

As an alternative to trip distribution models, growth factor models may be used. In this approach a base year matrix is needed (in Dutch: 'basismatrix'). Each cell of this matrix is multiplied by a growth factor. Growth factors may be computed in a number of ways, e.g. as the output of an economic model, a trend model, etc. However, in these course notes, we only discuss methods of computing growth factors based on trip generation modeling.
The base year matrix contains an estimate of the trips being made in the base year. Theoretically it is possible to directly observe a base year matrix using a travel survey. However, if the study area is decomposed into many zones, and the travel survey represents only a part of all travel, directly observing a base year matrix would result in a matrix mainly consisting of zero cells. This can be seen from the following example:

Consider a town with 100.000 inhabitants that produce 15.000 trips during a two hour peak period. If this town is divided in 50 zones, there are more than 2.000 OD cells, and the average number of trips per OD-cell is 15.000/2.000 = 6. If 10% of the population is surveyed (10% is a large number for an survey) on average 0.6 trips per cell are reported in the observed matrix. This means that at least 40% of all observed cells must be zero (note the distinction between observed zero and unobserved cell).

Applying a growth factor model to a base year matrix mainly consisting of zeros results in a predicted matrix mainly consisting of zeros. A better approach is to impose extra constraints on the base year matrix in order to supply a realistic value for the cells in which no trips are observed, for example by requiring that the base year matrix complies with the gravity model, i.e.:

$$T_{ij}^{0} = Q_i X_j f(c_{ij}) \tag{5.37}$$

where:

T_{ij}^0	base year matrix
X_i, Q_j	parameters in gravity model, calibrated in such a way that the expression $\left\|T_{ij}^0 - T_{ij}^{observed}\right\|$ minimized
$f(c_{ij})$	value of distribution function for OD-pair *i-j*

Note that imposing such a structure can also 'spoil' an observed matrix, for example by causing bias. When larger zones are used, determining an appropriate value for the travel costs, c_{ij}, that is representative for all trips between *i* and *j* becomes impossible. This might introduce inaccuracies in the approach described above.

Another source of information that can be used to estimate base year matrices is traffic counts. Dedicated estimation techniques are available to adapt a prior base year matrix to a set of traffic counts. These methods will be discussed in a separate chapter.

The growth factor methodology is pointed out in Figure 5.5. A trip distribution model is used to compute a current and future trip matrix. Combining these matrices results in a growth factor that is then applied to the base year matrix, resulting in a predicted trip matrix.

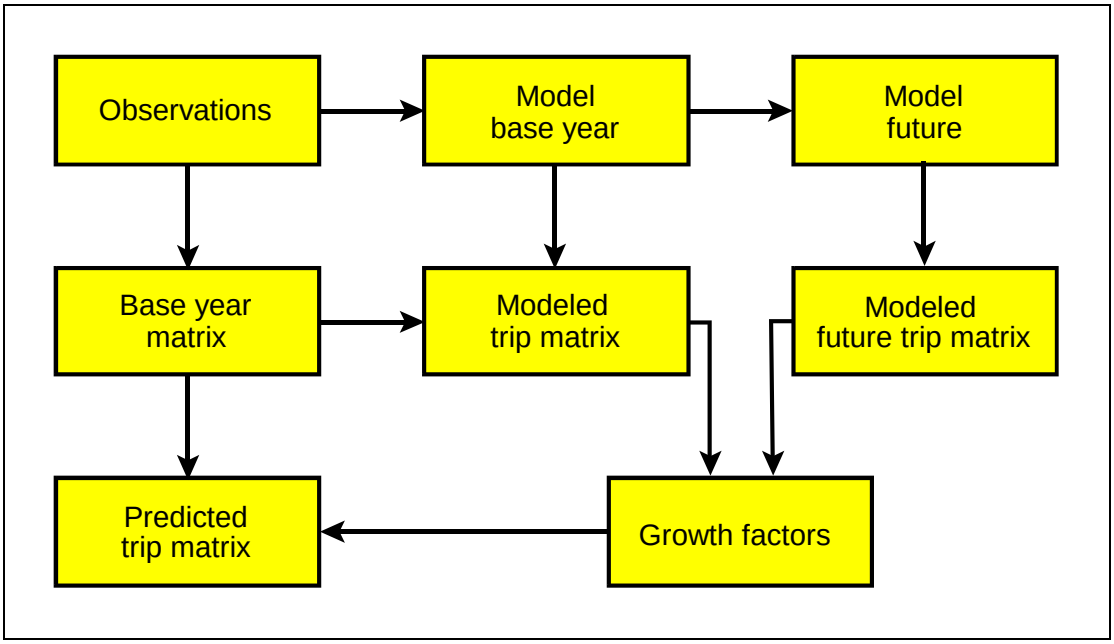


Figure 5.5: *Predicting future trip matrix using growth factor methodology*

The base year matrix is the best possible estimate of current origin-destination trip flows in the study area. In the growth factor methodology the base year matrix is the starting point for making predictions of future states. It is considered a better base for the future than using a trip distribution model on its own.
Reasons for this are among others:

- models cannot capture peculiarities in trip making that often can be found in study areas. In contrast, such peculiarities can be included in a base matrix to a large extent because it is based on observations.
- in planning practice it is necessary that parties involved in planning all agree on the fundamentals for planning. In this respect, a base matrix is a better tool to gain confidence in the fundamentals than a model because it is more understandable, it is verifiable, etc.

So, a widely accepted approach to prediction is to take a base matrix and adapt this using growth factors derived from models.

The methodology pointed out in Figure 5.5 also has a number of disadvantages of which the most important ones are:

- if new building sites are developed (e.g. VINEX) the resulting changes in trip distribution are difficult to capture in a growth factor model. This is because the travel behavior of the present inhabitants of these building sites is not representative for the future, especially when an agriculturally oriented environment changes to an urban environment.
- the base year matrix is influenced to a great extent by historical travel patterns. These patterns might fade away in a few decades time. This is particularly true if new cities have arisen as a result of suburbanization: initially the travel of suburbs is oriented toward the nearby town, but in time such strong historic ties vanish, and a more balanced trip making pattern arises. Of course, planners have to take account of these phenomena. Simply applying growth factors in this case would not lead to the desired result.

5.7.1 Computation of growth factors

Throughout this section we use the following notation:

T_{ij}^0 base matrix (known current OD-table)

T_{ij} future OD-table to be predicted

τ growth factor

$\hat{T}_{ij}^0$ model outcome of trip quantity (current situation)

$\hat{T}_{ij}$ model outcome of trip quantity (future situation)

We can distinguish various levels of updating a base matrix, ranging from simple to complex. In a first class of approaches, the growth factors do not reflect changes in the network and consider only changes in socio-economic conditions in the study area. The other approaches do reflect changes in interzonal accessibility.

A. Network independent base matrix updating.

A.1 General growth factor τ

$$T_{ij} = \tau T_{ij}^0 \quad \forall i, j$$

τ may be determined by general factors expressing growth in activities such as demographic growth, economic growth etc.

A.2 Origin or destination specific growth factors τ

$$T_{ij} = \tau_i T_{ij}^0 \text{ or } T_{ij} = \tau_j T_{ij}^0 \quad \forall i, j$$

Growth factors τ may be derived from trip end models applied to current and future conditions respectively.

A.3 Two sets of independently and exogenously determined growth factors for origins and destinations.

$$T_{ij} = \tau_i \tau_j T_{ij}^0 \quad \forall i, j$$

with constraints:

$$\tau_i \sum_j T_{ij}^0 = \sum_j T_{ij} \forall i$$
$$\tau_j \sum_i T_{ij}^0 = \sum_i T_{ij} \forall j$$
$$\sum_i (\tau_i \sum_j T_{ij}^0) = \sum_j (\tau_j \sum_i T_{ij}^0)$$

This updating problem can be iteratively solved by bi-proportional fitting. Alternatively, growth factors τ may be derived from trip end models applied to current and future conditions successively.

B. Network dependent base matrix updating.
In this case the growth factors reflect changes that are OD-relation specific. These changes can be calculated using trip distribution models applied to current and future conditions.

$$T_{ij} = \tau_{ij} T_{ij}^0 \text{ with } \tau_{ij} = \hat{T}_{ij} / \hat{T}_{ij}^0$$

It may even be considered to combine the approaches A.3 and B into one joint base year matrix updating.

5.8 Derived quantities; network performance

After trip distribution has been computed, various quantities can be derived. These quantities play a key role in judging a transport network. They may be derived in the following ways:

Notation
B_t = total travel time
B_k = total travel costs
B_l = total traveled distance

From link characteristics

$$B_t = \sum_a q_a t_a \tag{5.38}$$

$$B_k = \sum_a q_a c_a \tag{5.39}$$

$$B_l = \sum_a q_a l_a \tag{5.40}$$

q_a = flow on link a
t_a = travel time on link a
c_a = travel costs of link a
l_a = length of link a

From route-characteristics

$$B_t = \sum_i \sum_j \sum_r T_{ij}^r t_{ij}^r \tag{5.41}$$

$$B_k = \sum_i \sum_j \sum_r T_{ij}^r c_{ij}^r \tag{5.42}$$

$$B_l = \sum_i \sum_j \sum_r T_{ij}^r l_{ij}^r \tag{5.43}$$

T_{ij}^r = number of trips from i to j via route r

t_{ij}^r = travel time from i to j via route r

c_{ij}^r = travel costs from i to j via route r

l_{ij}^r = length of route r from i to j

In Figure 5.6 and Figure 5.7 these computational schemes are illustrated

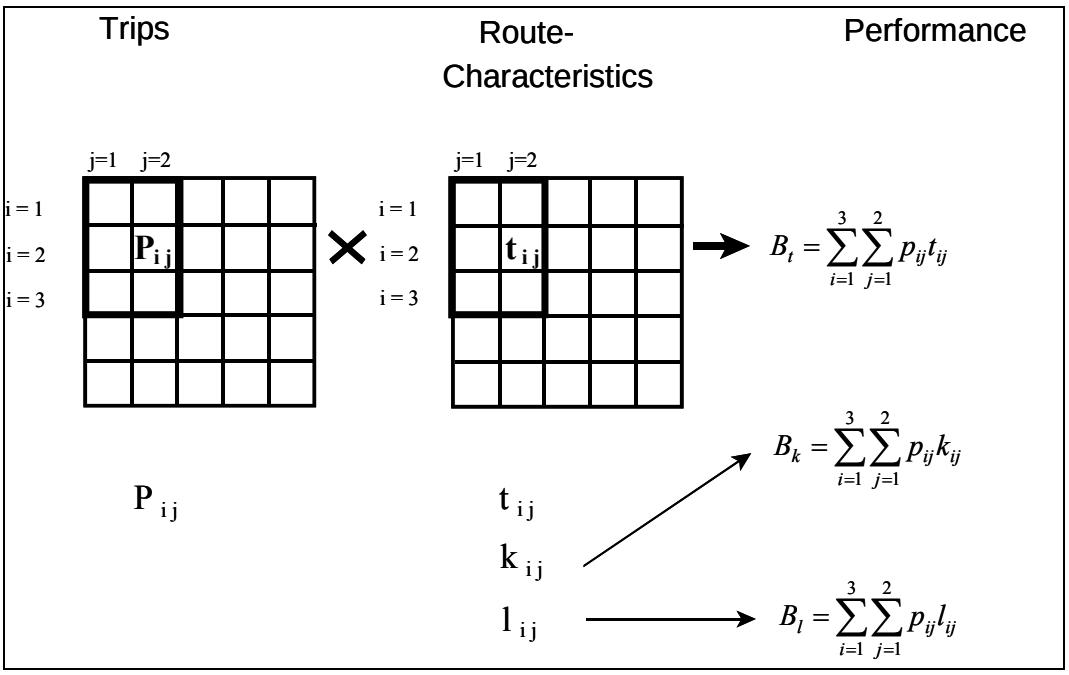


Figure 5.6: *Computing network performance index from trip distribution data*

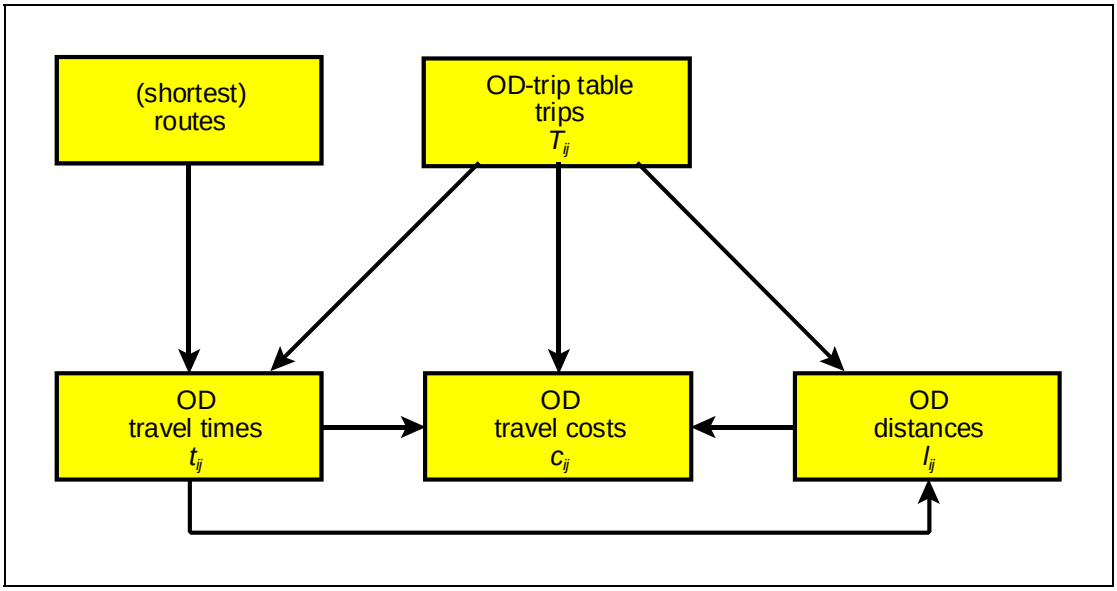


Figure 5.7: *Computing network performance index*

5.9 Departure time choice

The *time* of an activity influences the utility that is derived from it, and hence influences the utility of the trip that is needed for this activity. Employees usually have preferred times to start and end their daily job. Starting early or late brings about a certain disutility. This disutility is sometimes accepted if making the trip at the preferred time would bring about an even higher disutility, due to congestion on the roads or discomfort and irregularity in public transit. The phenomenon of travelers avoiding the peak hour is referred to as peak spreading.

Peak spreading is one of the pitfalls in transport planning. Not taking peak spreading into account leads to, among other things, an underestimation of travel demand in the peak hour: building new infrastructure in most cases leads to an inevitable 'back to the peak' effect.

A way to introduce departure time choice into the chain of transport models is to value early and late arrivals using the utility scale. If travelers start early to avoid congestion, they trade off the disutility of starting early against the disutility of incurring extra travel time due to congestion. Departure time choice (given travel mode) is usually modeled as a choice between a number of discrete time intervals, assuming that travelers maximize a utility that may be decomposed in the following components:
- utility of the activity (constant)
- disutility of the free flow travel time (constant)
- disutility of the travel time loss due to congestion (time dependent)
- disutility of arriving early (time dependent)
- disutility of arriving late (time dependent)

The way these components are valued differs per from person to person.

Figure 5.8 illustrates the (hypothetical) disutility function D given the preferred arrival time AT. Arriving late is penalized more than arriving early, and arriving much early is associated with a large disutility per time unit (see the slope in Figure 5.8).
New infrastructure causes changes in congestion levels. For example consider the opening of the Zeeburgertunnel in the Amsterdam beltway in the Netherlands. This may bring about large shifts in travel time choice. Usually the direction of these shifts is back to the peak as the preferred arrival time usually is in the middle of the peak period (especially when

commuters are considered). Departure time choice models make it possible to take travelers responses to changes in congestion levels into account when predicting future travel demand.

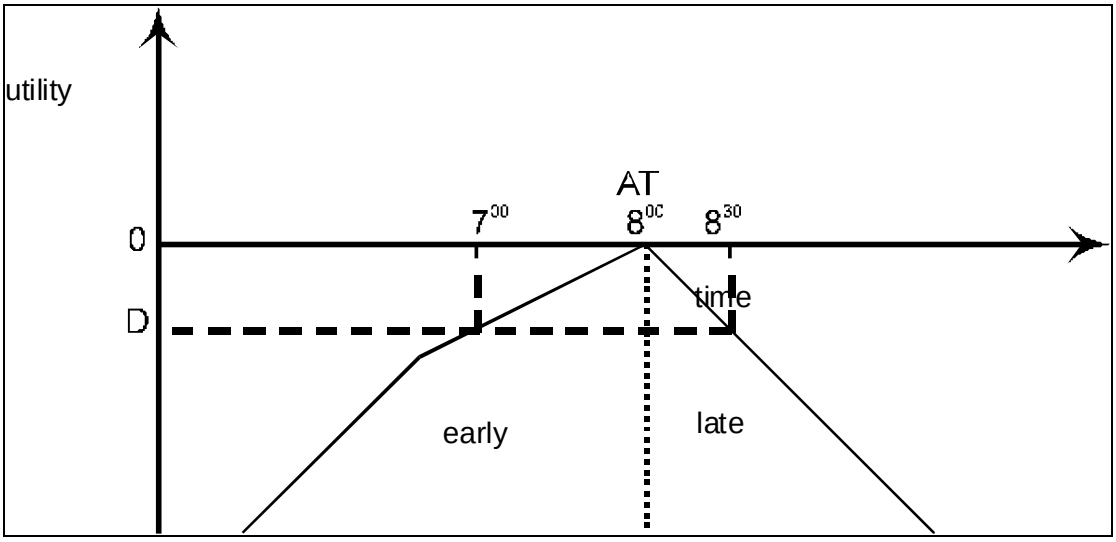


Figure 5.8: *(dis)utility as a function of time (AT: preferred arrival time)*

5.10 References

S. Erlander & N.F. Stewart
The gravity Model in transportation analysis: theory and extensions
Utrecht, VSP, 1990

A.S. Fotheringham & M.E. Kelly
Spatial interaction models: formulations and applications
Dordrecht, Kluwer Academic Publishers, 1989

A. Sen & T.E. Smith
Gravity models of spatial interaction behavior

6 Mode choice models

The choice of travel mode of travelers (e.g. car, public transit or non-motorized) depends on the availability of transport means, especially cars, and on the travel resistance for each mode from origin to destination. Apart from that, each travel mode has its specific advantages and disadvantages irrespective of travel time and travel costs. These are accounted for in the mode specific constants.

6.1 Sequential trip distribution modal split

6.1.1 General mode choice model

The traditional approach to transport planning is the four stage model: a model that distinguishes between trip generation, trip distribution, modal split and route choice. In this approach the mode choice is modeled in *sequence*, following the steps trip generation and trip distribution. The input to the mode choice model in this case is hence the total travel demand between each Origin-Destination (OD) pair.
A commonly used approach is to distribute the total travel demand for a given OD-pair over the available modes using the logit model:

$$\beta_{ijv} = \frac{\exp[bV_{ij}^{v}]}{\sum_{w}(\exp[bV_{ij}^{w}])} = \frac{\exp\sum_{k}\alpha_{k}^{v}X_{ijk}^{v}}{\sum_{w}(\exp\sum_{k}\alpha_{k}^{w}X_{ijk}^{w})} \qquad (6.1)$$

= proportion using mode *v* on OD-pair *ij*

$$T_{ijv} = T_{ij}\beta_{ijv} \qquad (6.2)$$

with:

b = variance (or scale) parameter in logit model

V_{ij}^{v} = observable part of utility of traveling between OD-pair *i-j* with mode *v*

X_{ijk}^{v} = the k^{th} explanatory variable for mode *v* on OD-pair *i-j* (e.g. travel time travel costs, mode specific constant);

α_{k}^{v} = weight parameter for the k^{th} explanatory variable for mode *v* on OD-pair *i-j* (e.g. the value of time).

Note that these weights α_{k}^{v} may depend on the travel mode. For example, the weight attached to travel distance by pedestrians differs from the weight attached to travel distance by car-drivers.

6.1.2 Mode specific constants

In Equation (6.1) the variables X_{ijk}^{v} denote the *observable* attributes of a trip using mode *v*.
However many factors that influence mode choice cannot be observed, or can only be observed at a cost which is excessively expensive. Examples are non-physical factors like comfort, status, image, and safety of the travel mode, but also the average waiting time at a

bus stop, the average walking distance to the bus stop, etc. The impact of these factors is usually summarized in a *mode specific constant*, i.e. (6.1) changes in:

$$\beta_{ijv} = \frac{\exp(C^v + \sum_k \alpha_k^v X_{ijk}^v)}{\sum_w \exp(C^w + \sum_k \alpha_k^w X_{ijk}^w)} \tag{6.3}$$

Example 6.1
As an example we use a logit model that models the proportions using the three travel modes: car (C), public transit (PT) and non-motorized traffic (NM) for home-work trips.

$$P_{NM} = \frac{\exp \sum_k \alpha_k^{NM} X_k^{NM}}{1 + \exp \sum_k \alpha_k^{NM} X_k^{NM} + \exp \sum_k \alpha_k^{PT} X_k^{PT}}$$

$$P_{PT} = \frac{\exp \sum_k \alpha_k^{PT} X_k^{PT}}{1 + \exp \sum_k \alpha_k^{NM} X_k^{NM} + \exp \sum_k \alpha_k^{PT} X_k^{PT}}$$

$$P_C = 1 - P_{NM} - P_{PT}$$

The explanatory variables in this model are:
- mode specific constant,
- distance from *i* to *j* [km]
- ratio travel time PT/travel time car on OD pair *ij*,
- ratio travel time NM/travel time car on OD pair *ij*

The parameter values are contained in Table 6.1.

	NM	**PT**
constant	3.753	1.681
distance *i j* [km]	-0.111	-0.002
travel time PT/car	0.084	-2.269
travel time NM/car	-1.109	0.477

Table 6.1: *parameter values α_k of utility function in logit model, source: [Maanen &Verroen, 1992]*

Note that in order to write the model in the form of the first line of (6.1), we must define the following utility functions:

$V_{NM} = V_C + 3.753 - 0.111\ L + 0.084\ T_{PT}/T_C - 1.109\ T_{NM}/T_C$
$V_{PT} = V_C + 1.681 - 0.002\ L + 2.269\ T_{PT}/T_C - 0.477\ T_{NM}/T_C$
V_C = constant

A first observation is that the utilities for the non-motorized and the public transit mode are defined relative to the utility of the car mode.
A second observation is that the utility of the car mode need not be known in order to apply the model. This can be a great practical advantage.

- end of example -

6.1.3 Purpose-specific mode choice model

Mode choice proportions heavily depend on the trip purpose, e.g. if the purpose is 'drinking' the travel mode 'driving' is less attractive.

A condition for applying a purpose specific mode choice model is that a purpose specific trip distribution should be available, e.g.:

$$T_{ij}^{p} = \mu^{p} Q_i^{p} X_j^{p} F_{ij}^{p} \tag{6.4}$$

$$F_{ij}^{p} = F^{p}(c_{ij}) \tag{6.5}$$

The purpose-specific trip distribution is assigned to the travel modes *v* using:

$$T_{ij}^{p} = T_{ij}^{p} \beta_{ijv}^{p}, \tag{6.6}$$

where the choice probability β_{ijv} of mode *v* (logit model) for purpose *p* satisfy:

$$\beta_{ijv}^{p} = \frac{\exp(f_p(c_{ijv}))}{\sum_{w} \exp(f_p(c_{ijw}))} \qquad 0 \le \beta_{ijv}^{p} \le 1 \tag{6.7}$$

where

c_{ijv} = travel resistance (generalized costs) of mode *v* between *i* and *j*

6.1.4 Trip distribution revisited: generalized travel cost

A consequence of applying trip distribution and modal split sequentially is that at the time of computing the trip distribution the travel costs for each mode are known but not the mode choice proportions. The obvious question that arises at this point is: 'which travel costs should be used when sequentially computing the trip distribution?'.

The general trip distribution model is:

$$T_{ij} = \mu Q_i X_j F_{ij}$$

with:

$$F_{ij} = f(c_{ij})$$

The following options for supplying a value to be used during trip distribution for cij may be considered:

1. $c_{ij} = \min_{v}\{c_{ijv}\}$ (minimum costs)
2. $c_{ij} = -\frac{1}{\alpha} \ln\left(\sum_{v=1}^{V} \exp(-\alpha c_{ijv})\right)$ (logit analogy)
3. $c_{ij} = 1/\left(\sum_{v=1}^{V} 1/c_{ijv}\right)$ (electric circuit analogy)

4. $c_{ij} = \sum_{v=1}^{V} c_{ijv} / V$ (average costs)

5. $c_{ij} = \sum_{v=1}^{V} \beta_{ijv} c_{ijv}$ (weighted average costs). In this case estimates of the mode choice proportions β_{ijv} need to be available.

In general the addition of an extra route or travel mode reduces the perceived travel costs between origin and destination. This is because the perceived travel costs for an individual equal the minimum of the costs of the travel alternatives that this individual has access to and is aware of. For example if a person is aware of two alternative travel modes between A and B with travel costs 2 and 10 respectively, this person will judge the travel time between A and B to be 2. Therefore options 4 and 5 seem not realistic. Still practitioners may have reason to use these options, for example if large groups of people are captive to a particular travel mode (usually public transit). In this case the car alternative with low generalized travel costs is only available to part of the travelers.

The following example illustrates the idea behind option 2.

Example 6.2
Suppose a person traveling from origin *i* has three travel alternatives:
- traveling to *j* by car at cost c_{ij1} (option 1)
- traveling to *j* by public transit at cost c_{ij2} (option 2)
- traveling to *k* by car at cost c_{ik1}(option 3)

In addition assume that the utility of performing an activity at *j* is U_j while the utility of performing an activity at *k* is U_k (see Figure 6.1), and that travelers are distributed over travel alternatives according to a logit model with scaling parameter α.

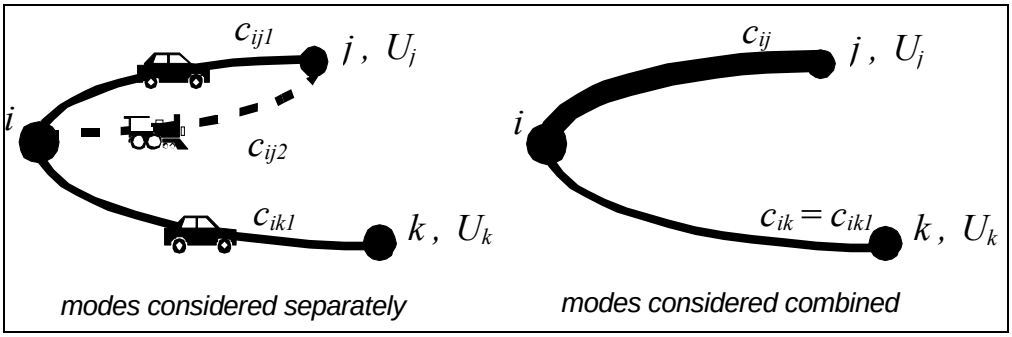


Figure 6.1: *Determine c_{ij} in such a way that the proportion of people traveling to destination j in the right network is equal to the proportion of people traveling to j in the left network. Assume travelers are distributed over alternatives according to the logit model*

As a result, the objective utility of each alternative is:
- Option 1: $U_j - c_{ij1}$
- Option 2: $U_j - c_{ij2}$
- Option 3: $U_k - c_{ik1}$

If each alternative may be considered as an independent option, the logit model is a plausible model to compute the proportion of travelers for each alternative. (Note that the first two options have their destination activity in common, which implies that the assumption of independence implies a simplification). According to the logit model the share of travelers opting for destination *j* is given by:

$$p_j = p_{j1} + p_{j2} = \frac{\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2})}{\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2}) + \exp\alpha(U_k - c_{ik1})}$$

However, to apply the sequential trip distribution modal split model we need to apply *network aggregation*, i.e. replace the option 1 and 2 with an hypothetical option that represents both. If we were to assign a travel cost c_{ij} to this option, the share of travelers opting for destination j would be given by:

$$p'_j = \frac{\exp\alpha(U_j - c_{ij})}{\exp\alpha(U_j - c_{ij}) + \exp\alpha(U_k - c_{ik1})}$$

We can now compute the value c_{ij} by solving $p_j = p'_j$. This yields the following sequence of expressions:

$$\frac{\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2})}{\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2}) + \exp\alpha(U_k - c_{ik1})} = \frac{\exp\alpha(U_j - c_{ij})}{\exp\alpha(U_j - c_{ij}) + \exp\alpha(U_k - c_{ik1})}$$

$$\Rightarrow [\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2})] \cdot [\exp\alpha(U_j - c_{ij}) + \exp\alpha(U_k - c_{ik1})]$$

$$= [\exp\alpha(U_j - c_{ij})] \cdot [\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2}) + \exp\alpha(U_k - c_{ik1})]$$

$$\Rightarrow [\exp\alpha(U_j - c_{ij1}) + \exp\alpha(U_j - c_{ij2})] \cdot \exp\alpha(U_k - c_{ik1}) = \exp\alpha(U_j - c_{ij}) \cdot \exp\alpha(U_k - c_{ik1})$$

$$\Rightarrow [\exp(-\alpha c_{ij1}) + \exp(-\alpha c_{ij2})] = \exp(-\alpha c_{ij})$$

$$\Rightarrow c_{ij} = -\frac{1}{\alpha}\ln[\exp(-\alpha c_{ij1}) + \exp(-\alpha c_{ij2})]$$

Note that this result corresponds to the logit analogy mentioned before.

- end of example -

The methodological difficulty with deciding which general travel costs, c_{ij}, to use in the sequential trip distribution modal split model, is one reason for the development of *simultaneous* trip distribution modal split models. These are applied especially if the difference in generalized travel costs between modes grows bigger, and if the proportion of alternative modes can not be ignored.

6.2 Simultaneous distribution/modal split model

In Dutch practice the computation of distribution and modal split usually happens simultaneously. This means that, given an origin i, every combination of destination j and mode v is considered as a separate travel alternative, with its corresponding distribution function value F_{ijv}. This is more realistic than the sequential model (every destination j is considered as a separate alternative with a distribution value F_{ij}) as utility maximizing travelers makes their choice of destination considering both the utility of the (activities at the) destination and the disutility of the travel to this destination. Following this line of thinking the simultaneous trip distribution- modal split model is formulated as follows:

Mode specific distribution model:

$$T_{ij}^{v} = \mu Q_i X_j F_{ijv} \tag{6.8}$$

v = travel mode
F_{ijv} = mode specific accessibility function = $f_v(c_{ijv})$
c_{ijv} = mode specific travel resistance values

Assumption of double constraints on P_i and A_j leads to:

$$\sum_j \sum_v T_{ij}^v = P_i \tag{6.9}$$

$$\sum_i \sum_v T_{ij}^v = A_j \tag{6.10}$$

$$\sum_j \sum_v \mu Q_i X_j F_{ijv} = P_i \tag{6.11}$$

$$\sum_i \sum_v \mu Q_i X_j F_{ijv} = A_j \tag{6.12}$$

$$Q_i = \frac{P_i}{\mu \sum_j (X_j \sum_v F_{ijv})} = \frac{a_i P_i}{\mu} \tag{6.13}$$

$$X_j = \frac{A_j}{\mu \sum_i (Q_i \sum_v F_{ijv})} = \frac{b_j A_j}{\mu} \tag{6.14}$$

The balancing factors a_i and b_j are now a function of the sum of mode specific travel resistance $f(c_{ijv})$.

$$T_{ij}^v = \frac{1}{\mu} a_i b_j P_i A_j F_{ijv} \tag{6.15}$$

Iterative solution is similar to that of a simple (single mode) doubly constrained trip distribution model (see Section 5.5). The parameter μ is absorbed in the estimated values for a_i and b_j.

Example 6.3
Compute trip distribution and modal split using the data given below. Trip ends are given. The travel modes considered are car and PT (Table 6.3). We assume that the distribution function for car trips is given as a discrete distribution function. The distribution function for PT is derived from the car distribution function by substituting the value of the next higher adjacent cost bin (see Table 6.2).

cost bin [min]	**2-5**	**5-7**	**7-10**	**10-15**	**15-20**	**20-30**
F_{ij}**-car**	200	67	26	19	7	0.7
F_{ij}**-PT**	67	26	19	7	0.7	0.7

Table 6.2: *distribution function F_{ijv}*

to zone	from zone					
	A		B		C	
	car	PT	car	PT	car	PT
A	6	8	12	17	9	12
B	12	17	6	8	15	8
C	9	12	15	8	6	8

Table 6.3: *travel time matrix (c_{ijv}) for car and PT [minutes]*

The corresponding distribution function values F_{ijv} are (Table 6.4):

to zone	from zone						$\Sigma_i F_{ij}$	P_i
	A		B		C			
	car	PT	car	PT	car	PT		
A	67	19	19	0.7	26	7	138.7	100
B	19	0.7	67	19	7	19	131.7	600
C	26	7	7	19	67	19	145	400
A_j	200		1200		600			2000

Table 6.4: *values of distribution function $F_{ijv}(c_{ijv})$ for car and PT*

The table shown in 6.4 initializes the balancing iterations. The operations are performed in analogy with the doubly constrained trip distribution model for one travel mode (see Chapter 5). Try solving this problem yourself!

- end of example -

6.3 References

Maanen, T. van, & E. Verroen, *Mobiliteitsprofielen revisited: een nadere analyse van de samenhang tussen bedrijfs-, lokatie- en mobiliteitskenmerken*, Delft Inro TNO, 1992

7 Route choice and traffic assignment

7.1 Introduction

Traffic assignment is the step in traffic analysis in which interzonal trips are assigned to the network. The interzonal trips constitute the traffic demand and are computed in the distribution/modal-split step. The traffic demand, as described in the origin-destination (OD) tables per trip purpose and per travel mode (and sometimes per period), is confronted with the infrastructure supply, which is a network of links and nodes having characteristics as capacity, maximum travel speed, one-way streets, tolls and other factors of resistance.

Traffic assignment involves computing one or more optimal (usually shortest) routes between each origin and destination and distributing travel demand over these routes. The sum of all trips along these routes over all OD pairs results in a traffic load on all links and nodes. Usually, there is a separate assignment for each mode, since the networks for each of the modes is very different. For the sake of simplicity, we restrict ourselves to assignments of individual road traffic (car, bike); the more complex assignments on public transportation networks will be discussed in another course.

Usually trip generation (Chapter 4) and trip distribution (Chapter 5) is computed separately per trip purpose, while traffic assignment is carried out for all trip purposes simultaneously, i.e. after combining the OD tables per trip purpose in one table.

Question:
Under which circumstances is it desired to do a separate assignment per trip purpose?

7.1.1 Purpose of traffic assignment

Traffic assignment has five functions in traffic planning:

- *Gaining insight in the characteristics of the network*
  By performing several different assignments, insight may be gained in the shortcomings of the existing network (missing links, capacity deficiency), misusage of functional classes and prevention of large detours. Also, ideas can be gained for solving existing problems. The same analysis is useful for planned network scenarios in the future.
- *Traffic forecasts*
  With regard to current and future network scenarios, a number of aspects are computed to be able to forecast the future traffic situation. These include, among others, link loads, travel time, speed, congestion, junction resistance and detour factors. All this information is used for the evaluation of alternative plans.
- *Computation of derived impacts*
  Also on the basis of computed traffic assignment, impacts are computed, such as noise levels, air pollution, energy consumption and traffic safety, which are of importance for the evaluation of plans.
- *Supply of design data*
  The evaluated network scenarios should be translated into a design of physical infrastructure. The computed link loads are an important aid for the technical design. So-called design flows are derived from the computed flows.
- *Supply of input data*
  In the other steps of traffic analysis such as distribution and mode choice, resistance characteristics of OD-pairs (travel times, distances, travel costs) are used which are computed in the assignment step as attributes of routes (see Section 1.4).

7.1.2 Input and output of the assignment computation

Necessary input for the assignment:
- an OD table of trips between the zones, usually all trip purposes combined;
- a (computer)representation of the network;
- characteristics of the network elements (links and nodes);
- a route choice model.

Direct output of the assignment computation:
- the routes (consecutive series of adjacent links and nodes);
- the route characteristics (travel times, distances, costs);
- route loads: the number of trips per route;
- link and node loads: the number of trips per unit time (flow) on each link and each turn at junctions.

Figure 7.1 summarizes the framework of the assignment.

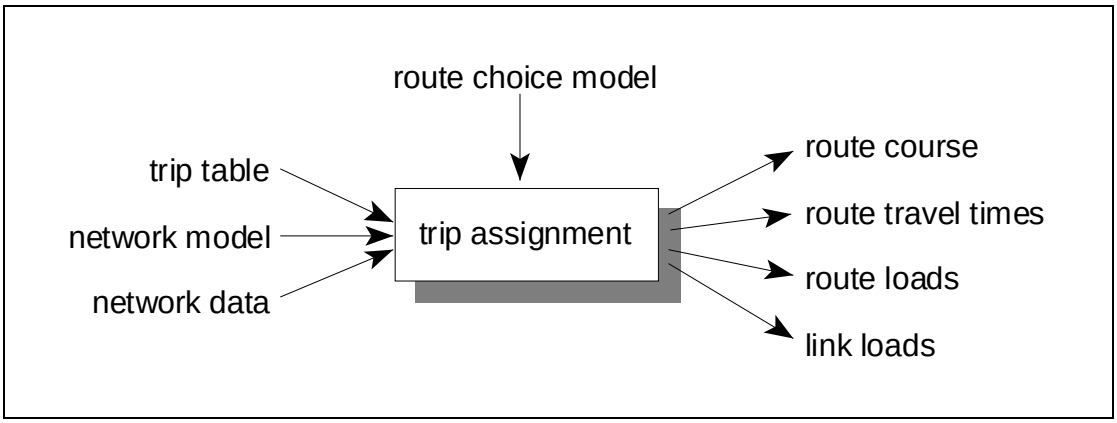


Figure 7.1: *Trip assignment to a network*

The route choice model distributes trips of an OD pair over the alternative routes: it describes the traveler choice behavior. The network data provide a basis on which the routes are determined and on which the trip distribution over the routes takes place (among others travel time and distance); furthermore, these data are necessary to describe possible congestion in the network. The trip table and the route choice model represent the demand side of the traffic system, while the network model and the network data describe the supply side.

7.1.3 Classification of assignment models

For the sake of completeness, we note that a distinction between static and dynamic models exists. Static models assume that the traffic demand and supply are time-independent, hence constant during the considered time period (stationary). Dynamic models use the more realistic assumption that OD demand and link characteristics are time varying and hence are more computationally demanding and pose extra requirements with respect to data. We restrict ourselves to static models. For a discussion of dynamic models, see lecture notes CT5804 (dynamic traffic management).
Within the static assignment models used in practice, one can differentiate on the basis of two main characteristics:
- differences in individual route choice behavior;
- congestion effects in the network.

This results in four types of models, see Table 7.1. The application of each of these models depends on the situation (wether or not congestion effects exist) and the level of accuracy.

		congestion effect modeled	
		NO	YES
Random utility route choice modeled	NO	all-or-nothing (Section 8.3)	deterministic equilibrium (Section 8.5)
	YES	stochastic (Section 8.4)	stochastic equilibrium (Section 8.6)

Table 7.1: *Classification of static traffic assignment techniques*

All-or-nothing assignment

Some methods, such as all-or-nothing (AON) assignment, ignore the fact that link travel times are flow-dependent (i.e., that they are a function of link volumes). These simple methods assign all trips between a zone pair to a single (optimal) route, for example the shortest route in time or in distance.

Deterministic equilibrium assignment

Equilibrium methods take account of the volume-dependency of travel times and result in the calculation of link flows and travel times that are mutually consistent. Equilibrium flow algorithms require iterating back and forth between assigning flows and calculating travel times. Despite the additional computational burden, equilibrium methods will almost always be preferable to other assignment models. These methods divide the total flow of trips between an OD pair over a number of routes.

The key behavioral assumptions underlying the user equilibrium (UE) assignment model are that each traveler has perfect information concerning the attributes of the network, each traveler chooses a route that minimizes his/her travel time or travel costs, and such that all travelers between the same O and D have the same travel time or cost. Wardrop was the first to propose the following condition for a UE: No individual traveler can unilaterally reduce his travel time by changing paths (Sheffi, 1985). A consequence of the UE principle is that all used paths for an O-D pair have the same minimum cost. Unfortunately, this is not a realistic description of loaded traffic networks (Slavin, 1995).

Stochastic assignment

In many urban areas, there are many alternative routes that could be and are used to travel from a single origin zone to a single destination zone. Often trips from various points within an origin zone to various points in a destination zone will use entirely different major roads to make the trip. Also, different individuals will judge each alternative in a different way. These mechanisms cause many paths between origin i and destination j to be used, even if link costs are assumed to be independent of link flows. Different criteria are described in the literature to enumerate all reasonable paths. However, in some instances, reasonable alternative routes may be so numerous so as to preclude their easy enumeration. For the traffic assignment model to be valid, it must correctly assign car volumes to these alternative paths, based on sound behavioral rules.

From a behavioral perspective, traffic assignment is the result of aggregating the individual route choices of travelers. Assignment models, not surprisingly, also differ in the assumptions made about how and which routes are chosen for travel. Individual route choice behavior of

travelers is expressed by the behavioral parameter Θ. This parameter determines the split of flow among alternative routes expressed by the split rates β.

Stochastic equilibrium assignment

An assignment that combines the properties of the stochastic assignment and the deterministic user equilibrium assignment was proposed by Daganzo and Sheffi (1977). Known as stochastic user equilibrium (SUE), this model is premised on the assumption that travelers have imperfect information about network paths and/or vary in their perceptions of network attributes.

At stochastic user equilibrium, no traveler believes that he or she can increase his/her expected utility by choosing a different path. Because of variations in traveler perceptions, this results into a spread of each OD flow over multiple routes. The route choice proportions depend on the travel times experienced. The travel times in turn depend on the link flows, while the link flows depend again on the route choice proportions. The system is said to be in SUE if route choice proportions and travel times are consistent.

Other assignment techniques

Another, however not mutually exclusive, approach to making assignment models more realistic entails multi-class and/or multicriteria models in which different groups of travelers value network attributes such as travel time or reliability differently (Dial, 1994). These models are particularly useful in multi-modal traffic assignment.

7.1.4 Notation

For simplicity reasons, in the following we will use the term ‘travel time’ as a representative for the more general expressions travel cost, or travel resistance, or travel disutility.

In the following of this chapter, we describe the core traffic assignment models for road traffic and discuss the primary advantages and disadvantages of each method.

In the mathematical explanation of the various assignment models, the following notation will be used:

Indices

i = origin
j = destination
r = route, path
k = optimal route
a = link

Variables

T_{ij} = number of trips from i to j (traffic flow)
T_{ijr} = number of trips from i to j on route r (route flow)
t_a = generalized travel time on link a
$\underline{t}_a$ = vector of all link travel times
t_{ijr} = travel time on route r
$\underline{t}_{ijr}$ = vector of all route travel times
α_{ijr}^a = link-route-incidence $= \begin{cases} 1 & \text{if link } a \text{ is on route } r \text{ from } i \text{ to } j \\ 0 & \text{otherwise} \end{cases}$
β_{ijr} = proportion of traffic flow from i to j via route $r \in [0,1]$ (route choice proportions)

q_a = traffic flow on link a (link flows)
$\underline{q}_a$ = vector of link traffic flows

Parameters
Θ = behavioral route choice parameter
$\underline{\Theta}$ = vector of parameters of the route choice model

7.2 The general network assignment problem

In general the inputs for the network assignment model are α_{ijr}^{a}, $\underline{\Theta}$, and T_{ij} (route composition, behavioral parameters and travel demand). The purpose of the network assignment is to determine the following quantities (see also Figure 7.1):

1. $\underline{t}_a$ - link travel times
2. β_{ijr} - route choice proportions
3. $\underline{t}_{ijr}$ - route travel times
4. T_{ijr} - route flows
5. $\underline{q}_a$ - link flows

The following relations hold. Due to their generality, the relations are referred to as definitional constraints.

Ad (1): Link travel time is a function of link flow:

$$t_a = t_a(q_a) \tag{7.1}$$

Note that we assume that the link travel time only depends on the flow on the link itself and not on flows on other links. This assumption is known as the separability assumption.

Ad (2): Route choice proportion β_{ijr} determines the distribution of flow among alternative routes. The route choice proportions are a function of the characteristics of the routes (expressed by the vector of travel times t_{ijr}) and the set of behavioral route choice parameters $\underline{\Theta}$:

$$\beta_{ijr} = \beta_{ijr}(\underline{\Theta}, t_{ijr}) = \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) \qquad \beta_{ijr} \in [0,1] \tag{7.2}$$

Ad (3): Route travel time t_{ijr} is the sum of link travel times t_a belonging to route r between i and j:

$$t_{ijr} = \sum_a \alpha_{ijr}^{a} t_a(q_a) \tag{7.3}$$

where indicator variable α_{ijr}^{a} indicates whether link a is part of route r between i and j. The link-route incidence matrix α defines the network structure.

Ad (4): The total number of trips from i to j multiplied by the proportion β_{ijr} of travelers choosing route r results in the route flows T_{ijr}:

$$T_{ijr} = \beta_{ijr} T_{ij} = \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) T_{ij} \tag{7.4}$$

Ad (5): Link flow is the sum of route flows traversing the link:

$$\begin{aligned} q_a &= \sum_i \sum_j \sum_r \alpha_{ijr}^a T_{ijr} \\ &= \sum_i \sum_j \sum_r \alpha_{ijr}^a \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) T_{ij} \end{aligned} \tag{7.5}$$

7.3 All-or-nothing assignment

7.3.1 All-or-nothing assignment as an optimization problem

In an All-Or-Nothing (AON) assignment, all traffic between an O-D pair is assigned to just one path (usually the shortest path) connecting the origin and destination. This model is unrealistic in that only one path between every O-D pair is utilized even if there is another path with the same or nearly the same travel time. Also, traffic is assigned to links without consideration of whether or not there is adequate capacity or heavy congestion; travel time is taken as a fixed input and does not vary depending on the congestion on a link.

In general, the AON model does not generate very realistic outcomes. The generated flows are too much concentrated on single routes. Further, it inhibits a high level of instability; small changes in the network can lead to major changes in the results (route choice, travel times, flows). Despite these disadvantages, the AON model is an important building block for computations with more complex assignment models.

Referring to the general network assignment problem (see previous section), the AON assignment is characterized by the following:

Assumptions:
t_a is constant (independent of flow level q_a);
$\Theta = 0$ (no flow diversion among routes);
route k_{ij} is the optimal route used for traveling from i to j.

The definitional constraints can be more specific now:

(1) t_a is given and constant

(2) $$\beta_{ijr} = \beta_{ijr}(\underline{t}_a) = \begin{cases} 1 \text{ if } r = k_{ij} \\ 0 \text{ otherwise} \end{cases} \tag{7.6}$$

(3) $$t_{ijr} = \sum_a \alpha_{ijr}^a t_a \tag{7.7}$$

(4) $$T_{ijr} = \begin{cases} T_{ij} \text{ if } r = k_{ij} \\ 0 \text{ otherwise} \end{cases} \tag{7.8}$$

(5) $$q_a = \sum_i \sum_j \alpha_{ijk_{ij}}^a T_{ij} \tag{7.9}$$

The concept of shortest path AON assignment is quite simple; it is defined as follows:

Shortest path all-or-nothing assignment:
The assignment in which for each OD pair the corresponding flow is assigned to a single path that, according to a fixed set of link costs, has minimum path costs (congestion effects are not taken into account).

It can be shown that solving this assignment problem is equivalent to solving an optimization problem. Two types of formulations are possible: a route based formulation and a link based formulation. The route based formulation uses route-level variables (t_{ijr} and T_{ijr} for route travel time and route flows, respectively), while the link based formulation uses link-level variables (t_a and q_a for link travel time and link flow). Both formulations lead to a linear programming (LP) problem.

Route based formulation:

$$\min_{T_{ijr}} Z = \sum_i \sum_j \sum_r t_{ijr} T_{ijr} \tag{7.10}$$

$$\text{subject to} \quad \sum_r T_{ijr} = T_{ij} \quad \forall i, j \tag{7.11}$$

$$T_{ijr} \geq 0 \quad \forall i, j, r \tag{7.12}$$

The objective function Z in the route based formulation (see (7.10)) expresses the total travel time experienced by all travelers. Constraint (7.11) implies that all travelers should be assigned to a route and constraint (7.12) is a nonnegativity constraint that applies to route flows. The total number of trips T_{ij} for each O-D pair is an input for this model. Note that we have to enumerate all the routes from each origin to each destination in order to use this model.

Link based formulation:

$$\min_{q_a} Z = \sum_a t_a q_a \tag{7.13}$$

$$\text{subject to} \quad q_a = \sum_{s \in S} q_a^s \quad \forall a \in A \tag{7.14}$$

$$T_{ms} + \sum_{a \in M^-} q_a^s = \sum_{a \in M^+} q_a^s \quad \forall s \in S \; \forall m \in N \setminus \{s\} \tag{7.15}$$

(conservation of route-flows directed at s)

$$q_a^s \geq 0 \quad \forall a \in A \; \forall \text{s} \in \text{S} \tag{7.16}$$

(non-negativity constraints)

where:

q_a^s	the flow on link a destined for node s
S	the set of destination nodes
M^-	set of incoming links to node m
M^+	set of outgoing links from node m
T_{ms}	The OD demand from node m to destination s (T_{ms} is only nonzero if m is an origin)
N	the set of all nodes

We will now demonstrate that, if it is assumed that the solutions to programs (7.10)–(7.12) and (7.13)–(7.16) are unique, the route based formulation (7.10)–(7.12) and the link based formulation (7.13)–(7.16) indeed yield identical link flows.

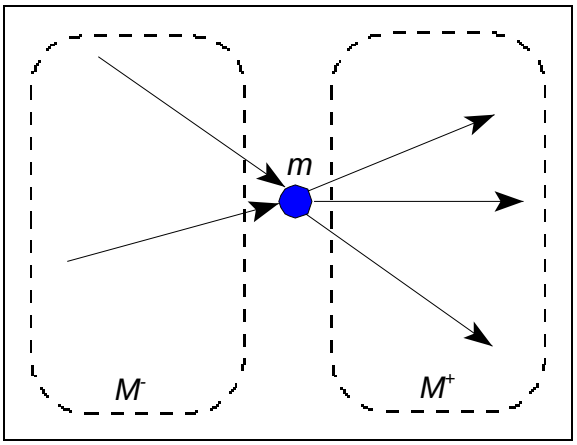


Figure 7.2: *Incoming and outgoing links*

Proof
In order to prove this we have to prove that objective functions (7.10) and (7.13) are identical and that conditions (7.11)–(7.12) are equivalent to conditions (7.14)–(7.16).

The first part of this proof is simple. Simply note that:

$$\sum_a t_a q_a = \sum_a t_a \sum_{i,j,r} T_{ijr} \alpha_{ijr}^a = \sum_{i,j,r} T_{ijr} \sum_a t_a \alpha_{ijr}^a = \sum_{i,j,r} T_{ijr} t_{ijr}$$

The second part is more complex, and falls apart in two parts:
-A- We prove that any set of route flows $\{T_{ijr}\}$ that satisfies (7.11)–(7.12) implies a unique set of link flows $\{ q_a^j \}$ that satisfies (7.14)–(7.16).

-B- We prove that for any set of link flows $\{ q_a^j \}$ that satisfies (7.14)– (7.16), a set of route flows $\{T_{ijr}\}$ that satisfies (7.11)–(7.12) can be constructed.

Ad A. If a set of route flows $\{T_{ijr}\}$ satisfies (7.11)–(7.12) then a set of link flows that satisfies (7.14)–(7.16) can be constructed with:

$$q_a^j = \sum_{i,j,r} \alpha_{ijr}^a T_{ijr}, \quad j = 1,2,\ldots; a = 1,2,\ldots$$

Ad B. Suppose a set of link flows $\{ q_a^j \}$ satisfies ((7.14),..(7.16)), then a set of route flows route flows $\{T_{ijr}\}$ that satisfies ((7.11), (7.12)) can be constructed as follows:
For all combinations *ij* and all sets of nodes N_{ijr} that form connecting paths between *i* and *j* define:

$$T_{ijr} = T_{ij} \prod_{m \in N_{ijr}} \beta_{ijr}^m$$

with :

$$\beta_{ij}^{rm} = \begin{cases} 0 & \text{if } \mathrm{M}^+(m) = \varnothing \\ \dfrac{\sum_{a \in M^+(m) \cap N_{ijr}} q_a^j}{\sum_{a \in M^+(m)} q_a^j} & \text{otherwise} \end{cases}$$

It can be verified that the set of route flows $\{T_{ijr}\}$ that is constructed in this way satisfies (7.11)–(7.12).

-A- and -B- imply that the optimal values that can be reached for objective functions (7.10) and (7.13) are equal. The proof is completed by observing that the optimal solutions $\{T_{ijr}\}$ and $\{q_a^j\}$ that lead to these optimal values are both unique and hence equivalent.

- end of proof -

In the link based formulation we have a similar objective function (see (7.13)) as in the route based formulation; it also expresses the total travel time experienced by all travelers. However, Z is written in terms of link-level variables in this case. The objective function implies minimizing the area indicated in Figure 7.3.

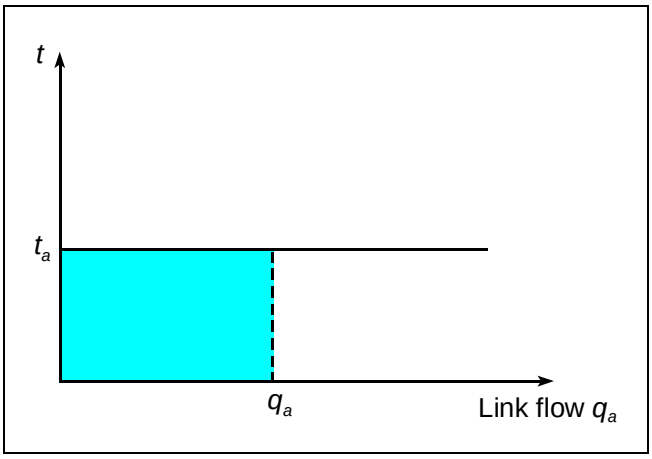


Figure 7.3: *Travel time function*

Constraints (7.15) are flow conservation constraints; for each direction, the sum of all incoming flows to a node must equal the sum of all outgoing flows at that node (Figure 7.2). Inequality (7.16) implies that all flows should be non-negative.
The optimal values of the objective functions of the route based model and the link based model are the same. If both solutions are unique, both models yield identical link flows. When for each OD pair a unique shortest path exists, the shortest path AON assignment is uniquely determined. In this case, both the route and link level based models generate output (route flows and link flows) corresponding to this assignment.
In case the shortest path AON assignment is not uniquely determined, i.e. when multiple routes corresponding to one OD pair have equal travel time, the models need not generate an AON assignment. It is possible that flow is distributed over multiple routes with equal travel time and hence is not AON. However, a computer implementation of either model will usually generate an AON assignment anyway due to the nature of the algorithms used.

7.3.2 Solving the AON assignment problem

Solving the AON assignment problem could be done in a straightforward fashion by solving the linear programs (7.10)-(7.12) or (7.13)-(7.16) directly using LP techniques. Due to the high dimensions of the problem, however, such an approach would soon become computationally infeasible. A more efficient way of solving it is by means of a shortest path algorithm. Shortest path algorithms exploit the properties of the problem that allow for a more efficient solution algorithm. For each origin a shortest path is determined to each destination, resulting in a shortest path tree for all origins (see Chapter 3). All trips are assigned along

these shortest paths which yields route flows on the network. By summation of all these route flows over all OD pairs (see Equation (7.9)), the link flows for the AON assignment are obtained. The algorithm is summarized in Figure 7.4.

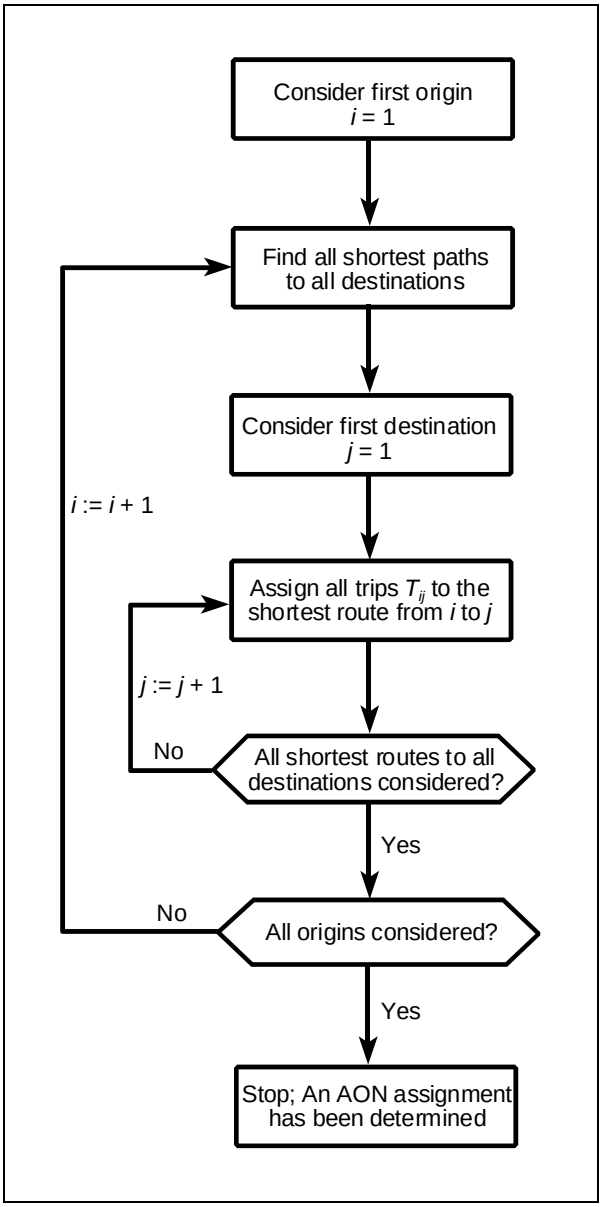


Figure 7.4: *AON assignment algorithm*

7.4 Stochastic assignment

7.4.1 Mathematical description of the stochastic assignment

There are several types of stochastic assignments. These assignments distribute the trips between two zones over several routes connecting those zones, making assumptions concerning route choice behavior. Since this model distributes the trips over several routes, a more equally spread flow pattern is obtained, compared with the AON assignment. This type of model is useful for analysis of bicycle traffic and car traffic during non-congested periods. In this section we will use the utility theory as a basis.

Again referring to the general network assignment problem (see Section 7.2), the stochastic assignment is characterized by the following:

Assumptions:
t_a is constant (independent of flow level q_a)
$\Theta>0$ (flow diversion among routes)

The definitional constraints simplify to:

(1) t_a is given and constant

(2) $$\beta_{ijr} = \beta_{ijr}(\underline{\Theta}, \underline{t}_a) \in [0,1] \tag{7.17}$$

(3) $$t_{ijr} = \sum_a \alpha_{ijr}^a t_a \tag{7.18}$$

(4) $$T_{ijr} = \beta_{ijr}(\underline{\Theta}, \underline{t}_a) T_{ij} \tag{7.19}$$

(5) $$q_a = \sum_i \sum_j \sum_r \alpha_{ijr}^a \beta_{ijr}(\underline{\Theta}, \underline{t}_a) T_{ij} \tag{7.20}$$

The stochastic assignment model assumes that the value of the generalized travel time that a traveler attaches to a route, known as the perceived or subjective generalized travel time, follows a probability distribution. The average of this probability distribution generally equals the generalized objective travel time. The standard deviation of the probability distribution is a measure for the behavioral differences between travelers. The difference between objective and subjective generalized travel time is due to differences in perceptions of the route attributes (such as cost and time), differences in weight of these attributes (value of time, time budget) and individual route preferences. In Figure 7.5 the situation of two alternative routes is shown.

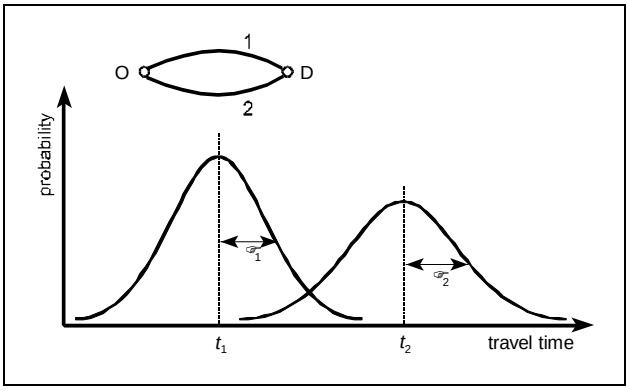


Figure 7.5: *Probability distributions route travel times*

Stochastic assignment can be defined as follows:

Stochastic assignment:
The assignment in which all travelers choose their perceived shortest path from the origin to the destination (congestion effects are not taken into account).

The subjective route travel time t_r can be formulated as follows (random variables are printed bold):

$$\boldsymbol{t}_r = t_r + \sqrt{(\Theta L_r)\boldsymbol{z}} \tag{7.21}$$

where:
- t_r = objective travel time of route r
- Θ = dispersion parameter
- L_r = measure for the length route r (e.g. distance)
- $\boldsymbol{z}$ is a random variable following some distribution.

Each route has a true travel time and a probability distribution over it for individual travelers. The trips are distributed to both routes based on the probability that a certain route is the shortest. The distribution can be depicted by a sigmoid curve (this is the probability distribution of the difference in subjective travel times between routes, see Figure 7.6). When both routes have the same objective route travel time t_r, the distribution will be fifty-fifty.

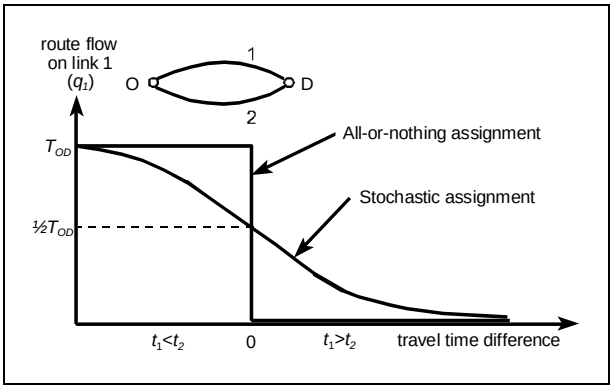


Figure 7.6: *Distribution of the flow between two routes in the stochastic assignment model and the AON assignment model*

Depending on the chosen probability distribution for the difference between objective and subjective generalized travel time, we obtain different models. Two types of probability distribution functions are frequently used. When the perceived attractiveness of the alternatives are assumed to be mutually independent, identically distributed Gumbel random variables, the *logit* model results. When these variables are multivariate normally distributed, the resulting model is called a probit model. In the latter case, z is standard normally distributed.

7.4.2 Solving the logit assignment

The logit model was discussed earlier in connection with mode choice. Therefore we only give the resulting formula for the route choice proportions in this context:

$$\beta_{ijr} = \frac{\exp(-\gamma t_{ijr})}{\sum_p \exp(-\gamma t_{ijp})} \tag{7.22}$$

Due to its computational simplicity, the logit model is the most popular model in stochastic assignment. It has however one major disadvantage: If more than two route are available of which two are largely overlapping (for example, see Figure 7.7), all routes are still considered as independent alternatives. This results in an overestimation of traffic on the overlapping routes.

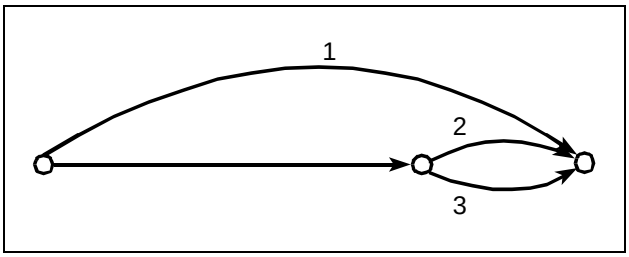


Figure 7.7: *Example of a three-route network in which two routes are largly overlapping*

7.4.3 Solving the probit assignment

In networks a solution for the probit assignment cannot be computed in an analytical fashion. Therefore, operational models use simulation instead. It can be shown that, under some reasonable assumptions, the stochastic formulation on route level (the actual behavior model) can be translated to an equivalent formulation on link level (see Bovy, 1990). The subjective link travel time t_a can for formulated as follows:

$$\mathbf{t}_a = t_a + \sqrt{(\Theta L_a)}\mathbf{z} \tag{7.23}$$

where:

$\mathbf{t}_a$ = objective travel time of link a
L_a = measure for the length of link a (e.g. distance)
$\mathbf{z}$ is a standard normally distributed random variable.

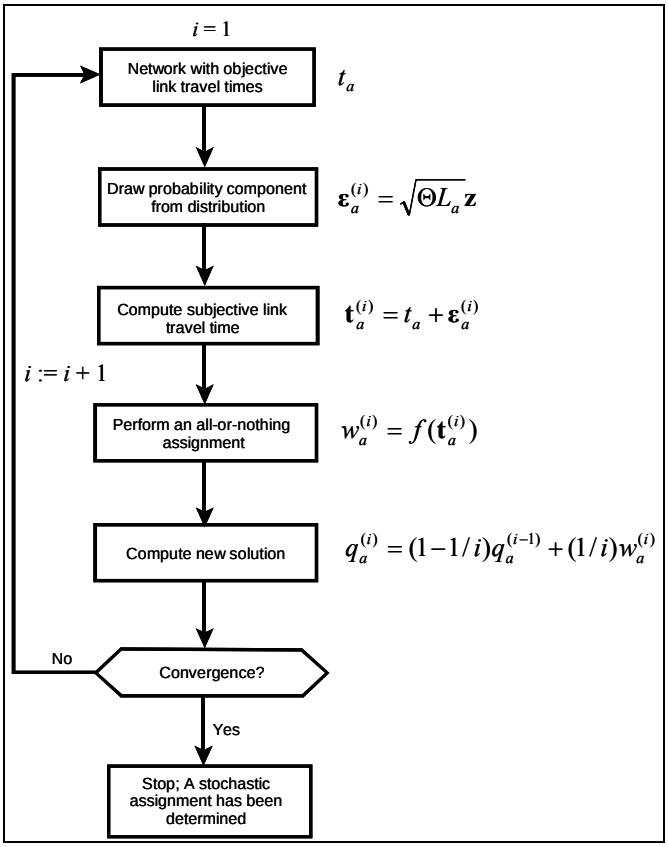


Figure 7.8: *Algorithm for stochastic assignment by simulation*

Check that (7.23) and (7.21) are equivalent. In the link based formulation of the probit model, shortest routes are being determined by a random sample of travel times from the link probability distributions. After each sample an all-or-nothing assignment is performed over the found shortest routes. This is repeated several times. Finally, the resulting flow on a route is the sum of all AON assignments divided by the number of iterations (see Figure 7.8). An advantage of the probit assignment is that it deals with overlapping route alternatives in a natural way. This is because error terms are defined on link level instead of route level.

Example 7.1: solution of the probit assignment by simulation
Given: Three parallel routes between A and B with objective travel times 20, 22 and 24 minutes, respectively (see Figure 7.9). There are 10,000 trips from A to B. The dispersion parameter has value 0.3 (obtained from observations). The subjective travel times are normally distributed around the objective travel times for each route.

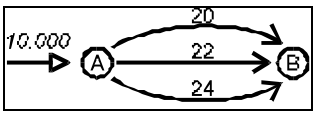


Figure 7.9: *Three parallel routes*

Question: How are the trips distributed over the three routes?

Solution: Using a random generator for z (available on every PC), a subjective route travel time is determined for each route (see Equation (7.21)). Then all trips are assigned (all-or-nothing) to the shortest subjective route (which results in a temporary route flow $w_r^{(i)}$ for iteration i). In the next iteration this process is repeated and the route flows, obtained from each iteration, are being averaged

recursively (which results in $T_r^{(i)}$). The convergence is measured by the sum of absolute deviations in the average route flows in consecutive iterations. Table 7.2 shows the results of the iterations.

Try to solve this example on your own PC! Remark: the results depend on the random generator you use.

After 50 iterations the trip distribution is 7800; 1400; 800. The absolute change in the final iteration is less than 1% (90/10000 to be precise). For a better understanding of the results it is recommended to draw the probability distributions of the subjective route travel times.

In the table the following equations are used:

(1) $\mathbf{t}_r^{(i)} = t_r + \sqrt{(\Theta L_r)}\mathbf{z}^{(i)} \quad \forall r$ (in this example we take $L_r = t_r$)

(2) $w_r^{(i)} = \begin{cases} T_{AB} & \text{if } \mathbf{t}_r^{(i)} = \min(\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \mathbf{t}_3^{(i)}) \\ 0 & \text{if } \mathbf{t}_r^{(i)} > \min(\mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \mathbf{t}_3^{(i)}) \end{cases} \quad \forall r$ (all or nothing assignment)

(3) $T_r^{(i)} = (1 - \frac{1}{i})T_r^{(i-1)} + \frac{1}{i} w_r^{(i)} \quad \forall r$

(4) $\Delta^{(i)} = \sum_r | T_r^{(i)} - T_r^{(i-1)} |$

	objective travel times			Θ			OD-flow			Σabsol. deviation sequent
iter.	20	22	24	0.3			10000			Δ [4]
	subjective travel times t_r [1]			temporary route flow w_r [2]			route flow T_r [3]			
i	route 1	route 2	route 3	route 1	route 2	route 3	route 1	route 2	route 3	
1	184	243	279	10000	0	0	10000	0	0	
2	211	229	251	10000	0	0	10000	0	0	0
3	182	222	271	10000	0	0	10000	0	0	0
4	214	182	205	0	10000	0	7500	2500	0	5000
5	204	208	222	10000	0	0	8000	2000	0	1000
6	203	221	197	0	0	10000	6667	1667	1667	3333
7	180	271	221	10000	0	0	7143	1429	1429	952
8	183	247	200	10000	0	0	7500	1250	1250	714
9	173	206	260	10000	0	0	7778	1111	1111	556
10	170	221	232	10000	0	0	8000	1000	1000	444
11	181	237	265	10000	0	0	8182	909	909	364
12	198	219	224	10000	0	0	8333	833	833	303
13	173	206	299	10000	0	0	8462	769	769	256
14	167	243	209	10000	0	0	8571	714	714	220
15	188	213	267	10000	0	0	8667	667	667	190
16	200	244	277	10000	0	0	8750	625	625	167
17	165	201	280	10000	0	0	8824	588	588	147
18	175	258	240	10000	0	0	8889	556	556	131
19	189	247	265	10000	0	0	8947	526	526	117
20	188	225	267	10000	0	0	9000	500	500	105
21	195	200	251	10000	0	0	9048	476	476	95
22	188	215	226	10000	0	0	9091	455	455	87
23	222	244	276	10000	0	0	9130	435	435	79
24	209	235	270	10000	0	0	9167	417	417	72
25	168	257	209	10000	0	0	9200	400	400	67
26	266	206	214	0	10000	0	8846	769	385	738
27	204	252	202	0	0	10000	8519	741	741	712
28	190	224	263	10000	0	0	8571	714	714	106
29	207	217	209	10000	0	0	8621	690	690	99
30	205	206	205	0	0	10000	8333	667	1000	621
31	236	197	208	0	10000	0	8065	968	968	602
32	224	225	246	10000	0	0	8125	938	938	121
33	236	220	219	0	0	10000	7879	909	1212	549
34	178	211	250	10000	0	0	7941	882	1176	125
35	158	237	199	10000	0	0	8000	857	1143	118
36	197	235	202	10000	0	0	8056	833	1111	111
37	204	266	274	10000	0	0	8108	811	1081	105
38	210	207	221	0	10000	0	7895	1053	1053	484
39	190	193	287	10000	0	0	7949	1026	1026	108
40	215	221	295	10000	0	0	8000	1000	1000	103
41	163	243	233	10000	0	0	8049	976	976	98
42	208	217	224	10000	0	0	8095	952	952	93
43	185	183	242	0	10000	0	7907	1163	930	421
44	183	256	218	10000	0	0	7955	1136	909	95
45	201	188	239	0	10000	0	7778	1333	889	394
46	197	266	250	10000	0	0	7826	1304	870	97
47	221	198	222	0	10000	0	7660	1489	851	370
48	212	254	224	10000	0	0	7708	1458	833	98
49	172	225	232	10000	0	0	7755	1429	816	94
50	240	291	250	10000	0	0	7800	1400	800	90

Table 7.2: *(Example) iterations of a stochastic assignment with three routes between A and B using the method in Bovy (1990).*

- end of example -

7.5 Deterministic equilibrium assignment

As opposed to stochastic assignment models, deterministic models do not take behavioral differences between travelers into account. Still a distribution of trips over multiple routes may arise if it is assumed that the traveler chooses the objective by shortest route, and level of

service attributes of links depend on link loads (see Figure 7.10). Models that take these conditions into account are known as equilibrium models.

Depending on the equilibrium definition, two types of deterministic equilibrium models can be distinguished; the deterministic user-equilibrium (UE) assignment and the deterministic system-equilibrium or system optimum (SO) assignment.

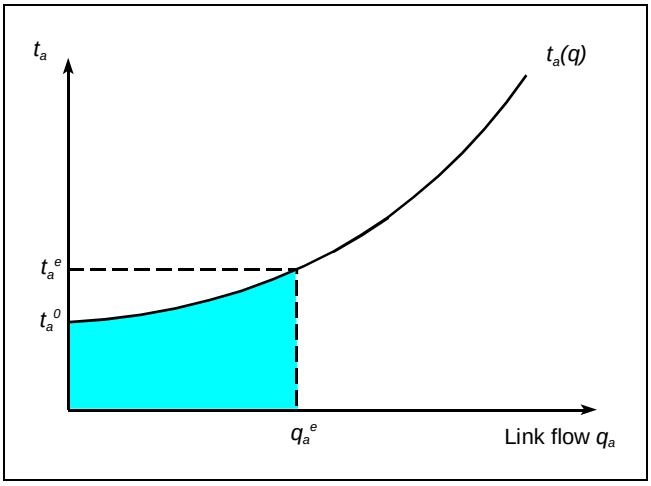


Figure 7.10: *Link travel time function*

7.5.1 Deterministic user-equilibrium assignment

7.5.1.1 Definition of the deterministic UE assignment

The deterministic user-equilibrium assignment can be defined as follows.

Deterministic user-equilibrium assignment:
The assignment which leads to an equilibrium in which no traveler can improve his travel time by unilaterally changing routes.

The formulation of the UE is attributed to Wardrop (1952) who showed that the deterministic UE conditions are equivalent to:

The travel time on any used route must be equal to the travel time on any other used route between the same origin and destination and no greater than the travel time on any unused route.

This principle is known as Wardrop's UE principle.

Example 7.2: solving for the deterministic user-equilibrium graphically.
In a case with three routes is illustrated. According to Wardrop, in user- equilibrium the following should hold:

$$t_a = t^e \text{ for all } q_a > 0, \text{ and } t_a > t^e \text{ if } q_a = 0 \tag{7.24}$$

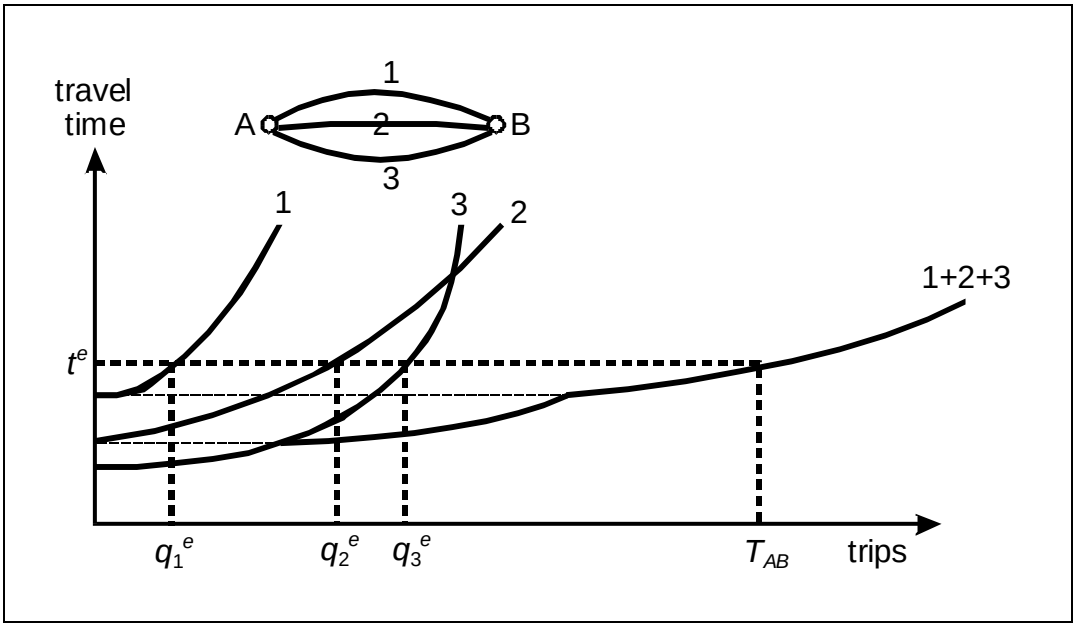


Figure 7.11: *Wardrop's principle*

Further, the sum of the route trips should equal the total number of trips from A to B:

$$q_1 + q_2 + q_3 = T_{AB} \tag{7.25}$$

This problem can be solved in a graphical way. Consider the link travel time functions $t_a(q_a)$ for link 1, 2 and 3 in Figure 7.11. Now assume that the equilibrium conditions are satisfied such that the travel time on all used routes is equal, say t^e. Then for each equilibrium travel time t^e the volumes on link 1, 2 and 3 can be read on the horizontal axis. Moreover, one may construct a combined link travel time function $t(q_1+q_2+q_3)$ for links 1, 2 and 3 by summing the link travel time functions horizontally (for each equilibrium travel time, the link volumes are added). Having constructed this combined travel time function, we look for T_{AB} trips on the horizontal axis and read t^e on the vertical axis. The equilibrium link loads are then found by reading the link load off the horizontal axis where the link travel time equals t^e, yielding q_1^e, q_2^e and q_3^e.

- end of example -

Although the above example shows an effective way to solve equilibrium problems involving only one OD-pair, such an approach cannot be applied to more complex equilibrium problems involving many OD-pairs and many partly overlapping routes. The problem should be solved mathematically in this case.

7.5.1.2 Mathematical description of deterministic user-equilibrium

Parallel to what was done in Sections 7.3.1 and 7.4.1, we will make the general assignment problem as stated in Section 7.2 more specific.

Assumptions:
$t_a = t_a(q_a)$ (travel time depends on the flow)
$\Theta = 0$ (no route split due to behavioral differences)

The definitional constraints simplify to:

(1) $t_a = t_a(q_a)$ (7.26)

(2) $\beta_{ijr} = \beta_{ijr}(\underline{t}_a(\underline{q}_a))$ (7.27)

(3) $t_{ijr} = \sum_a \alpha_{ijr}^a t_a(q_a)$ (7.28)

(4) $T_{ijr} = \beta_{ijr}(\underline{t}_a(\underline{q}_a))T_{ij}$ (7.29)

(5) $q_a = \sum_i \sum_j \sum_r \alpha_{ijr}^a \beta_{ijr}(\underline{t}_a(\underline{q}_a))T_{ij}$ (7.30)

Due to Beckmann *et al.* (1956) a technique has become available to solve equilibrium problems for more complex networks. Beckmann showed that solving the deterministic UE problem is equivalent to solving the following minimization problem:

$$\min_{\underline{q}_a} Z = \sum_a \int_0^{q_a} t_a(x)dx \quad (7.31)$$

$$\text{s.t. } \sum_r T_{ijr} = T_{ij} \quad \forall i,j \quad (7.32)$$

$$T_{ijr} \geq 0 \quad \forall i,j,r \quad (7.33)$$

$$q_a = \sum_i \sum_j \sum_r \alpha_{ijr}^a T_{ijr} \qquad \forall a \quad (7.34)$$

This minimization problem (known as Beckmann's formulation) can be described as follows: determine the link flows $\underline{q}_a$ such that the sum of the corresponding areas underneath the link travel time functions (see Figure 7.12) is minimal, taking into account the flow conservation constraints and nonnegativities.
The mathematical proof of the proposition that solving Beckmann's optimization problem and solving the deterministic UE problem are equivalent is omitted here. Instead, we make this proposition intuitively clear by illustrating it in a graphical way for a two link network (see Figure 7.12).

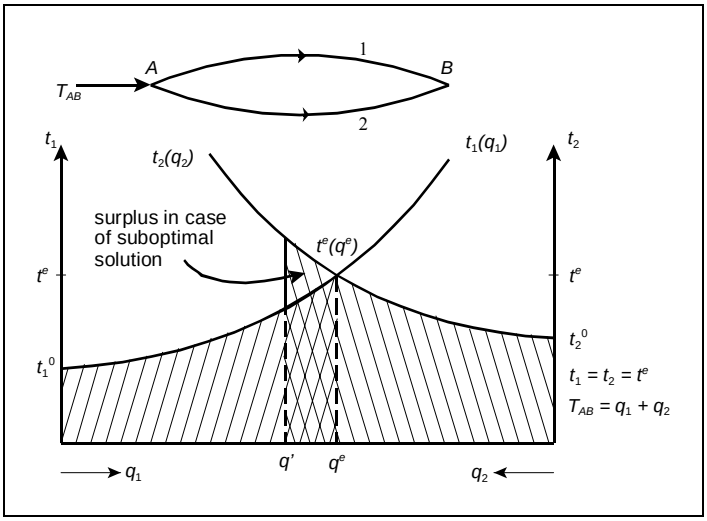


Figure 7.12: *Graphical derivation of deterministic user-equilibrium*

The horizontal axis measures exactly T_{AB} trips. This amount has to be split among links 1 and 2 such that travel times are equal (Wardrop-requirement). To derive this result graphically, the two travel time curves are shown opposite to each other. The point of equilibrium flows q is where the functions are equal. Minimizing the objective (7.31) function implies minimizing the summed areas underneath the link travel times from 0 to point q_a. Every other q-point left or right from the equilibrium point q^e (e.g. point q') would imply a larger surface under the travel time curves. Hence, solving the mathematical program (7.31)-(7.34) results the deterministic user-equilibrium.

As an alternative to the above program an equivalent minimization program exists solely expressed in terms of link level variables:

$$\min_{\underline{q}_a} Z = \sum_a \int_0^{q_a} t_a(x)dx \qquad \text{(Objective function)} \qquad (7.35)$$

$$\text{s.t. } q_a = \sum_{s \in S} q_a^s \quad \forall a \in A \qquad \text{(Flow conservation)} \qquad (7.36)$$

$$T_{ms} + \sum_{a \in M^-} q_a^s = \sum_{a \in M^+} q_a^s \quad \forall s \in S \; \forall m \in N \setminus \{s\} \qquad \text{(Flow conservation at nodes)} \qquad (7.37)$$

$$q_a^s \geq 0 \quad \forall a \in A \; \forall s \in S \qquad \text{(Non-negativity constraints)} \qquad (7.38)$$

see page 95 for an explanation of the symbols

7.5.1.3 Link performance functions

In all equilibrium assignment procedures the link performance (travel time) functions play a key role. These functions give a mathematical description of the relationship between (stationary levels of) travel time and link flow. The minimum value of this function corresponds to the free flow travel time, whereas the gradient of the function depends on the amount of interaction between vehicles. In general the amount of interaction and hence the gradient of the travel time function increase with flow, leading to a convex function. In theory travel time should approach infinity when flow raises to capacity. Link performance functions need not be defined for flows exceeding capacity because such flows do not exist in practice. However, there is a computational advantage in using travel-time functions that are also defined for flows exceeding capacity, because many algorithms compute user-equilibria in an iterative way. Intermediate solutions may violate the constraints imposed by link capacities.

Deriving link-performance functions
Link cost functions can be derived in different ways: using hydrodynamic theory (i.e. utilizing the analogy with fluids) and using queuing theory.
The *fundamental* diagram (see the lecture notes of Traffic Flow Theory and Simulation, CT4821) underlies the hydrodynamic theory as originally formulated by (Lighthill and Whitham,1955). The fundamental diagram defines a relation between vehicle speed and traffic density (traffic density is the number of vehicles per km of road), i.e. it defines at which speed a driver is willing to drive depending on the amount of *headway* available. Recall that speed, flow and density are interlinked through the relation ship: flow = speed * density. Hence a relationship between speed and density also implies relationships between speed and flow, and flow and density.
Queuing theory derives link performance in another way: it interprets travel time as the expected result of the interaction between a stochastic, traffic generating process and a fixed or stochastic capacity. As the rate of travel generation increases the probability of demand temporarily exceeding capacity increases, resulting into queues and hence increased travel time. For further reading see (Bell and Iida, 1955).

BPR function
The BPR (Bureau of Public Roads) formulation is one of the most-commonly used link travel time functions (see Figure 7.13). The BPR function relates link travel times as a function of the flow/capacity ratio of that link according to:

$$t_a(q_a) = t_a^0\left(1 + \alpha\left(\frac{q_a}{C_a}\right)^{\beta}\right) \tag{7.39}$$

where:
t_a = (congested) travel time on link a
q_a = flow on link a
t_0^a = free-flow travel time on link a (q_a=0)
C_a = capacity of link a
α,β are parameters.

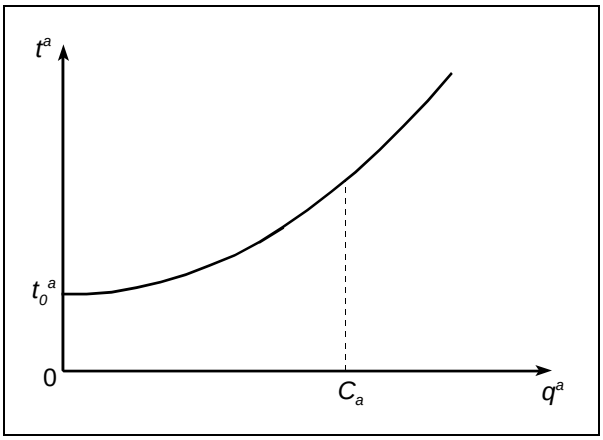


Figure 7.13: *BPR-function*

While a number of different formulations of such functions have been suggested over the years, the BPR function (Traffic Assignment Manual, 1964) is very well suited for use in conjunction with traffic assignment models. With a suitable choice of parameters, this function can represent a wide variety of flow-delay relationships (including those of many other flow-delay models) and it is used by the traffic assignment models in TRANSCAD 3.0.

Usual values for α and β are 0.15 and 4.0, respectively. However, different values can and should be used in many circumstances. For example, these parameters can be modified to include the approximate effect of intersection delay associated with a link.

Davidson function
Another well known link performance function is the Davidson function (see Davidson, 1966):

$$t_a(q_a) = t_a^0\left(1 + J\frac{q_a}{C_a - q_a}\right) \tag{7.40}$$

where J is a parameter that reflects the road type, design standard, etc. The Davidson function is not defined for values of q_a exceeding the road capacity C_a. Although in practice such flows are not feasible, they might turn up as an intermediate solution when solving equilibrium assignment problems in an iterative way (see next section). Therefore the contemporary

models use modified versions of Davidson's function, like the one proposed by (Tisato, 1991):

when $q_a \le \beta C_a$, $\beta \approx 0.9$:
see equation (7.40)

otherwise:

$$t_a(q_a) = t_a^0\left(1 + J\frac{\beta C_a}{C_a - \beta C_a} + JC_a\frac{(q_a - \beta C_a)}{(C_a - \beta C_a)^2}\right) \tag{7.41}$$

The modified Davidson function equals the original function for flows under 90% of capacity, but is a linear interpolation of Davidson above this value.

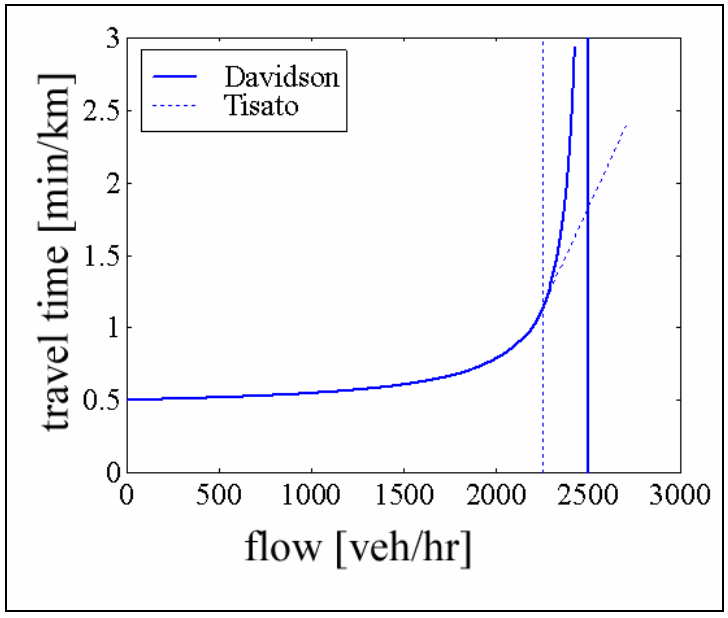


Figure 7.14: *Davidson and modified Davidson function*

Junction delays
A network is composed of links and nodes. Above travel time functions are defined for links. However, in practice nodes are often the main source of traffic load dependent delays. Examples are intersections in urban networks and merging sections at freeways. For this reason often all turning movements in an intersection are separately coded (see Chapter 3). Note that the travel times on a specific turning direction depends not only on the flow for that direction but also on the flows for conflicting directions, i.e. the travel time functions for the different turning directions of intersections are no longer separable.

7.5.1.4 Solving the deterministic user-equilibrium assignment problem

Wardrop presented the user equilibrium principle in 1952 and only 4 years later Beckmann *et al.* (1956) proposed a rigorous mathematical framework to express this principle as a mathematical program. Yet it took many years before suitable algorithms for practical implementation were proposed and tested.

Beckmann *et al.* compared the equilibrium assignment problem to equilibria problems encountered in theoretical mechanics. One characteristic of such problems is that they may be

expressed as extremum problems. He showed that, when the cost t_a on any link a is a function of the flow q_a on link a only, and the link performance functions are increasing, then the flows q_a satisfying Wardrop's (first) principle are unique and equal to those which minimize Equation (7.31) subject to Equations (7.32)-(7.34).

The method normally used to solve this minimization problem is the convex combination algorithm, originally suggested by Frank and Wolfe in 1956 as a procedure for solving quadratic programming problems with linear constraints; it is also known as the FW method. To efficiently solve this problem in networks, the following algorithm can be derived.

Frank-Wolfe Algorithm for solving the deterministic UE assignment problem:

1. [Initialization]
   Take i:= 1, and perform an AON assignment based on $t_a = t_a(0)$. This yields flow vector $\underline{q}_a^{(i)}$.

2. [Update link travel times]
   Compute $t_a = t_a(q_a^{(i)}) \ \forall a$.

3. [Determine descent direction]
   Perform an AON assignment based on t_a. This yields the auxiliary flow vector $w_a^{(i)}$.

4. [Determine step size]
   Find $\alpha^{(i)}$ that solves:

$$\min_{0\le\alpha\le 1} \sum_a \int_0^{q_a^{(i)}+\alpha(w_a^{(i)}-q_a^{(i)})} t_a(x)dx \qquad (7.42)$$

   A simpler approach to this step is choosing $\alpha^{(i)}=1/i$. This approach is called the Method of Successive Averages (MSA).

5. [Move]
   Set $q_a^{(i+1)} = q_a^{(i)} + \alpha^{(i)}(w_a^{(i)} - q_a^{(i)}) \ \forall a$.

6. [Convergence test]
   If a certain predetermined convergence criterion is met, then stop. Otherwise, set i:= i+1 and return to step 2.

The steady state UE problem is formulated as in (7.31)–(7.34) or (7.35)–(7.38). LeBlanc proved that this optimization problem is convex with respect to the link flows q_a. However, this problem is not convex with respect to path flows T_{ijr}. We now derive the algorithm for solving the UE problem.

Intermezzo: Mathematical underpinning of the FW method:

Outline of the method

The Frank-Wolfe method is applied to minimize the following function:

$$Z(\underline{q}_a) = \sum_a \int_0^{q_a} t_a(x)dx \qquad (7.43)$$

under the following conditions:

$$\text{(1)} \qquad \sum_r T_{ijr} = T_{ij} \qquad \forall i, j \tag{7.44}$$

$$\text{(2)} \qquad T_{ijr} \geq 0 \quad \forall i, j, r \tag{7.45}$$

$$\text{(3)} \qquad q_a = \sum_i \sum_j \sum_r \alpha_{ijr}^a T_{ijr} \qquad \forall a \tag{7.46}$$

Step 1: Determine any solution that satisfies the conditions (1), (2) and (3). Refer to this solution as the initial feasible solution $\underline{q}_a^{(0)}$. Set $i := 1$.

Step 2: Linearize the objective function $Z(\underline{q}_a)$ around the point $\underline{q}_a^{(i)}$. This results in the linear function $Z^{(i)}(q_a)$.

Step 3: Find the vector $\underline{w}_a^{(i)}$ that minimizes the linear function $Z^{(i)}(\cdot)$ under the conditions (1) – (3). Refer to this solution as the auxiliary solution.

Step 4: Find the point on the line $\underline{q}_a^{(i-1)}$ between $\underline{w}_a^{(i)}$ and in which Beckmann's objective function is minimized.

Step 5: Refer to this point as $\underline{q}_a^{(i)}$.

Step 6: Evaluate the convergence criterion. For example, check the difference between the shortest and the longest route in use. Stop if differences in travel time are sufficiently small.

Mathematical description of the method

Step 1: Check that by performing an AON assignment a feasible solution satisfying (1), (2) and (3) is found.

Step 2: Given at iteration step i a feasible flow vector $\underline{q}_a^{(i)}$ (a flow vector that satisfies the constraints in the mathematical program), a first order expansion of the objective function Z around $\underline{q}_a^{(i)}$ can be written as:

$$Z^{(i)}(\underline{w}_a) = Z(\underline{q}_a^{(i)}) + \sum_a \frac{\partial Z(\underline{q}_a^{(i)})}{\partial q_a}(w_a - q_a^{(i)}) \tag{7.47}$$

Step 3: In order to find a good direction in which to seek a decreased value of the original objective function, the following linear program must be solved:

$$\text{LP: } \min_{\underline{w}_a} Z^{(i)}(\underline{w}_a) = Z(\underline{q}_a^{(i)}) + \sum_a \frac{\partial Z(\underline{q}_a^{(i)})}{\partial q_a}(w_a - q_a^{(i)}) \tag{7.48}$$

under conditions (1), (2) and (3)

Note that

$$\frac{\partial Z(\underline{q}_a^{(i)})}{\partial q_a} = \frac{\partial}{\partial q_a}\int_0^{q_a} t_a(x)dx\Big|_{q_a=q_a^{(i)}} = t_a(q_a^{(i)}) \tag{7.49}$$

Further manipulations of equation (7.48) and the removal of all constant terms yield the following objective function:

$$\text{LP: } \min_{\underline{w}_a} \sum_a t_a(q_a^{(i)})w_a \tag{7.50}$$

subject to constraints on $\underline{w}_a$, equivalent to (1), (2) and (3).

Note the similarity between this LP problem and the one defined by (7.10) - (7.12), which was shown to correspond to an AON assignment. The solution of the LP defined above is hence given by an AON assignment based on $t_a(q_a^{(i)})$.

Solving the linear program (7.50) yields a solution vector $\underline{w}_a^{(i)}$, which is also a feasible solution of the original non-linear problem. The direction $d^{(i)} = \underline{w}_a^{(i)} - \underline{q}_a^{(i)}$ is a descent direction in which we seek a decreased value of the objective function *Z*.

Step 4: In order to find the best next point (the flow vector for the next iteration), *Z* has to be minimized along $d^{(i)} = \underline{w}_a^{(i)} - \underline{q}_a^{(i)}$:

$$\min_{0\le\alpha\le 1} \sum_a Z\left(q_a^{(i)} + \alpha(w_a^{(i)} - q_a^{(i)})\right) \tag{7.51}$$

Since the interval is closed, this one-dimensional minimization problem can be solved using any interval reduction method (for example, the bisection method), yielding the optimal stepsize $\alpha^{(i)}$.

Step 5: The next point can be calculated as:

$$q_a^{(i+1)} = q_a^{(i)} + \alpha(w_a^{(i)} - q_a^{(i)}) \tag{7.52}$$

Convergence criterion
The Frank Wolfe Algorithm consists of an iterative scheme. As with all iterative procedures, the stop criterion may have a considerable impact on the results. The stop criterium is the result of a trade off between computation time and accuracy. The following stop criteria may be applied:

- *Stop when results do no longer change*. This is risky because this is not a guarantee that the optimum has been reached. Continuing the iterations may result in a significantly different solution.
- *Stop after a fixed number of iterations*. This does not guarantee convergence at all, unless a very high number of iterations is selected. However, one may argue that applying a fixed number of iterations for each assignment makes the assignment results more comparable. Also, the convergence pattern of the Frank Wolfe algorithm is such that the first few iterations are the most cost effective. According to (Sheffi, 1985) only four to six iterations are usually sufficient to find the equilibrium flow pattern over large urban networks.

- *Stop if the difference between shortest and longest route is sufficiently small*. Routes having equal costs is one of the of the Wardop conditions. However this stopping criterion does not consider the usage of the routes.
- *Stop if duality gap is sufficiently small*. The duality gap is defined as the excess travel time relative to the minimum possible travel time based on the route costs computed in the present iteration divided by this minimum travel time:

$$DG = \frac{\sum_a q_a t_a - \sum_{i,j} T_{ij} t_{ij}}{\sum_{i,j} T_{ij} t_{ij}}$$

with:

t_{ij} the travel time on the shortest path from origin i to destination j
T_{ij} the travel demand on OD-pair i-j

Because the duality gap takes into the account the differences between route travel time and weighs these with the route flows it is an effective criterium to measure wether or not the Wardrop conditions are met.

- end of intermezzo -

Example 7.3: solution of the deterministic user-equilibrium assignment
Given: Three parallel routes between *A* and *B* with free-flow travel times 20, 25 and 30 minutes, respectively, and capacities 2000, 3000 and 5000 veh/hour respectively (see Figure 7.15).

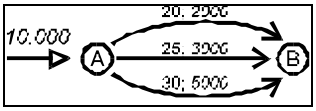


Figure 7.15: *Three parallel routes*

There are 10,000 trips from *A* to *B*. The travel time functions all have the form:

$$t_a = t_a^0 \left(1 + 0.15 \left(\frac{q_a}{C_a} \right)^4 \right) \tag{7.53}$$

Question: How are the trips distributed over the three routes and which travel times occur?

Solution:
Step 0: Start with an empty network ($q_a = 0$ for all links a).
Step 1: Calculate the corresponding travel times on each route, using Equation (7.50). (Note that each route only consists of one link, such that a link flow equals a route flow and the link travel time is that same as the route travel time). Assign all trips to the shortest route (AON assignment).This yields our first equilibrium flow pattern.
Step 2: Calculate the travel times which correspond to the equilibrium flows found in the previous step. Then perform an AON assignment to the shortest paths, which results in an auxiliary flow. The new equilibrium flow pattern is determined by a weighted average between the previous equilibrium flow and the auxiliary flow. In this example, the method of successive averages is used. Repeat this process until the equilibrium flows or equilibrium travel times no longer change significantly (convergence).

Table 7.3 shows the results of the iterations. Try to solve this example on your own PC!

After 50 iterations the route travel times are nearly equal (Wardrop-requirement) and the trip distribution is 2800; 3400; 3800. The absolute change in the final iteration is less than 3% (253/10000 to be precise).

What can we say about the traffic conditions (for example, congestion) on each of the three routes?

In the table the following equations are used:

1) see problem (7.50)

2) $\alpha^{(i)} = 1/i \times 100\%$

3) $T_r^{(i)} = (1-\alpha^{(i)})T_r^{(i-1)} + \alpha^{(i)} w_r^{(i)} \quad \forall r$

4) $\Delta^{(i)} = \sum_r \left| T_r^{(i)} - T_r^{(i-1)} \right|$

	free-flow travel times			capacity				OD-flow T_{AB}			Σabsol. deviation sequent.
	20	25	30	2000	3000	5000		10000			
iter. no.	route travel times t_r [1]			auxiliary flow w_r			α [2]	equilibrium flow T_r [3]			Δ [4]
i	R-1	R-2	R-3	R-1	R-2	R-3	MSA	R-1	R-2	R-3	
0				0	0	0		0	0	0	
1	20	25	30	10000	0	0	100	10000	0	0	
2	1895	25	30	0	10000	0	50	5000	5000	0	10000
3	137	54	30	0	0	10000	33	3333	3333	3333	6667
4	43	31	31	0	10000	0	25	2500	5000	2500	3333
5	27	54	30	10000	0	0	20	4000	4000	2000	3000
6	68	37	30	0	0	10000	17	3333	3333	3333	2667
7	43	31	31	0	10000	0	14	2857	4286	2857	1905
8	32	41	30	0	0	10000	13	2500	3750	3750	1786
9	27	34	31	10000	0	0	11	3333	3333	3333	1667
10	43	31	31	0	10000	0	10	3000	4000	3000	1333
11	35	37	31	0	0	10000	9	2727	3636	3636	1273
12	30	33	31	10000	0	0	8	3333	3333	3333	1212
13	43	31	31	0	10000	0	8	3077	3846	3077	1026
14	37	35	31	0	0	10000	7	2857	3571	3571	989
15	32	33	31	0	0	10000	7	2667	3333	4000	857
16	29	31	32	10000	0	0	6	3125	3125	3750	917
17	38	29	31	0	10000	0	6	2941	3529	3529	809
18	34	32	31	0	0	10000	6	2778	3333	3889	719
19	31	31	32	0	10000	0	5	2632	3684	3684	702
20	29	34	31	10000	0	0	5	3000	3500	3500	737
21	35	32	31	0	0	10000	5	2857	3333	3810	619
22	32	31	32	0	10000	0	5	2727	3636	3636	606
23	30	33	31	10000	0	0	4	3043	3478	3478	632
24	36	32	31	0	0	10000	4	2917	3333	3750	543
25	34	31	31	0	10000	0	4	2800	3600	3600	533
26	32	33	31	0	0	10000	4	2692	3462	3846	492
27	30	32	32	10000	0	0	4	2963	3333	3704	541
28	34	31	31	0	10000	0	4	2857	3571	3571	476
29	32	33	31	0	0	10000	3	2759	3448	3793	443
30	31	32	31	10000	0	0	3	3000	3333	3667	483
31	35	31	31	0	10000	0	3	2903	3548	3548	430
32	33	32	31	0	0	10000	3	2813	3438	3750	403
33	32	31	31	0	0	10000	3	2727	3333	3939	379
34	30	31	32	10000	0	0	3	2941	3235	3824	428
35	34	30	32	0	10000	0	3	2857	3429	3714	387
36	32	31	31	0	0	10000	3	2778	3333	3889	349
37	31	31	32	0	10000	0	3	2703	3514	3784	360
38	30	32	31	10000	0	0	3	2895	3421	3684	384
39	33	31	31	0	0	10000	3	2821	3333	3846	324
40	32	31	32	0	10000	0	3	2750	3500	3750	333
41	31	32	31	10000	0	0	2	2927	3415	3659	354
42	34	31	31	0	0	10000	2	2857	3333	3810	302
43	32	31	32	0	10000	0	2	2791	3488	3721	310
44	31	32	31	10000	0	0	2	2955	3409	3636	328
45	34	31	31	0	10000	0	2	2889	3556	3556	293
46	33	32	31	0	0	10000	2	2826	3478	3696	280
47	32	32	31	0	0	10000	2	2766	3404	3830	268
48	31	31	32	10000	0	0	2	2917	3333	3750	301
49	34	31	31	0	10000	0	2	2857	3469	3673	272
50	32	32	31	0	0	10000	2	2800	3400	3800	253

Table 7.3: *(Example) iterations of a deterministic user-equilibrium assignment with three routes between A and B.*

- end of example -

7.5.2 Deterministic system optimal assignment

7.5.2.1 Definition of the deterministic SO assignment

The deterministic system optimal assignment can be defined as follows:

Deterministic system optimal assignment:
The assignment which leads to the system state in which the total travel costs on a network is minimized.

Under SO assignment, no user can change routes without increasing the total travel costs on the system, although it is possible that the traveler could reduce his/her own travel costs. SO assignment can be thought of as a model in which congestion is minimized when drivers are told which routes to use. Obviously not a behaviorally realistic model, SO assignment can be useful in analyzing Intelligent Transport System (ITS) scenarios.

A SO assignment shows the most efficient performance of the transportation system. Because of this it can be used as a benchmark with which to compare other assignments in order to measure their relative performance. In addition, the results of an SO assignment can give clues how to alter the network to achieve an UE that is close to the SO performance.

7.5.2.2 Mathematical description of deterministic system-equilibrium

The model formulation and assumptions for the deterministic SO are equivalent to the formulation and assumptions made for the deterministic UE, except for the objective function. From the principle of the SO assignment, it is clear that the objective function to be minimized now is:

$$\min_{\underline{q}_a} Z(\underline{q}_a) = \sum_a t_a(q_a) q_a \tag{7.54}$$

Graphically, this means minimizing the sum of areas of the rectangles $t_a^e q_a^e$ (see Figure 7.16).

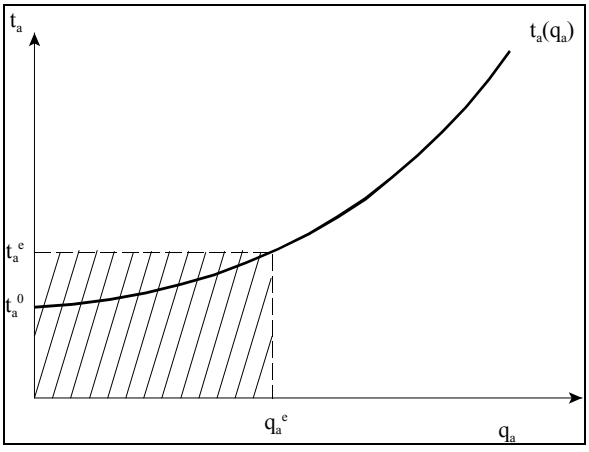


Figure 7.16: *SO-minimization formulation using average travel time function t_a*

7.5.2.3 Solving the deterministic system optimal assignment problem

It can be easily shown that the SO assignment problem can be solved with an equivalent approach to the one explained in the UE case (Section 8.5.1.4).

For a differentiable function $f(x)$ it holds that

$$f(x) = \int f'(x)dx \tag{7.55}$$

where $f'(x)$ is the derivative of $f(x)$. Since

$$\frac{d}{dq_a}[t_a(q_a)q_a] = t'_a(q_a)q_a + t(q_a) \tag{7.56}$$

we can write the SO objective function (7.54) in the form:

$$\min_{\underline{q}_a} Z(\underline{q}_a) = \sum_a \int_0^{q_a} [t'_a(x)x + t_a(x)]dx \tag{7.57}$$

where $t'_a(\cdot)$ is the derivative of the link travel time function.

Define

$$t^*_a(x) \equiv t'_a(x)x + t_a(x) \tag{7.58}$$

The function $t^*_a(\cdot)$ is referred to as the marginal link cost function. Now it follows that an SO can be found by solving:

$$\min_{\underline{q}_a} Z(\underline{q}_a) = \sum_a \int_0^{q_a} t^*_a(x)dx \tag{7.59}$$

In other words, an SO can be found by solving a UE assignment using the marginal link cost function! The methods for solving UE assignments have been discussed in Section 7.5.1.

The interpretation of the marginal link cost function $t^*_a(x)$ is that it defines the total extra system cost of one extra vehicle at traffic volume x. The marginal cost (I) can be decomposed in

- $t_a(x)$, the travel time that the extra driver on link a will incur (II).
- $t'_a(x)x$, the extra travel time that the other drivers will incur due to one extra trip on link a (III).

See Figure 7.17.

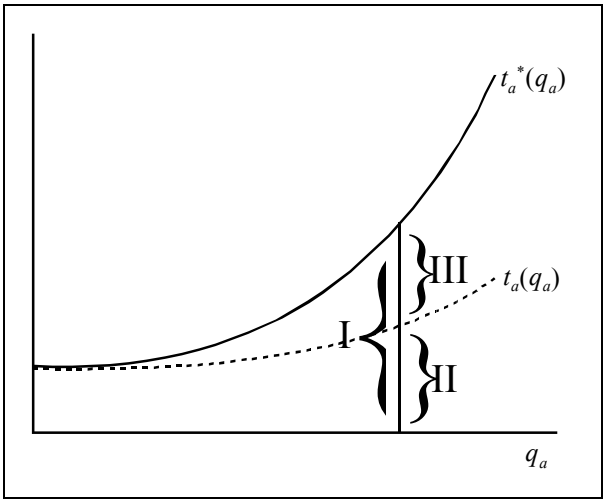


Figure 7.17: *Marginal link cost function*

Example 7.4: using the BPR function in an SO assignment
Suppose we use the well-known BPR function as the link travel time function:

$$t_a(q_a) = t_a^0\left(1+\alpha\left(\frac{q_a}{C_a}\right)^{\beta}\right) \tag{7.60}$$

To achieve an SO assignment, we have to use the following (marginal link) cost function:

$$t_a^*(q_a) = t_a^0\frac{\alpha\beta}{C_a}q_a\left(\frac{q_a}{C_a}\right)^{\beta-1} + t_a^0\left(1+\alpha\left(\frac{q_a}{C_a}\right)^{\beta}\right) = t_a^0\left(1+\alpha(\beta+1)\left(\frac{q_a}{C_a}\right)^{\beta}\right) \tag{7.61}$$

- end of example -

7.5.2.4 Congestion pricing

In the previous section we have seen that, given a set of link cost functions, an assignment that minimizes total system costs can be computed by performing a user optimum assignment using the marginal link cost functions. The implications of this are wider than only the computational aspect. It also means that the state of an actual traffic system can be directed to a system optimum by making sure the costs experienced by its users equal the marginal costs of their trips (where 'costs' is defined in terms of the cost function that one wishes to minimize). Such a strategy is referred to as congestion pricing (other related terms are road pricing and value pricing).

Example 7.5: Optimal congestion pricing
Problem:
Suppose a road operator is responsible for a transport network, and its mission is to minimize total travel time spent in the network. Apart from charging drivers for their use of the network, the road operator has no means to influence drivers. It is assumed that the OD-demand is not influenced by road pricing or congestion levels. So the only decision that can be influenced is the route choice decision. Road users are assumed to select those routes that minimize their generalized travel times. These are defined by:

$$c_r^g = c_r + VOT \cdot t_r \tag{7.62}$$

where:

c_r^g	the generalized travel time for route *r*
VOT	Value of Time (the monetary costs of one unit of travel time delay)
t_r	travel time of route *r*
c_r	charge for route *r*

At which level should the drivers be charged in order for the road operator to reach his goal?

Answer:
Drivers select their routes by minimizing generalized travel time. For a route, the generalized travel time is given by:

$$\begin{aligned} c_r^g &= c_r + VOT \cdot t_r = \sum_{a \in A_r} \left(c_a(q_a) + VOT \cdot t_a(q_a) \right) \\ &= \sum_{a \in A_r} c_a(q_a) + VOT \sum_{a \in A_r} t_a(q_a) \end{aligned} \tag{7.63}$$

where:

A_r	the links that constitute route *r*

The road pricing should be at such a level that the user equilibrium that results minimizes total system travel time. Therefore the generalized travel costs of each link equal the marginal costs of the cost component attributed to the travel time, i.e. a sufficient condition for system optimum is:

$$c_a(q_a) + VOT \cdot t_a(q_a) = VOT \cdot \left(t'_a(q_a) q_a + t_a(q_a) \right) \tag{7.64}$$

From this it follows that a flow dependent charging regime should be introduced, using the following level of charging:

$$c_a(q_a) = VOT \cdot t'_a(q_a) q_a \tag{7.65}$$

- end of example -

In practice the technology to apply a tolling strategy such as depicted in equation (7.65) is not (yet) available. At the moment the feasible alternatives for road pricing are:
- link based toll charging (variable or fixed),
- road pricing zones, every vehicle driving in a zone pays a fixed price,
- cordon based pricing, every vehicles that enters a zone is charged a fixed price,
- travel distance based pricing,
- time-based road pricing.

To find optimal road pricing strategies in these contexts requires solving a so-called *bi-level* problem: the *upper level* corresponds to the (system) objective function one wants to minimize, e.g. total travel time, while the *lower level* corresponds to the typical objective function that is minimized under user equilibrium conditions, see e.g. equation (7.59).

$$\begin{aligned} &\theta^{\text{optimum}} = \arg\min_{\theta} \sum_a q_a(\theta) t_a^{system}(q_a) \qquad \text{<upper level problem>} \\ &\text{sub:} \\ &q_a(\theta) = \arg\min_{q_a} \sum_a \int_0^{q_a} t_a^{user}(w, \theta) dw \qquad \text{<lower level problem>} \end{aligned} \tag{7.66}$$

where:

θ	the vector of steering parameters (e.g. coefficients in road pricing functions)

t_a^{system} system costs function, the contribution per vehicle to the total system objective function (e.g. emission, travel time, etc.)

t_a^{user} user costs function, the costs that each individual aims to minimize

Solving this kind of problems is not within the scope of this course. See e.g. the course CT5804 for further study on this subject.

7.5.2.5 Link performance functions revisited: optimizing policy objectives

The system optimum assignment is defined as the distribution over OD-demand over routes that minimizes the total travel costs. Intuitively costs are often associated with travel time, monetary costs or other sacrifices made by the driver. However, also other disadvantages and advantages maybe included in the systems costs, e.g.:

– Energy usage
– Pollutant emissions
– Noise
– Pedestrian exposure to risk
+ Accessibility
+ Reliability of transport system

Many of these can be incorporated in a link cost function in one way or the other. By subsequently computing the system optimal assignment, one can gain insight in which distribution of traffic over routes would be most favorable from the perspective of a societal objective.

Example 7.6: Optimal congestion pricing
Consider a freeway from A to B (route 1) with parallel to it a route over the secondary network (route 2). During the peak, the freeway is usually severely congested. Diverting some of the traffic to the secondary network would alleviate the congestion to some extent, but also implies many disadvantages. Suppose we only take into account total travel time and safety. It is clear that trips over route 2 induce a higher risk than trips over route 1. Which proportion of traffic should (be allowed to) divert to the secondary network, in order to minimize the overall system costs?

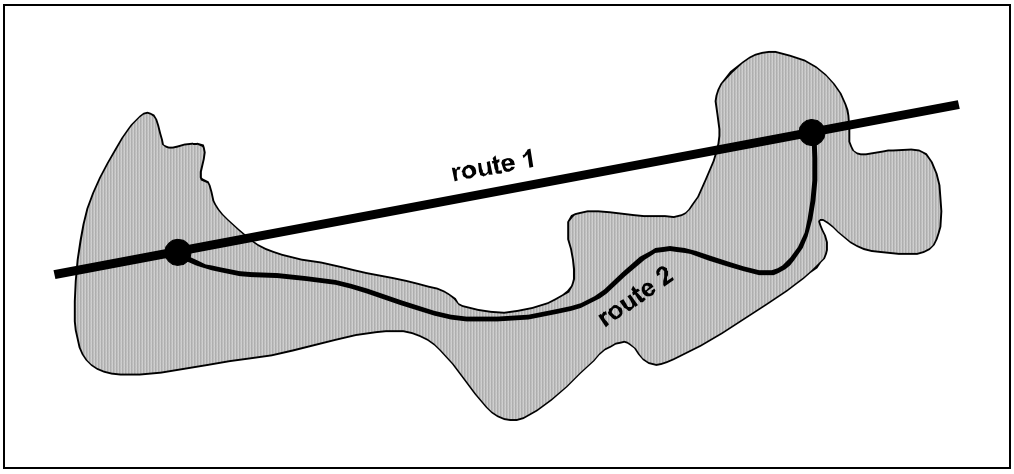


Figure 7.18: *Route alternatives over primary and secondary network respectively*

In order to answer this question we must first define an objective function. If only safety and travel time are taken into account, the objective function is:

$$c_r^{\text{TOTAL}} = c_r^{\text{TIME}} + c_r^{\text{RISK}} \tag{7.67}$$

where c_r^{TOTAL} refers to total costs for route *r*, c_r^{TIME} refers to costs as a result of time delays for route *r*, and c_r^{RISK} refers to the level of safety. To compute travel time costs, we multiply the total travel time (computed with the earlier defined BPR function) with the Value Of Time (VOT). To compute safety costs we multiply the total traveled distance with a risk factor that depends on the road type that is used, and a monetary value that expresses the amount of money one is willing to pay to avoid one casualty. We refer to this quantity as the Casualty Avoidance Value (CAV). From these assumptions, the following costs functions result:

$$c_r^{\text{TIME}}(q_r) = VOT \cdot t_r q_r = VOT \cdot q_r t_r^0 \left(1 + 0.15\left(\frac{q_r}{C_r}\right)^4\right) \tag{7.68}$$

$$c_r^{\text{RISK}}(q_r) = CAV \cdot SL_r \cdot q_r l_r \tag{7.69}$$

The marginal cost function corresponding to (7.67). Hence equals:

$$\begin{aligned} & c_r^{\text{TOTAL,marginal}}(q_r) = c_r^{\text{TIME,marginal}}(q_r) + c_r^{\text{RISK,marginal}}(q_r) \\ & = VOT \cdot t_r^0 \left(1 + 0.75\left(\frac{q_r}{C_r}\right)^4\right) + q_r VOT \cdot t_r^0 \left(\frac{4}{C_r} 0.75\left(\frac{q_r}{C_r}\right)^3\right) \\ & + CAV \cdot SL_r \cdot q_r \cdot l_r + q_r \cdot CAV \cdot SL_r \cdot l_r \\ & = VOT \cdot t_r^0 \left(1 + 5 \cdot 0.75\left(\frac{q_r}{C_r}\right)^4\right) + 2CAV \cdot SL_r \cdot q_r \cdot l_r \end{aligned} \tag{7.70}$$

The following shows the coefficients that were used in this example.

	Route 1	Route 2
t_r^0	9 min.	10 min.
C_r	2000 veh./hr.	1000 veh./hr
SL_r	$0.5*10^{-8}$ cas./km	$6.0*10^{-8}$ cas./km
l_r	10 km	10 km
VOT	10 fl./km	
CAV	$5*10^6$ fl./cas.	

System optimum assignment
The resulting marginal costs are plotted in the figure below:

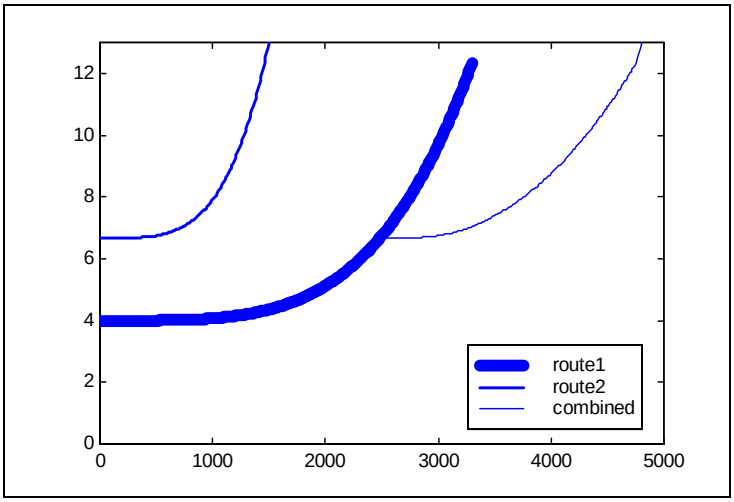


Figure 7.19: *Marginal system costs for route 1, route 2 and the combination of both, plotted against travel demand*

From this figure it can be concluded that it is not desirable to divert traffic over route 2, unless the total travel demand exceeds 2700 veh./hr.

User optimum assignment
If no further action is taken, drivers will select their optimal routes. Assuming that they will do this by minimizing their travel time, the user equilibrium can be determined from analyzing the user cost functions. The following figure shows these user costs:

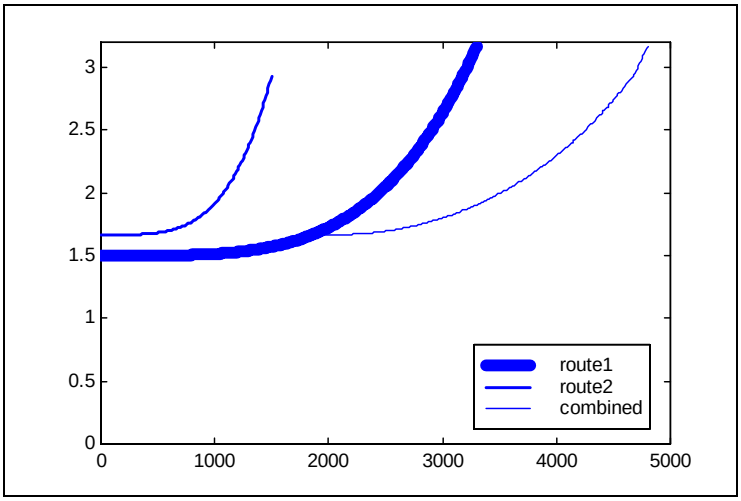


Figure 7.20: *User costs for route 1, route 2 and the combination of both, plotted against travel demand*

From this figure it can be seen that if total travel demand exceeds 2000 veh./hr, traffic will start 'rat running' via route 2.

This example shows once again that user optimal solutions need not be system optimal. If the road operator in this example wants to achieve a system optimal usage of the network, some action is needed if travel demand exceeds 2000 veh./hr.

- end of example -

7.6 Stochastic user-equilibrium assignment

The stochastic user-equilibrium (SUE) assignment model is a combination of the stochastic assignment model and the deterministic user-equilibrium assignment model; it takes both the behavioral differences between travelers and congestion effects into account.

The stochastic user-equilibrium assignment can be defined as follows.

Stochastic user-equilibrium assignment:
The assignment which leads to an equilibrium in which no traveler can improve his *perceived* travel time by unilaterally changing routes.

Because of the variations in traveler perceptions, the SUE assignment results in a spread of each OD flow over multiple routes. The route choice proportions depend on the travel times experienced. The travel times in turn depend on the link flows, while the link flows depend again on the route choice proportions. The system is said to be in SUE if route choice proportions and travel times are consistent.

SUE assignments produce more realistic results than the deterministic UE model, because SUE permits use of less attractive as well as the most attractive routes. Less attractive routes will have lower utilization but will not have zero flow as they do under UE.

7.6.1 Mathematical description of stochastic user-equilibrium

The stochastic user-equilibrium can be characterized by the following:

Assumptions:
$t_a = t_a(q_a)$ (travel time depends on the flow)
$\Theta > 0$ (flow diversion among routes)

In contrast to the AON, UE and stochastic assignment, the definitional constraints in the SUE reflect all dependencies formulated in the general network assignment problem (see Section 7.2):

(1) $$t_a = t_a(q_a) \tag{7.71}$$

(2) $$\beta_{ijr} = \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) \tag{7.72}$$

(3) $$t_{ijr} = \sum_a \alpha_{ijr}^a t_a(q_a) \tag{7.73}$$

(4) $$T_{ijr} = \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) T_{ij} \tag{7.74}$$

(5) $$q_a = \sum_i \sum_j \sum_r \alpha_{ijr}^a \beta_{ijr}(\underline{\Theta}, \underline{t}_a(\underline{q}_a)) T_{ij} \tag{7.75}$$

7.6.2 Solving the stochastic user-equilibrium assignment problem

The algorithm for solving the SUE is more or less identical to the algorithm for solving the deterministic UE. Instead of having deterministic link travel times, they are drawn from a travel time probability distribution as in the case of the stochastic assignment. SUE is computed in TRANSCAD using the Method of Successive Averages (MSA), which is the only known convergent method. Due to the nature of this method, a large number of iterations should be used.

7.7 Multi user-class traffic assignment

Until now we have not explicitly dealt with the distinction between different groups of road users when assigning traffic to a network. Yet there may be a strong practical need to do so, for example when a study is designed to answer questions with the following dimension:

- what is the impact of imposing tolls, assuming that drivers respond differently to these tolls?
- what is the effect of restricting the use of lanes to special groups of vehicles (e.g. freight, high occupancy vehicles, commercial vehicles)?
- what is the effect of equipping part of the vehicles with in car route guidance equipment?

For these and similar questions one would like to make use of multi user-class assignment models.

A multiclass model differs from a single class assignment model in one or more of the following ways:

- The modeling of route choice. In a multiclass assignment different assumptions on route choice may be used for different groups of travelers
- The modeling of supply demand interaction. In multiclass assignment one may use multiclass link cost functions, e.g. functions that depend on both the volume of trucks and the volume of cars, rather than the total number of vehicles
- Differences in driving speed. This only applies to dynamic models which model time-space trajectories of (groups of) vehicles explicitly. If different groups of vehicles drive at significantly different speeds, this may have an impact on the assignment results.

Because this course does not involve dynamic models we will only discuss the first two possibilities.

If only route choice differs over user classes, a single user-class assignment model can be converted into a multi user-class assignment with a few trivial modifications. For example multi user-class all-or-nothing (MUC-AON) assignment or stochastic multi user-class (SMUC) assignment is obtained by superposing the assignments for the different user-classes. Likewise MUC-DUE and MUC-SUE assignments can be computed using the MUC-AON as a component in an iterative procedure.

As an exercise, formulate the minimization problem that corresponds to a MUC-DUE assignment and outline an algorithm that can be used to compute this assignment. Only consider differences in route choice behavior over user-classes.

If also differences in supply demand interactions over user-classes need to be considered a more complex problem arises. However, it is possible to convert an assignment problem with MUC supply-demand interaction into a problem with only MUC route choice behavior. This is done by replacing each link in the network to which MUC supply demand interaction applies by two (parallel) links, each of which is accessible to only one user class. In this case,

the link cost functions that apply to these links are no longer *separable*: the link costs on one link depend on the link volumes on both links. This is illustrated in the Figure 7.21.

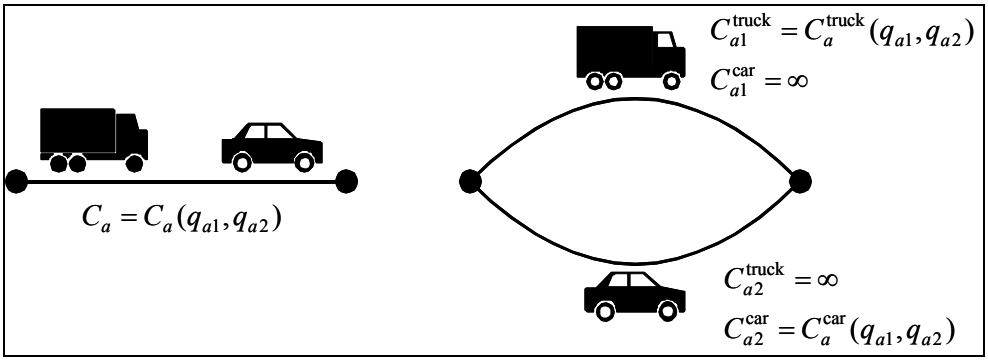


Figure 7.21: *Replacing a link with MUC supply demand interaction by two parallel links with non separable link cost function*

Once the problem is converted into a form in which only MUC-route choice applies it may be solved using the techniques discussed earlier in this chapter. However one should note that the separability of link cost functions is an assumption that was used to proof that unique solutions exist to the DUE and SUE assignment problems. This means that multiple route choice patterns may satisfy the requirements for MUC equilibria. As a consequence, the algorithms used to compute MUC equilibria need not always converge.

7.8 Assignment to public transit networks

7.8.1 Introduction

Until now we have silently assumed that the assignment problems we have dealt with apply to cars. However a large class of assignment problems deals with assigning OD-demand to public transit networks, i.e. buses and train services. For Public Transit (PT) assignment, many of the techniques can be used that were earlier applied to the assignment of car traffic. In this section we will analyze which particular differences between car and passenger assignment should be dealt with, before the techniques described in this chapter can be utilized for public transit assignment.

As a basis for this analysis we will distinguish the two behavioral components of traffic assignment:
- modeling the interaction between traffic volume and level of service
- modeling the route choice proportions

Interaction between traffic volume and level of service
As opposed to road traffic, congestion usually is not an issue in public transit. Noted exceptions are assignment to metro systems in large cities (e.g. London Underground) and some applications of modeling pedestrian assignment in train stations.

Route choice proportions
Route choice is usually modeled as a result of utility maximization. In *car traffic*, travel time and travel costs are generally considered to be the decisive attributes for route choice. These route attributes can be conveniently derived from link attributes by *adding* over all links that constitute a route. For this reason route choice proportions can be computed with the aid of efficient shortest path algorithms.

In *public transit* (PT) assignment route choice is also modeled as a result of utility maximization. However, the attributes travel time and travel costs only would give a poor explanation of the route choice behavior. First of all, a PT trip involves different types of actions/periods like the 'hidden' waiting time at home, the (walking) trip to a PT access point, a trip by bus to the train station, etc. Hence, many route attributes, like costs, frequency, waiting time, in vehicle time, etc., influence route choice. Although most of these attributes can be translated into time elements, not all of these time elements are appreciated equally by travelers as is shown in the example below.

Also, the way in which routes are represented differs from traditional assignment.
The different types of actions/periods in a PT trip mentioned above may be thought of as separate links, e.g. access links, transfer links, in vehicle links, etc. So not the physical network, but rather the network of services is referred to in the PT assignment models.

Example 7.7: Weighing time elements of transit routes
Van der Waard [1988] has investigated the impact of the various route attributes on route choice proportions of public transit users. This was done by estimating the coefficients of each attribute in the utility function of a multinomial logit-model. The following utility function was used:

$$\text{MODEL A: } U = a_1 T_1 + a_2 T_2 + a_3 T_3 + a_4 T_4 + a_5 T_5 + a_6 T_6 + a_7 NRC \tag{7.76}$$

where U denotes the utility of a route, T_i denote the time related attributes, NRC denotes the number of transfers and a_i denotes the coefficients. The following table summarizes the coefficients that were found by van der Waard [1988], based on a sample of 1095 public transit trips. The third column of the table shows the coefficients for MODEL A (see Equation **Error! Reference source not found.**)), while the fourth column describes the coefficients for an even more detailed model, model B.

Route attributes	Symbol	Model A	Model B
Access time (walking time to travel from trip origin to PT access point)	T_1	2.2	2.3
Waiting time at first stop	T_2	1.5	1.4
In-vehicle time (all types)	T_3	1	
In-bus time			1
In-tram time			1
In-rapid transit time			0.9
Egress time (time to travel from PT egress point to destination)	T_4	1.1	1.2
Walking time at interchange	T_5	2.3	2.2
Waiting time at interchange	T_6	1.3	1.2
Number of transfers	NRC	5.7	5.9

Source: Van der Waard (1988)

It should be noted that in this table, route costs are not mentioned. Although costs may be an important factor, it is difficult to observe the costs for an average traveler due to the complexity of PT tariff structures. Moreover, in many cases the route has little or no influence on the costs (e.g. if charges are based on a zone system).

- end of example -

We have now discussed the route choice behavior as a function of *route attributes*. Before this knowledge can be applied to compute choice proportions, two steps must be completed:
- Identifying feasible routes, i.e. determining which routes should be considered in the assignment.
- Computing route attributes

These tasks are closely interrelated. For example, in a shortest path algorithm such as proposed by Dijkstra (1959), see Chapter 3, a route and its total travel costs are determined simultaneously. This is only possible due to the additive structure of the cost function that is adopted. All assignment techniques that are described in this chapter, except for the logit

model, can be implemented without separately identifying feasible routes: they can be based on shortest path computations in one way or the other.

If such an additive cost structure is absent, it is necessary to first enumerate the routes before route choice proportions can be computed. Some of the attributes that are dominant in PT route choice, like frequency, waiting time and costs, lack such an additive property. Therefore from the modeling point of view, the identification of routes for PT networks is more complex then its corresponding problem in ordinary link networks.

After the routes have been identified, route attributes should be computed. This is by no means straightforward. This applies in particular to attributes like frequency, time spent waiting for connections, and costs.

- If two lines with frequencies $f1$ and $f2$ are parallel, they may be replaced with a *hypothetical* line with a frequency f, $\max[f1, f2] \leq f \leq f1 + f_2$. This example is still simple: what happens when two lines are parallel, but are characterized by (slightly) different travel time (e.g. a local train and an intercity)?.
- If two lines with frequencies $f1$ and $f2$ are in sequence, they may be replaced with a *hypothetical* line with a frequency f, $f \leq \min[f1, f_2]$
- If a route involves boarding on different lines, the waiting time at interchanges may be modeled at different levels of detail:
- only taking into account the frequencies, i.e. the waiting time only depends on the frequency of the next line,
- assuming that the two lines are synchronized (e.g. cross platform transfer), the waiting time at a transfer from one line to the synchronized other line is less than the one to be expected based on the frequency of that line,
- using the time table to compute the waiting time (see e.g. the 'Reisplanner' as published annually by the Dutch Railways). This is the most accurate, but also the most data demanding way of computing route attributes.
- The costs of a trip in public transit depends on a complex tariff structure. In order to build an accurate model, multiple classes of travelers should be distinguished depending on how they pay for their rides (examples are single trip, round trip, combined ticket, seasons ticket, zone ticket, etc).

7.8.2 Public transport network representation

For PT networks, there is a significant difference between the administrative coding of networks, and the *computational* coding of networks.

The administrative coding is oriented towards the physical network, and in addition specifies PT lines with their attributes, like vehicle type, traveling speed, stops, etc. Penalties for transferring are often specified globally, i.e. these penalties apply to transfers at all nodes. The administrative coding is aimed at specifying the network in unique and compact manner.

The computational coding is oriented towards the computational processes, like traffic assignment. This may mean that (potential) elements of a route are coded as link segments (see Example 7.8 for a simple example). All information that is not relevant for computations is deleted.

In modern transport planning software like Omnitrans or TransCAD the user is shielded from the conversion from an administrative to a computational coding for greater ease of operation. However, one should be aware of this issue because in specific instances one may be forced to directly supply a specific (part of a) computational network in order to model a specific practical problem, like a new multi-modal way of transport.

Example 7.8: Coding PT networks
Suppose one wants to travel from A to B en may choose between two routes:
a. Traveling with a local train from A to B
b. Traveling with a local train from A to 1, interchanging to an intercity train to 2, and traveling from 2 to B with a local train again.

This problem can be represented as a route choice problem. The interchanges from local to intercity services are then modeled with a hypothetical link, to which a cost is assigned that corresponds with the time it takes to interchange. The shortest route hence satisfies:

$$T_{AB} = T_{A1} + T_{2B} + \min[T_{12}, T_{\underline{11}} + T_{\underline{12}} + T_{\underline{22}}]$$

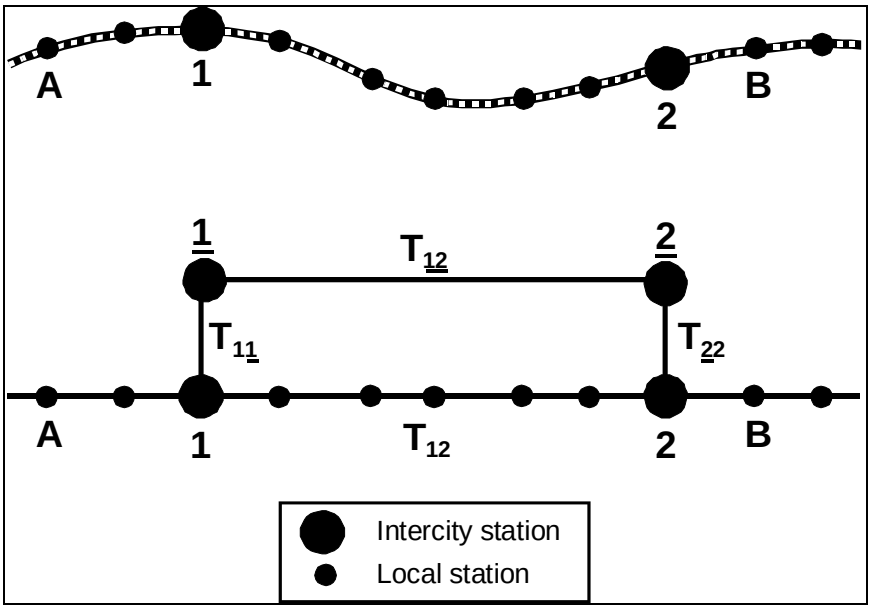


Figure 7.22: *Coding PT lines that use common infrastructures as separate links*

- end of example -

Public transport or transit networks have two characteristics that make modeling public transport more difficult than modeling private transport modes such as car: the time dimension, i.e. frequencies and schedules, and the concept of lines and thus the need for transfers. The adopted network representation is a balance between network size, network complexity and algorithmic complexity. Furthermore, the purpose of the study itself has a significant influence on the way of network modeling. For short-term studies a higher level of detail might be attained and required than for long-term studies. For a study focussing on the modal split a different level of detail may be necessary than for a study on the impact of ITS on vehicle occupancy.

In the past, network size was crucial. Furthermore, there was a strong preference for using shortest path algorithms that were available for modeling private modes, such as Dijkstra (1959), Moore (1959), or even Floyd (1962).

Dial (1967) proposed a network structure in which the public transport network is represented by “trunk line links” (Figure 7.23). These links have as attributes a travel time and the line numbers using the link. The typical public transport modeling problem can be illustrated by the fact that the shortest route to C depends on the final destination. For traveling to C line I might be interesting, while for trips to D line I is the obvious choice. For traveling to E, however, line II is the best choice. He adapted the shortest path algorithm of Moore in order

to account for transfers. The transfer penalty depends on the expected waiting time of the line that will be boarded. The waiting time is usually defined as half the headway.

An alternative approach is presented by Fearnside & Draper (1971), in which line specific nodes and links are introduced (Figure 7.24). This approach requires transfer links between stops related to the same physical location. A problem that is not dealt with, however, is the so-called common lines problem: when a traveler waiting at a stop might use different lines for reaching his destination, he has to decide which line to use. This might be a specific line, yielding the shortest travel time from the stop to the destination, or it might be the first that arrives. In this example, for instance, traveler traveling to A or to C might choose between lines I and II. An approach for dealing with the common lines problem is discussed in the next section.

An interesting approach for representing public transport networks was suggested by Last & Leak (1976). In their model they introduced "direct links" connecting links that might be reached, with or without transfers. Figure 7.25 gives an example of a network with direct links excluding transfers.

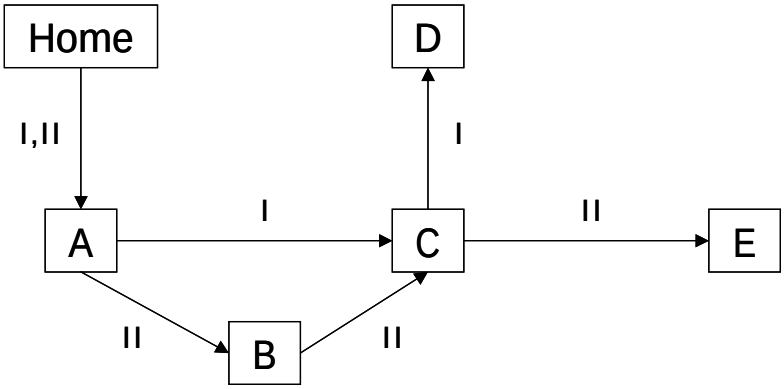


Figure 7.23: *Public transport network using trunk line links consisting of 2 PT lines (I, II)*

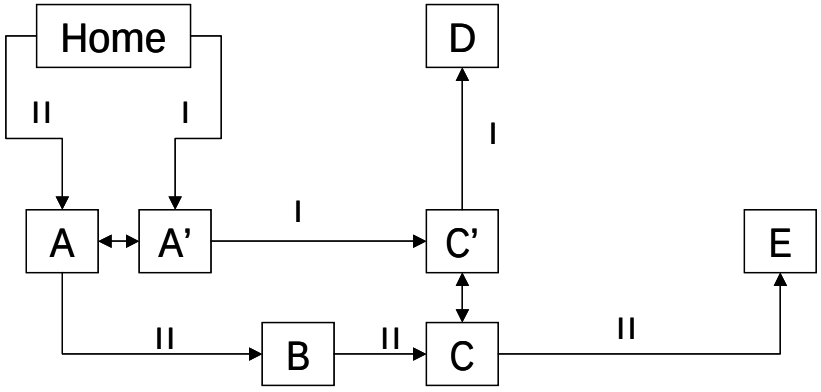


Figure 7.24: *Public transport network with node-specific nodes and links*

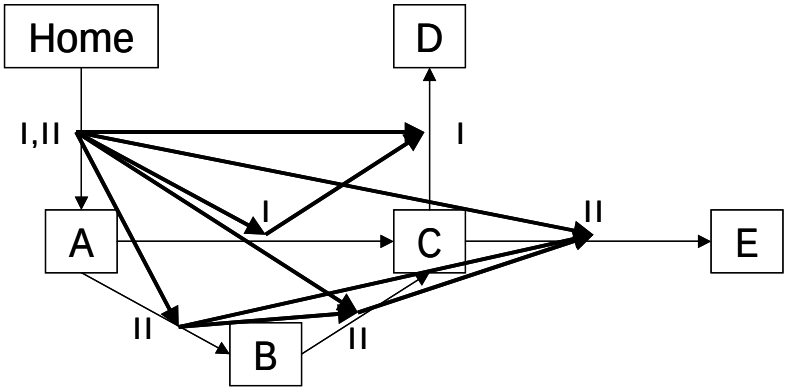


Figure 7.25: *Public transport network with directed links*

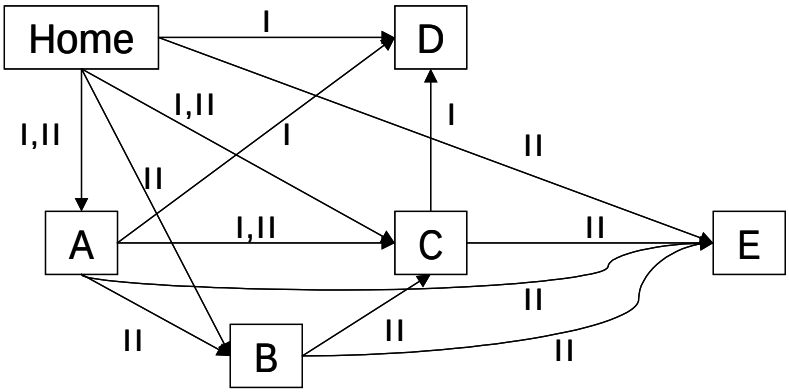


Figure 7.26: *Public transport network with route sections*

It is, of course, possible to use the same philosophy for all stops that are served by the same line, as in the approach suggested by De Cea & Fernandez (1993) (Figure 7.26). In the case of multiple lines serving the same pair of stops, an aggregate direct link might be created. Consequence of this approach is that the network size increases enormously.

All approaches discussed before still deal with lines having frequencies. A more detailed approach is proposed by Nuzzulo & Russo (1994), in which the individual runs are the basic components. They explicitly account for the time dimension, as can be seen in Figure 7.27. The main advantage is that this approach makes it possible to model transfers properly, which is certainly relevant for low-frequency networks.

An issue that has not yet been discussed is capacity. Congestion has become a relevant issue in public transport just as in car traffic. There are various ways to incorporate capacity limitations. In-vehicle time might become less attractive due to crowding or to standing instead of having a seat, which might be represented by a higher subjective in-vehicle time, for instance using kind of BPR-function. Furthermore, vehicles might be delayed, leading to bunched arrivals and thus irregular services and longer waiting times. Ultimately, travelers might not be able to board the vehicle, forcing them to wait until the next vehicle arrives.

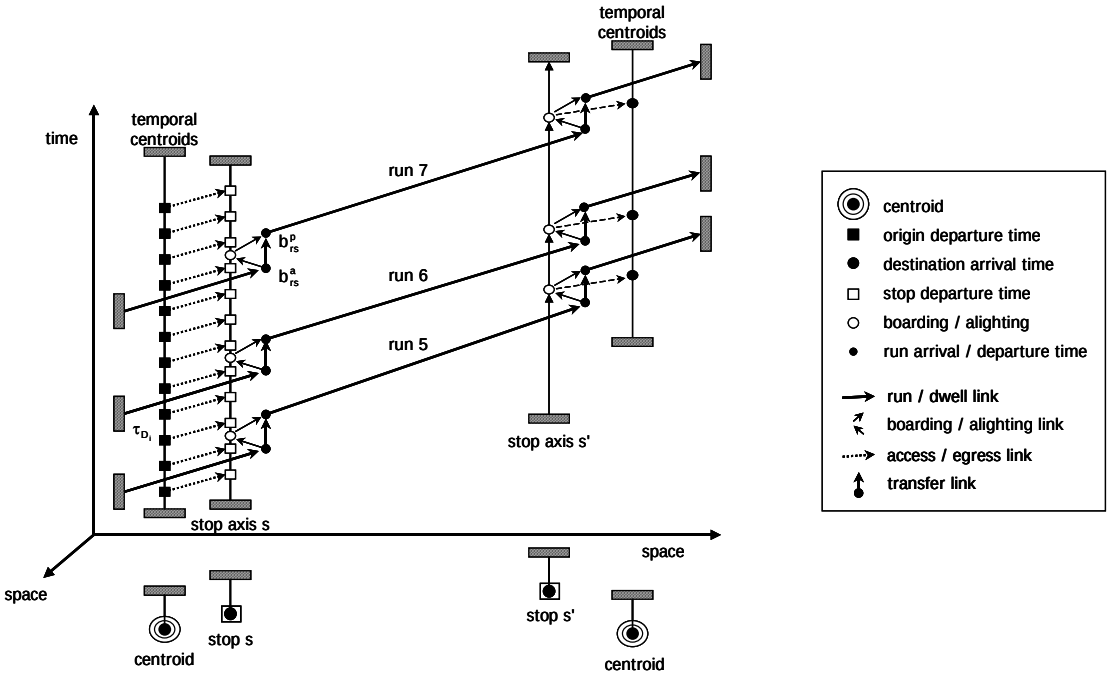


Figure 7.27: *Diachronic run-based representation of public transport services*

7.8.3 Public transport assignment approaches

The earlier models only use shortest path algorithms, which is clearly very simplistic. The first extension is to consider the possibility of parallel lines on the same link, which can also be extended to parallel services between the same stops. The most realistic situation, however, is that multiple paths are considered.

An example of a shortest path algorithm is for instance the adapted Floyd-algorithm. It is based on the route section representation (De Cea & Fernandez (1993)), which is stored in a matrix describing all connections between stops. At the start all connections without transfers are determined: travel time, boarding node (or back node), and frequency (Figure 7.28). Please note that in case of multiple services serving the same pair of stops aggregate values might be used.

	1	2	3	4	5	6
1	X	5 / 1	10 / 1	-	-	-
2	5 / 2	X	5 / 2	5 / 2	10 / 2	-
3	10 / 3	5 / 3	X	-	-	10 / 3
4	-	5 / 4	-	X	5 / 4	-
5	-	10 / 5	-	5 / 5	X	-
6	-	-	10 / 3	-	-	X

Figure 7.28: *Graphical representation of the matrix used in the Floyd algorithm: contents of each cell is travel time and back-node*

The algorithm is straightforward. For all combinations of stops it is checked whether a combination of another stop might results in s shorter travel time, than the current solution. This procedure is repeated for a number of iterations. In formula it would be:

For n = 1 to n_i

  For i = 1 to n_{stops}

    For j = 1 to n_{stops}

      For k = 1 to n_{stops}

        If $Z_{ij} > Z_{ikj} = Z_{ik} + P_k + Z_{jk}$

          $Z_{ij} = Z_{ikj}$

        End if

Where Z is the travel time or generalized costs en P is the transfer penalty. Interesting characteristic of this procedure is that the number of iterations is correlated with the number of transfers per trip. Disadvantage is of course that the algorithm is inefficient for large networks.

An assignment approach that explicitly acknowledges that travelers have multiple routes to reach their destination is presented by Spiess & Florian (1989). In the network shown in Figure 7.29 the shortest path from A to B would be to board line 2 and to transfer to line 3 at node X, that is looking at in-vehicle times only. The alternative approach is that travelers will board the first arriving vehicle that might bring them closer to their destination. Thus travelers might board both lines 1 and 2. Travelers boarding line 2 might transfer at node X, however, transferring at node Y offers better possibilities. Using again the strategy of boarding the first arriving vehicle, leads travelers boarding line 3 and line 4. The net result is a set of paths:

- Line 1
- Line 2 and line 3 via node Y
- Line 2 and line 4 via node Y

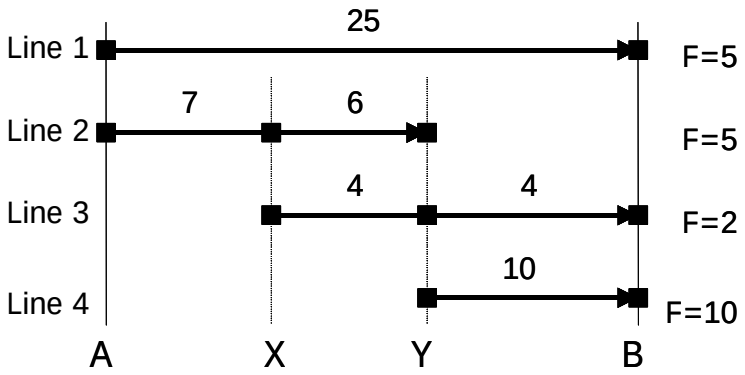


Figure 7.29: *Public transport network example*

Spiess and Florian use a network representation comparable to the Fearnside & Draper approach. Each line has it's own boarding and alighting nodes. Links have two attributes, that is a link travel time (boarding time, in-vehicle time) and a waiting time (related to the frequency). The resulting network can be simplified by eliminating all nodes having only two links. Figure 7.30 and Figure 7.31 give an example of the resulting network structures. Using this representation they apply a backward search (from destination to all origins), following from a linear programming framework, to determine the strategies in a first step while assigning the travelers in a second step (from origin to destination!)

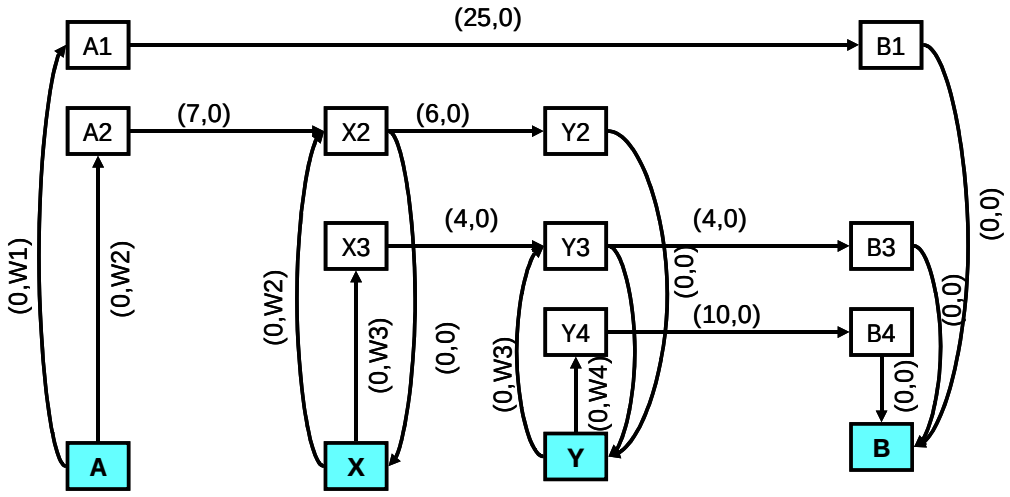


Figure 7.30: *Extended network representation*

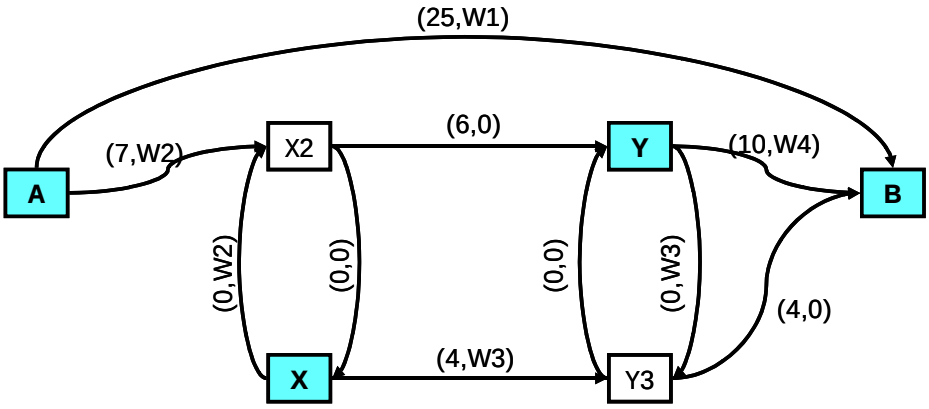


Figure 7.31: *Reduced version of the extended network*

This set of paths or strategies is also called a hyperpath. A hyperpath means that different routes are available, and that the attractiveness of these routes varies over time. The actual route choice thus depends on the moment the travelers arrives at the start of the hyperpath. In the approach of Spiess and Florian this concept is translated in 'boarding the first vehicle that arrives', that assigning the travelers according to the frequencies of the lines departing from the stop. Please note, that this assumption ignores all other path characteristics such as in-vehicle time and number of transfers.

Of course it is also possible to use more elaborate choice models for assigning the travelers to a set of paths. This might be done by considering each path individually or by a system of consecutive choices. The latter approach can be illustrated using the network in Figure 7.32.

The first question is to determine which paths are included in the choice set. A possible approach is to exclude those paths that do not lead to a further reduction of the overall travel time between origin and destination, for instance using a logsum (in this case the log of the denominator of the route choice logit model, that is the sum of the exponential of the utilities for all route alternatives). Next, a set of sequential choices can be defined:

- Which stop to access?
- Which line to board?
- Which stop to alight?

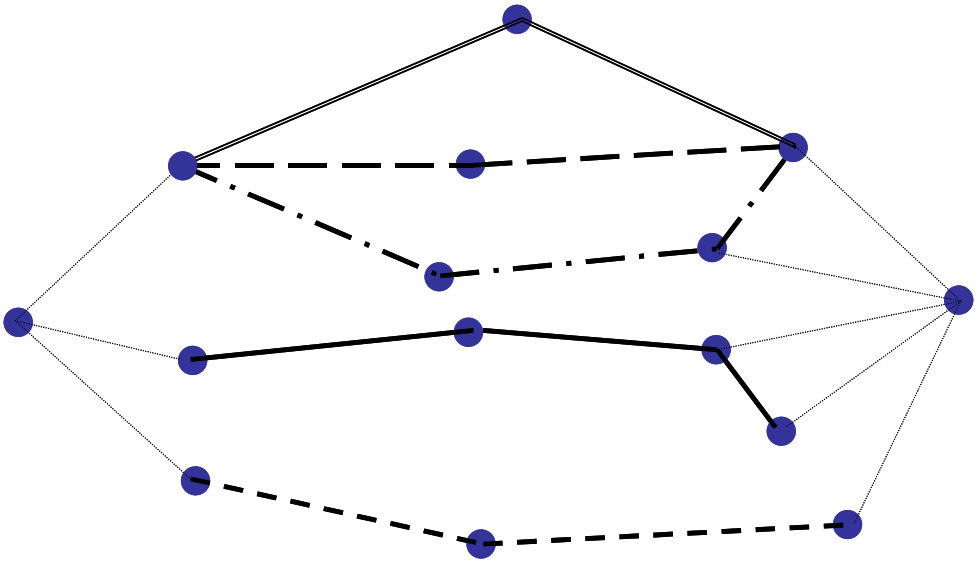


Figure 7.32: *Network example illustrating various route choice problems: relevant routes, boarding stop, available public transport services, alighting stops*

For each option a utility can be defined consisting of the characteristics of the route or routes related to the specific option. Here, again, a logsum approach might be used to model the characteristics of the remaining route options. A specific situation might be relevant when travelers also choose between transport techniques, for instance that travelers first choose between train and bus before considering the number of train and bus lines available. In that case, an additional choice model should be included.

Final remarks

- Public Transit-assignment can not be seen separately from trip distribution and mode choice. The OD-demand for public transport highly depends on the quality that is offered. If the quality is poor, travelers will divert to other modes or will select different destinations. Working with a fixed OD-table that is directly derived from socio-economic data, irrespective of the other modes, therefore leads to poor results.
- In the traditional four stage approach PT-assignment is a step that is preceded by trip-generation, trip distribution and modal split. The result of the modal split step is a matrix that is assumed to travel by public transport, i.e. using PT as *main* mode. However, to realistically assign these trips to the network, one should also model the access en egress part of the routes. This part of the route may be traveled by foot, PT (buses), bicycle or even cars. For greater realism one would want to specify the network for these modes in the vicinity of the PT access and egress points. However, this would lead to a new problem: if the network for other modes is specified, a part of the PT OD-demand as predicted by the modal split module might entirely be routed over this network and not use public transport at all, which is a contradiction with the earlier assumption.
- Relative to other modes, the volume of travelers using public transit on specific links of the network is only small. For planning purposes however, only a small relative error is allowed, say a standard deviation that amounts of 20% of the true volumes is acceptable. To predict PT volumes with such an accuracy is much more difficult than to predict the much higher car volumes with this accuracy. The ratio for this is the following: Many practical phenomena like noise terms or errors display a variation that is proportional to the quantities to which they are related. Therefore, in general (unbiased) estimates display a standard deviation that is proportional to the root of the quantity that is estimated. In these cases the relative standard deviation of an estimate of flow volume decreases if the actual flow volume increases.

7.9 Elasticity of travel demand

Until now we have assumed that the total travel demand that is to be assigned is fixed. However, in practice the total OD-demand is influenced by the quality that is jointly offered by the available routes. If this quality improves for a specific OD pair, the total OD demand will increase due to:

- diversion of other travel modes
- relocation of activities
- rescheduling of activities
- generation of new activities

The extra demand that is realized when a transport system is improved is known as *latent* demand. If OD-demand is considered as a consumer good, the *price-elasticity* is defined as the proportional change in travel demand that occurs as a result of one percent change in the travel costs. If the demand for travel at cost level c is given by $q(c)$ the price-elasticity is equal to $(\mathrm{d}q(c)/\mathrm{d}c)\cdot(c/q(c))$.

The demand curve can be plotted in the same figure as the cost functions, see e.g. Figure 7.33. The equilibrium between travel demand and supply can be determined from the intersection between the supply curve and the demand curve.
One of the effects of taking into account latent demand is that computing the benefits of new infrastructure is less straightforward then it used to be, because two effects need to be taken into account simultaneously:

1. The travel costs change as a result of the new infrastructure
2. The travel demand changes as a result of the change in travel costs.

If the total travel would have remained constant the benefits simply could have been computed by multiplying the change in route costs with the total travel demand. Because this is not the case, the analysis of the benefits of new infrastructure is more complex.

Example 7.9: Consumer surplus
Consider the example in Figure 7.33 Which are the total benefits from adding route B?
When only the existing route is available the equilibrium between supply and demand is at point 1. When a second route is added (cost function B), the cost function of the combination of route A and B can be constructed in the way as explained in Example 7.2. The new equilibrium between supply and demand now is at point 2 (try determining yourself which part of this total demand will use route A, and which part will use route B).

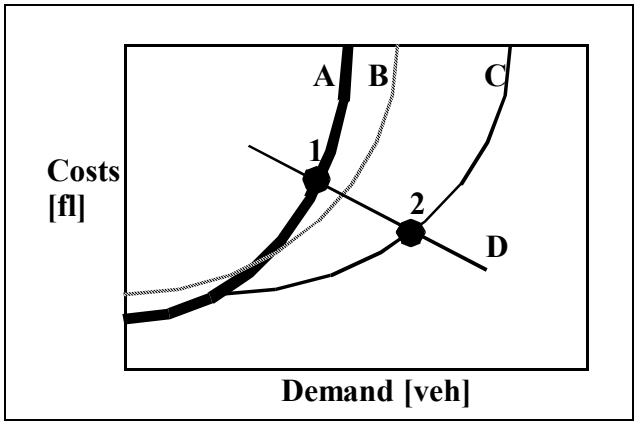


Figure 7.33: *Equilibrium between supply and demand, with:*

A: cost function
D: demand function
1: supply-demand

B: cost function for new alternative route
C: cost function for combined route system A+B
2: supply-demand equilibrium for new situation (route A and route B available)

To determine the benefit of adding the new route we need to introduce the notion *consumer surplus*.

At a given price c, $q(c)$ people are willing to travel. However some of these travelers would have been prepared to pay more then c. The difference between the total yields that would have been achieved if each traveler was charged at the maximum level that he or she is prepared to pay and the actual total amount that is charged to travelers is referred to as the *consumer surplus*.

Another way of viewing it is the following. If we interpret the difference between the value a traveler is *willing* to pay for the use of a particular service and the value a driver *has to* pay for that service as *profit*, the consumer surplus corresponds with the total economic value a particular service represents for all customers.

Given a demand function $q(c)$ and a price level $\tilde{c}$ the maximum total yields (MTY) can be computed as follows (see Figure 7.34):

$$\begin{aligned} MTY(\tilde{c}) &= \lim_{\Delta \downarrow 0} \sum_{k=0}^{\infty} (q(\tilde{c} + k\Delta) - q(\tilde{c} + (k+1)\Delta))(k + \frac{1}{2})\Delta \\ &= \lim_{\Delta \downarrow 0} \sum_{k=0}^{\infty} \frac{q(\tilde{c} + k\Delta) - q(\tilde{c} + (k+1)\Delta)}{\Delta}(k + \frac{1}{2})\Delta \cdot \Delta \\ &= -\int_{\tilde{c}}^{\infty} q'(c)c\mathrm{d}c = q(\tilde{c})\tilde{c} + \int_{\tilde{c}}^{\infty} q(c)\mathrm{d}c \end{aligned}$$

The consumer surplus (CS) hence is given by:

$$CS(\tilde{c}) = MTY(\tilde{c}) - q(\tilde{c})\tilde{c} = \int_{\tilde{c}}^{\infty} q(c)\mathrm{d}c$$

The societal benefits of introducing route B equal the increase of the consumer surplus:
$CS(\tilde{c}_1) - CS(\tilde{c}_2)$

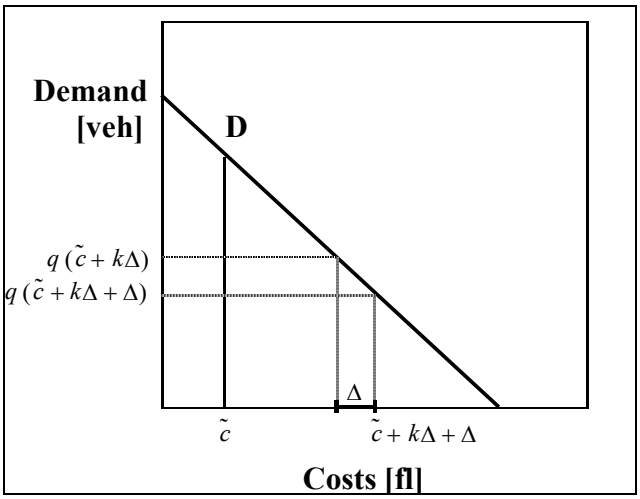


Figure 7.34: *Travel demand, plotted against costs*

Question: Can you graphically illustrate the increase of consumer surplus in Figure 7.33?

- end of example -

7.10 Some paradoxal examples of traffic assignment

Traffic assignment is aimed at simulating the complex reality of travel demand interacting with infrastructure supply. Sometimes, this leads to paradoxes: results that are at first sight unexpected, but have a logical explanation. This section discusses a number of famous paradoxes. The examples were derived from an article by [Arnott and Small, 1994].

The traffic we see does not represent the full demand for peak travel at the prevailing generalized costs, since congestion itself causes many potential rush-hour trips to be cancelled, diverted (for example, to mass transit, to car pools and to less-congested routes and destinations) or rescheduled. Any reduction in congestion resulting from capacity expansion encourages others to drive during hours or on routes they ordinarily would not use. So measures to relieve congestion may at least be partially undone by latent demand.
The other reason capacity expansion alone does not work is that the use of infrastructure during congested periods is mispriced. Because drivers do not pay for the time loss they impose on others by their presence during congestion, they make socially inefficient choices concerning how much to travel and what route to take.
The combination of latent demand and mispriced congestion may be so perverse that an expansion of capacity brings about no change in congestion, or even makes it worse, as the following examples will illustrate.

Example 7.10: Pigou-Knight-Downs paradox
Consider the two route system in Figure 7.35, and assume the cost- and demand functions:

$T_1 = 10 + 10\ F_1/C_1$, $T_2 = 15$, $F_1 + F_2 = 1000$

Scenario A: Suppose $C_1 < 2000$
And $T_1 = 10 + 10\ F_1/C_1 = 15 = T_2$
Then: $F_1 = ½\ C_1$, $T_1 = T_2 = 15$

Scenario B: When $C_1 > 2000$
$F_1 = 1000$, $F_2 = 0$, $T_1 = 10 + 10000/C_1$

Example:
Suppose $C_1 = 2500$
Then: $F_1 = 1000$, $F_2 = 0$, $T_1 = 10 + 10000/2500=14$
Everyone uses the bridge.

Increasing the bridge capacity to any value less than twice the traffic flow has no effect on travel time (T_1). Suppose that route 1, the route over the bottleneck (e.g. a bridge), takes 10 minutes with no traffic, but travel time rises linearly with the ratio of traffic flow (F_1) to bridge capacity (C_1). Route 2 always takes 15 minutes (T_2). There are 1000 travelers faced with the choice of route 1 or route 2. In scenario A, the bridge's capacity is defined as less than 2000. The traffic flow over the bridge adjusts to ½ C_1 so that travel time on routes 1 and 2 are equal at 15 minutes. In scenario B, the bridge capacity is increased to exceed 2000. In that case, everyone uses the route with the bridge, but the travel time decreases, as can be seen in the example where bridge capacity equals 2500.

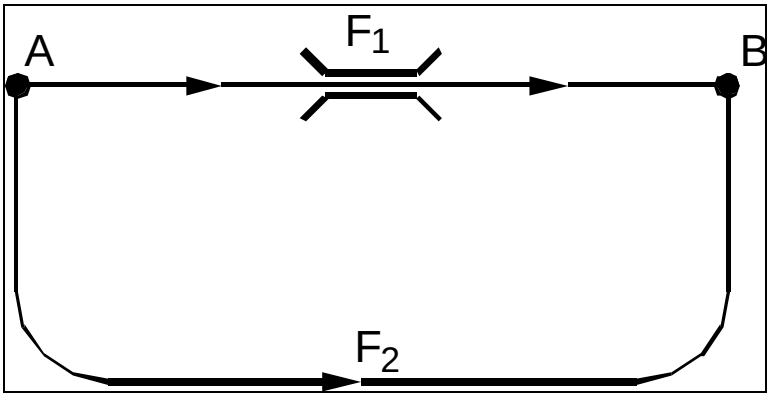


Figure 7.35: *Expanding road capacity creates its own demand, a phenomenon known as the Pigou-Knight-Downs paradox. Because route 1, over the bridge, is the most direct route from point A to point B, more people want to use it, and the resulting congestion makes route 1 take as long as the more circuitous route 2. Travel time on each route is 15 minutes. Expanding the capacity of the bridge over route 1 only attracts more users, and the travel time remains unchanged. The paradox disappears only if the bridge capacity exceeds twice the total travel flow.*

- end of example -

The above example is known as the Pigou-Knight-Downs paradox. The paradox occurs when total travel demand is 1000 and the bridge capacity, C_1, is less than 2000. In this case, travelers divide themselves across the two routes, such that travel time on each route is 15 minutes, which implies that traffic flow over the bridge is exactly half its capacity. Therefore, expanding the bridge's capacity to anywhere in the range from 0 to 2000 has absolutely no effect on anyone's travel time. Instead, it diverts more people from the route with spare capacity to the route crossing the bridge. In other words, the new bridge capacity generates its own demand.
Attempts to reduce congestion on the bridge by instead encouraging car-pooling, expanding mass transit or improving telecommunication facilities would likewise be frustrated unless total vehicular traffic were reduced to below half of the bridge's capacity. So long as any traffic remained on the second route, latent demand for the bridge would undermine these attempts to relieve its congestion.
The crux of the paradox lies in the distinction between the private and the marginal costs (also referred to as social costs) of a trip. The private cost is the cost the driver incurs. The marginal cost equals the private cost plus the external cost, which is the cost the driver imposes on other drivers by slowing them down. In the example, the social cost of traveling on the bridge exceeds the private cost because it is congested. Typically, drivers choose the route with the lower cost to them-the lower private cost. This results in an equilibrium in which private costs on the two routes are equalized. If, instead, drivers where distributed across the two routes as to equalize the social cost, the paradox would disappear; bridge expansion would relieve congestion. This suggests that conventional policies to relieve congestion would work better if each driver faced the social cost of his or her trip.

Example 7.11: Downs-Thomson paradox
Consider the two route system in Figure 7.36, and assume the cost- and demand functions:

$T_1 = 10 + 10\,F_1/C_1,\ T_2 = 20 - F_2/300\ ,\ F_1 + F_2 = 1000$

Scenario A: When $C_1 < 1000$
$T_1 = 10 + 10\,F_1/C_1 = 20 - (1000 - F_1)/300$
So: $F_1 = C_1/(1.5 - C_1/2000)$, $T_1 = 10 + 10/(1.5 - C_1/2000) = T_2$

Examples of equilibrium solutions:

If $C_1 = 250$ then $F_1 = 182$ and $T_1 = T_2 = 17.27$
If $C_1 = 750$ then $F_1 = 667$ and $T_1 = T_2 = 18.89$
If $C_1 \rightarrow 1000$ then $F_1 \rightarrow 1000$ and $T_1 = T_2 \rightarrow 20$

Scenario B: When $C_1 > 1000$
$F_1 = 1000$, $F_2 = 0$, $T_1 = 10 + 10000/C_1$

Example: Suppose $C_1 = 2000$. Then $T_1 = 15$

The Downs-Thomson paradox, expressed mathematically, shows how increasing road capacity in the situation in Figure 7.36 actually raises travel time, as long as the road capacity (C_1) is smaller than the number of travelers. Suppose that the equation for travel time by the congested highway route (T_1) is the same as in the previous example; that the maximum travel time by train (T_2) is 20 minutes; and that 10 minutes will be cut from the train trip for every 3000 travelers. Since there are 1000 total travelers, the number using the road (F_1), plus the number using the train (F_2) will equal 1000. In scenario A, at equilibrium, some of the travelers use the road, and others use the train. Under these conditions, as the road capacity approaches 1000, travel time for each route, T_1, and T_2, approaches 20 minutes. In scenario B, the road capacity is expanded to exceed 1000. All of the travelers use the road, so that the traffic flow over the road is 1000, whereas the traffic on the train is zero. Expanding the capacity of the road to 2000, for example, lowers travel time to 15 minutes.

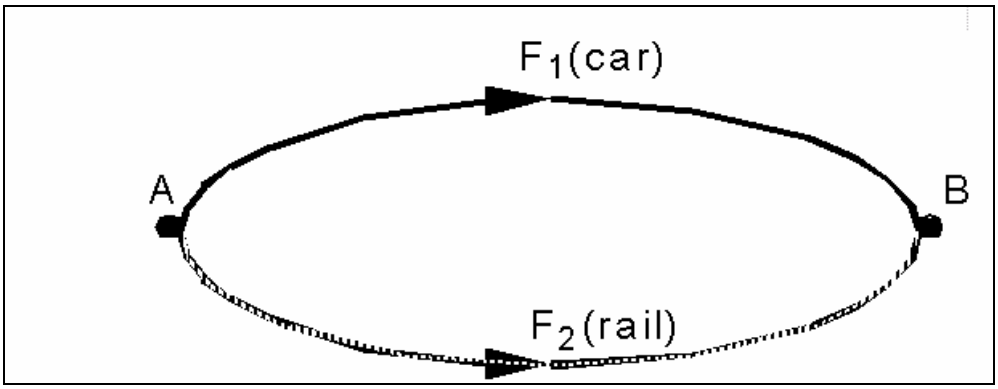


Figure 7.36: *Increased capacity leads to more, rather than less, congestion in the Downs-Thomson paradox. Here the second route, a passenger train, shows increasing returns with added flow because service quality improves as more travelers use it. Expanding road capacity draw people off the train, worsening train service. Equilibrium between the two routes dictates that road travel becomes worse as well, such that increasing road capacity actually increases travel time on both routes*

- end of example -

The above paradox, referred to as the Downs-Thomson paradox, is like the Pigou-Knight-Downs paradox, except the alternative to taking the congested route 1 is now a privately operated train line. The train operator breaks even financially by ensuring that all of the train cars are full. If more people take the train, then trains run more frequently, saving people some waiting time at the stations. In this case, let us say that the maximum travel time by train is 20 minutes, and that 10 minutes will be cut from the trip for every 3000 travelers. The intriguing feature of this situation is that now travel time increases with any increase in bridge capacity within the range from 0 to 1000. The reason is that, just as in the earlier example, capacity expansion diverts people to the congested road. But now the diversion causes train service to get worse, so equilibrium can occur only when congestion is worse also. Here, new capacity generates more than its own demand.
The reason this paradox is even more perverse than the previous one is that there is not only an external cost imposed by each automobile user, as before, but there is now an external benefit created by each user of the train as well. This is because using the train causes the

frequency of service to increase and hence reduces other users' waiting times. This is a technological property of all types of mass transit, including bus and even taxicab service. The same perverse result can be obtained if instead of expanding the road, well-intentioned planners entice some fraction of travelers away from both routes by providing some third alternative such as subsidized van-pools, telecommuting centres or even a new train service. If, in the example above, the bridge capacity is expanded to equal or exceed 1000, a very different thing happens. Capacity exceeds demand, and everyone starts using the road. The number of train users, F_2, drops to zero. Now, further increasing the bridge capacity does reduce the travel time. For example, increasing capacity to 1500 decreases the travel time to 16.67 minutes, the same time a train trip would take if all commuters traveled by train. Further increasing the road capacity lowers road travel time more, so the paradox disappears.

Example 7.12: Braess Paradox
Consider the system shown in Figure 7.37 with three routes from A to B. Compare the scenario that the diagonal link is not present, i.e. $F_3 = 0$ (scenario A) with the scenario that the diagonal link is present (scenario B). Assume the following cost functions:

Traffic on bridge A: $F_A = F_1 + F_3$, $T_A = F_A/100$

Traffic on bridge B: $F_B = F_2 + F_3$, $T_B = F_B/100$

$T_1 = 15 + T_A$, $T_2 = 15 + T_B$, $T_3 = 7.5 + T_A + T_B$, $F_1 + F_2 + F_3 = 1000$

Scenario A: Equilibrium with no diagonal link
$F_1 + F_2 = 1000$, $F_3 = 0$
$T_1 = T_2 = 15 + F_1/100 = 15 + (1000 - F_1)/100$
So: $F_1 = F_2 = 500$, $T_1 = T_2 = 20$

Scenario B: Equilibrium with diagonal link
$F_1 + F_2 + F_3 = 1000$
$T_1 = T_2 = T_3 = 15 + (F_1 + F_3)/100 = 15 + (F_2 + F_3)/100 = 7.5 + (F_1 + F_2 + 2F_3)/100$
So: $F_1 = F_2 = 250$, $F_3 = 500$. $T_1 = T_2 = T_3 = 22.5$

The example shows how travel time increases when the diagonal link is added. Both bridges are the congested points. (The time it takes to travel bridge A is expressed as T_1, bridge B as T_2; traffic flow on the bridges is expressed as F_1 and F_2) In scenario A, before the additional link, equilibrium is reached when the total time (T_1) to travel on route 1, over bridge A, is equal to the time (T_2) to travel over route 2, which uses bridge B. Under these circumstances, the traffic flow on route 1 (F_1) and the traffic flow on route 2 (F_2) are each equal to 500 (half of the 1000 travelers); the total travel time on each route is 20 minutes. In scenario B, the diagonal link has been added, so travelers have the choice of taking this additional route, which we call route 3. Route 3 takes traffic over both bridges A and B, so the bridges become even more congested than before. Equilibrium is reached when travel times on all three routes, T_1, T_2 and T_3, are equal. When this happens, traffic flow over route 3 (F_3) is 500 vehicles, and travel time on all three routes is 22.5 minutes.

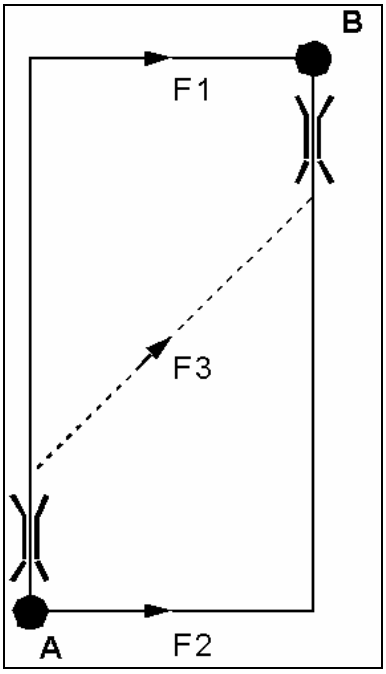


Figure 7.37: *Braess paradox example network; Adding the diagonal (dotted) link leads to an increase of the UE travel time*

- end of example -

The above example is known as the Braess paradox, named for a German operations researcher who in 1968 described an abstract road network in which adding a few links causes total travel time to increase. The paradox is explained by congestion externalities on the bridges; that is, because each traveler ignores the external cost he or she imposes by crossing a bridge, too many people choose the route 3, which crosses both bridges. The faster the diagonal link, the more people are enticed to take it, and the worse is their trip. If diagonal-traversal time were only 5 minutes, all 1000 would choose that route, and travel time would rise to 25 minutes. Only if the diagonal link speed were infinite would equilibrium travel time return to its original 20 minutes.

7.11 References

R. Arnott and K. Small
The economics of traffic congestion, American Scientist, Vol 82, 1994

M.J. Beckmann, C.B. McGuire & C.B. Winsten
Studies in economics of transportation
Yale University Press, New Haven, 1956

M.G.H. Bell
Capacity constrained transit assignment models and reliability analysis, In: Lam W.H.K. & M.G.H. Bell, *Advanced modeling for transit operations and service planning*, Pergamon, Amsterdam, 2003

M.G.H. Bell, Y. Iida
Transportation Network Analysis
Wiley, 1997

P.H.L. Bovy
Toedeling van verkeer in congestievrije netwerken
Rotterdam, RWS, 1990

P.H.L. Bovy & E. Stern
Route choice: wayfinding in transport networks
Dordrecht, Kluwer Academic Publishers, 1990

C.F. Daganzo and Y. Sheffi
On stochastic models of traffic assignment. *Transportation Science* 11, 253-274

K.B. Davidson
A flow travel-time relationship for use in transportation planning, *Proc 3rd Australasian Transport Research Forum* 21 (2) pp.599-616, 1966

J. De Cea and E. Fernández
Transit assignment for congested public transport system: An equilibrium model, *Transportation Science* 27(2), pp. 133-147, 1993

R.B. Dial
Transit pathfinder algorithm, *Highway Research Record* 205, pp. 83-111, 1967

E.W. Dijkstra
A note on two problems in connection with graphs, *Numer. Math.* 1, pp. 269-271, 1959

K. Fearnside and D.P. Draper
Public transport assignment - a new approach, *Traffic Engineering & Control* 12, pp. 298-299, 1971

R.W. Floyd
Algorithm 97, Shortest Path, *Communications of the Association of Computing Machinery* 5, pp. 345, 1962

A. Last and S.E. Leak
TRANSEPT: A bus model, *Traffic Engineering & Control* 17, pp. 14-20, 1976

M.J. Lighthill and G.B. Whitham
On kinematic waves, II. A theory of traffic flow on roads. *Proc of the Royal Society*, 229A, 317-345, 1955

E.F. Moore
The shortest path through a maze, *Proceedings of International Symposium on the Theory of Switching*, Harvard University, Cambridge, 1959

A. Nuzzolo
Transit path choice ans assignment approaches, In: Lam W.H.K. & M.G.H. Bell, *Advanced modeling for transit operations and service planning*, Pergamon, Amsterdam, 2003

A. Nuzzolo and F. Russo
An equilibrium assignment model for intercity transit networks, *Proceedings of the TRISTAN III Conference*, San Juan, Puerto Rico, 1994

M. Patriksson
The traffic assignment problem: models and methods
Utrecht, VSP, 1994

R. Thomas
Traffic assignment techniques
Avebure Technical, 1991

Y. Sheffi
Urban transportation networks: equilibrium analysis with mathematical programming methods
Englewood Cliffs (N.J.) Prentice Hall, 1985

H. Spiess H and M. Florian
Optimal strategies: A new assignment model for transit networks, *Transportation Research B*, Vol. 23, pp. 83-102, 1989

P.M.Tisato
Suggestions for an improved Davidson's travel time function, *Australian Road Research* 21 (2), pp.85-100, 1991

J. van der Waard
The relative importance of public transport trip-time attributes in route choice, *Proc. Of the annual PTRC meeting*, 1988

J.G. Wardrop
Some theoretical aspects of road traffic research, *Proceedings, Institute of civil engineering* II(1), 1952

8 Estimating origin-destination trip tables and distribution functions

8.1 Objective

In the earlier chapters of these course notes models have been presented for trip generation, trip distribution, mode choice and route choice. Within the framework of transport planning these models may be used in two ways:

- To compute the present travel demand. The present travel demand is summarized in a *base year matrix*. The cell values in this matrix can be directly observed or derived from observed data. However, in general there are multiple matrices that would fit a given set of observations. Hence the base year matrix cannot be uniquely determined. We refer to this phenomenon as *underspecification*. A model can in such case be used to narrow down the set of possible solutions to a unique 'best' one. We refer to this as OD-trip matrix estimation.

- A second possibility is to predict future travel demand. Usually this is done by extrapolation of model parameters that are calibrated for the present state. An example is prediction of future trip generation for, e.g., a new building site, while behavioral data such as the parameters in the distribution function are kept constant.

Not all models can be applied in both ways. For example, if a model contains many parameters that are not transferable to a future date or state, such a model is less suitable to predict future travel demand.

This chapter primarily considers the problem of the estimation of model parameters, in particular:
- the estimation of parameters in distribution functions;
- the estimation of base year matrices.

A future network can be analyzed, using a base year matrix expanded using growth factors, or by using a model whose behavioral parameters are calibrated.

An abstract representation of a model is the following:

$$y = f(\theta, x) \tag{8.1}$$

with:

y	vector of dependent variables (output)
x	vector of independent variables
θ	model parameters

Estimating the model means that the parameters are estimated using statistical theory and observations of x and y.

The estimation of a base year matrix $\{T_{ij}^0\}$ involves the following:
- a prior matrix (if available),
- a trip distribution model,
- a distribution function F_{ij} (if available)
- observed or estimated number of trip departures and trip arrivals (trip ends)
- other observations, e.g. traffic counts

The prediction of a future OD table involves the following:
- a base year matrix (if available),
- a trip distribution model,
- a distribution function F_{ij} (if available)
- predicted number of trip departures and trip arrivals (trip ends)

8.2 Types of data used in transport planning

Various types of input data can be used in transport planning. The most important sources of data are discussed below. Data stem from observations, for example by counting vehicles, or can be imported from other studies, for example by using model parameters estimated earlier.

- *Trip generation*. Trip generation (also referred to as trip ends) is the number of trips originating in a zone or destined for it. Various ways exist to estimate trip generation (see Chapter 4).

- Based on socio-economic data. For this purpose data on the socio economic contents of the zones are collected, e.g. the number of people living in the zone, the number of jobs, the number of cars parked in the street at night, etc. The next step is to use an earlier calibrated model to predict the trip ends based on these data. An exception is made for specific generators, such as hospitals and airports. The amount of traffic generated by these facilities has to be estimated separately.

- Another way to estimate trip ends, at least in the Netherlands, is to use the national travel survey (in Dutch: Onderzoek VerplaatsingsGedrag, OVG, collected by CBS). Because a minimum sample size is required to obtain a reliable prediction, the use of OVG for direct estimation of trip ends is limited to a high level of aggregation.

- *Prior OD Trip table(s)*. A prior trip table is a trip table/matrix that needs (further) updating, with new data. Ways to obtain a prior matrix are:
  - using the results of an old study
  - using survey data
  - applying a trip distribution model

- *Distribution function*. A distribution function represents the relative willingness to make a trip as a function of the generalized trip costs. This function is defined using a functional form (which may be continuous or discrete, see Chapter 5). This functional form defines the function except for a few parameters. Estimates for the parameters in the distribution function are often a by-product of computing travel demand. However, it is common practice to keep the distribution function fixed during the computation of travel demand. In such instances the distribution function parameters are imported from other studies. The OVG (mentioned earlier) is a good source of data for the estimation of parameters in distribution functions. Usually distribution functions are estimated separately for various trip purposes (e.g. work, business, other), traveler categories(e.g. car owners, income categories, level of urbanization) and travel modes (e.g. car, public transit, cycle/walk).

- *Car- or passenger counts*. Car and passenger counts are indirect observations: OD-cells are not observed, only linear combinations of them are. Traffic counts may be collected in an automated way by means of induction loops or by using pneumatic tubes. Incidentally, one can analyze video data to reconstruct the number of vehicles passing the observation point. Passenger counts on public transit can be automated by using a counting device in the steps leading to the passenger vehicles. A requirement for using

counts for the purpose of estimation of OD tables is that information on route choice is available. As this information is not always available, often screen line or cordon counts are used. These involve a combination of counting points, chosen in such a way that the study area is divided in two (screen line) or a certain area is surrounded (cordon).

– *Surveys*. Travel surveys are used to obtain direct information on the number of OD-trips. When using survey data one should distinguish between observed zero values and non-observed cells. In the first case no trips were made by the surveyed population for this OD-cells, in the second case no enquiries were made with respect to trips for a specific OD-pair. This may be due to the location on which the survey is held.
The main types of survey used in practice are home surveys, cordon surveys and screen line surveys. If the survey is organized in such a way that certain groups of cells are not observed, we refer to such a survey as *incomplete*.

– *License-plate surveys*. This is a special category of surveys in which registrations of passing vehicles are recorded at multiple locations in the network. Usually this is done manually. But during the last decade equipment has been under development to automate this task. Ideally all points of observations jointly make up a *complete survey* as discussed earlier. In reality this is usually not the case. Because many mistakes are made while recording registrations, the fact that license plate surveys are usually incomplete may cause significant bias in the data set, which need correction. For example, when at one end of a cordon survey a mistake is made writing down a registration, this may lead to the conclusion that instead of one trip, two trips have been observed: one trip that has ended in study area, and a second trip has originated in the study area. This mechanism hence leads to an overestimation of the total number of trips and an underestimation of the number of trips through the study area. Apart from simple corrections for mistakes such as switching two letters, two modifications in the experimental setup are possible that are aimed at reducing the influence of mis-interpretations:
  - selection based on colour or type. In this only vehicles of a certain colour or type are involved in the experiment. This reduces the burden of the observer and hence reduces the probability of error. On the other hand such a strategy might introduce new errors if the selection criterion is interpreted differently by different observers. Selection based on vehicle colour therefore is not recommended for OD surveys.
  - partial registration of vehicle licenses. In this case only a few digits of the registration are recorded, for example the last two numbers/digits. This reduces the probability of erroneous records. The registration of partial license plate numbers also introduces a new problem: that of spurious matches, two different vehicle with identical partial registrations. Specific statistical procedures have been designed to correct for these errors.

– *Observed Trip Length Distribution, OTLD, (in Dutch: Ritlengte frequentie verdeling)*. These are the number of trips observed in each distance category. An OTLD should not be confused with a distribution function. The latter represents the relative willingness to make a trip at a certain cost. The first relates to the actual trip making behavior. A difference between the two arises because the productions of origin zones and the attractions of destination zones influence the number of trips made within a specific distance category. A quantity that is derived from the OTLD is the mean trip length (MTL). This quantity is also frequently used during the estimation of parameters in trip distribution models.

Data can be used in various combinations when estimating travel demand. Depending on the data available and the parameters that need to be estimated (estimating distribution function parameters, estimating base year matrix), different estimation methods can be used. An overview of methods discussed in this chapter is given in Table 8.1.

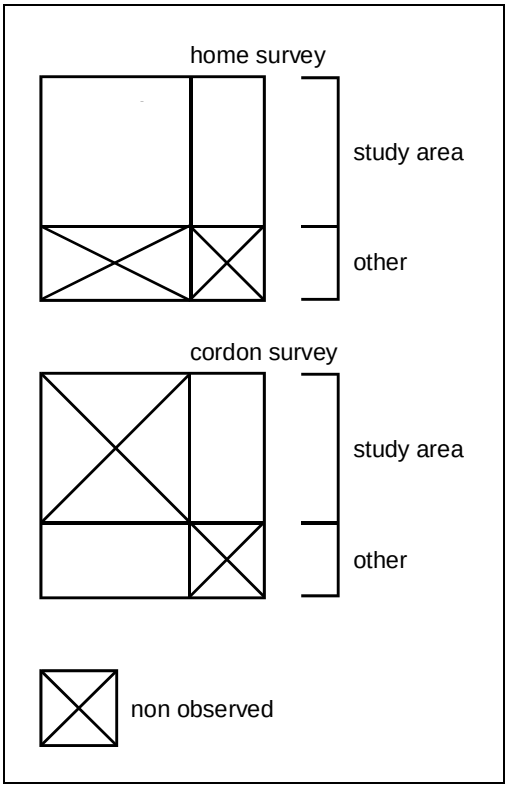


Figure 8.1: *Observed OD-pair with home interviews and cordon surveys respectively*

Result	Model applied	Data	Estimation methodology applied	See
OD-matrix, Parameters of distribution function	Gravity model + discrete distribution function	complete or incomplete survey	ML/ Poisson estimator	Section 8.4
OD-matrix, Parameters of distribution function	Gravity model + discrete distribution function	Trip ends OTLD	ML/ Poisson estimator	Section 8.4
OD-matrix	Gravity model	Trip ends, Distribution function	Balancing	Section 8.5
OD-matrix, Parameters of distribution function	Gravity model + exponential distribution function	Trip ends, MTL	Balancing + Iteration	Section 8.6
OD-matrix	-	Trip ends, Prior OD-matrix	Balancing (Furness)	Section 8.7
OD-matrix	-	Prior OD-matrix, traffic counts	Balancing/ Binary calibration	Section 8.8

Table 8.1: *Overview of estimation problems*

Depending on the data available different approaches are possible. Another way of looking at Table 8.1 is shown in Figure 8.2. This figure shows that the different procedures mentioned in the table can be combined to new ones.

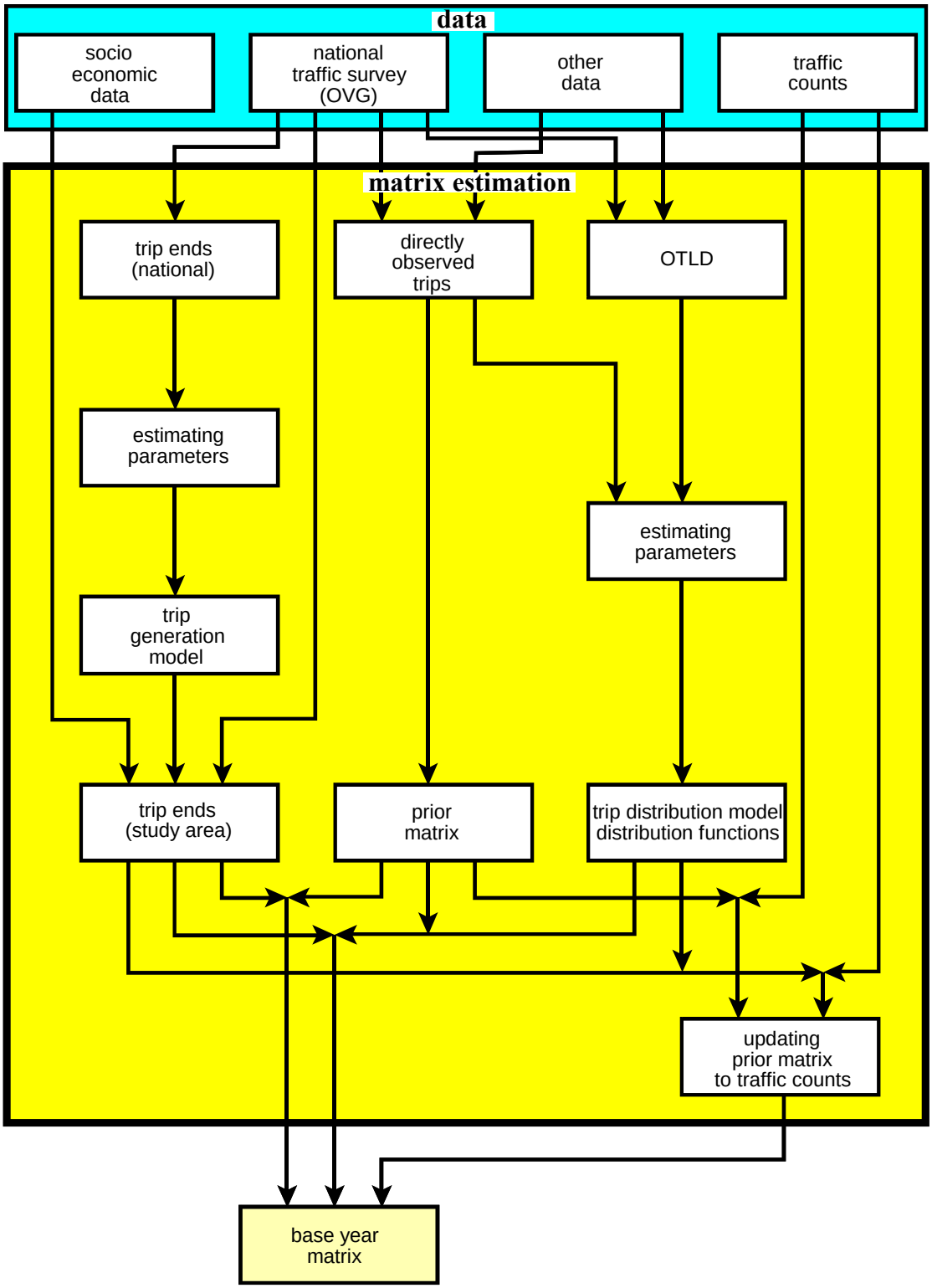


Figure 8.2: *Approaches for estimating base year matrices and parameters in distribution functions*

Example 8.1:
Consider a study area for which a base year matrix needs to be estimated (external zones are not considered). The following data are available:
- a network and its corresponding zones;
- the OVG travel survey;
- detailed socio-economic data on the study area;
- traffic counts on a number of strategically chosen links.

To compute a base year matrix we could for example follow the following strategy (see also Figure 8.2):

Step 1. Based on the OVG survey data, the parameters of a trip generation model are estimated. (Trip generation models were discussed in Chapter 4). Note that the OVG data that are used in this context may also refer to areas outside the study area, as long as the travel behavior in those areas is representative for the travel behavior in the study area. Utilizing these additional data prevents there being too few observations available for the estimation.

Step 2. Using the socio-economic data corresponding to the study area and the trip generation model calibrated in the first step, an estimate is made of the number of trip departures and trip arrivals in the zones of the study area.

Step 3. Based on observations in the OVG travel survey the parameters of the distribution functions are estimated. Sometimes this is done separately for each category of persons and each trip purpose (multi user-class) (see Section 8.4).

Step 4. The trip ends estimated in step 2 and the distribution function estimated in step 3 are combined to estimate an OD-matrix that is used as a first approximation to the base year matrix (see Section 8.5)

Step 5. The OD-matrix estimated in step 4 is adapted to the traffic counts that are available (see Section 8.8). The resulting matrix is the base year matrix.

- end of example -

8.3 The estimation and calibration of models

A model represents a number of assumptions which jointly define the relation between system variables,

$$y = f(\theta, x) \tag{8.2}$$

System variables y and x are 'real world' quantities, such as trip generation, OD-flows. Usually, a portion of the system variables can be directly observed (for example link flows) while another portion of the system variables (for example, OD-flows) can not be directly observed and need to be estimated using the model. We refer to the vector y as the dependent variable and the vector x as the independent variable in the model. When using a calibrated model, the x-variables are the model input while the y-variables are the model output.

Most models contain one or more parameters, represented by the vector θ. Parameters do not always have a real world interpretation - for example, consider the scale parameters in the logit route choice model, or the parameters in a distribution function. When a good model specification is used, the model parameters θ are transferable to other areas or time frames. This means that once parameters have been determined for a particular study, they can be used in other studies as well. To be able to produce useful output of the dependent variables y, model parameters that cannot be imported from other studies should be estimated.

Deterministic and stochastic models
A deterministic model is defined in terms of equalities, e.g. an all-or-nothing assignment model. A stochastic model is defined in terms of probability distributions. The equations in the model now contain error terms:

$$y = f(\theta, x) + \varepsilon \tag{8.3}$$

The probability distribution of the error terms ε is specified in the model.

Estimating model parameters using a distance criterion
The fact that a model is specified in a deterministic manner does not mean that the relations specified in the model are exact, but rather that no information is available on the statistical distribution of the model errors. Usually the calibration of the parameters θ in a deterministic model is based on the minimization of the difference between the observations y and the model predicted values $\hat{y}$ where y for example is a set of traffic counts. This difference is quantified using a distance criterion $D(y,\hat{y})$:

$$D(y, \hat{y}(\theta, x)) \tag{8.4}$$

where:

$y = [y_1, y_2, ...y_n]$	: observed values
$\hat{y}(\theta, x) = [\hat{y}_1(\theta, x), \hat{y}_2(\theta, x), ...\hat{y}_n(\theta, x)]$	: values predicted by the model
$\theta = [\theta_1, \theta_2, ...\theta_k]$	: model parameters
k	: number of (unknown) parameters
n	: number of observations

The optimal value for θ we are looking for (which minimizes D) is indicated with: $\hat{\theta}(y, x)$:

$$\hat{\theta}(y, x) = \underset{\theta}{\operatorname{argmin}}\, D(y, \hat{y}(\theta, x)) \tag{8.5}$$

In this expression 'argmin θ' represents the value of θ that minimizes the expression that follows it ($D(y,\hat{y}(\theta,x))$). The choice of distance criterion is usually inspired by practical reasons, such as the ease of implementation in a computer program. A proper distance measures is non-negative and is zero if and only if two elements are equal:

$$\begin{aligned} &D(a,b) \geq 0, \quad \forall a,b \\ &D(a,b) = 0 \Leftrightarrow a = b \end{aligned} \tag{8.6}$$

Some examples of distance measures are given below:

least squares: $$D(a,b) = \sum_{i=1}^{n} (a_i - b_i)^2 \tag{8.7}$$

weighted least squares: $$D(a,b) = \sum_{i=1}^{n} \frac{(a_i - b_i)^2}{w_i} \tag{8.8}$$

entropy (for two probability functions): $$D(a,b) = \sum_{i=1}^{n} a_i \log(a_i / b_i) - a_i + b_i \tag{8.9}$$

The trip generation models in Section 4.4 are regression models of which the parameters are estimated using weighted least squares. The entropy distance measure is used in Section 8.7.

Estimation using maximum likelihood
In a way, stochastic models are specified more completely then deterministic models are: apart from the relation between system variables, the probability function of modeling and observation errors are also specified. This makes it possible to underpin the estimation of model parameters in a statistical manner. The most common method for doing this is the *Maximum Likelihood* (ML) method. When applying the ML-method, observed data are considered as realizations of *Random Variables* (RV's). The probability of observing the outcomes ($y_1, y_2,...y_n$) is defined by a probability function. This probability function can be derived from the model except for a number of model parameters ($\theta_1, \theta_2,...\theta_k$):

$$\text{probability of observing } y = p(y_1, y_2,...y_n \mid \theta_1, \theta_2,...\theta_n) \tag{8.10}$$

The objective is to determine the parameters $\theta_1, \theta_2,...\theta_k$ that maximize the probability of observing $y_1, y_2,...y_n$. Because the observations are given while the parameters still need to be estimated, the expression that needs to be maximized can be considered as a function of the model parameters $\theta_1, \theta_2,...\theta_k$. We refer to this function as the Likelihood function, denoted with $L(\theta;y)$. The objective of estimating the model parameters is now equivalent to:

$$\text{maximize } L(\theta_1, \theta_2,...\theta_n; y_1, y_2,...y_n) = p(y \mid \theta) \tag{8.11}$$

When the observation $y_1, y_2,...y_n$ can be considered as *independent* realizations of an *identical probability distribution* (*iid*) the probability of jointly observing $y_1, y_2,...y_n$ equals the product of the probabilities of individually observing $y_1, y_2,...y_n$ respectively. In this case it holds that:

$$L(\theta_1, \theta_2,...\theta_n; y_1, y_2,...y_n) = \prod_{i=1}^{n} p(y_i \mid \theta_1, \theta_2,...\theta_k) \tag{8.12}$$

In case of independent observations one usually works with the *logarithm* of the likelihood function, the *loglikelihood*. The loglikelihood usually is easier to manipulate mathematically:

$$\log L(\theta_1, \theta_2,...\theta_n; y_1, y_2,...y_n) = \sum_{i=1}^{n} \log[p(y_i \mid \theta_1, \theta_2,...\theta_k)] \tag{8.13}$$

The maximization of the loglikelihood is equivalent to the maximization of the likelihood; The optimal value for θ is denoted with $\hat{\theta}$:

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}}\, L(\theta_1, \theta_2,...\theta_n; y_1, y_2,...y_n) \\ &= \underset{\theta}{\operatorname{argmax}} \log L(\theta_1, \theta_2,...\theta_n; y_1, y_2,...y_n) \end{aligned} \tag{8.14}$$

Frequently used probability distributions
The mathematical form of a likelihood function strongly depends on the choice of probability distribution that is assumed in the model. Below an overview is given of some frequently used probability functions in transport modeling:

Name	Type	Function	Parameters	Mean	Variance
Uniform	Continuous	$p(x) = \frac{1}{b-a} I_{[a,b]}(x)$	$-\infty < a < b < \infty$	$\frac{a+2}{2}$	$\frac{(b-a)^2}{12}$
Normal $N(\mu, \sigma^2)$	Continuous	$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$	$-\infty < \mu < \infty$ $\sigma^2 > 0$	μ	σ^2
Poisson	Discrete	$P(x) = \frac{\exp[-\lambda]\lambda^x}{x!}$	$\lambda > 0$	λ	λ
Multinomial	Discrete	$P(x_1, x_2, ...x_n) = \frac{N!}{x_1! x_2! ... x_n!} p_1^{x_1} p_2^{x_2} ... p_n^{x_n}$ where $N = \sum_i x_i$	$0 \le p_i \le 1$ $\sum_i p_i = 1$	$E[x_i] = Np_i$	$\text{var}[x_i] = Np_i(1-p_i)$ $\text{cov}[x_i, x_j] = -Np_i p_j$

Name	Examples	Properties
Uniform	Departure time	
Normal	Trip length, Travel time	The sum of two normally distributed rv's is a normally distributed rv
Poisson	Trip generation, traffic counts, number of accidents, Survey data	The sum of two Poisson distributed rv's is a Poisson distributed rv
Multinomial	Survey data	This distribution describes the probability of xi successes when performing N experiments with identical probability of success

Table 8.2: *Frequently used probability functions in transport modeling*

It should be noted that the Poisson and Normal distribution are almost identical if their expected values are greater than 12 ($E[x] > 12$) and their variances are equal ($\lambda = \sigma^2$). For this reason the choice of distribution used in a model in practice often depends on practical considerations, such as the ease of manipulation of the mathematical equations that arise from the choice of probability distribution.

For OD-survey data, the condition $E[x] > 12$ is not likely to be met. Usually the number of trips observed in an OD cell is small. In these cases the normal distribution cannot be used as a substitute for the more realistic Poisson and Multinomial distributions.

8.4 The Poisson estimator

The Poisson estimator is designed for the estimation of base year matrices and distribution function parameters using OD-survey data or observed trip ends. The method is based on the gravity model, and can be considered as a statistically underpinned method for determining the parameters in this model.

Stochastic specification of the Gravity Trip Distribution model
The trip distribution model that is used as a point of departure is:

$$\hat{T}_{ij}(Q_i, X_j, F(c_{ij})) = Q_i X_j F(c_{ij}) \tag{8.15}$$

with:

$\hat{T}_{ij}(Q_i, X_j, F(c_{ij}))$	model predicted OD cell value
Q_i	production ability
X_j	attraction ability
$F(c_{ij})$	discretized distribution function
c_{ij}	generalized travel costs

A discretized distribution function can be considered as a piecewise constant function; the travel cost axis is divided into a limited number of cost bins, and for each cost bin a distribution function value is assigned. These distribution function values are the parameters of the distribution function. Mathematically this is denoted as follows:

$$F(c_{ij}) = \sum_k F_k \Delta_{ij}^k, \quad \Delta_{ij}^k \in \{0,1\} \tag{8.16}$$

with:

F_k	value of the distribution function for cost bin k
Δ_{ij}^k	equals 1 when generalized travel costs c_{ij} are in cost bin k and 0 otherwise

The trip distribution model will not describe reality exactly. Therefore it is assumed that the OD-cells T_{ij} are Poisson distributed with the model predicted cell values as a mean:

$$T_{ij} \sim \text{Poisson}[Q_i X_j F^k] \tag{8.17}$$

In other words:

$$P[T_{ij}] = \frac{\exp[-Q_i X_j F^k] \cdot (Q_i X_j F^k)^{T_{ij}}}{T_{ij}!} \tag{8.18}$$

in which F^k is short for $\sum_k F_k \Delta_{ij}^k$.

Taking into account the interpretation of T_{ij} in transport planning, the choice of the Poisson distribution can be motivated as follows:
- The Poisson distribution does not allow negative values;
- The sum of two Poisson distributed random variables is again Poisson distributed. Therefore changing the aggregation level of the model, for example by joining two zones or trip purposes, does not lead to inconsistencies.

In addition to this, the Poisson distribution leads to mathematically tractable expressions.

Deriving the Likelihood expression for survey data
Survey data can be seen as a random sample drawn from a set of trips. The result of the survey is denoted with $\{n_{ij}\}$ with:

n_{ij}	the number of observed trips in OD-cell *i-j*
N	the total number of observed trips, $N = \sum_{i,j} n_{ij}$

Not all OD-cells need be represented in the survey. See e.g. Figure 8.2. An important difference hence exists between non-observed cells and observed zeros ($n_{ij} = 0$). If certain OD-pairs are not represented in the survey we refer to this survey as an incomplete survey. To indicate which OD-pairs are part of the survey we use the indicator matrix *S*. The matrix *S* has identical dimension as the OD-matrix for the study area. The cells of the matrix are defined as follows:

$S_{ij} = 1$ if OD-pair *i-j* is represented in the survey
$S_{ij} = 0$ if OD-pair *i-j* is not represented in the survey

When the condition is met that the total sample size *N* is small relative to the total number of trips made, we can assume that the sample data are independently Poisson distributed:

$$n_{ij} \sim \text{Poisson}[cQ_i X_j F(c_{ij})] \tag{8.19}$$

with:

$$\text{c} = \frac{N}{\hat{T}} \tag{8.20}$$

and:

$$\hat{T} = \sum_{i,j} \hat{T}_{ij} \tag{8.21}$$

Because the outcomes of the survey can be considered to be independent, the probability of outcomes $\{n_{ij}\}$ given model parameters $\{Q_i\}$, $\{X_j\}$ and $\{F_k\}$ is given by:

$$P[n_{ij} \mid Q_i, X_j, F^k] = \prod_{i,j|S_{ij}=1} \frac{\exp[-cQ_i X_j F^k] \cdot (cQ_i X_j F^k)^{n_{ij}}}{n_{ij}!} \tag{8.22}$$

The corresponding loglikelihood function can thus be written as:

$$\begin{aligned} &\log L[Q_i, X_j, F^k; n_{ij}] \\ &= \sum_i \sum_j S_{ij} \cdot [-cQ_i X_j F^k + n_{ij} \cdot \log(cQ_i X_j F^k) - \log(n_{ij})!] \end{aligned} \tag{8.23}$$

A condition that applies is that all model parameters are nonnegative:

$$\begin{aligned} Q_i &\geq 0 \quad \forall i \\ X_j &\geq 0 \quad \forall j \\ F^k &\geq 0 \quad \forall k \end{aligned} \tag{8.24}$$

Maximization of the loglikelihood
The loglikelihood function that has been described above can be considered as a continuous function of the model parameters Q_i, X_j, and F^k. In the maximum the following should apply:

$$\begin{aligned}
&\frac{\partial \log L}{\partial Q_i} = 0 \quad \forall i \\
&\frac{\partial \log L}{\partial X_j} = 0 \quad \forall j \\
&\frac{\partial \log L}{\partial F^k} = 0 \quad \forall k
\end{aligned} \tag{8.25}$$

The model parameters Q_i, X_j, and F^k should hence satisfy:

$$\begin{aligned}
&\sum_j (-cX_j F^k + \frac{n_{ij}}{Q_i}) S_{ij} = 0, \qquad \forall i \\
&\sum_i (-cQ_i F^k + \frac{n_{ij}}{X_j}) S_{ij} = 0, \qquad \forall j \\
&\sum_i \sum_j (-cQ_i X_j + \frac{n_{ij}}{F_k}) S_{ij} \Delta_{ij}^k = 0, \quad \forall k
\end{aligned} \tag{8.26}$$

Which is equivalent to:

$$\begin{aligned}
&Q_i = \sum_j S_{ij} n_{ij} \,/ \sum_j S_{ij} c X_j F^k, \qquad \forall i \\
&X_j = \sum_i S_{ij} n_{ij} \,/ \sum_i S_{ij} c Q_i F^k, \qquad \forall j \\
&F_k = \sum_i \sum_j S_{ij} \Delta_{ij}^k n_{ij} \,/ \sum_i \sum_j S_{ij} \Delta_{ij}^k c Q_i X_j, \quad \forall k
\end{aligned} \tag{8.27}$$

The number of equations are hence equal to the number of unknowns. However the model parameters Q_i, X_j, and F_k can not be directly derived from these equations: instead, they are implicitly defined by these equations. An iterative procedure is needed to solve the equations.

Solution algorithm
A method to solve for the model parameters is known as *Gauss-Seidel* iteration. In this method we solve for one group of parameters while another group of parameters is kept constant. When applied to the problem of maximizing the loglikelihood function above, the method reduces to four steps:

Step 1. Initialize the parameters with:

$$Q_i^{(0)} = 1, \quad \forall i, \quad X_j^{(0)} = 1, \quad \forall j, \quad F_k^{(0)} = 1, \quad \forall k \tag{8.28}$$

This implies that all matrix cells are set to 1.

Step 2. Determine the new model parameters with:

$$\begin{aligned}
&Q_i^{(n+1)} = \sum_j S_{ij} n_{ij} \,/ \sum_j S_{ij} X_j^{(n)} \Delta_{ij}^k F_k^{(n)}, \qquad \forall i \\
&X_j^{(n+1)} = \sum_i S_{ij} n_{ij} \,/ \sum_i S_{ij} Q_i^{(n+1)} \Delta_{ij}^k F_k^{(n)}, \qquad \forall j \\
&F_k^{(n+1)} = \sum_i \sum_j S_{ij} \Delta_{ij}^k n_{ij} \,/ \sum_i \sum_j S_{ij} \Delta_{ij}^k Q_i^{(n+1)} X_j^{(n+1)}, \quad \forall k
\end{aligned} \tag{8.29}$$

Step 3. Check the difference between the parameters computed in the last iteration and the parameters computed in the previous iteration. If there is a significant difference: repeat step 2.

Step 4. Stop

Conclusions
It catches the eye that the survey results n_{ij} are used only in an aggregated format. This has an interesting consequence when all OD-cells are the target of the survey ($S_{ij} = 1$ for all i and j). In this case only the trip ends and the observed trip length distribution (OTLD) are used to estimate the model parameters! Three remarks can be derived from this:

- The Poisson estimator can be applied effectively using only observed trip ends and an OTLD. In practice these data will certainly be easier to collect than a complete OD travel survey.

- A pitfall in this case might be that the observed trip ends are not always mutually consistent. When no correction is applied this might lead to the algorithm not converging. However a simple remedy is to *balance the trip ends* before applying the algorithm (Step 0).

- Survey costs can be saved when the target of the survey is reduced to certain groups of cells. In this case it is still possible to estimate all model parameters. An example of this is when traffic counts are used. However, depending on how counting locations are selected, traffic counts may be highly correlated, which can lead to poor convergence and even non unique solutions.

The computation method that is described above has many similarities with the balancing methods for doubly constrained trip generation models.

Finally, it can be concluded that the Poisson estimator has two main applications: Firstly parameters of a distribution function can be estimated using this method. Secondly, OD-matrices can be estimated from indirect observations such as trip ends and an OTLD.

Example 8.2:
As an example we carry out a computation with the Poisson estimator. The following is given:
- A model specification consisting of a gravity model with a discretized distribution function:

$$T_{ij} = Q_i X_j \sum_{k=1}^{6} \Delta_{ij}^{k} F_k \tag{8.30}$$

- The cost ranges corresponding to the distribution function (see table below)
- The travel cost between each OD pair
- The number of departures and arrivals for each zone

Note that all data are consistent. Therefore it is not necessary to balance the trip ends and the OTLD.

The objective is to estimate the base year OD matrix and the parameters in the distribution function.

distance range (minutes)	Parameter in distribution function	OTLD
1.0-4.0	F_1	365
4.1-8.0	F_2	962
8.1-12.0	F_3	160
12.1-16.0	F_4	150
16.1-20.0	F_5	230
20.1-24.0	F_6	95

Table 8.3: *Ranges corresponding to the cost bins of the distribution function*

from	1	2	3	4	P_i
1	3 (F_1)	11 (F_3)	18 (F_5)	22 (F_6)	400
2	12 (F_3)	3 (F_1)	13 (F_4)	19 (F_5)	460
3	15.5 (F_4)	13 (F_4)	5 (F_2)	7 (F_2)	460
4	24 (F_6)	18 (F_5)	8 (F_2)	5 (F_2)	702
A_j	260	400	500	802	1962

Table 8.4: *Travel costs in minutes with corresponding parameter in distribution function*

Computation

We start with the initial solution:

$$Q_i = 1, \quad X_j = 1, \quad F_k = 1, \quad \forall i, j, k \tag{8.31}$$

From this the initial tableau follows, including the scaling factors for the production abilities Q_i:

from/to	1	2	3	4	$\sum_j$	P_i	factor Q_i
1	1	1	1	1	4	400	100
2	1	1	1	1	4	460	115
3	1	1	1	1	4	460	100
4	1	1	1	1	4	702	175.5
X_j	1	1	1	1			
A_j	260	400	500	802		1962	

After scaling the production abilities Q_i the next tableau follows, from which the scaling factors for the attraction abilities X_j can be determined:

from/to	1	2	3	4	$\sum_j$	Pi	Q_i
1	100	100	100	100	400	400	100
2	115	115	115	115	460	460	115
3	100	100	100	100	400	460	100
4	175.5	175.5	1755	175.5	702	702	175.5
$\sum_i$	490.5	490.5	490.5	490.5			
A_j	260	400	500	802		1962	
factor X_j	0.53	0.82	1.01	1.63			

After scaling the attraction abilities X_j the next tableau follows:

from/to	1	2	3	4	$\sum_j$	Pi	Q_i
1	53	81.5	1019	163.5	400	400	100
2	61	93.8	117.2	188	460	460	115
3	53	81.5	1019	163.5	460	460	100
4	93	143.1	1789	287	702	702	175.5
$\sum_i$	260	400	500	802			
A_i	260	400	500	802		1962	
X_j	0.53	0.82	1.01	1.63			

From this tableau a Modeled Trip Length Distribution (MTLD) can be derived. By comparing this with the Observed Trip Length Distribution (OTLD) the scaling factors for the distribution function parameters can be determined:

distance range (min)	previous value F_k	MTLD	OTLD	scaling factor F_k	new value F_k
1.0-4.0	1	146.79	365	2.49	2.49
4.1-8.0	1	731.3	962	1.32	1.32
8.1-12.0	1	142.51	160	1.12	1.12
12.1-16.0	1	251.78	150	0.6	0.6
16.1-20.0	1	433.09	230	0.53	0.53
20.1-24	1	256.53	95	0.37	0.37

After scaling the distribution function parameters the first iteration is completed, with the following result:

from/to	1	2	3	4	Q_i
1	131.8	91.6	54.1	60.5	100
2	68.4	233.2	70	99.9	115
3	31.6	48.6	134.1	215.1	100
4	34.5	76	235.3	377.5	175.5
X_j	0.53	0.82	1.01	1.63	

This process can be repeated, for example to a total of 10 iterations. After the 10^{th} iteration the final result is:

from/to	1	2	3	4	Q_i
1	156.2	101.4	69	74.3	128.1
2	58.6	208.8	85.1	108.6	112.6
3	25.7	39.2	119.9	214.5	89
4	20.7	52.4	225	402.6	167
X_j	0.417	0.634	1.089	1.948	

distance range (min)	Parameter distribution function	Parameter value
1.0-4.0	F_1	2.9254
4.1-8.0	F_2	1.2374
8.1-12.0	F_3	1.2487
12.1-16.0	F_4	0.6942
16.1-20.0	F_5	0.4949
20.1-24	F_6	0.2976

In the final tableau the parameters Q_i are large relative to the other parameters. This is due to the order in which the groups of parameters are scaled. Because the initial tableau consists of only ones while the margins are no less then a few hundred, the first scaling operation has the largest impact on the parameter values. If the attraction abilities would have been the first group of parameters to be scaled, these would be the largest parameters. The values of the OD-cells are uniquely determined. However,

the model parameters that imply these values have two degrees of freedom. For example, doubling the production abilities Q_i while halving the attraction abilities X_j or the distribution function parameters F_k has no impact on the predicted OD-values. In practice it may be needed to take measures to prevent one group of parameters from growing very large while other parameters become very small. This may lead to problems with the representation of numbers within the computer (underflow or overflow).

The example above and similar computations can be programmed in a simple way using spreadsheets or other computational aids. As an illustration below the Matlab© source code is shown that is needed to estimate the parameters for the example above:

```
% Matlab source code for the calibration of distribution functions
% INPUT DATA
departures=[400 460 400 702]'
arrivals=[260 400 500 802]'
OTLD=[365 962 160 150 230 95]'% Vector with Observed Trip Length Distribution
dist_cat=[
1 3 5 6
3 1 4 5
4 4 2 2
6 5 2 2
]; %Distance category

% INITIALIZATION
Q=ones(size(departures));            % Vector of production abilities
X=ones(size(arrivals));                     % Vector of attraction abilities
F=ones(size(OTLD));                  % Vector of deterrence function values
MTLD=zeros(size(OTLD));              % Vector with modeled Trip Length Distribution

% Evaluate values for distribution function (note: the command 'reshape' converts a
vector into a matrix)
Fmat=reshape(F(dist_cat),size(dist_cat,1),size(dist_cat,2));
matrix = ones( size(dist_cat) );

% MAIN COMPUTATION
for iterate=1:10
       Q=Q.*departures./sum(matrix')';      % Modify Q & compute new matrix
       matrix=((Q*X').*Fmat);

       X=X.*arrivals./sum(matrix)';         % Modify X & compute new matrix
       matrix=((Q*X').*Fmat);

       for k=1:length(OTLD);         % Compute Modeled Trip Length Distribution
                     MTLD(k)=sum(matrix(dist_cat==k));
       end

       F=F.*OTLD./MTLD;              % Modify F

       Fmat=reshape(F(dist_cat),size(dist_cat,1),size(dist_cat,2));
                                     % Evaluate values for distribution function

       matrix=((Q*X').*Fmat);        % Compute new matrix
end

disp('end result:');                 % Plot end results
disp([[matrix Q];[X',0]])
disp(F)
```

- end of example -

8.5 Estimating a base year matrix using a fixed distribution function

The method described in Section 8.4 can be used in many types of computations related to trip distribution. In the previous section one of the results of the computation was the set of parameters of the (discretized) distribution function. However, in some cases one might want to estimate an OD-matrix at the lowest possible cost of data collection. In this case it is not necessary to estimate the parameters of the distribution function again, provided that these can be imported from another study. This is because the distribution function represents

behavioral parameters that are transferable from one study area to the other. Therefore if the distribution function estimated using data from, e.g., the 'onderzoek verplaatsings gedrag (OVG)', is available, it suffices to have knowledge of the trip ends for a study area to estimate an OD-matrix. The trip ends can be estimated using trip generation models described in the earlier chapters.

To estimate an OD-matrix using an imported distribution we begin with the formulation of the trip generation model:

$$T_{ij} = Q_i X_j F(c_{ij}) \tag{8.32}$$

where $F(c_{ij})$ is a given distribution function. This may be an exponential, discretized, or other shape of distribution function. Other boundary conditions are:

$$\begin{aligned} \sum_j T_{ij} &= P_i \\ \sum_i T_{ij} &= A_j \end{aligned} \tag{8.33}$$

To determine Q_i and X_j we use an identical scheme to the one used in the previous section, with the modification that step 1 is replaced with a step in which the matrix cells are initialized with the corresponding values of the given distribution function (instead of the unitary values), and that in step 2 the distribution function parameters are not modified;

Step 0. Balance the trip departures and trip arrivals. This can be done either by changing the trip departures or by changing the trip arrival or both. This step is needed to guarantee convergence of the algorithm.

Step 1. Initialize the parameters with:

$$Q_i^{(0)} = 1, \quad \forall i, \quad X_j^{(0)} = 1, \quad \forall j, \quad F_k^{(0)} = F(c_{ij}), \quad \forall k \tag{8.34}$$

This implies that all matrix cells are initialized with the values of the distribution that apply to these cells.

Step 2. Determine the new model parameters using:

$$\begin{aligned} Q_i^{(n+1)} &= P_i / \sum_j X_j^{(n)} F(c_{ij}), \quad \forall i \\ X_j^{(n+1)} &= A_j / \sum_i Q_i^{(n+1)} F(c_{ij}), \quad \forall j \end{aligned} \tag{8.35}$$

Step 3. Check whether or not the parameters computed in the current iteration substantially differ from the parameters in the previous iteration. If this is the case, repeat step 2.

Step 4. Stop

Because the iterations consist of scaling rows and columns alternatively, the final result of the iterations can be written as:

$$T_{ij}^{(n)} = \left[\prod_{p=1}^{n} \alpha_i^{(p)} \right] \left[\prod_{q=1}^{n} \beta_j^{(q)} \right] F(c_{ij}) \tag{8.36}$$

with

$$\alpha_i^{(p)} = [\text{scalefactor for row } i \text{ in iteration } p] = P_i / \sum_j T_{ij}^{(p)}$$
$$\beta_j^{(q)} = [\text{scalefactor for row } j \text{ in iteration } q] = A_j / \sum_i T_{ij}^{(q)} \quad (8.37)$$

with $T_{ij}^{(p)}$ and $T_{ij}^{(q)}$ the matrix cells that apply at the moment of scaling

After the iterations have been completed, the difference between model predicted trip ends and imposed boundary conditions are minimized. Also, it can be seen that the solution resulting still satisfies the general trip distribution model. This follows if we substitute:

$$Q_i = \left[\prod_{p=1}^{n} \alpha_i^{(p)} \right]$$
$$X_j = \left[\prod_{q=1}^{n} \beta_j^{(q)} \right] \quad (8.38)$$

8.6 The estimation of parameters in an exponential distribution function

In Section 8.4 it has been shown how parameters can be estimated for a discrete *distribution function* when trip ends and an OTLD are available. A similar procedure exists for the estimation of the parameter in an *exponential distribution function* based on these data. This method is known under the name *Hyman's method*. Similar procedures can also be derived for the estimation of parameters in other continuous distribution functions.

An exponential distribution function is defined by the following formula:

$$F(c_{ij}) = \exp[-\alpha c_{ij}] \quad (8.39)$$

In which α is a parameter that needs to be estimated. The choice of α has a large impact on the estimated OD-matrix when a procedure such as described in Section 8.5 is used. Because the distribution function represents the relative willingness to make a trip as a function of the travel costs, a large value of α implies a large number of short trips and vice versa.

The parameter α needs to be chosen in such a manner that the observed trip length distribution is reproduced as well as possible by the model predicted OD-matrix. In practice this is only possible to a certain extent, given the fact that only one parameter can be varied. We use the goal that the mean travel cost for the estimated OD matrix should match the observed mean travel cost. It can be shown that this is also the maximum likelihood solution. The optimal value of α can be found by applying the following iterative scheme:

Step 1. Choose as an initial solution for a the value $\alpha^{(0)} = 1/\bar{c}$, with $\bar{c}$ the mean observed trip length.

Step 2. Compute the trip distribution using the method described in Section 8.5 using the observed trip ends and the estimated value $\alpha^{(n)}$. This results in an estimated OD-matrix from which a modeled mean trip length $c^{(n)}$ can be derived. If the observed mean trip length $\bar{c}$ is approached sufficiently close the iterations are stopped.

Step 3. When $n = 1$, a better value for α can be estimated using:

$$\alpha^{(1)} = (c^{(1)} / c^{(0)})\alpha^{(0)} \tag{8.40}$$

When the iteration counter $n > 1$, α is estimated in the following manner:

$$\alpha^{(n+1)} = \frac{(\bar{c} - c^{(n-1)})\alpha^{(n-1)} - (\bar{c} - c^{(n)})\alpha^{(n)}}{c^{(n)} - c^{(n-1)}} \tag{8.41}$$

Step 4. Repeat steps 2 and 3 until convergence has been reached.

8.7 Updating OD-matrices to trip end totals (growth factor models)

When a historic matrix t is available, this matrix may be updated to meet current or predicted trip end totals, for example, as is the case with growth factor models. The procedure that is applied in this case is much like the one needed for the estimation of OD-matrices from trip ends and a given distribution function. The most significant difference is that the iterations are initialized with the historic matrix as a first approximation of T. Step 1 is hence replaced with:

Step 0. Balance the trip ends, by modifying A_j, P_i or both.

Step 1. Initialize with:

$$\begin{aligned} &T_{ij}^{(0)} = t_{ij}, \quad \forall i, j \\ &Q_i^{(0)} = 1, \quad \forall i, \quad X_j^{(0)} = 1, \quad \forall j \end{aligned} \tag{8.42}$$

This implies that OD-cells are set to the values of the prior matrix.

Step 2. Determine the new factors Q_i and X_j with:

for $\forall i$:

$$Q_i^{(n+1)} = \frac{P_i}{\sum_j X_j^{(n)} t_{ij}} = Q_i^{(n)} \frac{P_i}{\sum_j T_{ij}^{(n)}}, \qquad \text{with: } T_{ij}^{(n)} = Q_i^{(n)} X_j^{(n)} t_{ij} \tag{8.43}$$

for $\forall j$:

$$X_j^{(n+1)} = \frac{A_j}{\sum_i Q_i^{(n+1)} t_{ij}} = X_j^{(n)} \frac{A_j}{\sum_j T_{ij}^{(n)}}, \quad \text{with: } T_{ij}^{(n)} = Q_i^{(n+1)} X_j^{(n)} t_{ij} \tag{8.44}$$

Step 3. Check whether or not the parameters computed in the current iteration substantially differ from the parameters computed in the previous iteration. If this is the case, repeat step 2.

Step 4. Stop

This procedure is known under various names, such as Furness balancing and biproportional fitting. It is easy to see that after repeatedly scaling rows and columns to meet boundary values a matrix is obtained that satisfies:

$$T_{ij} = Q_i X_j t_{ij}, \quad \forall i, j \tag{8.45}$$

The final result hence only satisfies the general trip distribution model with a distribution function if the historic matrix that was used as a seed for this procedure does too. This does not mean there is no underpinning of the method. The most common interpretation is that of *entropy maximization*.

The solution that results minimizes an expression that is known as the entropy:

$$S[T \mid t] = -\sum_{i,j} (T_{ij} \log T_{ij} / t_{ij} - T_{ij} + t_{ij}) \tag{8.46}$$

Maximization of the entropy is equivalent to the minimization of the entropy distance measure (see Section 8.3). It can be shown that in approximation the maximization of $S[T|t]$ is equivalent to the minimization of:

$$S_2[T \mid t] = \sum_{i,j} \frac{0.5(T_{ij} + t_{ij})^2}{T_{ij}} \tag{8.47}$$

This is the weighted squared difference between historic and modified matrix. The method can hence be interpreted as a way of finding a matrix that satisfies the boundary conditions and at the same time is as close as possible to the historic matrix, according to a specific distance measure.

For additional information on this subject we refer to Ortúzar and Willumsen (1990), page 162.

8.8 Updating an OD-matrix to traffic counts

An important topic related to the estimation of OD-tables is updating OD-matrices to traffic counts. As opposed to collecting survey data, traffic counts can be obtained at low costs, while they contain important (indirect) information about OD-tables.

The relation between OD-tables and traffic counts
In order to utilize traffic counts for the estimation of OD-tables, one should know to which extent OD-cells contribute to traffic counts. The map **B** that assigns OD-cells to route flows (the route choice map) is given by:

$$\begin{aligned} &\beta_{ij}^r = \text{the proportion of OD flow } i\text{-}j \text{ that uses route } r \\ &0 \leq \beta_{ij}^r \leq 1 \end{aligned} \tag{8.48}$$

The map that defines the relation between route flows and link flows (the route-link incidence map) is denoted by the symbol **A** and is given by:

$$\begin{aligned} &\alpha_r^a = 1 \text{ when link } a \text{ is on route } r \\ &\alpha_r^a = 0 \text{ when link } a \text{ is not on route } r \end{aligned} \tag{8.49}$$

The relation between the OD-matrix and link flows q_a is hence given by:

$$q_a = \sum_i \sum_j \sum_r \beta_{ij}^r \alpha_r^a T_{ij} = \sum_i \sum_j T_{ij} \left[\sum_r \beta_{ij}^r \alpha_r^a \right] \tag{8.50}$$

Often the *route choice* map **B** and the *route link incidence map* **A** are combined in an *assignment map* denoted with the symbol **P**, with:

$$\begin{aligned} &\pi_{ij}^a = \text{the proportion of OD-flow } i\text{-}j \text{ that uses link } \alpha \\ &0 \leq \pi_{ij}^a \leq 1 \\ &\pi_{ij}^a = \sum_r \beta_{ij}^r \alpha_r^a \end{aligned} \tag{8.51}$$

Updating OD-matrices under the assumption of AON assignment
When one uses all or nothing assignment as a point of departure, the assignment map **P** reduces to a collection of ones and zeros. This is because all OD-flows are assigned to single routes.

The procedure that is used to update an existing matrix to a set of given traffic counts resembles the methods discussed earlier; it consists of a sequence of scaling operations. In this case the imposed boundary conditions do not relate to trip ends (columns or rows of the matrix), but to blocks of cells in the matrix.

Step 1. Initialize the matrix with cells of the prior matrix:

$$T_{ij} = t_{ij}, \quad \forall i, j \tag{8.52}$$

Step 2. Update the OD-matrix by scaling groups of cells until they fit the traffic counts qa. Repeat this for all available traffic counts (a in A):

initialize an auxiliary solution $T = [T_{ij}]$

$$T_{ij} = T_{ij}^{(n)}, \quad \forall i, j \tag{8.53}$$

update to traffic counts:
for all *a*
(determine scaling factor)

$$c(a,n) = \frac{q_a}{\sum_{i,j} T_{ij} \pi_{ij}^a} \tag{8.54}$$

(scale all cells that contribute to link *a*)
for all *i* and *j* with $\pi_{ij}^a = 1$:

$$T_{ij} := c(a,n) T_{ij} \tag{8.55}$$

set $T^{(n+1)}$ equal to the auxiliary solution

$$T_{ij}^{(n+1)} = T_{ij}, \quad \forall i, j \tag{8.56}$$

Step 3. Check whether or not the matrix computed in the current iteration substantially differ from the parameters in the previous iteration. If this is the case, repeat step 2.

Step 4. Stop

From the above it follows that the final solution satisfies:

$$T_{ij} = t_{ij} \prod_a X_a^{\pi_{ij}^a} \tag{8.57}$$

Just like the case where a matrix is updated using trip ends, it can be shown that the resulting trip matrix T_{ij} maximizes the entropy.

The procedure described above can converge only is the traffic counts that are available are *consistent*. In practice this means that traffic counts, usually need to be made consistent before they can be used.

A theoretical weakness of the method is that the traffic counts are considered as quantities without error, while in practice traffic counts contain error due to:
- physical counting errors;
- differences in the period over which counts are collected at various points (this applies especially if counts are used from different sources);
- Sometimes estimated counts are used instead of actual traffic counts.

Another source of error is the assignment map. This map may imply the use of routes that in reality are not, or not exclusively used.

8.9 Discussion

In this chapter a number of algorithms are presented to estimate OD-matrices and parameters in distribution functions. A theoretical underpinning that applies to the Poisson estimator. In a similar way the other methods can be underpinned, however this is left as an exercise. Because all methods consist of a sequence of scaling operations, OD-cells that do not contribute to any count or margin (A_j or P_i) are left unchanged. This may be an undesirable effect, for example when an old matrix is being updated to a new base year. In this case one would like to express the general growth of traffic with a scaling factor that applies to all cells, and apply these scaling before scaling smaller groups of cells. Modifications like this to the methods described in this chapter are used frequently in practice for a variety of reasons.

9 Engelse-Nederlandse woordenlijst

Access ..... *voortransport*
Accessibility ..... *bereikbaarheid*
Arcs ..... *schakels*
Assessment criteria ..... *beoordelingscriteria*
Assignment ..... *toedeling*
Backnode ..... *voorlaatste node*
Base year matrix ..... *basismatrix*
Captives ..... *captives(i.t.t. keuzereizigers)*
Car ownership ..... *autobezit*
Centroids ..... *voedingsknoop*
Coefficient of determination ..... R^2 *(in regressie)*
Congestion pricing ..... *rekening rijden*
Connector links ..... v*erbindingsschakels*
Consumer surplus ..... c*onsumenten surplus (in micro-economische analyse)*
Cross-classification ..... c*ross classificatie*
Destination ..... *bestemming*
Deterrence function ..... *distributie functie*
Disutility ..... *disnut*
Duality gap ..... *bepaald type* c*onvergentie criterium*
Egress ..... *natransport*
Fares ..... *tarieven*
Friction factor ..... *weerstands coefficienten*
Fundamental diagram ..... *basisdiagram*
Gender ..... *geslacht*
Generalized cost ..... *algemene kosten*
Habitual behavior ..... *gewoonte gedrag*
Headway ..... *volgafstand of -tijd*
Heterocedasticity ..... *heteroskedasticiteit (in regressie)*
Household composition ..... *huishoud samenstelling*
Impedance matrix ..... *weerstandsmatrix*
Intercept ..... *het snijpunt tussen de y-as en de regressie lijn*
Observed trip length distribution ..... *ritlengte verdeling*
Occupation ..... *beroep*
Origin ..... *herkomst*
Perceived travel time ..... *beleefde reistijd*
Queuing theory ..... *wachttijd theorie*
Ramp metering ..... *toerit dosering*
Ridership ..... *voertuigbezetting*
Road pricing ..... *rekening rijden*
Spatial ..... *ruimtelijk*
Study area ..... *studiegebied*
Subjective utility maximization ..... *subjectieve nuts maximalisatie*
Transfer ..... *overstap*
Transport modes ..... *modaliteiten*
Travel purposes ..... *reismotieven*
Trip frequency ..... *rit frequentie*
Trip generation ..... *rit generatie*
Trip production and attraction ..... *herkomst- en bestemmingsgebonden ritgeneratie*
Trip purpose ..... *motief*
Unimodal ..... *unimodaal*
Value of time ..... *reistijdwaardering*

Value pricing ..... *rekening rijden*
Variable message signs..... *dynamische route informatie panelen (drip's)*
Variant .....
Construction .....*bouw var.*
Current.....*huidige situatie*
Do-nothing ..... *nul var.*
Environmental ..... *milieu var.*
Public transport .....*ov var.*
Zero base ..... *nul var.*

10 Register